Novel Transfer Learning Approaches for Improving Brain Computer Interfaces

Ahmed Mohamed Azab

Automatic Control and System Engineering Department

University of Sheffield

A thesis submitted for the degree of

Doctor of Philosophy

2019

Acknowledgements

My foremost gratitude goes to my supervisor Dr.Mahnaz Arvaneh, who has guided me on this journey. Without her excellent supervision, constructive suggestions and encouragement in my research and writing, I would never have gained this view of science and how it can be achieved. Her advice and insight have strengthened me throughout every stage of my PhD journey, not only with my studies demands but also my family life.

I would also like to express my most sincere gratitude to my secondary supervisor Prof.Lyudmila Mihaylova for her constructive comments and suggestions regarding my work and publications throughout my PhD and always being available for guidance.

Next, I must thank my beloved wife for her unconditional love and support that keeps me strong throughout this journey. She never complains about not being at home most of the times in the last three years. Not only that but also at the same time she does an excellent job by taking care of our children. I also feel incredibly blessed for the love and support that I continuously receive from my children Omar, Isel, and Yassin whose smiles let me forget what happens through the day.

For my friends, Ahmed Abo-Elfadl and Moustafa Senousi, thank you all for your valuable support, and I was fortunate to have wonderful friends like you.

Lastly, but certainly not least, the work reported in this thesis would not have been possible without the financial support of Military technical college and the Egyptian Ministry of defence, and support from Col. Mohamed Ismail, Col.Dr.Ahmed Zyada, and Col.Hatem Ezz-Eldin to which I am grateful.

Abstract

Despite several recent advances, most of the electroencephalogram (EEG)-based brain-computer interface (BCI) applications are still limited to the laboratory due to their long calibration time. Due to considerable inter-subject/inter-session and intra-session variations, a time-consuming and fatiguing calibration phase is typically conducted at the beginning of each new session to acquire sufficient labelled training data to train the subject-specific BCI model.

This thesis focuses on developing reliable machine learning algorithms and approaches that reduce BCI calibration time while keeping accuracy in an acceptable range. Calibration time could be reduced via transfer learning approaches where data from other sessions or subjects are mined and used to compensate for the lack of labelled data from the current user or session. In BCI, transfer learning can be applied on either raw EEG, feature or classification domains.

In this thesis, firstly, a novel weighted transfer learning approach is proposed in the classification domain to improve the MI-based BCI performance when only few subject-specific trials are available for training.

Transfer learning techniques should be applied in a different domain before the classification domain to improve the classification accuracy for subjects whom their subject-specific features for different classes are not separable. Thus, secondly, this thesis proposes a novel regularized common spatial patterns framework based on dynamic time warping and transfer learning (DTW-R-CSP) in raw EEG and feature domains.

In previous transfer learning approaches, it is hypothesised that there are enough labelled trials available from the previous subjects or sessions. However, in the case when there are no labelled trials available from other subjects or sessions, domain adaptation transfer learning could potentially mitigate problems of having small training size by reducing variations between the testing and training trials. Thus, to deal with non-stationarity between training and testing trials, a novel ensemble adaptation framework with temporal alignment is proposed.

Contents

Abbreviations

AAWE Adaptive Accuracy-Weighted Ensemble

Adaptive-C3AL Adaptive Combined-CCA

AL Active Learning

ALS Amyotrophic Lateral Sclerosis

ATL Active Transfer Learning

BAA Broad Agency Announcement

BCI Brain Computer Interface

CAD Computer Aided Design

CCA Canonical Correlation Analysis

CCSP Composite Common Spatial Patterns

CSP Common Spatial Patterns

DASVM Domain Adaptation Support Vector Machine

DTW Dynamic Time Warping

DTW-CSP DTW-based CSP

DTW-Ensemble Framework The proposed DTW-Ensemble Framework

DTW-RCSP DTW-based regularized CSP

DTW-RCSP-CV DTW-based regularized CSP using cross-validation

DTW-RCSP-Off DTW-based regularized CSP using offline method

DTW-RCSP-On DTW-based regularized CSP using online method

EEG Electroencephalogram

EEG Electroencephalography

ELM Extreme Learning Machine

ERP Error Related Potential

ERS Event-Related Synchronisation

fNIRS Functional Near Infra-red Spectroscopy

GNMF Group Non-negative Matrix Factorisation

HCI Human Computer Interaction

IPTO Information Processing Technology Office

IWLDA Importance-Weighted LDA

KL Kullback-Leibler

KMM Kernel Mean Matching

LDA Linear Discriminant Analysis

LTL Logistic Regression-based Transfer Learning Algorithm

MEG Magnetoencephalography

MI Motor Imagery

MLLin Multi-Task Learning-based Classifcation Algorithm- Linear Model

MLLog Multi-Task Learning-based Classifcation Algorithm- Logistic Model

mSDAs Marginalized stacked denoising autoencoder

NMF Non-negative Matrix Factorisation

PET Positron Emission Tomography

PCA Principal Component Analysis

SCPs Slow Cortical Potentials

SDS Source Domain Selection

SIITAL Selective Informative Instance Transfer with Active Learning

SITAL Selective Instance Transfer with Active Learning

SMLLin Supervised Weighted Multi-Task Algorithm Linear Model

SMLLog Supervised Weighted Multi-Task Algorithm Logistic Model

SS Subject Specific Classification

ssCSP stationary subspace CSP

SSEP Steady State Evoked Potentials

SSVEP Steady State Visually Evoked Potentials

STIG Spectral Transfer using Information Geometry

SVM Support Vector Machines

S-wLTL supervised weighted logistic regression-based transfer learning

TL Transfer Learning

UMLLin Unsupervised Weighted Multi-Task Algorithm Linear Model

UMLLog Unsupervised Weighted Multi-Task Algorithm Logistic Model

Us-wLTL Unsupervised weighted logistic regression-based transfer learning

VEP Visual Evoked Potential

wAR Weighted Adaptation Regularisation

WML Weighted Multi-Task Algorithm

List of Figures

List of Tables

Chapter 1

Introduction

Brain-computer interface (BCI) provides a direct communication between a person's brain and an electronic device without the need for any muscle control [1,2]. The first BCI research has been proposed in 1973 [3]. It is shocking to know that the field of BCI has been in research for more than fifty years now, even though not many people outside academic life realize it is any more than a mystery. Till now most of the BCI applications and especially non-invasive ones (e.g. EEG-BCIs) are still limited to the laboratories. Although, these applications are working well in labs but bringing them to daily real-life scenarios is challenging. There are two critical questions here. What is the importance of developing a BCI system that can be used in real-life applications?and how can we move towards this step?

It is easy to answer the first question. For people with various neurological conditions, it is not easy to communicate with the world. For those people, it would be acceptable if they were granted a reliable communication method with minimum movements that they can afford such as eye-tracking applications [4,5], movement-based systems [6], and smart typing systems which are used by many disabled people in communications [7]. Even so, there remain patients, for whom none of this is applicable such as Amyotrophic Lateral Sclerosis (ALS) patients who can not afford any form of muscle activity [8]. For these people, the BCI system is the only way to express themselves. Moreover, there are other applications of BCI such as: BCI- based rehabilitation applications, BCI-based games, security and authentications, educational, and smart environment [9]. Thus, accurate, reliable and efficient brain-based communication is highly demanded.

My research aims to answer the second question partially. As in order to use BCI on daily basis out of the laboratory, many challenges need to be addressed. Generally, in BCI, EEG is the most widely used brain signals since it is measured non-invasively with a high temporal resolution [2,10]. Different neurophysiological patterns of EEG have been used to operate BCIs, such as steady state visual evoked potentials, P300, readiness potentials and motor imagery [11]. Among them motor imagery (MI) has attracted increased attention, as unlike many other types of BCI, MI-based BCI does not require any external stimuli and can be used in a self-paced way which is closer to a natural and intuitive control [12]. One of the major limitations of EEG-BCI, specially Motor imagery-based BCI, is its long calibration time. The BCI system has to learn the user's brain patterns and calibrate the system accordingly for every new session. Due to inter-sessions/inter-subjects and intra-session variations in the properties of brain signals, a large amount of training data needs to be collected at the beginning of each session to calibrate the parameters of the BCI system for the target user. Typically, this calibration phase could take up to 20 - 30 minutes for each new session, which is time-consuming, fatiguing, and leaving a reduced amount of time for real BCI interactions [13, 14].

The reasons for having a long calibration could be as follows: First, EEG signals are high dimensional and very noisy. Therefore, it is hard to estimate probability distributions of the features, especially when few trials are available for training, of high dimensional noisy EEG signals where outliers will have tremendously adverse effects. Second, EEG signals are highly non-stationary. This non-stationarity could be caused by many factors such as the users' mental and psychological states variations, fatigue and miss-concentration; also it may be affected by various measurements circumstances. Therefore, the classifier usually performs poorly on a new session data if trained using the features extracted from data of the previous sessions recorded on another day. Third, uniqueness of brain patterns for every person. Typically, there is a great change of the properties of EEG signals when transferring from one subject to another subject. Thus, it is not straightforward to calibrate BCI model for a new subject from EEG data collected from previous subjects or sessions.

One promising approach to deal with the problem of long calibration time can be transfer learning, where data from different users or sessions can be used to compensate the lack of labelled data from the new BCI user [15]. Transfer learning is a machine learning technique used to improve the accuracy of a model trained from one domain by transferring useful information from other domains [16]. When there is a limited supply of training data from the target domain, and it is not easy to collect more, the need for transfer learning appears. Transfer learning has been successfully applied in different machine learning applications such as text, image, and human activity classification [16]. However, in BCI this is relatively a new field of research that needs to be further explored. Transfer learning aims at learning characteristics that are consistent across sessions and subjects and at the same time adjusting those characteristics to the available target subject's few training trials. For BCI, transfer learning can be applied on either raw data, feature or classification domains.

1.1 Motivation

BCI research area has gained more interest in the last decade. However, many challenges need to be addressed to develop accurate and reliable BCI systems that can be used in a daily basis. These challenges could be considered at different levels, e.g., at the neuroscience level [17, 18], by finding more reliable neurophysiological markers, at the human level by developing more advanced and successful user training techniques [19, 20], or at the signal processing level to build more robust approaches which could be calibrated with the least possible data.

The main motivation of this thesis is providing novel transfer learning-based machine learning approaches leading to a better BCI system with less calibration time and improved accuracy.

To have a practical and reliable BCI system for daily use basis, the robustness and the precision of the designed BCI systems have to be particularly considered and improved. Moreover a BCI system with much shorter calibration time is required. For this purpose, this thesis focuses on proposing transfer learning approaches, with the goal of making EEG-based BCI system more accurate and reliable using less calibration data. These improvements will consequently lead to a more intuitive and pleasing interface. This thesis focuses on motor imagery-based BCIs, however, the proposed approaches can be potentially applied to other types of EEG-based systems.

1.2 Aims and objectives

This thesis aims to improve the usability of BCI as a future technology by reducing its calibration time. Fig.1 summarizes the aim, the related challenges, and

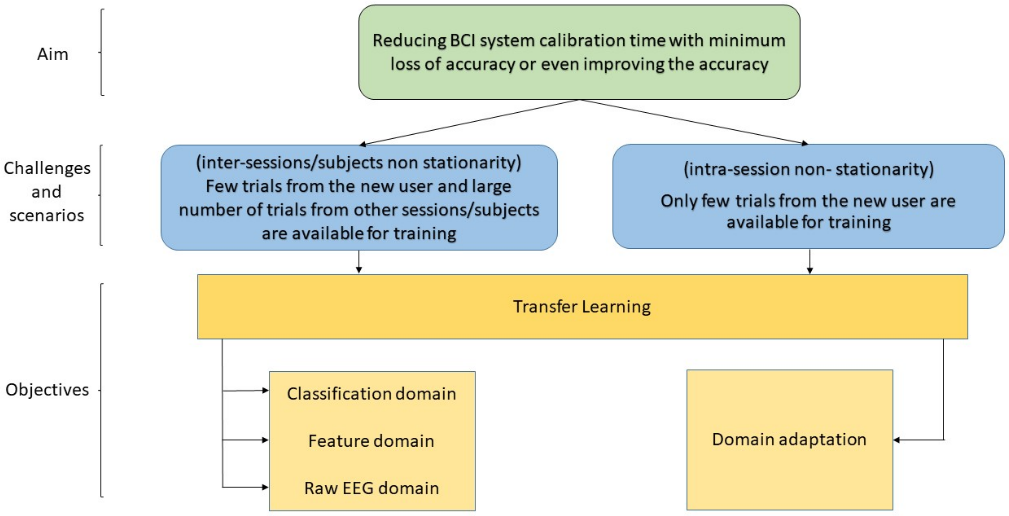

Aim
Reducing BCI system calibration time with minimum loss of accuracy or even improving the accuracy
Challenges and scenarios
(inter-sessions/subjects non stationarity) Few trials from the new user and large number of trials from other sessions/subjects are available for training
(intra-session non- stationarity) Only few trials from the new user are available for training
Transfer Learning
Objectives
Classification domain
Feature domain
Raw EEG domain
Domain adaptation

Figure 1.1: The aim, challenges, and objectives that have been addressed in this thesis

the objectives that have been addressed in this research. As there is typically a trade-off between calibration time and performance of the system, my goal is to reduce this calibration time as much as possible without losing performance and even with improving it. Calibration time could be reduced by either minimizing inter-session/subject or intra-session non-sataionarity through developing novel transfer learning algorithms. Using transfer learning, previously recorded data are mined, processed and reused to improve the BCI model trained for new subjects, hopefully resulting in a reduction of calibration time for new subjects and increasing the accuracy of the system.

To achieve the aim of this project, the following objectives are addressed:

- Developing novel transfer learning algorithms on classification domain to improve the MI-based BCI performance when only a few subject-specific trials are available for training by reusing the classifiers parameters learnt from other subjects or sessions to aid better classification of the target subject new data.

- Developing novel transfer learning algorithms on feature space to improve feature extraction/selection. The proposed algorithms will explore the common information across subjects/sessions to find more robust features that can enhance the model trained by a small training data.

- Developing novel transfer learning algorithms that can be applied on raw EEG directly where data from other subjects/session are transformed to be similar to the few available trials from the target subject. These trials can be used for training a better model for the target subject.
- Developing novel domain adaption transfer learning algorithm to reduce calibration time when there are no available trials from other other subjects or sessions through dealing with non-stationarties that happen over the time between training and testing trials.

The developed algorithms are carefully analyzed and their performance is evaluated across different groups of users. The advantages and disadvantages of these algorithms are discussed in terms of accuracy and computational time and their complementary benefits are considered.

1.3 Thesis overview

This manuscript describes the work carried out in order to address the mentioned objectives. The detailed contents are listed as follows:

In chapter 2, a general introduction to the field of BCI research is given. The reviewed topics include the available techniques for brain activity measurement, different types of brain signals that can be used in BCI-based applications. It also describe how useful features can be extracted from these neurophysiological signals using signal processing methods and then how these features can be converted to become control commands for an external device. After that a general introduction to the field of transfer learning is given. The reviewed topics for transfer learning start with transfer learning definition. Then, the available categories for transfer (inductive transfer learning, transductive transfer learning, and unsupervised transfer learning) are described. After that, transfer learning approaches and categorises are explained. Afterward, the transfer learning methods applied on BCI are reviewed. Then, the challenges and limitations of the available transfer learning approaches in BCI are discussed and some possible future research directions are suggested. Finally, dynamic time warping definition and how it can be used to reduce temporal variations between time domain signals are explained followed by a discussion on how it can be applied to improve BCI performance.

In chapter 3, four weighted multi-task transfer learning algorithms are proposed in the classification domain to reduce the calibration time without sacrificing the classification accuracy of the BCI system. The classification parameters of multiple subjects are learnt jointly such that the average errors as well as dissimilarities between the parameters of the different classifiers get minimized. Dissimilarity is minimized by giving higher weights to previous subject's data that are more similar to the target subject's data and less weights to data that are less similar. Two versions of weighted multitask learning are proposed, namely supervised and unsupervised.

Chapter 4 introduces the proposed weighted transfer learning algorithms in the classification domain when only few subject-specific trials are available for training. In the proposed approach, the classification parameters of each available subject with relatively large number of trials are calculated independently by minimizing the subject-specific classification error. Then, the classification parameters of the new target subject with few labelled trials are calculated such that not only the classification error is minimized but, also the classification parameters of this target subject get as close as possible to the classification parameters of other existing subjects. To further improve the proposed transfer learning approach, different weights are assigned to the previous subjects based on their similarities with the new subject in terms of feature distributions.

Chapter 5 proposes a novel dynamic time warping-based transfer learning for improving common spatial patterns in BCI. Common spatial patterns (CSP) is a popular algorithm for motor imagery EEG feature extraction in the context of brain-computer interfaces (BCIs). However, CSP is computed using sample-based covariance-matrix estimation. Hence, its performance deteriorates if the number of training trials is small. To address this problem, this chapter proposes a novel regularized covariance estimation framework for CSP (i.e. DTW-RCSP) based on dynamic time warping (DTW) and transfer learning. The proposed framework combines the subject-specific covariance matrix estimated using the few available trials from the new subject, with a novel DTW-based transferred covariance matrix estimated using previous subjects trials. In the proposed DTW-based transferred covariance matrix, the available labelled trials from the previous subjects are temporally aligned to the average of the few available trials of the new subject from the same class using DTW. This alignment aims to reduce temporal variations and non-stationarities between previous subjects trials and the available few trials from the new subjects. Moreover, to tackle the problem of regularization parameter selection when only few trials are available for training, an online method is proposed, where the best regularization parameter is selected based on the confidence scores of the trained classifier on upcoming first few labelled testing trials. Impressively, our results show that successful BCI interactions could be achieved with a calibration session as small as only one trial per class.

In chapter 6, a novel domain adaptation transfer learning framework is proposed to reduce calibration by minimizing temporal intra- and inter-session non-stationarity when there are only few trials available for training for the target subject and no trials are available from other subjects or sessions. The proposed framework composed of two main parts, training and testing parts. In training part, a novel dynamic time warping (DTW)-based approach to improve common spatial patterns (CSP) covariance matrix estimation and hence improve feature extraction is proposed. The proposed approach reduces within class temporal variations and non-stationarity by aligning the training trials to the average of the trials from the same class. Using DTW, the available trials from the same class get as close as possible to the mean of this class and also to each other. The new aligned trials are used to calculate the CSP covariance matrices. However, it is found that even using the proposed robust CSP-based DTW to achieve significant improvement for feature extraction does not guarantee a perfect BCI system. The problem might be related to the testing trials, especially, when the BCI users start to feel fatigued or being distracted. Thus, in the testing part, DTW is used to reduce the dissimilarity between testing and training trials and then an ensemble decision making is used to predict the test trials labels with the option of rejecting them.

Chapter 7 summarizes the main conclusions reached and contributions achieved through this thesis. Suggestions for the further research work and potential application areas are also proposed.

1.4 Publications based on this thesis

1- A conference paper was published about improving control of EMG-based assistive device by applying the learnt transfer learning techniques.
Azab, A. M., Arvanch, M., and Mihaylova, L. S., "Estimation of joint angle based on surface Electromyogram signals recorded at different load levels". In the proceedings of the 39th IEEE Annual International Conference of the Engineering in Medicine and Biology Society (EMBC), 2017 (pp. 2538-2541).
2- A book chapter about transfer learning in brain computer interface was published based on the review in chapter2.
Azab, A. M., Toth, J., Mihaylova, L. S., and Arvaneh, M., "A review on transfer learning approaches in braincomputer interface". In Signal Processing and Machine Learning for Brain-Machine Interfaces 2018 (pp. 81-101). Institution of Engineering and Technology
3- A conference paper was published on what proposed in chapter 3 is published
Azab, A. M., Mihaylova, L. S., and Arvaneh, M., "Weighted Multi-task Learning in Classification Domain for Improving Brain-Computer Interface". In the proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC), Miyazaki, Japan, 2018, pp. 1093-1098.
4- A journal article based on what proposed in chapter 4 was published in IEEE Transactions on Neural Systems and Rehabilitation Engineering.
Azab, A. M., Mihaylova, L. S., Ang, K. K., and Arvaneh, M., "Weighted Transfer Learning for Improving Motor Imagery-Based BrainComputer Interface". IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2019, 27(7), 1352-1359.
5- A journal article based on what proposed in chapter 5 has been submitted to the Journal of Neural Engineering Engineering.
Azab, A. M., Mihaylova, L. S., H. Ahmadi and Arvaneh, M., "Dynamic Time Warping-based Transfer Learning for Improving Common Spatial Patterns in Brain-computer Interface".
6- A conference paper was published on what proposed in chapter 6.
Azab, A. M., Mihaylova, L. S., H. Ahmadi and Arvaneh, M., "Robust Common Spatial Patterns Estimation Using Dynamic Time Warping to Improve BCI Systems". In the proceedings of IEEE International Conference on Acoustics,

Speech and Signal Processing (ICASSP), Brighton, United Kingdom, 2019, pp. 3897-3901.

Chapter 2

Introduction to Transfer learning in Brain-Computer Interfaces

2.1 Introduction

A Human Computer Interaction (HCI) which using a mouse or keyboard as an interface to communicate between human and computer is very common [21]. Unfortunately, disabled people who are unable to generate the necessary muscular movements cannot use these standard HCIs. So, in order to help people, Brain-Computer Interfaces (BCIs) needed to be developed. A brain-computer interface (BCI) is a device that allows communication without muscular movement. People can communicate via thoughts only. As BCIs do not require movement, they may be the only possible communication system for users with severe disabilities who cannot speak or use mice, keyboards, or other interfaces [22]. The majority of BCI studies focus on how to help disabled people interacting. However, some studies have initiated BCI-based games for healthy people [23], and other research groups are developing or discussing BCIs for other applications [9].

It is hoped that using BCI applications lead to improve the quality of life of people severe motor disabilities as well as healthy people. BCIs can be a way to improve or recover the mobility of patients with severe motor disorders, e.g. amyotrophic lateral sclerosis (ALS), brain-stem stroke, cerebral palsy or spinal cord injury [24]. A wheelchair can be controlled with motor imagery [25], a P300-speller leads to word spelling [26], and also can be used to control a house environment; opening windows, turning off lights, etc [27, 28]. Recently, BCI has been studied to be used for stroke rehabilitation in order to restore the impaired motor function [29]. Moreover, BCI/neurofeedback systems have been recently studied to enhance central nervous system (CNS) functions such as perception, action, cognition, or emotion [30, 31]. Although, there are many BCI applications, there are many challenges need to be addressed to bring these applications to real life situations. The future may have options to bypass injured sections of the spinal cord, allowing normal movement of the disabled limbs with only the motor imagery based BCI [32], or BCIs can be used for human-machine collaboration, or early disease prediction.

Despite several recent advances, most of the BCI applications are still limited to the laboratory due to their long calibration time. As the literature shows [33–35], due to considerable inter-subjects/sessions or intra-session variations, a reliable machine learning model that performs well across all sessions and subjects has not been feasible yet. Thus, developing reliable methods and approaches that reduce calibration time while keeping accuracy in an acceptable range is highly desirable in BCI research [13,34,36]. One potential approach to reduce the calibration time is transfer learning, where data from other sessions or subjects are mined and used to compensate the lack of labelled data from the current target user [37,38].

This chapter is a comprehensive overview of brain-computer interfaces and transfer learning in BCI. First, a general introduction to the field of BCI research is given. The reviewed topics include the main components of a BCI system main, types of BCIs, the available techniques for brain activity measurement (where EEG is empathised as it will be the core techniques in this thesis), different EEG signal types that can be used in BCI-based applications, and BCI challenges. After-that, transfer learning definitions, techniques, and transfer learning applications in BCI are presented. Finally, the transfer learning challenges in BCI are discussed.

2.2 BCI system components

The whole block diagram of a BCI system is shown Fig.2.1. The BCI system's input is the user brain signals. BCI outputs can be letter or icon selection on a computer screen to improve communication, a wheelchair control, neuroprosthesis guided by functional electrical stimulation (FES) to improve rehabilitation and motor restoration etc [39–42]. In order to translate the input into command signals to control the output device, each BCI uses its' specific algorithms. The main components of a BCI system are as follows:

1. Data acquisition unit: This part record brain activities using different types of

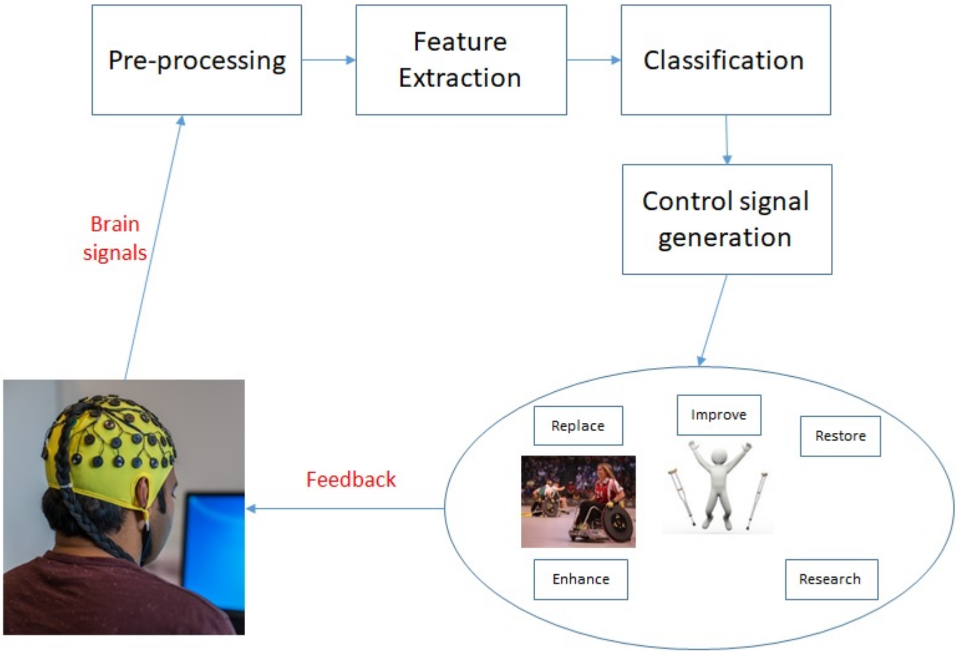


Figure 2.1: BCI system components

sensors. The recorded brain signals serve as BCI inputs after amplification and digitization.

2. Preprocessing unit: This unit reduces noise and artifacts present in the brain signals.

3. Feature extraction: This part generates features related to the underlying neurological states using the pre-processed signals . These features are used by BCI to control the output device.

4. Classification unit: Classification part is used to identify the user's intention from the extracted features.

5. Output device: This could be a wheelchair, a computer, or a prosthesis device etc. The classifier's output is used as a command to control the output device.

6. Feedback: BCI should be a closed loop system, where the system output can be shown to the user can after processing the brain signals. This feedback can be in visual, auditory or tactile form. Using this feedback help the user to better control his brain activities which probably will enhance the BCI performance.

2.3 Types of BCIs

Brain computer interface types can be categorized from different points of view for example (i) exogenous or endogenous, (ii) synchronous (cue-paced) or asynchronous (self-paced), and (iii) invasive or non-invasive [43]. From the point of view of signal acquisition, the electrical activity within a persons brain can be detected by a number of methods. For example the acquisition method can be categorized in two main approached,

- Invasive BCI
- Non Invasive BCI

2.3.1 Invasive measurements

Brain signals are recorded by implanting electrode arrays into the patients cortical tissues, recording extra-cellular voltages from neurons. Two techniques are used to record brain signals invasively included Intracortical neural recordings, and Electrocorticography (ECoG). This recording has high spatial resolution, which require a large amount of small electrodes to be implanted inside the brain. But this method is prone to failure when brain tissue reacts with the implants and so it is not suitable for long-time performance stability [44]. In addition a requirement for a highly skilled surgical team to attach the acquisition device, and also it's high cost.

2.3.2 Non-invasive measurements

Non-invasive BCI, is a technique that measuring signals from the surface of the skull. The main advantage that there is no surgery to implant the electrodes, but at the same time it has the disadvantage of signals being deformed and deflected by the bone tissues of the skull, creating noise and making it harder for a computer to interpret [22]. The non-invasive techniques are preferred in the BCI systems than invasive measurement techniques [45]. There are different non-invasive BCI techniques as follows:

- Electroencephalography (EEG)

- Functional Near Infra-red Spectroscopy (fNIRS)
- Functional Magnetic Resonance Imaging (fMRI)
- Magnetoencephalography (MEG)
- Positron Emission Tomography (PET)

Among all these different techniques, EEG signals are widely used in BCI applications. EEG recording is easy, portable, safe, relatively low cost. Also it has high temporal resolution and can provide many degrees of freedom when used as a control signals [46]. So more information about EEG will be detailed in the next section.

2.4 Electroencephalography (EEG)

The ability of disabled and paralyzed people to communicate again with local environment and make contact with other people can be achieved, by acquiring signals from the scalp surface in a non-invasive way. These signals are in the magnitude of micro-volts and can be detected by very sensitive electrodes which is part of an equipment called an electroencephalograph, which was discovered and developed by Hans Berger a German psychiatrist in 1928 [47]. He announced how to read and interpret person's thoughts by analyzing the EEG waveforms with mathematical processing.

EEG is the measure of the electrical activity of billions of neurons in our brain when they communicate with each other, and this communication is done by generation and propagation of action potentials. These action potentials induce current, and create an electric field that can be measured by the surface electrodes. These potential differences are measured outside the skull, representing the synchronous activity of these neurons [48].

EEG recording system contains electrodes, amplification stage, analogue to digital converter, and a device to record the signal. These signals were acquired using the electrodes, then amplification of the analog signal to increase the signals amplitude in a way that the A/D converter can convert it accurately [43].

There are two types of electrodes which are used to detect EEG: wet electrodes and dry electrodes. When wet electrodes are used, a conductive gel is applied for better transmission of charge between scalp and electrode. In comparison to dry electrodes, which don't use a gel, wet electrodes are less pleasant for the user due to the sticky properties and the long application process to be fixed [49].

The electrodes are placed on the head at fixed locations according to the international 10-20 system, based on standard locations of the skull [50]. The signals recorded by the electrodes are called brain waves. Each person has it's unique EEG signals, but it can be changed according to age, sensory stimuli, brain disease and the mental state of the person. These brain waves divided into four categories, based on their frequency content [43, 51].

- Delta waves (<4 Hz): high amplitude waves seen in babies and in adults during deep sleep or as movement artifacts.
- Theta waves ($\geqslant$ 4 and $<$ 8 Hz): more common in children and very rare in normal awake adults. Sometimes seen in adults when concentrating or during cognitive process.
- Alpha waves ($\geqslant$ 8 and $<$ 12 Hz): relatively regular, rhythmic, low-amplitude waves when in a relaxed state. It is also related to memory brain functions, visual processing and mental effort.
- Beta waves ($\geqslant$ 12 and $\leqslant$ 30 Hz): less regular than alpha waves and occur when mentally alert, focusing on a problem or visual stimulus.
- Gamma waves ($>$ 30 and $<$ 100 Hz): can be obtained during deep concentration and during certain motor functions or perceptions. Gamma waves are not commonly used in BCI systems, because it easily can be affected by artifacts from electromyography (EMG) or electrooculography (EOG).

EEG is a non-invasive way to detect the electrical activity of brain with high temporal resolution, being in the millisecond range, which make it the best method for real-time applications. However, it is still has some drawbacks as EEG has a poor spatial resolution due to the space between the electrodes and the neurons [52]. This leads to that the measured signal is the result of the activity of thousands of neurons, making it hard to distinguish exactly where the activity came from. Moreover, during acquisition of EEG there is a high probability to pick other signals from other sources rather than the brain, called artifacts. These may arise from power line signals (50 or 60 Hz), eye blinking, muscle movement, chewing or heartbeats [53, 54].

2.5 Types of EEG signals for controlling BCI

There are different types of EEG signals that can be used in BCI depending on the task that the system will do. These signals can be divided into two main categories [55].

- Evoked signals
- Spontaneous signals

2.5.1 Evoked Potentials

A visual evoked potential (VEP) is initialized using a visual inspiration, such as an alternating pattern on a screen. Recording and observation of responses are done using electrodes placed on the back of skull during EEG recording. The recorded responses usually originate from the occipital cortex, where the brain involved should receives and interprets the visual signals [56].

With comparison to other sources for BCI (e.g. motor imagery, and slow cortical potentials), VEP based BCIs offer higher information transfer rates with shorter calibration time, less number of electrodes, and low cost [57, 58]. There some limitations for using VEP as a control signal such as: visual fatigue, false positive, and some possibility of causing a seizure [59]. The steady state evoked potentials (SSEP) and P300 [55] are two well-known signals belonging to evoked potentials.

Steady state evoked potentials (SSEP)

SSEP appear as brain potentials when a periodic stimulus is perceived by the subject such as a flickering picture or a sound modulated in amplitude. SSEP are defined as an increase in the power of the EEG signals in the frequencies being equal to the stimulation frequency or being equal to its harmonics and/or subharmonics [20, 23]. In an SSEP-based BCI application, there are several stimuli simultaneously flickering at different frequencies. Each stimulus is corresponding to a task. To activate a task, the user should draw his continual attention to the corresponding stimulus [60].

P300

This signal is a positive wave peak at around 300 ms after task-relevant stimulus. This signal is evoked by different paradigms,the frequency of stimulus occurrence and task relevance are the major factors that influence it. This signal has been shown to be stable in locked-in patients. This finding makes it possible to be used as a control signal for locked-in patients [61]. One of the major application of P300 in BCI is The P300-speller which is used for spelling words or sentences by flashing rows and columns on a screen [62].

2.5.2 Spontaneous signals

Sensorimotor rhythms are the most commonly used signals among all the spontaneous signals. However, other neurophysiological signals are also used in BCI such as slow cortical potentials. These signal types require a sufficient amount of training work before operation of the BCI. During the training phase, the brain signals are recorded while the user is performing mental tasks (e.g. motor imagery).

Slow Cortical Potentials (SCPs)

”SCPs are shifts in the depolarization level of the upper cortical dendrites which are caused by intracortical and thalamocortical afferent inflow to neocortical layers” [63]. There are two types of SCPs. Negative SCPs which are the summation of synchronized ultra slow excitatory postsynaptic potentials from the apical dendrites. SCPs positive type results from synchronized inflow reduction to the apical dendrites or may be caused by inhibitory activity or by excitatory outflow from the cell bodies. Behavioral and cognitive performance are improved after increasing the negativity of the SCP is learnt by subjects, while cognitive behavioral performance is reduced during positive cortical potentials [63].

Sensorimotor Rhythms (Motor Imagery)

Sensorimotor rhythms are EEG rhythms that change with movement or the imagination of movement and do not require any specific stimuli for their occurrence [64]. When planning for movement, this leads to two actions: amplitude modulation, named event-related de-synchronisation (ERD), after that amplitude enhancement, which is called event-related synchronisation (ERS). In the mu band, the de-synchronisation starts 2.5 seconds before movement-onset, peaks after movement-onset and recovers back to baseline within a few seconds [64]. In the beta band, the de-synchronisation is only short-lasting, followed by synchronisation reaching it's maximum in first second after the movement. In the gamma band, synchronisation reaches a maximum right before movement-onset, but these gamma oscillations are rarely found in the human EEG [65].

Sensorimotor rhythms (motor imagery) have been investigated extensively in BCI research [66, 67]. Motor imagery has attracted increased attention, as unlike many other types of BCI, MI-based BCI does not require any external stimuli and can be used in a self-paced way which is closer to a natural and intuitive control [12]. Thus, motor imagery will be the main paradigm in this thesis.

2.6 BCI systems difficulties

2.6.1 Synchronizations

Most BCI studies are cue-paced systems, which means that the time intervals in which communication will be possible, is paced by the BCI system [68]. The EEG-signal can be analysed in predefined time windows, but this severely limits the autonomy of the user, allowing only one thought per time window. Asynchronous BCIs on the other hand, allow the user to communicate at any time, like real time life situations, but this requires continuous analysis, and classification which is a very challenging task [69].

2.6.2 Inter-subject variability

This is one of the most important issues that have been studied in the last few years, as not everyone has the exact same brain, or has the same capability to develop thoughts. Thus, BCI performance depends on the user. A number of algorithms have been developed so that the BCI can automatically detect its' current user and adapt parameters to maximise BCI system performance, making it faster to initialize the BCI [70].

2.6.3 Inter-sessions non-sataionarity

As EEG signals differ from person to another, it also can differ from session to session for the same person. Variations may occur, due to fatigue, task involvement, changes in motivation, or may be due to a slightly different placing of the cap.

Due to inter-sessions non-sataionarity, a long calibration time is needed before each use of the BCI system [13]. This leads the necessary for a large amount of EEG training trials for each subject to be collected for each session to train the classifier. Acquiring these EEG trials are a time consuming and a stressful job for anyone who needs to use the BCI system. So, calibration period reduction has a significant effect to achieve daily life BCI systems applications.

2.6.4 Intra-session non-stationarity

Using large training data sets does not guarantee a good BCI performance as testing trials might be very different from training trials. This problem is more pronounced especially when the BCI user starts feeling fatigued or is distracted [71].

2.6.5 Noise and outliers

High dimensionality of brain signals leads to that brain states related components are often revealed by the background noises. Also outlires due to muscles or eyes movements during recording process. To predict brain states accurately, a large number of training trials are required for feature extraction and classifier training. Besides, when only few data trials are available for training, it is hard to estimate the classifier parameters. As, noises and outliers will negatively affect the extracted features.

2.7 Transfer learning

In order to use BCI on daily basis out of the laboratory, many challenges need to be addressed. One of the main problems is the need for recalibrating the system for every new session/subject. Using machine learning methods for every new session, the BCI system has to learn the user's brain patterns and calibrate the system accordingly. Typically, the calibration could take up to 20 - 30 minutes for each new session, which is an exhausting and tiring amount of time that the patient has to undergo before the system is fully functional [13, 14].

The reasons for having a long calibration in EEG-based BCI can be as follows: The first reason is the high dimensionality of EEG signals which are very noisy as well [72]. In order to predict the right brain states, features need to be extracted from the training EEG data to train the classifier. It is hard to estimate probability distributions of the features from a few trials of high dimensional noisy EEG signals where outliers will have tremendously adverse effects. Second, EEG signals are highly non-stationary [73]. Many factors lead to this non-stationarity such as the variations of users' mental and psychological states, miss-concentration and fatigue; also it may be affected by various measurements circumstances, i.e. changes in the impedance of the electrodes due to sweating [74]. So, the classifier trained on the features extracted from data of the previous sessions usually performs poorly on a new session data. Third, each person has unique brain patterns. The properties of EEG signals typically change when transferring from one subject to another subject. Thus, it is not straightforward to calibrate BCI model for a new subject from EEG data collected from previous subjects.

In order to address the mentioned problem, recent studies try to reduce calibration time based on different methods and algorithms while keeping accuracy in an acceptable range [13, 70, 75, 76]. One promising approach to deal with this problem can be transfer learning, where data from different users or sessions can be used to compensate the lack of labelled data [15].

Transfer learning is a machine learning technique used to improve the accuracy of classifier trained from one domain by transferring useful information from other domains. Machine learning methods have granted remarkable success within different engineering research fields. However, most machine learning methods work well only when data for training and testing purposes is extracted from the same feature space with a fixed distribution. Hence, if any changes happen to this distribution, most statistical models need to be reassembled by collecting new data for training. In many daily life applications, it is expensive and time-consuming to recollect the required data for retraining the model each time we need to use the system. Moreover, in some scenarios, we have access to insufficient labelled data. In such cases, transfer learning between task domains would be a potential solution to reduce the model recalibration efforts. Transfer learning has been successfully applied in different machine learning applications such as text, image, and human activity classification. For brain-computer interface this is relatively a new field of research that needs to be further explored.

2.7.1 History of transfer learning

Machine learning algorithms predict labels of newly coming data by using models that are learnt using available labelled (supervised learning) or unlabelled training data (unsupervised learning) [77]. Also if there are few labelled samples and a large number of unlabelled samples, semi-supervised techniques can be applied [78]. Most of the machine learning algorithms assume that the labelled and unlabelled data have the same distribution, whereas transfer learning allows the domains, tasks, and distributions used in training and testing to be different, which is more related to real-world situations.

The fundamental motivation for transfer learning in the machine learning research was firstly discussed in a Learning to Learn NIPS-95 workshop [79]. Since then much more attention has been paid to transfer learning. In 2005, the Broad Agency Announcement (BAA) of Information Processing Technology Office (IPTO) set a formal definition for transfer learning as: ”the ability of a system to recognise and apply knowledge and skills learnt from previous tasks to novel tasks” [37]. Here, the goal of transfer learning is finding usable information in different tasks of different sources and using it to better deal with the target task. Thus, transfer learning is different from multitask learning where both learnings of the source and target tasks happen at the same time.

Transfer learning approaches have been applied efficiently and successfully in many real-world applications, such as learning text data between different domains [80], image classification problem [81], Wifi localization [82], computer-aided design (CAD) applications [82], and cross-language classification [83]. Also, transfer learning techniques were applied in some biomedical engineering studies such as human activity, muscle fatigue, drug efficacy, and human activity classification [84].

Transfer learning solutions have been first implemented for multi-language processing and image processing classifications, the majority of these transfer learning algorithms could be applied to the different application rather than the one it was implemented for. This property opens the door for transfer learning to be used in other different areas such as attentiveness of drivers, analysing social media reactions, atmospherics data classification [85].

One of the promising applications of transfer learning could be BCI to enhance the overall system accuracy and reduce calibration time. Different studies have tried to apply different transfer learning types in BCI. These studies will be explained in detail.

2.7.2 Transfer learning definition

A domain D is defined by its feature space $\mathcal{X}$ and its marginal probability distribution $P(\mathbf{X})$, where $\mathbf{X} = \{\mathbf{x}_1; ...; \mathbf{x}_n\} \in \mathcal{X}$. Subsequently, given a specific domain, $D = \{\mathbf{X}, P(\mathbf{X})\}$, its' task consists of two terms: a label space $\mathbf{Y}$ and an objective classification function $f(.)$ (denoted by $T = \{\mathbf{Y}, f(.)\}$), which can be learnt using available training data. Thus for a pair of $\{\mathbf{x}_i, y_i\}$, where $\mathbf{x}_i \in \mathbf{X}$ and $y_i \in \mathbf{Y}$, classification of the labels of new trials is done using $f(.)$.

Generally, when there are two different domains, they have either different feature space, different marginal probability distributions or both. Similarly, two different tasks have either different label space, different classification function or both. In this chapter for simplicity, source domain and target domain will be referred as D_S, and D_T respectively.

Definition 2.7.1 *"Given D_S, T_S, D_T, and T_T , transfer learning aims to help improve the learning of the target classification function $f_T(.)$ in D_T using the knowledge in D_S and T_S, where $D_S \neq D_T$ or $T_S \neq T_T$". As $D_S \neq D_T$ means $P_S(\mathbf{X}) \neq P_T(\mathbf{X})$ or/and $\mathbf{X}_S \neq \mathbf{X}_T$. Also, $T_S \neq T_T$ means $\mathbf{Y}_S \neq \mathbf{Y}_T$ or/and $P_S(\mathbf{Y}|\mathbf{X}) \neq P_T(\mathbf{Y}|\mathbf{X})$" [37].*

When using transfer learning, we need to know which types of knowledge we need to transfer and which types we should not transfer. Moreover, it is important to know how and when to transfer them. To address these questions, different types of transfer learning categories and approaches have been proposed in the literature. We will discuss a number of them in the following sub-sections.

2.7.3 Transfer learning categories

Based on the relationship between source and target domains and tasks, transfer leaning can be divided into three main categories. These categories are:

- Inductive transfer learning
- Transductive transfer learning
- Unsupervised transfer learning

Inductive transfer learning

The purpose of inductive transfer learning algorithms is to improve estimation of the target classification function $f_T(.)$ in target domain when the target and source tasks are different ($T_T \neq T_S$). It does not matter if the source and target domains are the same or not [37]. It is noted that in the inductive transfer learning, we assume some target domain labels are available to train $f_T(.)$.

Subsequently, depending on the availability of labelled and unlabelled trials from the source domain, two types of the inductive transfer learning are shown here:

- A large amount of source domain labelled data are available. This case is the most common type of the inductive transfer learning. It is noted that the multitask learning setting also deals with the same situation (i.e. having a large amount of source domain labelled data available). However, in multitask learning, the learning of both target and source tasks are done at the same time while in the inductive transfer learning the target task is learnt based on knowledge transferred from the source task [86].
- There are no labels available in the source domain. Here, it is similar to the self-taught learning [87].

Transductive transfer learning

The goal of transductive transfer learning algorithms is to improve estimation of the target classification function $f_T(.)$ in the target domain when the target and source tasks are similar, but the target and source domains are different [37]. It is noted that in the transductive transfer learning we assume no or few labelled trials are available in the target domain whereas a large amount of labelled trials are available in the source domain.

Subsequently, due to different situations between the source and target domains, transductive transfer learning techniques can be divided into two situations:

- When the feature spaces are different between both the target and source domains, i.e. $\mathbf{X}_S \neq \mathbf{X}_T$. This is also called heterogeneous transfer learning [88].

- When both the source and target domains have the same features space, $\mathbf{X}_S = \mathbf{X}_T$, but the features have different marginal probability distributions, $P_S(\mathbf{X}) \neq P_T(\mathbf{X})$. This kind of transfer learning is related to domain adaptation such as covariate shift method [89]. This is also called homogeneous transfer learning [88].

Unsupervised transfer learning

This type of transfer learning tries to solve the learning problem when there are no labelled trials available in both the source and target domains during training. In the unsupervised transfer learning, while both the source and target tasks are different, there is a relation between them. Unsupervised transfer learning algorithms can be applied to problems involving clustering and dimensionality reduction [37].

2.8 Transfer learning approaches

Based on the type of information that needs to be transferred, the transfer learning algorithms can be categorised into four different approaches, explained as follows:

2.8.1 Instance-based transfer learning

This approach is based on the assumption that the entire source domain cannot be used directly. However, some parts of the source domain data can be reused for learning the target domain function. The estimation of the target classification function is improved by combining the few target labelled data with some instances from the source domain where re-weighted is done if needed [84]. The well-known techniques using this approach are instance re-weighting and importance sampling [37, 90].

2.8.2 Feature-representation transfer learning

This approach focuses on improving the construction of the feature space for the target domain using the data from the source domains, such that the performance of the target task is improved by minimising the classification errors.

Depending on the amount of labelled data available in the source domain, the feature-representation transfer learning can be either supervised or unsupervised [37]. The feature-representation transfer learning can also be formulated in two different types, namely asymmetric and symmetric feature-based transfer learning. The former aims to transform the source features of the source domain in a way to be closer to the target domain. The latest tries to discover the underlying representative structures between both domains to find common latent features that have a same marginal distribution across the source and target domain [88].

2.8.3 Classifier-based transfer learning

This approach focuses on improving the construction of the classification function (classifier) of the target domain using the classification functions of source domain subjects/sessions. Parameter-based transfer learning assumes that some parameters and prior distributions are shared between the individual functions of the source and target tasks. So these shared parameters or priors can be transferred to the target classification function such that the classification errors are reduced. As an example, classifier-based transfer learning can be done by combining multiple source classifiers (ensemble learners) to form an improved target classifier [91].

2.8.4 Relational-based transfer learning

Different from other approaches, this approach deals with problems that the source and target data are not independent and identically distributed (non-i.i.d) and can be presented in many relations. So, this approach aims to find the relational patterns between the source and target domains, and then transfer the knowledge in the source domain to the target domain based on statistical relational learning techniques [92].

2.9 Transfer learning methods used in BCI

As stated before, BCI applications are obstructed by the long calibration time required at the beginning of each session. Transfer learning is a promising approach that can potentially avoid this limitation. Transfer learning can transfer information from different domains (raw data, features, or classification domain) to compensate the lack of labelled data from the subject.

Typically in BCI, two types of information could be transferred; i.e. either discriminative or stationary [93]. Transferring discriminative information aims at constructing more discriminative systems (e.g. by focusing on features, classifier, filters, etc). This approach has been successfully applied to scenarios where the available data samples are few to avoid over-fitting and when the source and target domains are very similar. However, this approach may fail when the target and source domains are not very similar. In this case, transferring stationary information which aims at constructing more invariant systems is more successful as it focuses on common information across domains [94].

In this section, we will review the transfer learning methods applied on BCI from point of view of transfer learning approaches mentioned in section 2.8.

2.9.1 Instance-based transfer learning in BCI

Two well-known techniques using this approach in BCI: importance sampling, where certain values of the input variables have more impact on the parameter being estimated than others and instance re-weighting which aims to weight certain parts of the source domains to be reused in the target domain [37, 90].

Importance sampling instance-based transfer learning

In [95] a method called Bagged importance-weighted LDA (Bagged IWLDA) based on covariate shift adaptation has been proposed to reduce non-stationarities between sessions. Covariate shift adaptation is used to overcome the problem of the supervised learning process which requires a big amount of labelled test samples under the assumption that training and test samples follow the same distribution [96]. However, this basic assumption is mostly violated in real life applications for BCI. Indeed covariate shift adaptation is applied under more realistic assumption where the training and testing samples have different distributions, and at the same time, the conditional distribution of output labels are unchanged. So in [95] a random subset was chosen from the available data to train the classifier. The proposed IWLDA classifier was presented to be an extension of LDA classifier based on the concept of importance sampling as under covariate shift normal LDA is not stable. Importance was calculated as the ratio between the test and training input densities. Then the proposed classifier was learnt using what was randomly chosen as a training subset. N repetitions of this step were held to compute N number of IWLDA classifiers. In the end, the final predictor was obtained based on the average of these N classifiers.

Another approach based on the principle of covariate shift adaptation has been proposed in [97] to reduce non-stationarity between sessions. Marginalized stacked denoising autoencoder (mSDAs) was used to calculate the importance weights. The calculated importance weights were used in the learning algorithm to minimise the mismatch between different sessions. The authors assumed, by proposing this algorithm, to overcome the limitation of traditional techniques that were used to calculate the importance weights (i.e.Kernel Mean Matching (KMM) [96], Kullback-Leibler Importance Estimation Procedure [96] and Unconstrained Least Squares Importance Fitting [98]). These methods were used to calculate the importance weights under the assumption that training and testing data must be available, which is not a practical assumption.

The authors of [99] have proposed a method of transferring selective instances based on an improved active transfer learning (ATL) algorithm. Active learning (AL) is used to find the most informative samples to be chosen for labelling so that a higher performance learning process can be achieved with less labelling effort. For example, if there are two classes, A and B, with unknown distributions and a few labelled trials and a large amount of unlabelled trials are available from each class. AL task here is how to select the most informative unlabelled samples (say two) to be labelled and added to the training data to enhance the classifier training process. There are different method that can be used for this purpose (i.e. least confident, margin and entropy, query by committee, expected error reduction, variance reduction, and density weighted method [100]). AL was used previously in BCI to select the most informative samples from the unlabelled target domain samples to be labelled and added to the classifier training data [55]. In this research, active learning was applied to the source domain labelled samples to choose the samples that were close to target domain labelled samples, and could be added to the training domain.

The authors proposed two algorithms, the first algorithm they proposed was called selective instance transfer with active learning (SITAL), which aimed to enhance the accuracy of direct transfer learning problem that could lead to a negative transfer. The negative transfer happens when the source and target data have great dissimilarity. So, to find data in source subjects domain that is similar to the target domain data, a similarity finding solution was added with trials that are correctly classified using the new subject-specific classifier. This subject-specific classifier was trained using the few trials available in the target domain, were selected for instance transfer. They also proposed another algorithm called selective informative instance transfer with active learning (SIITAL). Active learning was used not only to select the most informative samples, after selection of correctly classified trials from other subjects (in SITAL), but also it checked the normalised entropy of the selected samples and chose samples with higher entropy from these selected trials for instance transfer.

Results showed that SITAL and SIITAL almost had gained slightly higher classification accuracy compared to baseline approaches for some subjects, but not for all subjects when fewer samples were available for training. These algorithms have some drawbacks which might be due to the class imbalance problem during random selection of instances to be labelled in SITAL and SIITAL. That might be because the criteria of instances selection by informativeness when using SIITAL reduced the number of functioning trials for some subjects.

Recently, a domain adaptation with source selection framework has been proposed in MI-BCI system. A deep network is trained using EEG from the target subject and some selected other subjects. These subjects are selected based on how their EEG characteristics are similar to the new subject characteristics using power spectral density in resting-state EEG signals [101].

Re-weighting instance-based transfer learning

There is no obvious re-weighting instance-based transfer learning application in BCI, that select some subsets from the source domain data and re-weight them to be used in the target domain, till now based on our knowledge. But there is one application which applied this approach but for the whole available source domain data.

Transfer learning in [71] was used to reduce non-stationarity between sessions in BCI, where a data space adaptation technique has been proposed to linearly transform the EEG data from the target space to the training space in a way to minimise the distribution difference between the two spaces. Two versions of the EEG data space adaptation were proposed using the Kullback-Leibler (KL) divergence, based on the availability of labels in the testing session: when labelled data were available it was called the supervised version, and when labelled data were not available it was called the unsupervised version. The results showed that concerning classification accuracy, the proposed algorithm for both versions significantly outperformed the results without adaptation even when applied for subjects with poor BCI performance.

2.9.2 Feature-representation transfer learning in BCI

As mentioned before, feature-representation transfer learning focuses on improving the construction of feature space using some knowledge from source domains. Multiple BCI transfer learning studies used spatial filters to learn the new feature representation. There are different algorithms to compute spatial features; among them, Common Spatial Patterns (CSP) is the most commonly used algorithm for extracting discriminative features from EEG signals. However, when there are only a few trials available for training, CSP tends to over-fit. So, different improved approaches for CSP were proposed to overcome this limitation. So from this point, we can categorise feature-representation transfer learning in BCI into two main subcategories based on which method is used to extract the information to be transferred. One category deals with approaches using CSP to extract features and the other category include other methods that can be used to extract EEG features. However, before going through these application an overview of CSP will be introduced.

Common spatial patterns (CSP)

CSP linearly transforms the data from the original EEG channels into new channels to better differentiate between two conditions by maximizing the variance of one condition while minimizing it for the other [102]. The CSP filters are calculated based on assigning new weights for each channel depending on the projection matrix. This projection matrix will have as many filters as the number of channels where each filter carries the weights to make linear combinations of the original EEG channels to decide which EEG channels carry the most useful information. The first half of the projection matrix will maximize the variance for class one and minimize it for class two, while the second half of the projection matrix will maximize the variance for class two and minimize it for class one under the assumption that the signal is band-pass filtered [103]. Based on the number of features needed a number filter pairs are selected. The following equations show how feature extraction based on CSP works.

Let us consider, $\mathbf{X}_i \subset \mathbb{R}^{V\times h}$ is the i^{th} band passed trial, where V is the number of channels and h is the number of EEG samples respectively. Whereas, $\mathbf{W} \subset \mathbb{R}^{V\times V}$ is the projection matrix of CSP, and $\mathbf{Z}_i \in \mathbb{R}^{h\times V}$ is the trial after spatial filtration which is calculated as follows:

$$\mathbf{Z}_i = \mathbf{X}_i^T\mathbf{W}. \tag{2.1}$$

Let $\mathbf{C}_1 \subset \mathbb{R}^{V\times V}$ and $\mathbf{C}_2 \subset \mathbb{R}^{V\times V}$ be covariance matrices of the two classes for EEG signal $\mathbf{X}$ and can be computed by [102]:

$$\mathbf{C}_{(\mathbf{c})} = \frac{1}{n_c}\sum_{i=1}^{n_c}\mathbf{X}_i \times \mathbf{X}_i^T, \qquad c \in [1,2] \tag{2.2}$$

where n_c is the total number of trials for class c. The CSP matrix $\mathbf{W}$ can be computed by:

$$\mathbf{C}_1 \times \mathbf{W} = (\mathbf{C}_1 + \mathbf{C}_2) \times \mathbf{WD}, \tag{2.3}$$

where, eigenvalues for $\mathbf{C}_1$ formed the $\mathbf{D}$ diagonal matrix. Normally, classification is done using m pairs of filters from $\mathbf{W}$. In this chapter, we use the first three and the last three rows of $\mathbf{W}$ which will be named $\mathbf{W}^* \subset \mathbb{R}^{V\times 2m}$ to acquire the spatial filtered signal $\mathbf{Z}_i^* \subset \mathbb{R}^{h\times 2m}$ [102].

$$\mathbf{Z}_i^* = \mathbf{X}_i^T \mathbf{W}^*. \tag{2.4}$$

Thereafter, the most relevant features are extracted so the feature vector $\mathbf{f}_i \subset \mathbb{R}^{2m}$ can be computed by calculating logarithm of variance of $\mathbf{Z}_i^*$ [102].

$$\mathbf{f}_i = log(var(\mathbf{Z}_i^*)). \tag{2.5}$$

These features are used as the input to train the classifier, and hence the trained classifier is used to estimate the labels of unlabelled trials.

CSP-based feature-representation transfer learning

This category can be implicitly divided also into two subcategories based on how transfer learning is applied. For some application modification on how CSP covariance matrix is estimated and for other applications, the modification can be done within the CSP optimisation function or the algorithm.

In [104], Lotte et al. used data from other subjects to improve CSP and Linear Discriminant Analysis (LDA) algorithms. More precisely, it has been proposed that using the data from a subset of source subjects could lead to improve the estimation of CSP covariance matrix and the proposed covariance matrix $\widehat{\mathbf{C}_t}$ was computed as follows:

$$\widehat{\mathbf{C}_t} = \mathbf{C}_t + \lambda(\frac{1}{|S_t(\Omega)|}\sum_{i\in S_t(\Omega)} \mathbf{C}_i), \tag{2.6}$$

where $\mathbf{C}_t$ denotes the estimated covariance matrix using the target subject's data; Ω is the set of subjects with previously collected trials, λ is the regularization parameter ($0 < \lambda < 1$) which was calculated heuristically, $S_t(\Omega)$ is a subset of subjects from Ω, and $\mathbf{C}_i$ is the estimated covariance matrix calculated using data from the subject i. This regularisation aimed to obtain a more stable covariance matrix estimation using covariance matrices of a subset of other available subjects. This subset of available subjects was selected using a sequential algorithm and based on their performance for labelling of the available few trials in the target domain. The results showed that the proposed covariance matrices led to enhance the classification accuracy when few trials were available from the target subject.

Authors of [105] have proposed a CSP algorithm for subject-to-subject transfer using a linear combination of covariance matrices of the source and target subjects to estimate a composite covariance matrix. Consequently, the composite CSP could transfer discriminative information from other domains to overcome the CSP limitation when only few samples are available for training. The composite covariance matrix was calculated using one of the two following proposed methods: Method 1 focused more on covariance matrices calculated using the data from subjects who had large number of trials, where a tuning parameter biased the estimates towards the source domains. Method 2 calculated covariance matrices using subjects' data which were similar to the target domain data, where similarity was calculated using KL-divergence. The general formula to calculate the composite covariance matrix $\widehat{\mathbf{C}_c^k}$ for subject k for both methods was as follows:

$$\widehat{\mathbf{C}_c^k} = (1-\lambda)w_{kk}\mathbf{C}_c^k + \lambda\sum_{j\neq k}^{K} w_{jk}\mathbf{C}_c^j, \tag{2.7}$$

Where λ $(0 < \lambda < 1)$ is the tuning parameter which controls the importance of the covariance matrix from the new subject related to covariance matrices of other subjects. K is the total number of available subjects. For each $c \in \{+,-\}$, $\mathbf{C}_c^{k,orj}$ is the covariance matrix for subject k, orj and class c. Weights for method 1 were computed as follows:

$$w_{jk} = \begin{cases} \frac{N_c^k}{\sum_{j\neq k}^{K} N_c^j} & \text{for } j = k \\ \frac{N_c^j}{\sum_{j\neq k}^{K} N_c^j} & \text{for } j \neq k \end{cases}, \tag{2.8}$$

where N_c^x is the total number of trials belonging to class c and subject x. Weights for method 2 were calculated as follows:

$$w_{jk} = \begin{cases} 1 & \text{for } j = k \\ a_{jk} & \text{for } j \neq k \end{cases}, \tag{2.9}$$

Where a_{jk} are weights for subjects have similar characteristics and can be computed by calculating KL divergence between subjects k and j.

Delvaminck et al. have modified CSP objective function by constructing a shared spatial filter between different subjects by dividing the subject's spatial filter $\mathbf{w}_s$ into a global part and a subject-specific part [106].

$$\mathbf{w}_s = \mathbf{w}_0 + \mathbf{v}_s, \tag{2.10}$$

where $\mathbf{w}_0 \in R^d$ represented the global spatial filter which was shared and learnt over all subjects and $\mathbf{v}_s \in R^d$ represented the subject-specific part of the filter. The number channels was denoted by d. An optimisation framework was described to couple these two parts using a regularised parameters that were used to make a trade-off between these two parts.

$$\underset{\mathbf{w}_0, \mathbf{v}_s}{max} \sum_{s=1}^{S} \frac{\mathbf{w}_s^T \mathbf{\Sigma}_s^{(1)} \mathbf{w}_s}{\mathbf{w}_s^T \mathbf{\Sigma}_s^{(2)} \mathbf{w}_s + \lambda_1 ||\mathbf{w}_0||^2 + \lambda_2 ||\mathbf{v}_s||^2}, \tag{2.11}$$

where $\mathbf{\Sigma}_s^{(1)}$ and $\mathbf{\Sigma}_s^{(2)}$ are the covariance matrices of the trials for the available two classes 1 and 2 respectively for subject s. The parameters λ_1 and λ_2 are a trade-off between the global and the specific parts of the filter, by choosing one of them high and the other one is zero, this leads to force the filter to be specific or more global.

Samek et al. [93] has proposed an extension for CSP using the same idea described above by dividing CSP into two parts. The proposed algorithm was called stationary subspace CSP (ssCSP) where stationary information across multi-subjects instead of discriminative information was transferred by learning a stationary subspace. At first common invariant information between the available subjects were extracted, for each subject eigenvectors decomposition of the difference between the test and training session covariance matrices was computed. Then for each subject, a predefined number of eigenvectors with the largest absolute values were selected; subsequently, all vectors from all subjects were aggregated in one matrix. After that principal component analysis (PCA) was used to reduce the dimensionality of this matrix and extract the most common non-stationarities directions. Finally, in order to construct invariant features for the new subject, CSP filters of this subject were regularised towards the orthogonal complement of the most common non-stationary directions extracted in the previous step. The proposed algorithm is promising when there is a significant change between the training and test data, as it was suggested to reduce the shift between the training and test data.

Also, Samek et al. has introduced a general spatial filter computation framework based on divergence maximisation (divCSP) in [107]. They showed that CSP algorithm could be solved as a divergence maximisation optimisation function. The authors proved that the CSP filters project data to a subspace with the maximum discrepancy, measured by symmetric KL divergence. So instead of calculating spatial filters using CSP, they obtained another solution based on KL divergence by solving the following regularised objective function which consists of two parts, the CSP term and regularisation term.

$$\ell(\mathbf{V}) = (1-\lambda)D_{KL}(\mathbf{V}^T\mathbf{\Sigma}_1\mathbf{V})||(\mathbf{V}^T\mathbf{\Sigma}_2V) - \lambda\Delta, \qquad (2.12)$$

where λ is the regularisation parameter which is used to make a trade-off between the two parts, and it was obtained here by cross-validation. D_{KL} is the symmetric KL divergence between the whitened covariance matrices of data from the two classes for the new subject. $\mathbf{\Sigma}_1$ and$\mathbf{\Sigma}_2$ are covariance matrices of data from the two classes for the new subject. Δ is the regularisation term and can be computed by:

$$\Delta = -\frac{1}{K}\sum_{k=1}^{K} D_{KL}(\mathbf{V}^T\mathbf{\Sigma}_1^k\mathbf{V})||(\mathbf{V}^T\mathbf{\Sigma}_2^k\mathbf{V}), \qquad (2.13)$$

where K is the number of the available subjects. Using this approach, information from other subjects were used. Whereby, by introducing regularisation into the optimisation function solution, it led to the design of a novel spatial filtering algorithm. Thus, by jointly optimising the divergence problems of different users, a subject-independent feature space could be extracted.

Non CSP-based feature-representation transfer learning

There are some other feature-representation-transfer applications where information can be transferred using different strategies, such as transfer stationary information using PCA based covariate shift adaptation as in [108]. Authors aimed to minimise the non-stationaries effect by proposing a new covariate shift adaptation method based on PCA. The most important non-stationary principal components were extracted and normalised by shifting a window over the data to reduce the effect of non-stationarity. Each feature normalisation was done individually rather than normalising the combination of all features. This method is beneficial when the number of dimensions is more than the number of trials and enhance the accuracy of CSP-based methods that work with stationary information. There are different applications using covariate shift for transfer learning proposed in [109, 110].

Another method also has been presented called a non-negative matrix factorisation (NMF). This method was shown to be useful in capturing discriminative information without using the concept of cross-validation in motor imagery EEG tasks [111]. However, direct application of NMF to EEG data of different subjects captures only intra-subject variations. In [112] authors applied NMF in a way called group non-negative matrix factorisation (GNMF) where discriminative information was transferred from multiple subjects. Given EEG data measured from several subjects under the same conditions, the goal was to estimate common task-related bases in a linear model capturing intra-subject variations and at the same time inter-subject variations.

2.9.3 Classifier-based transfer learning in BCI

Classifier-based transfer learning in BCI can be divided into two subcategories: domain adaptation and ensemble learning of classifiers. In domain adaptation, source domains classifiers are reused by adjusting classifier parameters concerning the target domain. Commonality between the source and target domains are compulsory to apply this type of transfer learning. This domain adaptation is commonly used to transfer the discriminative and stationary information between sessions. In the ensemble learning of classifiers, different classifiers trained from different source domains are combined to acquire better classification accuracy of the target domain samples [15].

Domain adaptation in classifiers

In order to solve the problem of EEG non-stationarity between sessions, Bamdadian et al. have proposed an algorithm named an adaptive extreme learning machine (ELM) [113]. At first, the ELM algorithm was trained using previous sessions EEG data trials, and then the trained ELM classifier was used to label test session data. Then these labelled data trials from test session were added to the training set to be used in the training process of the final ELM classifier.

A domain adaptation algorithm called domain adaptation support vector machine (DASVM) has been proposed in [114]. This algorithm was detailed in three steps: 1) using source domain to initialise the discriminative function; 2) replace samples from source domain with samples from target domain to adjust the discriminative function; 3) final discriminative function was learnt using only data from the target domain. Samples replacement were determined based on different settings that can be found in [114].

Ensemble learning of classifier

An example of ensemble leaning of the classifier has been proposed in [36] as a framework for subject transfer. This framework consisted of three parts: 1) two sparse filter sets called robust filter bank and adaptive filter bank were learnt for each subject using the subject's CSP filter bank; 2) two classifiers models were trained for each subject based on these two filter banks; 3) finally, a two-level ensemble strategy was applied to integrate all classifiers from the robust ensemble models and adaptive ensemble models into one robust ensemble learner and one adaptive ensemble learner. Then at the second level these two learners were combined into one final ensemble classifier using a tuning parameter for controlling the balance between adaptiveness and robustness.

When the test sample $\mathbf{x}_i$ was to be classified, the robust models of all subjects $\mathbf{M}_{rj}(j = 1, ..., K+1)$ were used to construct a robust ensemble learner as follows:

$$\mathbf{RE}(\mathbf{x}_i) = \sum_{j}^{K+1} \mathbf{W}_{rj} \times \mathbf{M}_{rj}(\mathbf{x}_i), \tag{2.14}$$

where $\mathbf{RE}(\mathbf{x}_i)$ denotes the robust ensemble result of test sample $\mathbf{x}_i$, $\mathbf{M}_{rj}(\mathbf{x}_i)$ is the result of the robust model of subject j and $\mathbf{W}_{rj}$ is the weight of the model M_{rj}. The adaptive ensemble learner was computed using the following equation:

$$\mathbf{AE}(\mathbf{x}_i) = \sum_{j}^{K+1} \mathbf{W}_{aj} \times \mathbf{M}_{aj}(\mathbf{x}_i), \tag{2.15}$$

where $\mathbf{AE}(\mathbf{x}_i)$ denotes the adaptive ensemble result of test sample $\mathbf{x}_i$ and $\mathbf{M}_{aj}(\mathbf{x}_i)$ is the result of the adaptive model of subject j. $\mathbf{M}_{aj}(j = 1, ..., K + 1)$ are the adaptive models of all subjects.

A dynamic ensemble strategy based on the classification consistency with the neighbourhood structure of the test example surrounded by subject j samples was applied to assign different weights for distinct test samples. Then the final ensemble learner was defined as follows:

$$E(\mathbf{x}_i) = (1 - \lambda)\mathbf{RE}(\mathbf{x}_i) + \lambda\mathbf{AE}(\mathbf{x}_i), \tag{2.16}$$

where $\lambda \in [0, 1]$ represents the tuning parameter which was calculated by cross validation.

The authors of [55] have proposed a novel application of transfer learning (TL) for online calibration of a single-trial error related potential (ERP) classifier. First, labelled training trials from the new subject only were used to train a support vector machines (SVM) classifier. Then, data available from each other subject was combined with the few labelled trials available from the new subject to train an SVM classifier for each subject. After that, the final classifier C_{new} was constructed by combining classifiers from all these subjects as follows:

$$C_{new} = C_i + \sum_{m=1}^{M} C_m * a_m, \tag{2.17}$$

where the subject-specific classifier of the new subject, C_i, had a unit weight, and the weight of the classifier of each subject m, C_m was assumed to be the average of the cross-validation accuracy a_m which is how accurately the few available trials of the testing subject were labelled. This iteration was repeated ten times, and each time two new labelled trials from the target domain were added to the training domain. Selection of these two trials was done either randomly or using AL. The authors attempted to enhance the classification accuracy by integrating AL with TL. They selected some unlabelled samples from the target domain to be labelled and added to the training domain and they named this algorithm ATL. The data instances that had the greatest amount of uncertainty were selected to be the most informative, as these samples had the most disagreement within the trained classifiers [55]. Results of the proposed methods, when compared with two baseline approaches using SVM classifier showed that TL and ATL almost outperformed baseline approaches when there were few trials available for training from the new subject. ATL mostly outperformed TL, and this could be because during TL the random selection of trials may lead to class trials imbalance when there are only a few labelled trials available.

In [115] the same authors again have proposed another ensemble classifier approach named weighted adaptation regularisation (wAR), which used data from other subjects to reduce the amount of labelled data required in the offline single-trial classification of ERPs. The proposed model explicitly handles class-imbalance problems which are common in many real-world BCI applications. They also aimed to reduce the computational cost of wAR by proposing a source domain selection (SDS) approach which selects the closest source domains (i.e. existing subjects) to the target domain. Thus, SDS was performed to select the closest source domains, and then wAR was applied on selected source domain separately to obtain the best classifier parameters for that specific source domain. The final classification was a weighted average of these individual classifiers, with the weight being calculated based on the training accuracy of the corresponding wAR.

Besides, an adaptive accuracy-weighted ensemble (AAWE) approach has been proposed in [116] to allow tracking of non-stationarities in EEG signals using data from other subjects. AAWE combines different individual classifiers, and each classifier trained using data recorded from each individual subject, the weight given to each classifier was initialised based on the accuracy of classifying calibration data for the new subject. After that, the weights were updated using ensemble learning within feedback phase, when there were no true class labels available in the classification domain.

Another approach used multi-task techniques to transfer information between session or subjects was proposed in [54]. In this algorithm, a parametric probabilistic approach that used shared priors was employed to calculate the classification parameters of a new session/subject by defining the relation between this session/subject parameters and the shared priors of available sessions/subjects. These shared parameters were used to compute the classifier parameters of a new subject when there were only a few trials available from this subject.

This algorithm works as follows: $s = \{1, ..., S\}$ is the multiple subjects or recording sessions with n_s trials.The class label of a new trial can be predicted by

$$y_s^{i+1} = sign(\mathbf{w}_s^T \mathbf{x}_s^{i+1}), \tag{2.18}$$

where $\mathbf{w}_s$ is the classification parameter being used to predict the class label for subject/session s trials, x_{i+1} denotes the feature vector extracted from new trial of subject s. y_s^{i+1} presents the classes, for example: the left or right hand movement motor imagery is performed in trial i at session/subject s is presented by $y_s^i \epsilon \{-1, 1\}$. So using the available data sets and labels, the objective is to determine the best $\mathbf{w}_s$ which lead to the best labels classification of the trials for each subject/session such that the number of errors in this dataset D_s is reduced.

The authors claimed that for a BCI problem, each subject/session is treated as one task, where $(\mu, \mathbf{\Sigma})$ shared structure can be presented respectively by the mean vector and covariance matrix of $\mathbf{W}$ where $\mathbf{W} = \{\mathbf{w}_1, ..., \mathbf{w}_s\}$. So the goal of this model is how to calculate these shared parameters from all the tasks jointly in a way that these $\mathbf{w}_s$ reduce the error and also are close together, and this can be achieved by solving the following optimisation problem:

$$minLP(\mathbf{W}; D_s, \mu, \mathbf{\Sigma}, \lambda) = min(1/\lambda) \sum_s ||(\mathbf{X}_s \mathbf{w}_s - y_s)||^2 + \sum_s \Omega(\mathbf{w}_s; \mu, \mathbf{\Sigma}) + C, \tag{2.19}$$

where the first term of this optimisation problem is the sum of the losses from each task, and by minimising it all sessions are ensured to be well fitted together. The divergence of each task model from shared structures is controlled using the second term. Finally, by solving this optimisation problem with respect to W and holding $(\mu, \mathbf{\Sigma})$ fixed this yields the following equation to update $\mathbf{w_s}$:

$$\mathbf{w}_s = ((1/\lambda)\mathbf{\Sigma}\mathbf{X}_s^T\mathbf{X}_s + I)((1/\lambda)\mathbf{\Sigma}\mathbf{X}_s^T y_s + \mu) \tag{2.20}$$

For fixed $\mathbf{W}$, solving the optimisation problem yields the update equations of μ and $\mathbf{\Sigma}$, which as also how the multi-task algorithm works until finding the new subject classification parameter [86].

In [117], a method for unsupervised transfer learning named spectral transfer using information geometry (STIG) was proposed. This process aimed to rank and combine unlabelled classifications from individual different subjects ensemble classifiers. Authors claimed that the proposed method significantly outperformed the existing techniques of classifying ERPs when few trials are available for training.

The within-session and subjects differences can be understood as geometric transformations of the covariance matrices using the Riemannian framework.

Riemannian geometry presents an optimum method for looking over the transfer learning problem because of the affine invariance property of the Riemannian distance and mean.

In [118] authors aimed to make EEG data of different subjects/session comparable by the affining transform of the spatial covariance matrices of the EEG signals of every session/subject. Authors assumed that covariance matrices shifts concerning a reference (resting) state could happen due to different source configurations and electrode positions. Using a reference covariance matrix, the covariance matrices of every session/subject were placed with respect to the reference, so that only the displacement with respect to the reference state was observed when there was a new task. For every session, there was a reference matrix estimation, but different subjects. Then, a congruent transformation was performed using the available data and this reference matrix. Although, there were different reference matrices within sessions and subjects, but the reference matrix was chosen accurately, different sessions/subjects data could be compared.

The proposed procedure was tested in a classification problem, where data from different sessions (subjects) were used to estimate the class parameters that needed to classify new trials.

2.9.4 Unsupervised transfer learning

Unsupervised adaptive transfer learning approach has been proposed in [119]. This approach provided robust class separation in the feature space of the target subject by learning steady state visually evoked potentials (SSVEP) templates for this subject which led to enhanced classification accuracy. By using an extended version of Canonical correlation analysis (CCA) called Adaptive Combined-CCA (Adaptive-C3A) that used a set of reference signals consisting of sinuses and cosines at the fundamental and harmonic frequencies of the SSVEP stimuli to define linear multivariate projections in EEG data. After that a simple matching classifier template was selected to predict the target class label by allocating the frequency label to the EEG segment which best coincided with the corresponding template frequency.

Recently, a novel approach has been proposed in the Euclidean space where EEG trials from different subjects are aligned to make them more similar, and hence improve the learning performance for a new subject. Results on two publicly available MI datasets showed the effectiveness of the purposed approach for some subjects. However, there are some limitations due the dataset shift among different subjects which need to be compensated [120].

2.10 Challenges and discussion

Through this chapter, most of the state of the art BCI-based transfer learning algorithms have been reviewed. However, it is very difficult to define a dominant algorithm that can be used in all scenarios. There is a lack of comprehensive comparisons between previously proposed transfer learning algorithms in the field of MI-BCI even either within the same domain or across domains. Previously proposed algorithms are different in the signal processing techniques, feature extraction methods, data sets used, or the number of trials used for transfer learning. Thus, in order to conduct comparative analysis or meta analysis to choose the best transfer learning based BCI algorithm, a huge number of parameters need to be considered. These parameters include settings related to the acquisition system such as: the amplifier model, cap model, type of electrodes, recorded channels, and analyzed channels. It also should include dataset and its related settings such as: number of subjects, subjects gender, subjects age, right handed or left handed, motor imagery task description, number of trials, feature extraction methods, feature selection, classifiers, results and analysis. Therefore, it would be very beneficial for the BCI community to conduct a systematic comparison between different transfer learning algorithms on the same datasets in the near future.

Although several studies focused on transfer learning in BCI, there are still many open questions that need further investigation. These include what to transfer, when to transfer and how to transfer. Current studies focused on transferring either discriminative or nonstartionary information. However, how to identify them are not reliably investigated across all subjects. Future studies are needed to better identify these subsets of information. Moreover, finding the subsets of information that satisfy both being stationary and discriminative would lead to better results. In addition, exploring information that more specifically reflect the mental activity performed by the user might better address the problem.

Last but not least, identifying different clusters of previous subjects and using only a subset that is similar to the target subject might be useful in improving transfer learning in BCI. Below, we will further discuss the limitations of different transfer learning approaches in BCI.

2.10.1 Instance-based transfer learning in BCI

Instance-based transfer learning is still a developing research area; different aspects need to be further investigated. For example, data from some subjects that are similar to the target subject can be chosen instead of using all the available data from all subjects. Selection of subjects and trials can be made based on the few trials available from the target subject. There is a need for algorithms that can accurately identify which parts of information should be transferred and what is the most suitable approach to transfer them, and how to re-weight these selected data if required. In addition transfer learning algorithms should be able to properly deal with the unbalanced class trials, as in a real scenarios the user should not be stressed to perform equal balanced class actions.

2.10.2 Feature-representation transfer learning in BCI

From what was described previously it is shown that existing traditional CSP-based methods calculate covariance matrices on a subject-specific basis. When there are only a few trials for training available, the performance of CSP methods on a subject-specific basis is degraded as the estimated covariance matrices are over-fitted. Different modifications were applied to traditional CSP algorithm to overcome this limitation as stated before. However, there are still some subjects that may not gain from these modifications. Moreover, finding the optimum regularisation parameter is still a challenge. Fo many cases, the regularisation parameter is calculated using cross validations over a number pre-defined values which requires a long computational time. For some other cases, the regularisation parameter is calculated empirically which is not optimum.

2.10.3 Classifier-based transfer learning in BCI

Since classifier-based transfer learning is just focused on construction of the classifier, it might not be useful for subjects who have non-separable features, as changing the parameters of the classifiers does not add any separability to the feature space. Thus, classifier-based transfer learning might be better to be coupled with either instance-based or feature-based transfer learning approaches in order to be useful for all subjects including those with initially poor BCI performance.

2.11 Summary

Within this chapter a general introduction to the brain computer interface and transfer learning in BCI was described. First, a general introduction to the field of BCI research was given. The reviewed topics include BCI system main components, types of BCIs, the available techniques for brain activity measurement, where EEG was empathised as it will be the core techniques in this thesis, different EEG signal types that can be used in BCI-based applications, and then BCI challenges. Moreover, transfer learning definitions and techniques were explained. Then, some of the available transfer learning applications in the brain-computer interface were explored to better identify the suitable approaches that can be used to reduce calibration time and at the same time increase the accuracy of the BCI-based system. These approaches could be summarised as follows:

- Transfer learning algorithms on classification domain that reuse classifiers learnt from other domains to aid better classification of new data.

- Transfer learning algorithms on feature space to improve feature extraction/selection. The investigated algorithms explore the common information across subjects/sessions to find more robust features that can enhance the model trained by a small training data.

- Transfer learning algorithms that can be applied on raw EEG directly to mine and reuse certain parts of data from other subjects/session for training a better model for a new subject.

Finally, the challenges and limitations of the available transfer learning approaches in BCI were discussed and some possible research directions were suggested.

Chapter 3

Weighted Multi-Task Learning in Classification Domain For Improving Brain-computer Interface

3.1 Introduction

A major challenge in brain-computer interface (BCI) is that everyone has unique brain signals [14]. Using machine learning techniques, BCI has to learn the user's brain signals, but this training takes time. In order to accurately classify the thoughts, the BCI system needs a calibration session to adapt its parameters to the user's signals. Generally for MI-based BCIs, this calibration session could take up to 20 - 30 minutes for each new session, which is an exhausting and tiring amount of time that the user has to undergo before the system is fully functional [65]. There are different reasons for having such a long calibration session, 20-30 minutes, in MI-based BCI. As mentioned before this can be because of the high dimensionality of EEG signals which are very noisy as well or/and because of highly non-stationarity of EEG signals. In the case when there are only few training trials available, it is hard to estimate probability distributions for high dimensional noisy EEG signals specially if these few trials contains outliers. Moreover, due to non-stationarity of EEG the classifier which uses the features extracted from previous sessions data usually performs poorly on the new session. Though, to mitigate the mentioned problem, recent studies try to reduce the calibration time based on different methods while keeping accuracy in an acceptable range [65, 70, 75, 76].

Achieving zero calibration time is the optimum case for a real time BCI system that can be used in daily life tasks. The main objective of this thesis is to reduce the calibration time to be as minimal as possible, for example: 10 training trials which are 2-3 minutes or even less. Transfer learning techniques can be used to reduce the calibration time. In BCI, there are some studies that applied transfer learning-based approaches on raw EEG [121], feature extraction [70, 93, 105] and classification domains [75, 86] and represent some improvements in reduction of calibration time.

Recently, a multi-task learning-based algorithm on the classification domain has been proposed to reduce calibration time in BCI for a new subject [86, 122]. Multitask learning is a sub-field of transfer learning where multiple classification tasks are learnt jointly. The classification parameters of multiple subjects are learnt jointly such that the average errors as well as dissimilarities between the parameters of the different classifiers get minimized. However, the proposed algorithm did not consider the similarity/dissimilarities between the data from the new subjects and the existing data from other subjects during the learning process. To address this problem and improve the BCI classifier trained for a new subject, this chapter proposes a novel weighed multi-task learning algorithm,where previously recorded data are mined, processed and reused in a way that higher weights are given to the data that are more similar to the new data and less weights to data that are less similar. A new similarity measure based on the kullback-leibler divergence (KL) is used to measure similarity between two feature spaces obtained using CSP. Two versions of weighted multitask learning are proposed, namely supervised and unsupervised. The proposed algorithms are evaluated using BCI Competition IV dataset 2a which was recorded from 9 subjects during a motor imagery paradigm. The experimental results showed that our proposed algorithms outperform the baseline algorithms not only by reducing the calibration time but also by enhancing the classification accuracy for some subjects.

The rest of this chapter is structured as follows. Section 3.2 introduces the baseline algorithms used throughout this chapter, then the proposed weighted multi-task model is presented. After describing the dataset used to evaluate the models in Section 3.3, Section 3.4 covers the results and discussions. Finally, Section 3.5 concludes this work with a short summary and future work suggestions.

3.2 Methodology

3.2.1 Baseline algorithms

Two main baseline training algorithms will be explained in this subsection. The first algorithm is the commonly used subject-specific BCI training model where the support vector machines (SVM) classifier is trained independent from other subjects using features extracted from the common spatial patterns (CSP) algorithm for the target subject. The second baseline algorithm is the standard multi-task learning-based classification algorithm. CSP algorithm has been chosen as it is the most commonly used subject specific algorithm in BCI. Although, filter-bank CSP algorithm has been used in several BCI applications, it tends to overfit when the available subject specific training trials are few. We have applied filter-bank CSP on the subjects of 2a BCI Competition IV 2008 when only 5 trials per class were used for training. We found that there was a huge loss in classification accuracy compared to CSP. The second baseline algorithm is the standard multi-task learning-based classification algorithm. This algorithm has two versions, the first one is the linear regression-based multi-task linear proposed in [86] and the second one is the logistic regression-based multi-task proposed in [123]. These multi-task algorithms have been chosen as baseline algorithms as it is the closest algorithms to the proposed algorithms in this chapter and the only multi-task algorithms applied on MI-BCI.

Subject Specific Classification (SS)

In this algorithm, subject-specific training trials with known labels are used to train an SVM classifier based on CSP features. The classical motor imagery-based BCI subject-specific model, used in this thesis, consists of the following parts: bad trials removal, band-pass filtering, common spatial filtering, extraction of log band power features and SVM classifier. These parts are described as follows: First, a threshold test is applied to remove bad trials due to blinks or any unintended motion, then a band pass filter within the band 8 to 30 Hz is used on EEG data to remove brain activities that are out of the range known for motor imagery [104,124,125]. Next, CSP, the commonly used spatial filtering algorithm in EEG, is applied for spatial filtering [103,126]. The importance of spatial filtering arises due to the poor spatial resolution of EEG measurements. CSP linearly transforms the data from the original EEG-channels into new channels to better differentiate between two conditions by maximizing the variance of one condition while minimizing it for the other condition [102]. Thereafter, normalized log band power of CSP filtered EEG signals are extracted as features. Finally, the extracted features are used to train a SVM classifier. This trained classifier is used to classify the labels of the test trials.

Multi-Task Learning-based Classification Algorithm- Linear Model (MLLin)

Alamgir et al. have proposed a framework for multi-task learning in BCI [86]. In this framework, each BCI subject/session was defined as one task. A parametric probabilistic approach that uses shared priors was employed to calculate classification parameters of a new subject by defining the relation between this subject's parameters and shared priors from the available subjects/sessions [86, 122].

This algorithm works as follows: $s \in \{1, ..., S\}$ is the multiple subjects or recording sessions. For each subject/session, the n_s EEG trials are presented as $d_s = (\mathbf{x}_s^i, y_s^i)_{i=1}^{n_s}$, where $\mathbf{x}_i$ denotes the feature vector extracted from the i^{th} trial of subject s, and y_s^i presents the class label of the i^{th} trial. Thus, $\mathbf{X} = \{\mathbf{x}^1, ..., \mathbf{x}^{n_s}\}$ is the feature matrix for each subject/session with labels presented as $y_s^i \epsilon \{-1, 1\}$.

By assuming the classification model as a linear model with a noise term η which is distributed as $\sim \mathcal{N}(0, \sigma^2)$, the label of any trial can be modelled as

$$y_s^i = \mathbf{w}_s^T \mathbf{x}_s^i + \eta, \tag{3.1}$$

where the classification parameters $\mathbf{w}_s$ refers to the individual features weights being used to classify the class labels of the trials belonging to the subject/session s. Thus, when a new test trial, $\mathbf{x}_s^{i+1}$, arrives for the subject/session s, the class label can be classified by

$$y_s^{i+1} = sign(\mathbf{w}_s^T \mathbf{x}_s^{i+1}). \tag{3.2}$$

Typically, when there is no prior information available about the distribution of the model's parameters, using the available labelled trials in the dataset, the objective is to determine the best w_s that minimizes the classification error in the dataset d_s. The loss function for calculating $\mathbf{w}_s$ can be defined using negative log-likelihood as follows:

$$L_1(\mathbf{w}_s) = \min_{\mathbf{w}_s} [1/\sigma^2 \sum_{i=1}^{n_s} (y_s^i - \mathbf{w}_s^T \mathbf{x}_s^i)^2] \tag{3.3}$$

When prior information about $\mathbf{w}_s$ is available and assumed to be Gaussian distributed with $\mathcal{N}(\boldsymbol{\mu},\, \boldsymbol{\Sigma})$, a regularization term R can be added to the loss function leading to reduce the complexity of the system and hence to prevent over-fitting. Thus, R is defined as:

$$R(\mathbf{w_s}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (1/2)([(\mathbf{w}_s - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{w}_s - \boldsymbol{\mu})] + log|\boldsymbol{\Sigma}|); \tag{3.4}$$

From this point of view the authors in [122] assumed that for a BCI problem, each subject/session is treated as one task, where the shared structure, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ can be presented respectively by the mean vector and covariance matrix of $\mathbf{W}$ where $\mathbf{W} = \{\mathbf{w}_1, ..., \mathbf{w}_S\}$. This model calculates these shared parameters from all the tasks jointly in a way that the $\mathbf{w}_s$ calculated for different subjects reduce the total classification error and also are close together, and this can be achieved by solving the following optimization problem:

$$L_2(\mathbf{W}) = \min_{\mathbf{W}} [1/\sigma^2 \sum_s ||(\mathbf{X}_s \mathbf{w}_s - y_s)||^2 + \sum_s R]. \tag{3.5}$$

Finally, solving this optimization problem with respect to $\mathbf{W}$ and holding μ and $\boldsymbol{\Sigma}$ fixed yields the following equation:

$$\mathbf{w}_s = ((1/\boldsymbol{\sigma}^2)\boldsymbol{\Sigma}\mathbf{X}_s^T \mathbf{X}_s + I)((1/\sigma^2)\boldsymbol{\Sigma}\mathbf{X}_s^T y_s + \boldsymbol{\mu}) \tag{3.6}$$

For fixed $\mathbf{W}$, solving the optimization problem yields to identify the update equations of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ as the following equations. Thus, optimum $\mathbf{w}_s$ can be calculated in an iterative manner by iteratively updating $\mathbf{w}_s$ and ($\boldsymbol{\mu}^*$ and $\boldsymbol{\Sigma}^*$) until convergence. Finally, σ^2 is calculated using cross validation.

$$\boldsymbol{\mu}^* = (1/S) \sum_s \mathbf{w}_s \tag{3.7}$$

$$\boldsymbol{\Sigma}^* = \frac{\sum_s (\mathbf{w}_s - \boldsymbol{\mu})(\mathbf{w}_s - \boldsymbol{\mu})^T}{Tr(\sum_s (\mathbf{w}_s - \boldsymbol{\mu})(\mathbf{w}_s - \boldsymbol{\mu})^T)} + \epsilon I \tag{3.8}$$

Multi-Task Learning-based Classification Algorithm- Logistic Model (MLLog)

The authors of [123] modified the previously presented MLLin algorithm by using logistic regression instead of linear regression. Assumptions on the distribution of the dependent variables in logistic regression model could be more suitable for a binary classification problem than those in linear regression [127].

The MLLog algorithm aims at minimizing the following optimization problem:

$$L_3(\mathbf{W}) = \min_{\mathbf{W}} - \sum_{s} \sum_{i=1}^{n_s} H(\mathbf{w}_s, y_i, \mathbf{x}_i) + \sum_{s} R, \tag{3.9}$$

where H is the point wise cross-error function, and R is the regularization term as defined in (3.4). By calculating the optimum $\mathbf{w}_s$ in (3.9), the classification of the labels of a given trial is then calculated as:

$$P(y_s^i|\mathbf{x}_s^i) = \frac{1}{1 + exp(-\mathbf{w}_s^T \mathbf{x}_s^i)}. \tag{3.10}$$

Similar to MLLin, L_3 should be minimized with respect to $\mathbf{W}$ in order to obtain the parameters of the classifiers across subject. However, unlike the MLLin algorithm, there is no closed form solution for $\mathbf{w}_s$ in this optimization problem. However, gradient based optimization procedures [128] could be applied to obtain the optimal $\mathbf{w}_s$ given the shared parameters ($\boldsymbol{\mu}, \boldsymbol{\Sigma}$). Following the same steps that were presented in the MLLin algorithm, the shared parameters were calculated using standard Gaussian sample statistics from the optimal weights $\mathbf{w}_s$ as in (3.7, 3.8) respectively. Iterative optimization should be then applied to update $\mathbf{w}_s$ and $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ iteratively until convergence.

3.2.2 Proposed weighted multi-task algorithm (WML)

The MLLin and MLLog algorithms treat all the subjects similarly so that the similarities/dissimilarities between the new subject and previous subjects are not considered in the learning process. The proposed WML algorithm addresses this limitation by giving each subject a different weight based on how the features distribution of this subject/session is close to the features distribution of the new subject. Thus, instead of updating shared parameters by giving the same weights to all subjects/sessions, they are weighted by taking into account similarities/dissimilarities of each subject with the new subject.

Fig.3.1 presents the proposed WML algorithm used to calculate the classification parameters of the new subject. As shown in Fig.3.1 the proposed WML algorithm consists of two parts. In the first part, the best $\mathbf{W} = \{\mathbf{w}_1, ..., \mathbf{w}_s\}$ for the previous subjects are calculated in away that the total classification error is reduced for these subjects and at the same time their classification parameters are close to their weighted average which is calculated by assigning weights to the subjects based on their similarities to the new subject. In the second part, weighted shared priors $(\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w)$ obtained from the previous part are used with the new subject few trials to obtain this new subject classifier parameters. Optimum $\mathbf{w}_{new}$ is calculated in an iterative manner aiming to reduce the classification accuracy error for the new subject while the defined regularization makes it close to the weighted shared priors.

There are two main differences between the proposed weighted algorithms and the baseline multitask algorithms. Firstly, three different methods for covariance matrix calculation are examined, and a comparison between these methods is held to choose the best method based on the best classification accuracy results. The first method to calculate a covariance matrix is referred to as cov1(size) and calculated as below:

$$\boldsymbol{\Sigma} = \frac{\sum_s(\mathbf{w}_s - \boldsymbol{\mu})(\mathbf{w}_s - \boldsymbol{\mu})^T}{size((\mathbf{w}_s - \boldsymbol{\mu})(\mathbf{w}_s - \boldsymbol{\mu})^T)} + \epsilon I. \tag{3.11}$$

The second method, called cov2 (trace), is calculated as:

$$\Sigma = \frac{\sum_s(\mathbf{w}_s - \boldsymbol{\mu})(\mathbf{w}_s - \boldsymbol{\mu})^T}{Trace((\mathbf{w}_s - \boldsymbol{\mu})(\mathbf{w}_s - \boldsymbol{\mu})^T)} + \epsilon I, \tag{3.12}$$

and the third method is called cov3 (diagonal) and its equation is as follows:

$$\boldsymbol{\Sigma} = \frac{diag\sum_s(\mathbf{w}_s - \boldsymbol{\mu})(\mathbf{w}_s - \boldsymbol{\mu})^T}{Trace(\sum_s(\mathbf{w}_s - \boldsymbol{\mu})(\mathbf{w}_s - \boldsymbol{\mu})^T)} + \epsilon I, \tag{3.13}$$

where $size$ refers to the total number of elements in the covariance matrix, $Trace$ refers to the trace of the matrix, $diag$ refers to the diagonal elements of the matrix, and $\epsilon = 0.0001$ is used to ensure the stability of the equation when the first part of the equations gets equal to zero, i.e to prevent creation of singular (non-invertible) covariance matrix

Algorithm 1: The proposed weighted multi-task algorithm

```
part 1
    Input  : $d = \{d_1, ....., d_S\}$, $\sigma^2$, KL weights($\alpha_s$)
    Output: $\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w$
Set $[\boldsymbol{\mu}, \boldsymbol{\Sigma}] = [0, I]$
Repeat
update $\mathbf{W} = \{\mathbf{w}_1, .., \mathbf{w}_s\}$
update $\boldsymbol{\mu}$ using weights (3.16)
update $\boldsymbol{\Sigma}$ using weights (3.17:19)
Until convergence
return $\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w$ weighted shared priors
part 2
    Input  : $d_{new}$, $\sigma^2_{new}$, $\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w$
    Output: $\mathbf{w}_{new}$
Set $[\boldsymbol{\mu}, \boldsymbol{\Sigma}] = [\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w]$
Repeat
calculate $\mathbf{w}_{new}$
Until convergence
return $\mathbf{w}_{new}$
```

Figure 3.1: The proposed weighted multi-task algorithm, where σ^2 and σ^2_{new} are selected using cross-validation

The second main difference is the weight that is defined for each subject to represent the similarity between this subject and the new subject. Kullback-Leibler (KL) divergence is used to calculate these weights [71]. The KL divergence between two gaussian distributions $\mathcal{N}_0(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ and $\mathcal{N}_1(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ has the following closed form [129],

$$\begin{aligned} \mathrm{KL}[N_0||N_1] = 0.5[(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}_1^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \\ + \mathrm{trace}(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_0) - \ln\left(\frac{\mathrm{d}et(\boldsymbol{\Sigma}_0)}{\mathrm{d}et(\boldsymbol{\Sigma}_1)}\right) - K], \end{aligned} \tag{3.14}$$

where det and k denote the determinant function and the dimensionality of the data, respectively. Therefore, in the proposed weighted algorithm, (3.14) is used to calculate the distance between the feature distributions of each subject and the new subject. It is noted that we use CSP features in this study. CSP features are normalized log variance of CSP-filtered EEG data, thus the assumption of having Gaussian distribution can be valid.

If labelled trials from the new subject are available, supervised KL distance is computed for each class and the total distance is the sum of the distances for the two classes. When there are no labelled trials available for the new subject, the KL distance between the two subjects is calculated without considering the class labels and it is called unsupervised KL. Considering these two weighted distances, the proposed algorithms can be supervised weighted multi-task (SMLLin, and SMLLog) and unsupervised weighted multi-task (UMLLin, and UMLLog), where Lin and Log are referring to the applied regression method. The weight between the subject s and the new subject, α_s, is calculated using the following equation:

$$\alpha_s = \frac{(1/(\bar{\mathrm{KL}}[d_t, d_s] + \epsilon)^4)}{\sum_{i=1}^{m}(1/(\bar{\mathrm{KL}}[d_t, d_i] + \epsilon)^4)}, \tag{3.15}$$

where $\bar{\mathrm{KL}}$ is the total divergence calculated using the features distributions of the few available training trials of the target subject d_t (i.e. 10, 20 or all trials per class depending on how many trials are defined as available) and the available trials from the previous subject/session d_s. In (3.15), $\epsilon = 0.0001$ is used to ensure the stability of the equation when $\bar{\mathrm{KL}}[d_t, d_s]$ gets equal to zero due to having two compared distributions completely similar. Although, this is a very rare event, we must take into account the possibility of unseen events. Based on the obtained weight for each subject, α_s, the new equation to update the weighted $\boldsymbol{\mu}$ is:

$$\boldsymbol{\mu}_w = \sum_s \alpha_s \mathbf{w}_s. \tag{3.16}$$

Similarly, the weighted $\boldsymbol{\Sigma}$ is calculated using the following modified equations for cov1 (size), cov2 (trace), and cov3 (diagonal) respectively:

$$\boldsymbol{\Sigma}_w = \frac{\sum_s (\alpha_s \mathbf{w}_s - \boldsymbol{\mu}_w)(\alpha_s \mathbf{w}_s - \boldsymbol{\mu}_w)^T}{size((\alpha_s \mathbf{w}_s - \boldsymbol{\mu}_w)(\alpha_s \mathbf{w}_s - \boldsymbol{\mu}_w)^T)} + \epsilon I \tag{3.17}$$

$$\boldsymbol{\Sigma}_w = \frac{\sum_s (\alpha_s \mathbf{w}_s - \boldsymbol{\mu}_w)(\alpha_s \mathbf{w}_s - \boldsymbol{\mu}_w)^T}{Trace((\alpha_s \mathbf{w}_s - \boldsymbol{\mu}_w)(\alpha_s \mathbf{w}_s - \boldsymbol{\mu}_w)^T)} + \epsilon I \tag{3.18}$$

$$\boldsymbol{\Sigma}_w = \frac{diag \sum_s (\alpha_s \mathbf{w}_s - \boldsymbol{\mu}_w)(\alpha_s \mathbf{w}_s - \boldsymbol{\mu}_w)^T}{Trace((\alpha_s \mathbf{w}_s - \boldsymbol{\mu}_w)(\alpha_s \mathbf{w}_s - \boldsymbol{\mu}_w)^T)} + \epsilon I \tag{3.19}$$

3.3 Experiments

In order to validate the proposed algorithms and compare them with the baseline algorithms, all the algorithms were applied to dataset 2a BCI Competition IV 2008 [130]. This dataset consists of EEG data from 9 subjects performing 4 classes of motor imagery task. In this chapter only data from right and left hand motor imagery are used. Two sessions on different days were recorded for each subject. Each session is comprised of 6 runs, each run consists of 12 trials for each class.

EEG signal was recorded using 22 electrodes. EEG signals were sampled at 250 Hz, and were bandpass-filtered between 0.5 Hz and 100 Hz. Moreover, a 50 Hz notch filter was applied to remove power line noise. In this chapter, 2 s window of EEG data starting 0.5 s after the cue is used to calculate the features for each trial. The proposed algorithms and the baseline algorithms were applied only on the trials recorded on the second day by dividing it to two sessions one for training (consists of the first 42 trials recorded per class) and one for testing (consists of the last 30 trials recorded per class). This was done to establish a practical case that new subject data is coming from the same session. Where, the BCI user uses the first few trials, 2-3 minutes, to train the system before using it, which is supposed to happen in daily life tasks. For the new subject, different training sizes were examined (i.e. 10, 20 and 42 trials per class). In each multitask learning algorithm, the train data of each new subject and the other 8 other subjects were used for calculating classification parameters.

3.4 Results and discussion

As mentioned before, in this section the multitask learning algorithms were applied based on three different covariance matrix calculation methods and two regression approaches (i.e. Linear and Logistic). All algorithms were evaluated using different number of training trials from new subjects (i.e 20, 40, all 84 trials from both classes).

To identify the most effective method of calculating covariance matrices, first a comparison between the three different covariance matrix calculation methods was held across different number of training trials for new subjects, regression approaches and all the applied multitask learning algorithms. Subsequently, a 3 (Number of trials)×6 (Algorithms)×3 (covariance calculation methods) repeated measure ANOVA test was performed on the results followed by post-hoc analysis.

Fig. 3.2 compares the classification results obtained by the different methods of calculating covariance matrices using 20 trials from the new subjects. These results include the classification accuracies of all the different multitask learning algorithms in both linear and logistic approaches. According to the average accuracies shown in the Fig. 3.2, cov3(diagonal) yielded higher classification accuracies than cov1(size) and cov2(trace). Indeed, conducting a repeated ANOVA test revealed that using different covariance matrix calculation methods had a main effect on the classification accuracy results with ($p = 0.009$). Based on the post-hoc analysis cov3(diagonal) significantly outperformed cov1(size) and cov2(trace) with the p values equal to 0.03 and 0.025 respectively. Thus, for the rest of this chapter, all the calculations and comparisons of multitask algorithms will be done using only cov3(diagonal).

Another comparison between the linear regression and the logistic regression approaches was conducted. As shown in Table 3.1, on average the logistic approach outperformed the linear approach in all the considered multitask learning algorithms when 40 trials used from the new subjects for training. Although not presented in the table, the results of using 20 or all the trials from new subjects

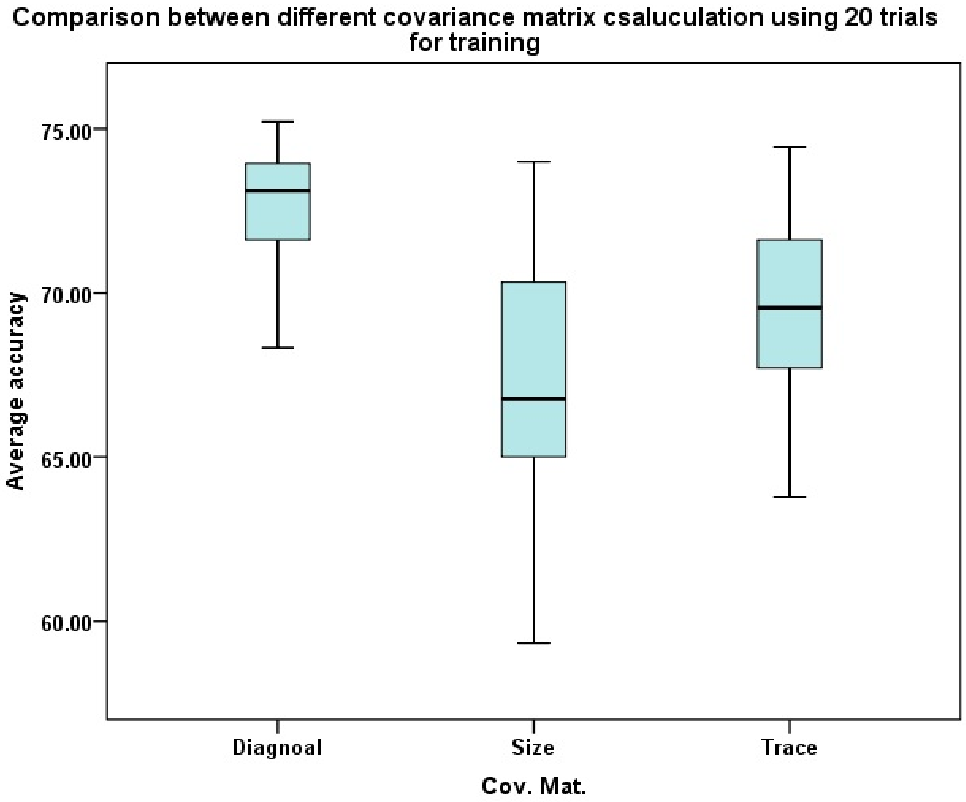


Figure 3.2: Comparison between different covariance matrix calculation methods when 20 trials from the new subjects are used for training. The average accuracy calculated include results obtained by MLLin, SMLLin, UMLLin, MLLog, SMLLog, and UMLLog.

also showed that the logistic regression approach worked better than the linear one for the majority of the subjects.

Finally, comprehensive comparisons were conducted based on the classification results of the 7 algorithms (i.e. SS, MLLin, MLLog, proposed SMLLin, proposed SMLLog, proposed UMLLin, and proposed UMLLog), followed by a 3 (Number of trials)×7 (Algorithms) repeated measure ANOVA test.

Fig. 3.3 shows that all the proposed weighted multitask learning algorithms outperformed the subject specific algorithm (SS) when there are only 20 trials available for training from the new subjects. When the number of the training trials from the new subject increased to 40 and all, still the majority of the proposed weighted multitask learning algorithms out performed SS. Besides the proposed

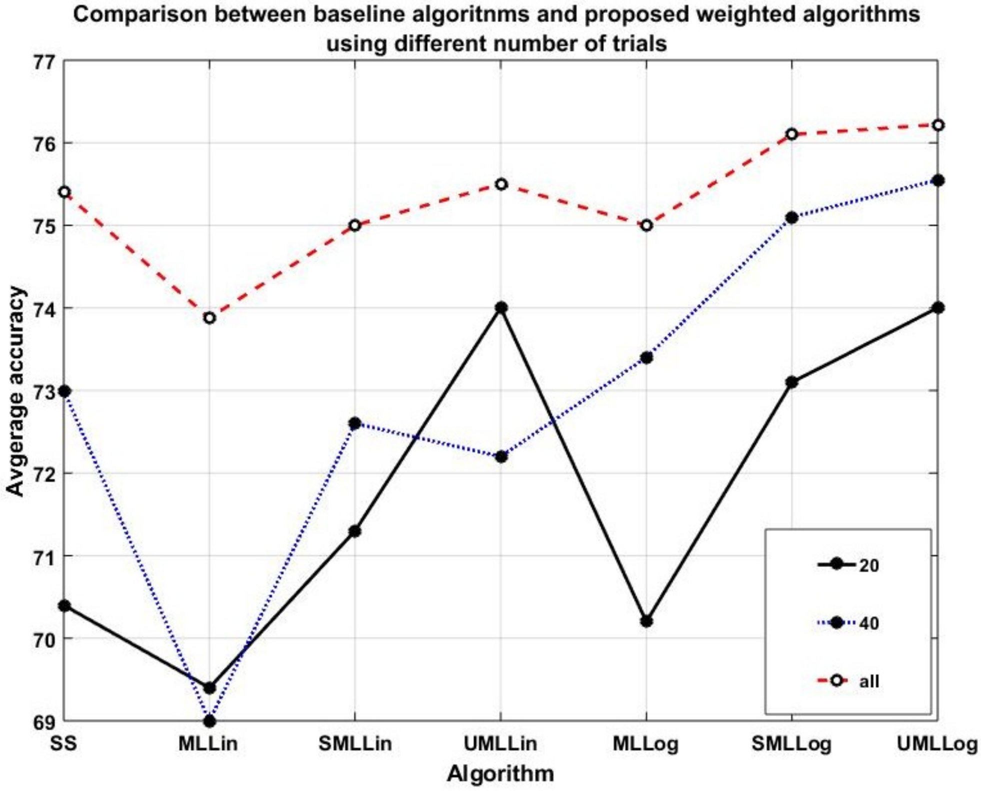


Figure 3.3: Comparison between the proposed algorithms (SMLLin, UMLLin, SMLLog, and UMLLog) and the baseline algorithms (SS, MLLin, and MLLog) using different number of trials for training (20, 40, and all trials) from new subject based on the average accuracy calculated over the nine subjects for each algorithm. UMLLog is the best algorithm when using any number of trials.

algorithms outperformed the baseline linear and logistic multi-task algorithms when using 20, 40, and all trials from the new subjects for training.

Based on the statistical tests, neither MLLin and MLLog significantly outperformed the state of art SS algorithm nor any of the proposed algorithms. Importantly, the classification accuracy of the proposed UMLLog algorithm tended to be significantly better than the SS algorithm results. Moreover, the proposed UMLLog algorithm significantly outperformed the baseline MLLog algorithm with the p value of 0.045, whereas SMLLog tended to be significantly better than MLLog with the p value of 0.078. Interestingly, when using diagonal matrix calculation method with the baseline logistic multi-task algorithm, the modified logistic algorithm was significantly better than MLLog with $P = 0.021$. Moreover, statistical

Table 3.1: Classification accuracies calculated using the baseline algorithms (SS, MLLin, and MLLog) and the proposed algorithms (SMLLin, UMLLin, SMLLog, and UMLLog) for each individual subject when there are 40 trials available for training from the new subject, showing that logistic algorithms outperform linear algorithms

Algorithm	sub1	sub2	sub3	sub4	sub5	sub6	sub7	sub8	sub9	Avg.
SS	85	53	98	66	55	56	73	86	86	73
MLLin	85	52	97	57	52	55	67	97	60	69
SMLLin	72	58	98	63	55	53	70	98	78	**72.6**
UMLLin	72	57	98	63	55	53	70	95	87	**72.2**
MLLog	90	48	97	67	52	52	75	97	83	73.4
SMLLog	90	50	98	63	58	55	77	98	87	**75.1**
UMLLog	95	50	97	63	58	55	78	97	87	**75.6**

tests showed that using different number of trials did not have a main effect on classification results. This finding strengthens the outcome of this work which is reducing the calibration time without altering the overall accuracy of the system.

Fig. 3.4 and Fig. 3.5 show the classification results calculated for each subject using the proposed and baseline algorithms for linear and logistic approaches respectively. The results were obtained when when 20 trials were available for training from the new subject. As can be seen, besides reducing the calibration time, the proposed algorithms outperformed the baseline algorithms for 7 subjects out of 9 in linear regression case and for 5 subjects out of 9 in the logistic regression.

In summary, average classification accuracy results suggest that the novel proposed unsupervised weighted logistic multi-task learning algorithm (UMLLog) outperformed all the other algorithms. The proposed UMLLog not only reduced the required calibration time but also enhanced the average classification accuracy. Although, there is no significant difference between UMLLog and SMLLog, it is preferable to use UMLLog as it doesn't require the availability of labelled trials from the BCI user.

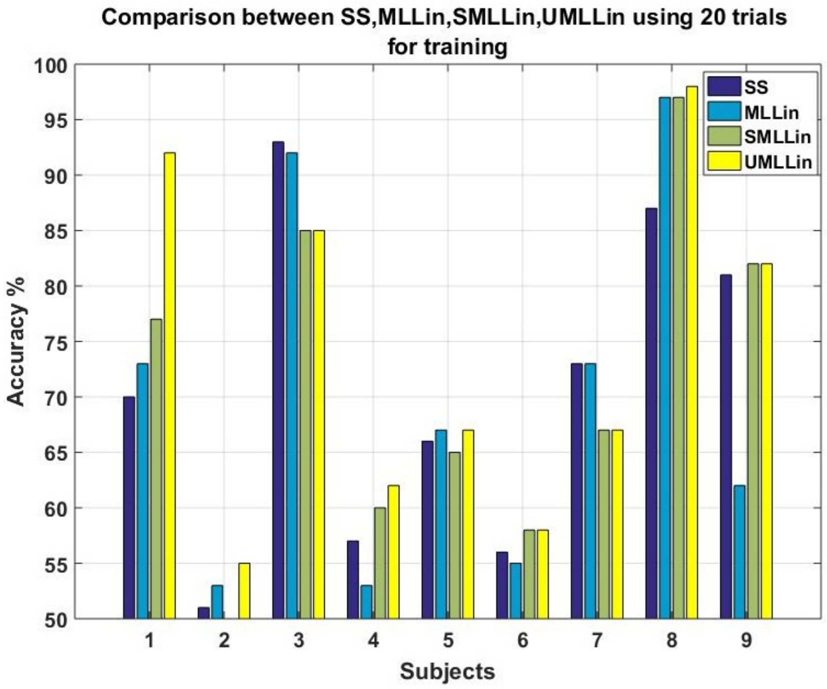


Figure 3.4: Comparison between the classification accuracies calculated using the proposed weighted linear multi-task learning algorithms (SMLLin, and UMLLin) and the baseline algorithms (SS, and MLLin) for all subjects individually when 20 trials are available for training from the new subjects. As can be seen in addition to the calibration time reduction, 7 subjects out of 9 gained an increase in the accuracy when the proposed algorithms are used.

3.5 Conclusion

The aim of this chapter was to develop novel weighted multi-task learning algorithms to reduce the calibration time for MI-based BCI systems and at the same time to enhance the overall accuracy of the system. Previous approaches on multi-task learning in BCI ignored the similarity/dissimilarities between the data from the new subjects and the existing data from other subjects during the learning process. In this chapter, novel weighted multi-task learning algorithms have been presented to address this problem. The main finding of this chapter suggested that applying the proposed weighted multi-task learning algorithms in classification domain led to reduce the calibration time and enhanced the average classification accuracy of the MI BCI-based systems. However, multi-task learning requires a very high computational cost to learn the classification parameters

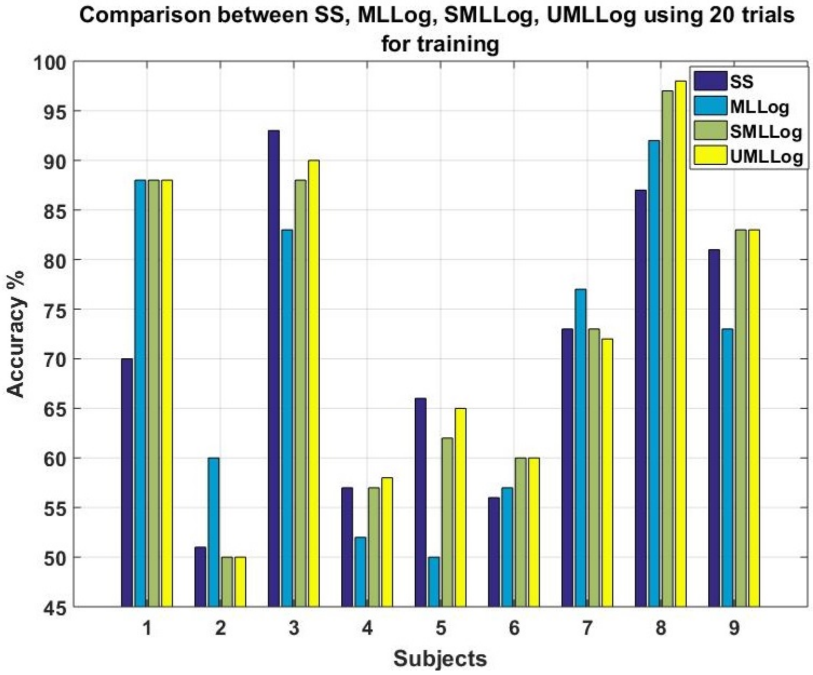


Figure 3.5: Comparison between the classification accuracy calculated using the proposed weighted logistic multi-task algorithms (SMLLog, and UMLLog) and the baseline algorithms (SS, and MLLog) for all subjects individually when 20 trials are available for training from the new subjects. In addition to the calibration time reduction 5 subjects gain an increase in the accuracy when the proposed algorithms are used

of the available subjects jointly. This computational cost will be even higher when the number of subjects (tasks) increases. This limitation is going to be addressed in chapter3 by applying transfer learning approaches on the classification domain using different techniques.

Chapter 4

Weighted Transfer Learning for Improving Motor Imagery-based Brain-computer Interface

4.1 Introduction

As mentioned in the previous chapter, most of the MI-based BCI applications are still limited to the laboratory due to their long calibration time. This calibration phase is time consuming and fatiguing, leaving a reduced amount of time for real BCI interactions [13]. Thus, developing reliable methods and approaches that reduce calibration time while keeping accuracy in an acceptable range is highly desirable in MI-based BCI research [13, 34, 36].

Applying transfer learning techniques could be a possible solution to reduce the calibration, where data from other sessions or subjects are mined and used to compensate the lack of labelled data from the current target user [38]. Indeed, how to do transfer learning is not a trivial task, due to the non-stationarity inherent in EEG signals [37, 38].

In MI-based BCIs, transfer learning can be applied on either raw EEG, feature or classification domains. Domain adaptation techniques [131–133] and ensemble learning of classifiers [36, 134] have been adapted in many existing MI-based BCI transfer learning algorithms on the classification domain. In the domain adaptation, the source domain classifier is used for a target domain while its parameters are adjusted with respect to the target data. Different from the domain adaptation, ensemble learning of classifiers combines different classifiers trained from different source domains to acquire better classification accuracy on the target domain.

Recently an application of multi-task learning has been proposed in BCI [86, 122] where the classification parameters of multiple subjects were learnt jointly such that the average total errors as well as dissimilarities between the parameters of the different classifiers were minimized. Despite success to some

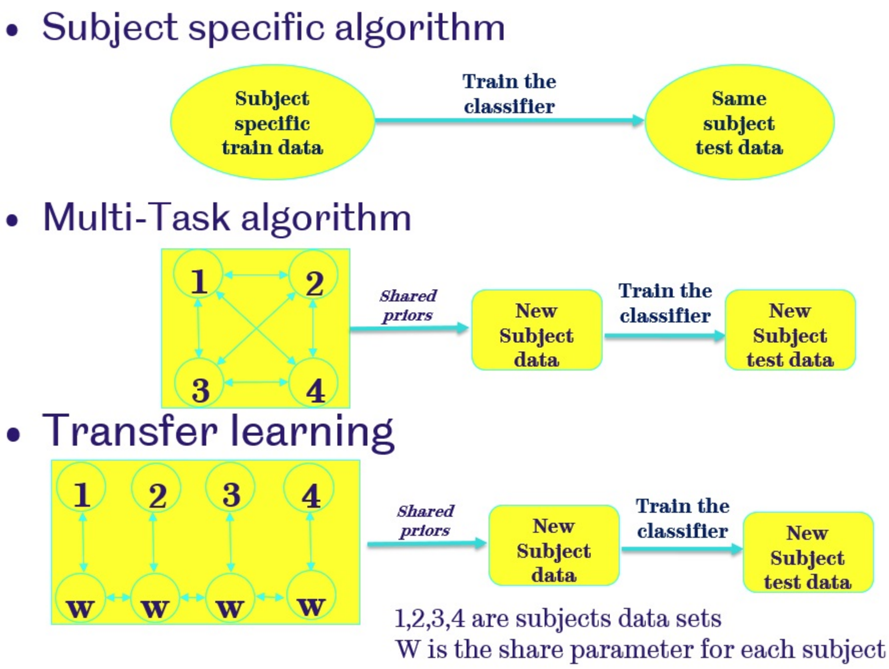


Figure 4.1: Simplified block diagrams for subject specific, Multi-task, and Transfer learning algorithms

extent, the proposed algorithm does not consider similarities/dissimilarities between the data from the new subject and the existing data from other subjects during the learning process. This issue has been considered in chapter2 by assigning different weights to previous subjects' data based on their similarities to the new subject's data. Moreover, multi-task learning algorithm is computationally expensive as a big number of parameters need to be optimized simultaneously. As shown in the block diagram of multi-task algorithm in Fig. 4.1 where the shared parameters of different subjects (tasks) need to be optimized jointly at the same time. This requires a huge computational time especially when the number of subjects (tasks) increases. Thus, in this chapter, transfer learning techniques on classification domain are applied to reduce computational time which lead to reduce the total time that BCI system requires.

This chapter proposes a novel transfer learning approach in the classification domain to improve the MI-based BCI performance when only a few subject-specific trials are available for training. In the proposed approach, the classification parameters (shared parameters) of each available subject with relatively large number of trials are calculated independently by minimizing the subject-specific classification error as shown in the block diagram of transfer learning algorithm in Fig. 4.1. To cope with the problem of having small train data for a new subject, we hypothesize that there is some common information across the subjects performing the same mental tasks (i.e. MI). Following this assumption, the classification parameters of the new target subject with few labelled trials are calculated such that not only the classification error is minimized but also the classification parameters of this target subject get as close as possible to the classification parameters of other existing subjects. This is achieved by adding a regularization term into the classification objective function making a trade-off between minimizing the classification error of the new subject and dissimilarities with the classification parameters of previous users.

It is important to consider that the above-mentioned transfer learning approach may not be very precise for MI-based BCIs that use CSP features, since using the subject-specific CSP for feature extraction leads to different feature spaces for different subjects. To address this issue, we assume, with a fixed coordinate of electrodes, these feature spaces are still relevant as EEG signals are originated from roughly the same areas of the brain for the same motor imagery task leading to nearly similar CSP weights for corresponding channels. Consequently, to transfer the classification parameters across different CSP feature spaces, we link the features of different subjects with the features of the target subject through a new similarity measure obtained using KL divergence. Therefore, the proposed transfer learning approach is further improved by assigning different weights to the previous subjects based on the similarities between their features and the features of the new subject.

The proposed approach is applied on a logistic regression classifier with and without considering similarity weights. The proposed classifiers are evaluated using three datasets with large, moderate, and small number of subjects. The performance of the proposed classifiers are also compared with the results of two state-of-the-art algorithms.

Our results suggest that the proposed weighted transfer learning approach could significantly reduce the required calibration time and also enhance the average classification accuracy, particularly when there are enough previously recorded EEG sessions available for transfer learning. Moreover, the obtained results showed that the proposed weighted transfer learning algorithms significantly outperformed the baseline algorithms.

The remainder of this chapter is structured as follows. Section 4.2, will describe our proposed transfer learning approach. The experimental setup is shown in Section 4.3. Evaluation results are analyzed in Section 4.4. Section 4.5 contains the results discussion. Finally, conclusions are drawn in Section 4.6.

4.2 Methodology

In this chapter, we assume that multiple EEG sessions previously recorded from different subjects or from the same subject are available. Given $s \in \{1, ..., m\}$ as one of the previously recorded sessions, the set of labelled EEG trials from session s can be presented as $d_s = (\mathbf{x}_s^i, y_s^i)_{i=1}^{n_s}$, where $\mathbf{x}_s^i$ and y_s^i respectively denote the feature vector and the class label of the i^{th} trial, and n_s refers to the total number of the trials. Thus, the feature matrix for the session s is presented as $\mathrm{X}_s = [\mathbf{x}_s^1, \mathbf{x}_s^2, ..., \mathbf{x}_s^{n_s}]$, where $\mathrm{X}_s \in \mathbb{R}^{v \times n_s}$ and v is the number of features per trial. Subsequently, the label vector is presented as $\mathrm{Y}_s = [y_s^1, y_s^2, ..., y_s^{n_s}]$, where $y_s^i \in \{0, 1\}$.

This chapter assumes that previously recorded sessions have sufficiently large numbers of labelled trials, whereas the new target subject has only few labelled trials available. Typically, a classifier function, $f(.)$, is trained using the available subject-specific training features to classify the labels of the unlabelled trials. However, when only few labelled trials are available for training, the estimation of the joint distribution $P(\mathrm{X}_s, \mathrm{Y}_s)$ may not be sufficiently accurate. Hence, the classifier function trained using few trials is often not optimal. This chapter proposes a number of transfer learning algorithms to improve the estimation of the classifier function of the new target subject using previously recorded EEG data. Indeed, how to do transfer learning is not a trivial task, due to the non-stationarity inherent in EEG signals $P(\mathrm{X}_s, \mathrm{Y}_s) \neq P(\mathrm{X}_t, \mathrm{Y}_t)$, where t refers to the new target subject.

4.2.1 Proposed logistic regression-based transfer learning algorithm (LTL)

A logistic regression model provides probabilistic predictions by transforming a linear model through a logistic sigmoid function as [135]:

$$P(y_s^i{=}1|\mathbf{x}_s^i;\mathbf{w}_s) = \frac{1}{1+exp^{-(\mathbf{w}_s^T\mathbf{x}_s^i)}}, \tag{4.1}$$

where s denotes the session s, and $\mathbf{w}_s \in \mathbb{R}^{v\times 1}$ refers to the classification parameters being used to classify the class labels of the trials X_s. The obtained probabilistic prediction is then used to classify the class label.

The proposed LTL algorithm consists of two main steps. In the first step, for every previously recorded session, $\forall d_s \in \{d_1, d_2, ..., d_m\}$, the classification parameters, $\mathbf{w}_s$, are calculated using the following objective function [136]:

$$\mathrm{L}_1(\mathbf{w}_s) = \min_{\mathbf{w}_s} \left(\sum_{i=1}^{n_s} \mathrm{H}(\mathbf{w}_s; y_s^i, \mathbf{x}_s^i) + \lambda_s ||\mathbf{w}_s||_2^2 \right), \tag{4.2}$$

where H and $||.||_2$ denote the cross-entropy and 2-norm functions respectively. In fact, in $\mathrm{L}_1(\mathbf{w}_s)$, the cross entropy aims at finding $\mathbf{w}_s$ that minimizes the error rate while the 2-norm penalizes large values of $\mathbf{w}_s$ to reduce the risk of over-fitting. The subject-specific regularization parameter λ_s is used to control the degree of penalization. Cross entropy function H is also called negative log-likelihood where its minimization is equivalent to maximizing the log likelihood [137], as follows [138]:

$$\begin{aligned}\mathrm{H}(\mathbf{w}_s; \mathbf{x}_s^i, y_s^i) = -y_s^i \log P(y_s^i{=}1|\mathbf{x}_s^i;\mathbf{w}_s) - (1-y_s^i) \\ \log(1 - P(y_s^i{=}1|\mathbf{x}_s^i;\mathbf{w}_s)),\end{aligned} \tag{4.3}$$

where $P(y_s^i{=}1|\mathbf{x}_s^i;\mathbf{w}_s)$ is calculated using (4.1). The objective function $\mathrm{L}_1(\mathbf{w}_s)$ does not have a closed form solution. However, it has a unique minimum that can be found using simple and effective iterative approaches such as the gradient descent or Newton's methods [135].

Despite being sufficiently effective for sessions with large training data sizes, the objective function $\mathrm{L}_1(\mathbf{w}_s)$ may fail in estimating the classification parameters of the new subject since few available subject-specific trials typically are not able to accurately represent the distributions of the features. Thus, to estimate the classification parameters of the new subject, $\mathrm{L}_1(\mathbf{w}_s)$ is modified such that not only the error rate is minimized, but also the estimated classification parameters get as close as possible to the classification parameters of the other existing sessions. In other words, in addition to the discriminative parameters, we are interested in parameters that are similar to the classification parameters of the other sessions with this assumption that there is some common information across the sessions performing the same mental tasks (i.e. motor imagery).

Given the above-mentioned assumption, after calculating the classification parameters of the previously recorded sessions using (4.2), in the second step, the classification parameters of the new target subject, $\mathbf{w}_t$, is calculated using the following objective function:

$$\mathrm{L}_2(\mathbf{w}_t) = \min_{\mathbf{w}_t} \left(\sum_{i=1}^{n_t} \mathrm{H}(\mathbf{w}_t; y_t^i, \mathbf{x}_t^i) + \lambda_t \mathrm{R}_{TL}(\mathbf{w}_t) \right), \tag{4.4}$$

where R_{TL} is the regularization term penalizing dissimilarities between $\mathbf{w}_t$ and the previously calculated $\mathbf{w}_s$, $\forall d_s \in \{d_1, d_2, ..., d_m\}$. The regularization parameter λ_t is making a trade-off between minimizing the error and dissimilarities between the new target subject and previous sessions in terms of the classification parameters. The term R_{TL} is calculated by taking into account the prior distribution of the existing classification parameters and comparing them with $\mathbf{w}_t$ as [122]:

$$\mathrm{R}_{TL}(\mathbf{w}_t) = 0.5[(\mathbf{w}_t - \boldsymbol{\mu})^T \boldsymbol{\Sigma}_{TL}^{-1} (\mathbf{w}_t - \boldsymbol{\mu}) + \log(|\boldsymbol{\Sigma}_{TL}|)], \tag{4.5}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}_{TL}$ are respectively calculated as follows:

$$\boldsymbol{\mu} = (1/m) \sum_{s=1}^{m} \mathbf{w}_s, \tag{4.6}$$

$$\boldsymbol{\Sigma}_{TL} = \frac{\mathrm{diag}(\sum_{s=1}^{m} (\mathbf{w}_s - \boldsymbol{\mu})(\mathbf{w}_s - \boldsymbol{\mu})^T)}{\mathrm{trace}(\sum_{s=1}^{m} (\mathbf{w}_s - \boldsymbol{\mu})(\mathbf{w}_s - \boldsymbol{\mu})^T)}. \tag{4.7}$$

As can be seen in 4.7, $\boldsymbol{\Sigma}_{TL} \in \mathbb{R}^{v \times v}$ only includes the normalized diagonal elements of the covariance matrix, where diag and trace give the diagonal elements and the sum of the diagonal elements of a matrix, respectively. Indeed, in this study, only diagonal elements are used to reduce the optimization complexity. Subsequently, in (4.5), $\boldsymbol{\Sigma}_{TL}$ is used to normalize the divergence of each parameter of $\mathbf{w}_t$ from the average of the corresponding parameters of the other classifier.

4.2.2 Proposed weighted logistic regression-based transfer learning algorithm

The proposed LTL algorithm attempts to improve the estimation of the classification parameters of a new subject by incorporating the data from previously recorded sessions. However, it treats different feature spaces from the previous sessions similarly, whereas the distribution of EEG signals can be different from session to session and from subject to subject, leading to different subject-specific CSP feature spaces. Thus, depending on the distributions of EEG signals, the EEG features of the new subject might be similar to the EEG features of some of the previously recorded sessions while very different from those of some others. Thus, taking into account these differences might further improve the estimation of the classification parameters for a new subject. To address this issue, in the proposed weighted logistic regression-based transfer learning algorithm different weights are allocated to the previously recorded sessions to represent similarities between these sessions and the new subject in terms of distributions of the features.

Kullback-Leibler (KL) divergence is frequently used in the literature to calculate similarities between two sets of EEG features [129]. Since in MI-based BCIs the features are typically normalized log-power of CSP filtered EEG data, they are commonly assumed normally distributed [107]. Thus, in this chapter, the KL divergence between two normal distributions are used to measure divergence between EEG features.

Given two normal distributions presented as $\mathcal{N}_0(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ and $\mathcal{N}_1(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, the KL divergence has the following closed form [129],

$$\begin{aligned}\mathrm{KL}[N_0||N_1] = 0.5[&(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}_1^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \\ &+\mathrm{trace}(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_0) - \ln\left(\frac{\mathrm{d}et(\boldsymbol{\Sigma}_0)}{\mathrm{d}et(\boldsymbol{\Sigma}_1)}\right) - K],\end{aligned} \tag{4.8}$$

where det, T and K denote the determinant function, transpose of the matrix, and the dimension of the data, respectively. In this chapter, the total divergence between the features of two EEG sessions, $\bar{\mathrm{KL}}$, can be calculated in two ways, namely supervised and unsupervised. In the supervised case, the total divergence is calculated by averaging the KL divergences calculated for each class separately. On the other hand, in the unsupervised case, the total divergence equals to the

KL divergence between the two sessions without considering the class labels. Subsequently, the similarity weight α_s between the feature sets of the target subject d_t and the feature sets of each of the previous sessions/subjects d_s, is calculated as:

$$\alpha_s = \frac{(1/(\bar{\mathrm{KL}}[d_t, d_s] + \epsilon)^4)}{\sum\limits_{i=1}^{m}(1/(\bar{\mathrm{KL}}[d_t, d_i] + \epsilon)^4)}, \tag{4.9}$$

where $\bar{\mathrm{KL}}$ is the total divergence calculated using the features distributions of the few available training trials of the target subject d_t (i.e. 10, 20 or all trials per class depending on how many trials are defined as available) and the available trials from the previous subject/session d_s. In (4.9), $\epsilon = 0.0001$ is used to ensure the stability of the equation when $\bar{\mathrm{KL}}[d_t, d_s]$ gets equal to zero due to having two compared distributions completely similar. Although, this is a very rare event, we must take into account the possibility of unseen events. The power of 4 is applied to the inverse of KL between the distribution of two feature sets to give larger weights to more similar distributions and smaller weights to less similar distributions. This results in an increased sparsity in the similarity weights α_s. Finally, the similarity weight, proposed in (4.9), is normalized by dividing it by the sum of all similarity measurements between the feature sets of the new target subject and all other available subjects.

The proposed weighted logistic regression-based transfer learning algorithm has the same steps as the proposed LTL. However, instead of equal weights, different weights are assigned to the previously recorded sessions using (4.9). Accordingly, the new weighted $\boldsymbol{\mu}$ is obtained as [139]

$$\boldsymbol{\mu}_w = \sum_{s=1}^{m} \alpha_s \mathbf{w}_s. \tag{4.10}$$

Likewise, the weighted $\boldsymbol{\Sigma}_{TL}$ is calculated as

$$\boldsymbol{\Sigma}_{TL_w} = \frac{\mathrm{diag}(\sum_{s=1}^{m}(\alpha_s \mathbf{w}_s - \boldsymbol{\mu}_w)(\alpha_s \mathbf{w}_s - \boldsymbol{\mu}_w)^T)}{\mathrm{trace}(\sum_{s=1}^{m}(\alpha_s \mathbf{w}_s - \boldsymbol{\mu}_w)(\alpha_s \mathbf{w}_s - \boldsymbol{\mu}_w)^T)}. \tag{4.11}$$

Finally, R_{TL} in (4.5) is calculated by replacing $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}_{TL}$ with $\boldsymbol{\mu}_w$ and $\boldsymbol{\Sigma}_{TL_w}$ respectively. Considering the two above-mentioned ways to calculate the similarity weights, the proposed weighted algorithms are referred to as either supervised weighted logistic regression-based transfer learning (S-wLTL) or unsupervised weighted logistic regression-based transfer learning (Us-wLTL) in the remaining parts of this chapter.

4.3 Experiments

4.3.1 Data description

In order to evaluate the proposed algorithms, three datasets with different number of subjects were used.

1-Dataset 1 [140]: EEG was collected from 19 healthy subjects using 27 channels. For each subject, EEG data were collected without feedback in two sessions conducted on separate days. In this chapter, we used only motor imagery data recorded in the first session. This was done to establish a practical case that new subject data is coming from the same session. Where, the BCI user uses the first few trials, 2-3 minutes, to train the system before using it, which is supposed to happen in daily life tasks. This MI part of the dataset consisted of two runs of EEG recording where the subjects were instructed to perform MI of the chosen hand versus background rest condition. Each run comprised of 40 trials of MI and 40 trials of background rest condition. Thus, in total, there were 160 trials per subject recorded without feedback.

2-Dataset 2 (Dataset 2a from BCI Competition IV) [130, 141]: This dataset consists of EEG data recorded from 9 subjects using 22 electrodes. During the recording sessions, the subjects were instructed to perform one of the four following motor imagery tasks: left hand, right hand, foot or tongue. Two sessions on different days were recorded for each subject with a total of 288 trials per session. In this chapter, only data from right and left-hand motor imagery were used. Moreover, only data recorded from the second day were used due to the practical assumption that the training and the testing data of a new subject are recorded on the same day.

3-Dataset 3 (Dataset IVa from BCI Competition III) [142]: This dataset includes EEG signals from five subjects. EEG was recorded using 118 electrodes. It contains data from two classes of right hand and foot imagery. In total, there are 280 trials per subject all recorded on the same day without receiving feedback.

4.3.2 Data processing

A single elliptic bandpass filter from 8 to 30 Hz was used for filtering the EEG data as recommended in [104, 124, 125], since this single frequency band includes the range of frequencies that are mainly involved in performing motor imagery. Then, CSP were computed for each previous subject independently. Similarly, for the new subject, the CSP filters were calculated only using the available subject-specific training trials. After that, the spatially filtered signals were obtained using the first and the last three spatial filters of CSP as recommended in [102]. Finally, the normalized log band power of the spatially filtered signals were obtained as the features.

For each subject of the three datasets the first 80 trials were considered as the training set and the remaining trials were used as the testing set. To assess the performance of the proposed transfer learning algorithms, three different numbers of training trials were examined for the new subjects; i.e. the first 10 and 20 training trials per class as well as all the training trials were used in order to form the subject-specific training set. Besides, all the available training trials of the other subjects from the same dataset were used for transfer learning. The regularization parameters, λ_s and λ_t, were selected from 21 values which satisfy e^i, where $i \in \{-1, -0.9, ..., 0.9, 1\}$. 5-fold cross-validation was performed for each subject using the available training trials to select the best regularization parameters. In this chapter, 2 s window of EEG data starting 0.5 s after the cue is used to calculate the features for each trial.

The results of the proposed transfer learning algorithms were compared with two baseline algorithms. The first algorithm is the commonly used subject-specific (SS) BCI model where the support vector machine (SVM) classifier is trained independent from other subjects using features extracted from CSP algorithm similar to what suggested in [38, 143]. This algorithm is abbreviated as (SS) in the rest of the chapter. logistic regression classifier was not included as a classifier for the subject-specific baseline algorithm in this chapter as it performed significantly worse than SVM classifier, specially when few subject-specific trials were available for training. The second baseline algorithm is the multi-task learning-based logistic regression classifier (Mt-L) proposed in [123]. This algorithm has been applied on the classifier domain similar to the proposed transfer learning algorithms.

Table 4.1: Classification accuracies calculated using the baseline algorithms (SS, and Mt-L) and the proposed algorithms (LTL, S-wLTL, and Us-wLTL) when only 10 trials per class were available for training from the new subject. The results of all datasets show that the proposed weighted logistic transfer learning algorithms (S-wLTL and Us-wLTL) outperformed the rest.

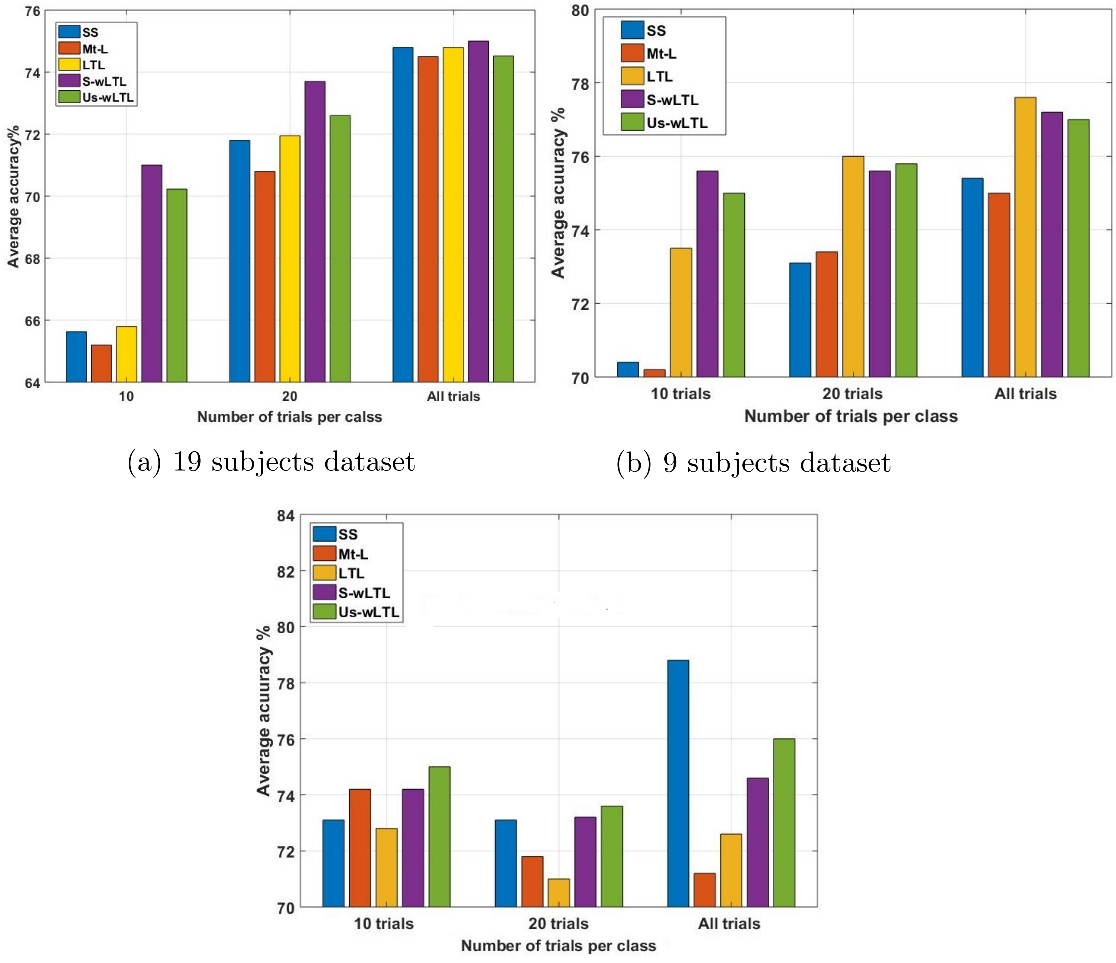
	Dataset 1																			Overall	
Algorithm	sub1	sub2	sub3	sub4	sub5	sub6	sub7	sub8	sub9	sub10	sub11	sub12	sub13	sub14	sub15	sub16	sub17	sub18	sub19	Mean	Std
SS	64	55	55	60	69	72	47	**90**	81	52	48	84	54	76	50	**64**	58	80	88	65.6	14
Mt-L	65	55	55	62	69	68	45	**90**	81	50	**48**	82	54	75	49	58	63	84	86	65.2	14.2
LTL	65	55	55	60	69	72	50	**90**	80	50	**48**	80	54	81	50	58	66	80	84	65.6	13.6
S-wLTL	**67**	**70**	60	**68**	69	**78**	**60**	**90**	**86**	**55**	**48**	79	54	**86**	**74**	58	68	**86**	**93**	**71**	13.3
Us-wLTL	66	57	**61**	65	**72**	**78**	**60**	**90**	82	53	**48**	**88**	**56**	**86**	73	55	**70**	85	**93**	70.3	14.2

	Dataset 2									Overall		Dataset 3					Overall	
Algorithm	sub1	sub2	sub3	sub4	sub5	sub6	sub7	sub8	sub9	Mean	Std	sub1	sub2	sub3	sub4	sub5	Mean	Std
SS	70	51	**93**	57	66	56	73	87	81	70.4	14.5	67.5	93.5	61	**66**	77.5	73.1	17
Mt-L	88	**60**	83	52	50	57	**77**	92	73	70.2	15.9	70	94	59	58	**90**	74.2	17
LTL	83	57	87	58	67	**60**	75	**98**	75	73.6	14.3	69	94	59	57	85	72.8	15
S-wLTL	**90**	55	**93**	**60**	**68**	**60**	73	**98**	**83**	**75.6**	16	69	**95**	**63**	56	88	74.2	15
Us-wLTL	88	53	**93**	**60**	67	**60**	73	**98**	**83**	75	16.2	69	94	**63**	61	88	**75**	16.6

4.4 Results

Table 4.1 presents the classification results of the proposed transfer learning algorithms (LTL, S-wLTL, and Us-wLTL) as well as the baseline algorithms (SS, Mt-L) when the new subjects had only 10 trials per class for training. Based on the results obtained from all the three datasets, the proposed LTL outperformed the results of SS and Mt-L by an average of 1% and 0.8% respectively. Importantly, the proposed S-wLTL algorithm achieved the highest average results with 3.9% and 3.7% higher than SS and Mt-L respectively. On average S-wLTL performed slightly better than Us-wLTL (0.2%). Looking deeper in Table I reveals that in the dataset 1, where data from 18 subjects were used for transfer learning, the proposed S-wLTL outperformed the baseline algorithms SS, and Mt-L by 5.4% and 5.8 % respectively. Whereas, the proposed Us-wLTL outperformed SS and Mt-L by 4.7% and 5.1% respectively. Moreover, S-wLTL and Us-wLTL improved the classification accuracy for 16 out of 19 subjects from this dataset. Interestingly, for sub2, sub7 and sub15 the proposed S-wLTL yielded 15%, 13%, and 24% improvements compared to the corresponding SS results. For the dataset 2, where data from 8 other subjects were used for transfer learning, the proposed weighted transfer learning algorithms, S-wLTL and Us-wLTL, outperformed SS in 7 subjects out of 9 by an average of 5.2% and 4.6%. Compared to Mt-L, S-wLTL and Us-wLTL outperformed in 7 subjects out of 9 by an average of 5.4% and 4.8% respectively. Interestingly, for sub1 and sub8, the proposed S-wLTL yielded 20% and 11% improvements compared to the corresponding SS results. Finally, in the dataset 3, where data from only 4 subjects were available for transfer learning, still the proposed weighted algorithms (S-wLTL and Us-wLTL) improved the results of SS in 4 out of the 5 subjects. Based on the average values, S-wLTL outperformed SS by 1.1% and yielded similar results as Mt-L, whereas Us-wLTL outperformed SS and Mt-L by an average of 1.9% and 0.8% respectively.

Fig. 4.2 presents the classification results of the different algorithms when 10, 20 and all subject-specific training trials per class were available from the new target subject. As shown in Fig. 4.2(a) all the proposed transfer learning algorithms outperformed SS and Mt-L algorithms when 10 and 20 trials per class were available for training whereas, only S-wLTL outperformed the baseline algorithms when all trials were available for training. Specifically, the improvement



(a) 19 subjects dataset

(b) 9 subjects dataset

(c) 5 subjects dataset

Figure 4.2: Comparison between the average classification accuracy calculated using the proposed logistic transfer learning algorithms (LTL, S-wLTL, and Us-wLTL) and the baseline algorithms (SS and Mt-L) when 10, 20, and all trials per class were available for training from the new subjects. From left to right, the sub-figures present the classification results of a) dataset 1, b) dataset 2, and c) dataset 3 respectively. This figure shows that the proposed S-wLTL and Us-wLTL algorithms outperformed the baseline algorithms, particularly when a small number of subject-specific train trials from the target subject, and/or a medium to large number of previously recorded sessions from different subjects were available.

was more pronounced when only 10 subject-specific trials per class were available for training. However, in Fig. 4.2(b) all the proposed transfer learning algorithms outperformed SS and Mt-L algorithms across all the above-mentioned different number of subject-specific training trials. Again, the improvement was more pronounced when only 10 subject-specific trials were available. Interestingly, on average the proposed weighted transfer learning algorithms when only 10 trials per class were available for training outperformed the subject-specific algorithm when all trials were available for training. These outcomes support our aim to reduce the calibration time and at the same time increase the classification accuracy.

Learning from few examples typically leads to an ill-posed optimization problem. That was why we applied transfer learning to overcome this problem when only few trials were available for training. Since dataset 3 contains only data from 5 subjects, transfer learning had been done using only the available data from 4 subjects. As shown in Fig. 4.2(c), despite having such a small pool of data for transfer learning, the proposed transfer learning algorithms still had superior results compared to the baseline algorithms when a few subject-specific trials were available for training. When only 10 training trials per class were available from the new subject, Us-wLTL outperformed baseline algorithms while S-wLTL outperformed only the SS algorithm. Moreover, when 20 trials per class were available for training from the new subject, both of the proposed S-wLTL and Us-wLTL outperformed the baseline algorithms. Increasing the number of subject-specific training trials from the new subject led to a decrease in the improvement, such that the SS algorithm outperformed the proposed transfer learning algorithms when all subject-specific trials (i.e. 80 trials) were available. Thus, with larger amounts of target training data, transfer learning became ineffective.

Concerning statistical significance, the Shapiro-Wilk test was used to make sure that our classification accuracy results were normally distributed. Based on the Shapiro-Wilk test results, we rejected the alternative hypothesis and concluded that our classification results came from a normal distribution and hence ANOVA test could be used to compare the classification accuracy between different algorithms at a different number of trials. A 3 (Number of trials)×5 (Algorithms) repeated measure ANOVA test was performed on the results of each dataset separately followed by post-hoc analyses. For dataset 1 Statistical results revealed that using different algorithms had a main effect on the classification

Table 4.2: Overview of the results when 10 trials per class were available for training from the new subject. Grouping was performed based on SS algorithm classification error rate.

Error Rate [%]	0-10	10-30	>30
SS (Mean)	93.3	80	57.9
Mt-L (Mean)	87	81.7	56.4
S-wLTL(Mean)	94	85.8	62.2
Us-wLTL(Mean)	93.5	86	61.4
$p - value$(SS versus S-wLTL)	0.5	**0.01**	**0.023**
$p - value$(SS versus Us-wLTL)	0.5	**0.003**	**0.038**
$p - value$(Mt-L versus S-wLTL)	0.258	0.069	**0.003**
$p - value$(Mt-L versus Us-wLTL)	0.314	0.056	**0.004**

accuracy with (p=0.001). Based on the post-hoc analysis, S-wLTL (Us-wLTL) significantly outperformed SS and Mt-L with the P values equal to 0.001 and 0.0001 (0.011 and 0.003) respectively. Similarly, for dataset 2, the use of different algorithms also had a main effect on the classification accuracy with (P = 0.035). Based on the post-hoc analysis, S-wLTL (Us-wLTL) significantly outperformed SS and Mt-L with the P values equal to 0.031 and 0.025 (0.035 and 0.04) respectively. Finally, for dataset 3, as expected, there was no significant difference between any of the proposed and the baseline algorithms.

Another comparison was done where results from the three datasets were combined together. A 3 (Number of trials)×5 (Algorithms) repeated measure ANOVA test was conducted. Results showed that using different algorithms significantly affected the classification accuracy with P=0.0001. Post-hoc multiple comparisons revealed that S-wLTL was significantly better than SS and Mt-L with P values of 0.002 and 0.001 respectively. Besides, Us-wLTL was significantly better than SS, and Mt-L with P-values of 0.032 and 0.01 respectively. Moreover, there was no significant difference between Mt-L and SS.

To gain a better insight into the performance of the proposed weighted transfer learning algorithms, the subjects from all datasets were categorized to three groups based on their error rates obtained using the SS algorithm. Table 4.2 presents the results when 10 subject-specific trials per class were available for training. The first four rows of this table compare the average classification accuracies of the different groups obtained by the baseline algorithms (SS, and Mt-L) and the proposed weighted transfer learning algorithms (S-wLTL, and Us-wLTL) respectively. As shown in these four rows, both S-wLTL and Us-wLTL outperformed the baseline algorithms in all the three groups. Subsequently, the last four rows show the statistical paired t-test results between the baseline and the proposed weighted transfer learning algorithms for the different groups. As shown in the fifth and sixth rows, the proposed weighted transfer learning algorithms were more effective when the error rate obtained by the SS algorithm was medium and high. On the other side, the subjects who performed well with the SS algorithm benefited less from applying the proposed transfer learning approach. This makes sense since these subjects already have well-separated features obtained using the standard CSP filters and the subject-specific classifier. Thus, there is not that much room for improvement of the performance for these subjects. In contrast, changing the classifier parameters through the proposed transfer learning approach improved the accuracy of the subjects with poor and medium BCI performance. Finally, the last two rows of Table II show that there was a significant difference between Mt-L and the proposed algorithms for poor subject-specific BCI performance and tends to be significant with medium performance subjects. Again, there was no significant difference between Mt-L and any of the proposed weighted algorithms at the low error rate.

4.5 Discussion

The KL divergence measurement requires estimation of the covariance matrices. The estimation of the covariance matrices could be very inaccurate when only few EEG trials are available [144] as those few trials may not well represent the entire distribution of the data. Despite this limitation, our results showed that even using a few trials from the target subjects the proposed KL-based weights were successful in enhancing the classification accuracy. To further improve the classification results, in the future work, we aim to improve the estimation of the KL divergence in the proposed similarity weight formula by applying robust methods of estimating the covariance matrices (such as [145] where the negative impact of having few trials are mitigated).

Another issue to discuss is the use of the power of 4 for KL in (9). In fact, in (9), power 4 was applied on KL rather than power 1 to increase sparsity between similarity weights and to give larger weights to subjects with similar feature distributions and smaller weights to subjects with dissimilar features. In a number of investigations using different values for power on the subjects from dataset 2, we noticed when using the power of 1, fairly similar weights were obtained for many different subjects. Subsequently, compared to LTL, the proposed Sw-LTL algorithm with KL power of 1 did not yield better results. On the other hand, the S-wLTL classification results were greatly enhanced when KL power was increased to 4 in (9) and then decreased when increasing the power value more than 4. For example, in dataset 2, when only 10 subject-specific trials per class were available, the Sw-LTL algorithm with the KL power of 4 significantly outperformed the Sw-LTL algorithm with the KL power of 1 by an average of 2.6% (p=0.0478). Future work could be extended to estimate the optimum KL power for each subject individually.

Regarding the calibration and computational complexity, the time required for collecting the calibration trials was reduced from around 15 minutes when using the trials of a full session to 2.83 minutes when using only 10 trials per class for training. In order to compare the proposed algorithms and SS from the computational time point of view, we need to note that the proposed algorithms can be divided into two parts. The first part, where the classification parameters of the previous subjects and share priors are calculated using equations (2) to (7), can be done offline without using any data from the target subject. The second part, where the classification parameters of the target subject are calculated using the few available trials of the target subject and the previous subjects shared priors (i.e. $\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_{TLw}$) needs to be done online. This part is the part that should be compared to the SS algorithm in terms of computational time. This computation time was considerably incomparable with the time needed for collecting calibration trials. Using MATLAB 2016b and an Intel Core i5-6500 CPU @ 3.20 GHz, the proposed algorithm required 0.14 sec more time for training the classification model compared to the SS algorithm. Thus, in summary, compared

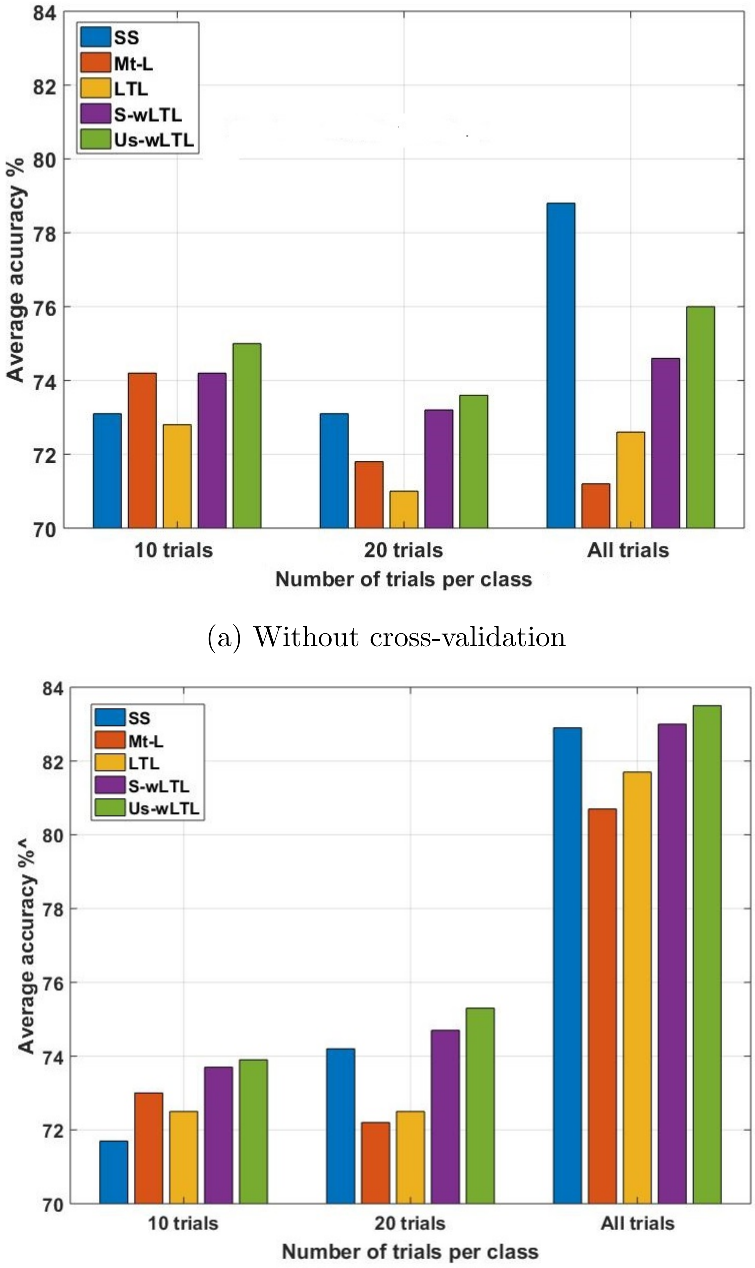
Table 4.3: Comparison of classification accuracy between SS with subject specific trials and SS using subject-specific trials and previously available trials from other subjects without any meaningful transfer learning of subjects from dataset 2

Algorithm	sub1	sub2	sub3	Sub4	Sub5	Sub6	Sub7	Sub8	Sub9	average
SS(10 trials)	70	51	93	57	66	56	73	87	81	70.44
SS (10 trials)+previous trials from other subjects	53	51	95	63	52	52	68	92	53	64.33
SS(20 trials)	85	53	98	66	55	56	73	86	86	73.11
SS (20 trials)+previous trials from other subjects	67	45	73	60	53	53	57	95	83	65.11

to the SS algorithm, the proposed approach remarkably reduced the calibration time, while it just required an added 0.14 S to train the classification model.

Another point to discuss if we directly combined the subject-specific trials and previously available trials from other subjects without any meaningful transfer learning to train the classifier. Table 4.3 shows the results of SS with subject specific trials and SS using subject-specific trials and previously available trials from other subjects at different number of trials. As can be seen, SS algorithm has better results when there are 10 or 20 subject-specific trials per class available for training compared to the classifiers trained using 10 or 20 subject-specific trials per class combined with previously available trials. Since these results are far worse than the SS results, we did not include them in the chapter even though they use the same number of train trials as the proposed transfer learning use. Subject-specific SS classifier results are better than results using subject-specific trials with previous trials from other subjects to train the classifier without transfer learning algorithms.

Finally, we have applied 5 fold cross-validation on the dataset with a small number of subjects (i.e. dataset 3). Fig. 4.3b shows that using cross-validation, the proposed weighted transfer learning algorithms outperformed the baseline algorithms at all three different number of trials which was not the case without applying cross-validation. Moreover, Fig. 4.3b also shows that increasing the number of training trials leads to an increase in the classification accuracies of SS, and the proposed S-wLTL and Us-wLTL algorithms. These findings were not observed in Fig. 4.3a when the cross-validation was not applied. This might be



(a) Without cross-validation

(b) With cross-validation

Figure 4.3: Comparison between the average classification accuracy of subjects from dataset 3 calculated with 5 fold cross-validation and without cross validation

due to the artifact of training or testing trials among other reasons such as user fatigue.

Despite getting better results, we decided not to use cross-validation in calculating our classification results in the manuscript. The main reason behind this decision was to make sure that our results are reflecting a real scenario. In a real scenario, we do not have access to future trials to do cross-validation (worth to say if we had, there was no need to do transfer learning). Our chapter targets the scenarios where we have only a few training trials available from the target subject where some of these trials might be possibly artifact corrupted. Testing trials are coming in future after the training trials, and the training trials are the first trials that the the user performs whatever their quality.

In summary, our results suggested that the proposed S-wLTL and Us-wLTL could improve the classification accuracy particularly when a small subject-specific training data was available. Importantly, when there were sufficient previously recorded subjects/sessions available, the proposed S-wLTL and Us-wLTL algorithms not only reduced the required calibration time but also enhanced the classification accuracy for many subjects. The classification results obtained by S-wLTL and Us-wLTL were on average very similar. Although, the main advantage of Us-wLTL against S-wLTL was that Us-wLTL did not need any labelled data for calculating the weights, it is not easy to define a dominant algorithm that can be used in all situations. However, we can suggest that when only 10 trials per class are available from the new BCI user and enough data from previous subjects/sessions are available for transfer learning, e.g. as in dataset 1 and dataset 2, the proposed S-wLTL algorithm is preferred to be used. Contradictory, when only 10 trials per class are available from the new BCI user and few data from previous subjects/sessions are available for transfer learning, e.g. as in dataset 3, the proposed Us-wLTL algorithm is preferred to be used.

4.6 Conclusion

This chapter proposed a novel weighted transfer learning approach on classification domain to improve MI-based BCI systems. Our results suggested that applying the proposed weighted transfer learning algorithms could lead to reducing the calibration time to 10 trials per class with significantly less sacrifice in the average accuracy of the MI-BCI systems. The results obtained showed that the proposed weighted algorithms significantly outperformed the subject-specific BCI algorithms and the multi-task learning algorithm.

The proposed transfer learning approach is not limited to the logistic regression classifier. It can be applied on any classifier with a mathematically defined objective function. Moreover, in this chapter similarity weights were calculated using KL-divergence as a similarity measurement. It is good to note that in the future other similarity measures could be used and their performance could be compared to what we proposed.

Interestingly, the proposed weighted transfer learning algorithms yielded a remarkable increase in the classification accuracy for most of the subjects that initially performed BCI with poor or medium accuracy. However, the observed improvement for a few subjects with initially low BCI performance was not pronounced. It was shown that changing the parameters of classifiers for these subjects was not effective since their feature spaces for different classes were not separable. These findings suggest that to increase the accuracy of these subjects with poor subject-specific BCI, transfer learning approaches should be applied in a different domain before the classification domain. Thus, in the next chapter we will investigate the transfer learning in raw EEG and feature domains to address these challenges.

Chapter 5

Dynamic Time Warping-based Transfer Learning for Improving Common Spatial Patterns in Brain-computer Interface

5.1 Introduction

Common spatial patterns (CSP) is a popular algorithm for motor imagery EEG feature extraction in the context of brain-computer interfaces (BCIs). Despite popularity and effectiveness of CSP, most of the CSP-based BCI applications are still limited to the laboratory [33, 35]. This is due to the fact that CSP requires estimation of the covariance matrices which could be very inaccurate when only a few EEG trials are available for training leading to CSP overfitting [107, 144].

For CSP to be usable in practice, it must be optimally robust across sessions and subjects, and with less possible calibration times. These challenges could be considered at different levels, e.g., at the neuroscience level, or at the human level, or at the signal processing level. Regarding EEG signal processing, developing reliable CSP-based algorithms that reduce calibration time without sacrificing the classification accuracy is highly desirable in MI-based BCI research [13, 70]. One potential approach to reduce the calibration time is transfer learning [38].

To the best of our knowledge, none of previous studies considered the temporal variations between EEG trials of a new subject and those of previous subjects to reduce between-subjects non-stationarity during transfer learning. Moreover, most of the proposed algorithms in the feature domain require calculating multiple regularization parameters which is computationally expensive.

To deal with the previously mentioned, in the previous chapter, problem of the subjects with initially low BCI performance who didn't gain much benefit from improving the classification parameters as their feature spaces for different classes were not separable. This chapter proposes a novel transfer learning framework in raw EEG and feature domains, called DTW-based regularized CSP (DTW-RCSP). At first, in the raw EEG domain, we transform previous subjects' trials to be more similar to the target subject's few training trials using a novel alignment method in time domain based on DTW, and hence use these aligned trials to form the transferred covariance matrix. Then, in the feature domain, we propose a novel regularization between the subject-specific and the transferred covariance matrices to improve the CSP covariance matrix estimation. The output of our proposed DTW-RCSP framework is a new regularized CSP matrix which is a combination of the subject-specific covariance matrix and the transferred covariance matrix from other subjects. Finally, to tackle the problem of regularization parameter determination when very few training trials are available, we propose an online method based on the upcoming first few labelled testing trials, where some predefined regularization parameters are evaluated based on the confidence scores of the trained classifier.

The proposed DTW-RCSP framework is evaluated across different number of subject-specific training trials using three datasets with small, moderate, and large number of subjects. The performance of the proposed DTW-RCSP is compared against two state of the art algorithms, standard CSP and Composite CSP (CCSP) [105]. Results show that DTW-RCSP significantly outperformed the baseline algorithms in various testing scenarios, particularly, when only a few trials are available for training. Impressively, our results show that successful BCI interactions could be achieved with a calibration session as small as only one trial per class.

5.2 Methodology

This section explains the proposed framework and the baseline algorithms.

5.2.1 Dynamic time warping-based transfer learning regularized CSP framework (DTW-RCSP)

This subsection presents our proposed transfer learning framework (DTW-RCSP) to improve the CSP features of EEG signals, when few trials from the target subject and a group of previously recorded trials from other subjects are available.

In our proposed DTW-RCSP framework, the previously recorded EEG trials from other subjects and sessions are pooled together as one single session s, and referred to as the source domain. Subsequently, the source domain is presented as $d_s = (\mathbf{X}_s^i, y_s^i)_{i=1}^N$, where $\mathbf{X}_s^i$ and $y_s^i \in \{-1, 1\}$ respectively denote the EEG instance matrix and the class label of the i^{th} trial, and N refers to the number of the trials. In each trial $\mathbf{X}_s^i \subset \mathbb{R}^{h \times V}$, h is the number of EEG samples and V is the number of channels. Similarly, the set of labelled trials of the target subject t is denoted as $d_t = (\mathbf{X}_t^i, y_t^i)_{i=1}^M$, where M is the number of the available subject-specific trials.

Dynamic Time Warping-based Transfer Learning Regularized CSP Framework (DTW-RCSP)

To improve CSP covariance matrix estimation when few trials are available for training, regularization based transfer learning techniques could be used. Regularized CSP works by specifying a trade-off between the estimates obtained using few target subject-specific trials and informative estimates obtained using previously recorded trials from other subjects/sessions [33]. In our proposed DTW-RCSP framework, the average CSP covariance matrix $\mathbf{\Sigma}_{\mathrm{TLR_c}}$ for each class c is calculated as follows:

$$\mathbf{\Sigma}_{\mathrm{TLR_c}} = (1 - r)\mathbf{\Sigma}_{\mathrm{SS_c}} + r\mathbf{\Sigma}_{\mathrm{DTW_c}}, \tag{5.1}$$

where r is the regularization parameter $(0 \leqslant r \leqslant 1)$. $\mathbf{\Sigma}_{\mathrm{DTW_c}}$ is the proposed DTW-based transferred average covariance matrix of the aligned trials of class c from other subjects which will be explained in 5.2.1. $\mathbf{\Sigma}_{\mathrm{SS_c}}$ is the average covariance matrix of the few subject-specific trials of class c from the target subject. $\mathbf{\Sigma}_{\mathrm{SS_c}}$ is calculated as

$$\mathbf{\Sigma}_{\mathrm{SS_c}} = \frac{1}{m_c} \sum_{i=1}^{m_c} \frac{\mathbf{X}_t^{i\top} \mathbf{X}_t^i}{\mathbf{tr}(\mathbf{X}_t^{i\top} \mathbf{X}_t^i)}, \tag{5.2}$$

where m_c is the number of trials per class c, $\top$ is the matrix transpose function, and $\mathbf{tr}$ is the trace function.

The regularization parameter r shrinks the subject-specific covariance matrix towards the DTW-based transferred covariance matrix to neutralize the possible estimation bias due to the availability of few training trials from the target subject. In fact, $\mathbf{\Sigma}_{\mathrm{DTW_c}}$ represents the information on how the covariance matrix for the considered intellectual condition should typically be. Finally, the DTW-RCSP filters $\mathbf{W}_{\textsc{dtw-rcsp}}$, replacing W mentioned in (5.2.1), are calculated by maximizing the following objective function using joint diagonalization [125]:

$$\mathbf{W}_{\textsc{dtw-rcsp}} = \mathop{\mathbf{arg\,max}}_{\mathbf{W}} \frac{\mathbf{W}\ \mathbf{\Sigma}_{\mathrm{TLR}_1}\mathbf{W}^{\top}}{\mathbf{W}(\mathbf{\Sigma}_{\mathrm{TLR}_1}+\mathbf{\Sigma}_{\mathrm{TLR}_2})\mathbf{W}^{\top}}. \tag{5.3}$$

From (5.1), the classical CSP can be considered as a special case of DTW-RCSP framework, when $r=0$.

Estimation of the Dynamic Time Warping Transferred Covariance Matrix

DTW has been initially proposed to solve the time deformation problem between two time series in speech recognition problems in a non-linear fashion. DTW finds an optimal alignment between two given sequences under certain restrictions to compensate the timing differences between them [146]. Subsequently, DTW has been applied to other problems such as object recognition, motion analysis, classification and clustering of time domain signals including EEG, and ECG [147,148]. For EEG, DTW is typically used as a measure of dissimilarity between two EEG segments after being optimally aligned. In our published paper, DTW has been used to reduce subject-specific temporal variations between two EEG segments [149].

In this thesis, DTW is used for the purpose of transfer learning. Unlike the previous EEG-based studies, the goal is to align a collection of EEG trials from other subjects or sessions to the average of the few available trials from the new target subject. Thus, to calculate $\mathbf{\Sigma}_{\mathrm{DTW}_c}$, the DTW-based transferred average covariance matrix, the following steps are taken.

First the average of the available few trials of the target subject from class c is computed as follows:

$$\bar{\mathbf{X}}_{t_c} = (1/m_c)\sum_{i=1}^{m_c}\mathbf{X}_t^i, \tag{5.4}$$

where $\bar{\mathbf{X}}_{t_c}$ and m_c respectively refer to the average and the total number of the target trials of class c.

Next, each trial from the source session gets aligned to the average of the few target trials from the same class, $\bar{\mathbf{X}}_{t_c}$, using DTW. To align $\mathbf{X}_s^i \subset \mathbb{R}^{h\times V}$ to

$\bar{\mathbf{X}}_{t_c} \subset \mathbb{R}^{h\times V}$, we construct a distance matrix $\mathbf{D}^{h\times h}$, where $\mathbf{D}(a,b)$ is the Euclidean distance between the EEG signals of two time instances of a and b from $\mathbf{X}_s^i$ and $\bar{\mathbf{X}}_{t_c}$ respectively,

$$\mathbf{D}(a,b) = \sqrt{\sum_{v=1}^{V} (\mathbf{X}_s^i(a,v) - \bar{\mathbf{X}}_{t_c}(b,v))^2}. \tag{5.5}$$

Thereafter, the elements of $\mathbf{X}_s^i$ and $\bar{\mathbf{X}}_{\mathbf{t_c}}$ are mapped through the matrix D by finding an optimum warping path, whereby the cumulative distance between the two above-mentioned EEG trials is minimized. Generally, a warping path, P, defines a mapping between $\mathbf{X}_s^i$ and $\bar{\mathbf{X}}_{t_c}$, and its elements are presented as

$$\mathbf{P} = [p(1),..,p(k),...,p(K)] \quad h \le K < 2h-1 \tag{5.6}$$

where $p(k)$=$\mathbf{D}(a_k, b_k)$. a_k and b_k belong to $\{1,2,...,h\}$, and remap the time indices of $\mathbf{X}_s^i$ and $\bar{\mathbf{X}}_{t_c}$ respectively. A warping path requires to be subject to the following constraints:
1- Boundary conditions: $p(1) = \mathbf{D}(1,1)$ and $p(K) = \mathbf{D}(h,h)$. In other words, $a_1 = b_1 = 1$ and $a_K = b_K = h$.
2- Continuity and monotonicity: $0 \le a_k - a_{k-1} \le 1$ and $0 \le b_k - b_{k-1} \le 1$.

There are exponentially many warping paths that satisfy the above-mentioned conditions. However, we are interested in the optimum warping path, $\mathbf{P}^*$, which has the shortest non-linear alignment between $\mathbf{X}_s^i$ and $\bar{\mathbf{X}}_{t_c}$, as

$$\mathbf{P}^* = \underset{\mathbf{P}}{\mathbf{arg\ min}} \ \left(\frac{1}{K}\sqrt{\sum_{k=1}^{K} p(k)}\right). \tag{5.7}$$

where $\mathbf{p}$ has defined in 5.6. To reduce the computational time, $\mathbf{P}^*$ can be found using dynamic programming to evaluate the following recurrence [148], where the cumulative distance $\boldsymbol{\gamma}(a,b)$ is defined as the distance between two time instances a and b from $\mathbf{X}_s^i$ and $\bar{\mathbf{X}}_{t_c}$, $\mathbf{D}(a,b)$, and the minimum of the cumulative distances of the adjacent elements:

$$\boldsymbol{\gamma}(a,b) = \mathbf{D}(a,b) + min[\boldsymbol{\gamma}(a-1,b-1), \boldsymbol{\gamma}(a-1,b), \boldsymbol{\gamma}(a,b-1)] \tag{5.8}$$

Given $\mathbf{P}^*$, $\mathbf{X}_s^i$ is aligned to $\bar{\mathbf{X}}_{t_c}$ as:

$$\mathbf{X}^i_{s_{aligned}} = \begin{bmatrix} \mathbf{X}^i_s(a_1^*,1) & \mathbf{X}^i_s(a_1^*,2) & \cdots & \mathbf{X}^i_s(a_1^*,V) \\ \mathbf{X}^i_s(a_2^*,1) & \mathbf{X}^i_s(a_2^*,2) & \cdots & \mathbf{X}^i_s(a_2^*,V) \\ \vdots & & \ddots & \vdots \\ \mathbf{X}^i_s(a_K^*,1) & \mathbf{X}^i_s(a_K^*,2) & \cdots & \mathbf{X}^i_s(a_K^*,V) \end{bmatrix} \tag{5.9}$$

where $[a_1^*, a_2^*, ..., a_K^*]$ are the time indices of $\mathbf{X}_s^i$ forming the minimum warping path $\mathbf{P}^*$. These time instances are the instances that will make $\mathbf{X}_s^i$ to be as much similar to $\bar{\mathbf{X}}_{\mathbf{t_c}}$ as possible given the above constraints. Subsequently the covariance matrix of $\mathbf{X}^i_{s_{aligned}}$ is calculated as follows:

$$\boldsymbol{\Sigma}^i_{\mathrm{s_{aligned}}} = \frac{(\mathbf{X}^i_{s_{aligned}})^\top \mathbf{X}^i_{s_{aligned}}}{\mathbf{tr}((\mathbf{X}^i_{s_{aligned}})^\top \mathbf{X}^i_{s_{aligned}})}. \tag{5.10}$$

Finally, the proposed DTW-based transferred average covariance matrix of the aligned trials from previous subjects/sessions for each class c is computed as

$$\boldsymbol{\Sigma}_{\mathrm{DTW_c}} = (1/n_c) \sum_{i=1}^{n_c} \boldsymbol{\Sigma}^i_{\mathrm{s_{aligned}}}, \tag{5.11}$$

where n_c is the overall available trials of class c from other subjects/sessions.

Regularization Parameter Selection

Typically, regularization parameter is selected from a set of predefined values by applying cross-validation on the training data [150]. However, cross-validation becomes ineffective and in some cases impossible when we have very few training trials available from the target subject. Here, we address the above-mentioned challenge by selecting the best regularization value using the classifier scores (i.e confidence scores) rather than the accuracy.

We propose using the classification scores to select the best regularization value in two different ways, namely referred to as offline and online. The offline method is applicable if we have sufficient training trials available from the new target subject. The offline method applies cross-validation on the training trials and selects the regularization value that yields the highest summation of classification scores multiplied by the true class labels of the corresponding evaluation target trials over the 10-fold validations. Please see our algorithm in Fig. 5.1 for more details.

In the online method, the few upcoming testing trials with known labels will be used for selecting regularization value. Thus, among a set of predefined values, the selected regularization value is the one which yields the highest summation of the classification scores multiplied by the true classification labels of the upcoming few testing trials. Fig. 5.2 provides more details on the proposed online regularization parameter selection method. The proposed online method can be used for any available number of training trials, while the proposed offline method is not applicable if less than K training trials are available from the new target subject where K is the number of cross-validation folds.

5.3 Experiments

5.3.1 Data description

In order to evaluate the proposed transfer learning framework, three datasets with 5, 9 and 17 subjects were used.
1) Dataset IVa from BCI Competition III (small dataset) [142]: This dataset includes EEG signals from five subjects who performed right hand and foot motor imagery. EEG was recorded using 118 electrodes. In this thesis, we use only data from the 22 electrodes similar to those used in the next dataset to reduce the complexity and computational time of calculating DTW. In total, there are 280 trials per subject all recorded on the same day.

2) Dataset 2a from BCI Competition IV (medium dataset) [130]: This dataset consists of EEG data recorded from 9 subjects using 22 electrodes. During the recording sessions, the subjects were instructed to perform one of the four following motor imagery tasks: left hand, right hand, foot or tongue. Two sessions on different days were recorded for each subject with a total of 288 trials per session. In this thesis, only data from right and left-hand motor imagery were used (i.e. 144 trials). Moreover, only data recorded from the second day were used due to the practical assumption that the training and the testing data of a new subject are recorded at the same day.
3) Dataset from [140] (large dataset): EEG was collected from 19 healthy subjects using 27 channels. For each subject, EEG data were collected without feedback

Algorithm 2: Offline method

```
Input: $\mathbf{\Sigma}_{\mathrm{DTW_c}}$, $\mathbf{\Sigma}_{\mathrm{SS_c}}$ for each class $c$, $A$ predefined values of $r$, $K$
       cross-validation folds, and $n_{eva}$ evaluation trials from the target
       subject
Output: Regularization parameter $r^*$
for $r = r_1$ to $r_A$ do
    for $k = 1 : K$ do
        for $c$ do
            calculate $\mathbf{\Sigma}_{\mathrm{TLR_c}}$ using (1)
        calculate the corresponding DTW-RCSP features using (3)
        train the classifier
        for $tr = 1 : n_{eva}$ do
            calculate the classifier score $CS$ for each $tr$
            $score_{tr}$= $CS_{tr} * label_{tr}$
        $score_k$=$\sum_{t_r=1}^{n_{eva}} score_{tr}$
    $score_r$=$\sum_{k=1}^{K} score_k$
Score* = arg max $score_r$
Return: $r^*$ assigned to the highest Score*
```

Figure 5.1: The proposed offline method to select the regularization parameter based on the confidence scores of the classifier on the training trials from the target subject

Algorithm 3: Online method

```
Input: $\mathbf{\Sigma}_{\mathrm{DTW_c}}$, $\mathbf{\Sigma}_{\mathrm{SS_c}}$ for each class $c$, $A$ predefined values of $r$, and $T$
       upcoming labelled test trials from the target subject
Output: Regularization parameter $r^*$
for $r = r_1$ to $r_A$ do
    for $c$ do
        calculate$\mathbf{\Sigma}_{\mathrm{TLR_c}}$using (1)
    calculate the corresponding DTW-RCSP features using (3)
    train the classifier
    for $tr = 1 : T$ do
        calculate the classifier score $CS$ for each $tr$
        $score_{tr}$= $CS_{tr} * label_{tr}$
    $score_r$=$\sum_{t=1}^{T} score_{tr}$
Score$^*$= arg max $score_r$
Return: $r^*$ assigned to the highest Score$^*$
```

Figure 5.2: The proposed online method to select the regularization parameter based on the classifier confidence scores of the upcoming few labelled testing trials

in two sessions conducted on separate days. Each session consisted of two runs of EEG recording where the subjects were instructed to perform a chosen hand MI versus background rest condition. Each run comprised of 40 trials of MI and 40 trials of background rest condition. Thus, in total, there were 160 trials per subject recorded without feedback. In this chapter, we used data from subjects who performed right hand motor imagery (17 subjects). We did that to ensure the data used for transfer learning were neurologically relevant. Moreover, we used only data recorded in the first session.

5.3.2 Data processing

A bandpass filter from 8 to 30 Hz was used for EEG data filtering, since the range of frequencies that are mainly involved in performing motor imagery are included in this single frequency band. After that, the spatially filtered signals were obtained using the first and the last three spatial filters of CSP/CCSP/DTW-RCSP as recommended in [102]. Thereafter, the normalized log band power of the spatially filtered signals were obtained as the features. Finally, Linear support vector machine (SVM) was used as the classifier.

For each subject, the investigated trials were divided into 3 sets, namely training, validation , and testing. The testing set consisted of the last 190 trials for the small dataset, the last 50 trials for the medium dataset, and the last 70 trials for the large dataset. For all datasets, the validation trials are the 10 trials immediately before testing trials, and the training set consisted of the remaining trials. Validation trials will be used in the proposed online method for selecting the regularization parameter. To assess the proposed DTW-RCSP framework's performance, different scenarios have been considered when different numbers of training trials from new target subjects were available. Besides, all the available training trials of the other subjects from the same dataset were used for DTW-based transfer learning covariance matrix estimation. The optimum regularization parameter was selected from the predefined set of $r \in \{0, 0.1, \cdots, 1\}$.

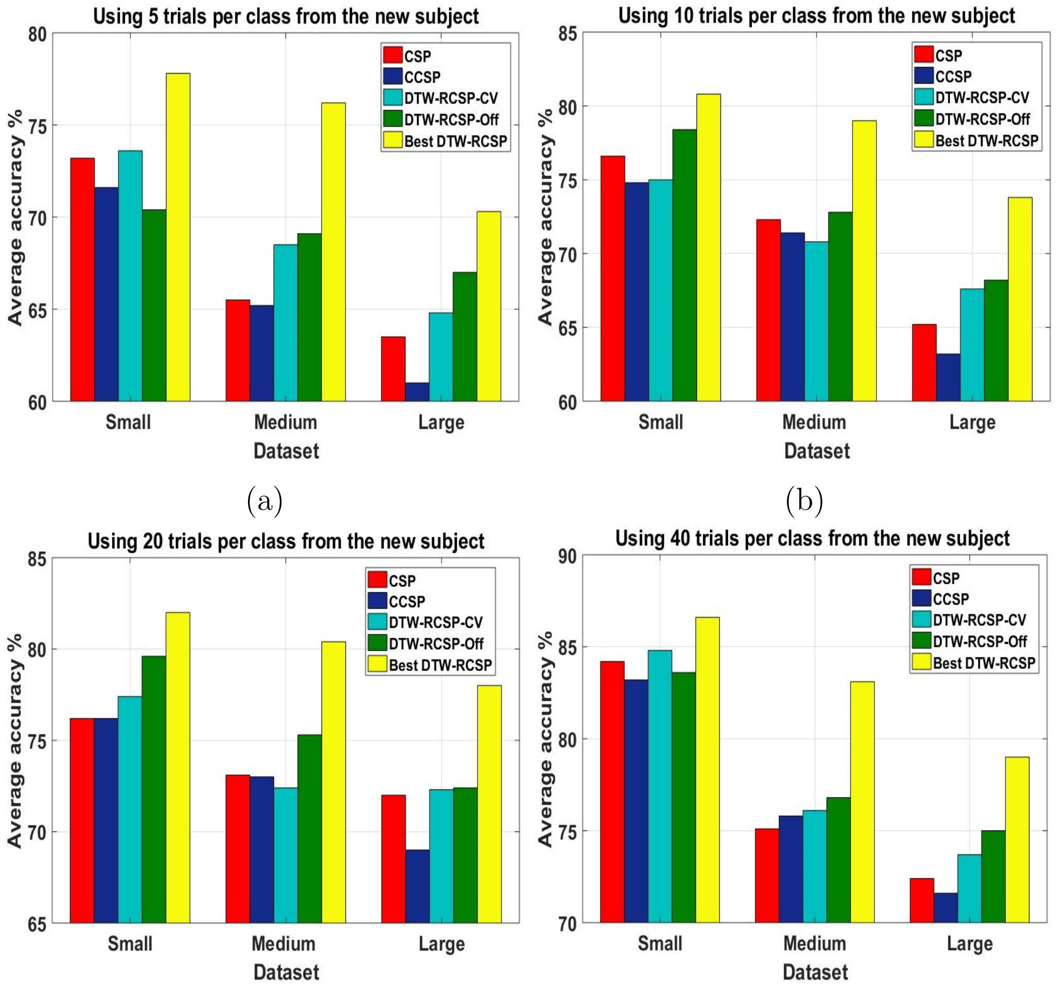
The three proposed transfer learning-based regularized CSP algorithms (namely DTW-RCSP-CV, DTW-RCSP-Off, and DTW-RCSP-On) were evaluated. These algorithms differ in terms of how the regularization parameter is selected. For DTW-RCSP-CV, the optimum regularization parameter is selected using 10 fold cross-validation on training data of the target subject based on the classification accuracy. For DTW-RCSP-Off and DTW-RCSP-On, the regularization parameter is selected using the proposed offline and online methods respectively. The results of the proposed algorithms were compared with the results of two baseline algorithms, i.e. the commonly used subject-specific CSP algorithm, and CCSP (the first method proposed in [105]). The regularization parameter in CCSP is selected using cross-validation for the available training trials from the target user. In fact, if DTW alignment is omitted from the proposed DTW-RCSP-CV, it gets identical with CCSP.

5.4 Results

The first part of this section presents the results when 5 or more trials per class were available from the target subject. Thus 10-fold cross-validation and our proposed offline method could be used to select the regularization parameter using the available subject-specific training trials.

Fig. 5.3 compares the average classification accuracies of the baseline algorithms (CSP, and CCSP) with the results of the proposed DTW-RCSP-CV, DTW-RCSP-Off and Best-DTW-RCSP. Best-DTW-RCSP represents the classification accuracy if the best regularization parameter yielding the highest test accuracy could have been selected from $\{0, 0.1, \ldots, 1\}$. As shown in Fig. 5.3, for all datasets the proposed DTW-RCSP-Off algorithm outperformed the CSP and CCSP algorithms using most number of training trials. Interestingly, DTW-RCSP-Off was more successful than DTW-RCSP-CV in selecting regularization parameters yielding a higher average test classification accuracy.

Statistical paired t-tests revealed that for the large dataset using DTW-RCSP-Off was significantly better than CSP when 10 trials were available for training from the target subject ($P = 0.04$) and tended to be significantly better when 5 trials were available ($P = 0.09$). Besides, DTW-RCSP-Off was significantly better than CCSP when 5 trials were available with P value equal to 0.015. Moreover, DTW-RCSP-CV was significantly better than CCSP when 10 and 20 trials were available with P values equal to 0.04 and 0.017 respectively. These statistical results suggested that our proposed transfer learning algorithms performed significantly better than the baseline algorithms if a large number of previously recorded data from other subjects were available. Nevertheless, comparing the



(a)

(b)

(c)

(d)

Figure 5.3: Comparison of the average classification results between the baseline algorithms (CSP, and CCSP), the proposed DTW-RCSP-CV, and DTW-RCSP-Off algorithms, and the DTW-RCSP results if the best regularization parameter yielding the highest test classification accuracy was selected (i.e. best DTW-RCSP). The classification results were calculated for different number of trials available for training from the new target subject.

Best-DTW-RCSP results with those obtained by DTW-RCSP-CV and DTW-RCSP-Off revealed that if better regularization parameters could have been selected, the proposed DTW-RCSP algorithm could yield much higher significant improvements.

Although the proposed DTW-RCSP-Off algorithm improved the average classification accuracy, the Best-DTW-RCSP results showed that there was still room for improvement. Moreover, DTW-RCSP-Off with 10-fold cross validation for selecting the regularization parameter could not be viable if the number of the available training trials from the target subject is less than 5 trials per class. Therefore, in such cases our proposed DTW-RCSP-On could be used where the first few testing trials (referred to as the validation set in this study) were employed to select the regularization parameter. Apart from the benefits mentioned above, using the first few testing trials for selecting the regularization parameter could possibly reduce the negative impact of non-stationarity between the training and testing trials.

Fig. 5.4 shows the results of DTW-RCSP-On. The average classification accuracy across all subjects from each dataset was reported when the subject-specific training trials were as few as 1, 2, and 5 trials per class. The proposed DTW-RCSP-On, when different number of testing trials were used to select the regularization parameter, was compared to CSP and DTW-RCSP with $(r = 1)$ (i.e. only Σ_{DTW} was used for obtaining features). It is shown that using the proposed DTW-RCSP-On algorithm greatly improved the average classification accuracy. Impressively, when only 1 subject-specific trial per class was available for training, the proposed DTW-RCSP-On outperformed CSP by an average of 5.2%, 5.8%, 7.2%, 8.6%, and 9% for dataset1, 3.7%, 5.2%, 6.4%, 8.1%, and 8.7% for dataset 2, and 8.1%, 2.9%, 4.9%, 3.7%, and 4.2% for dataset 3 when using 2, 4, 6, 8, and 10 validation trials for selecting the regularization parameter respectively. Moreover, in case of having only either 1 or 2 subject-specific trials per class, the classification results of DTW-RCSP with $(r=1)$ outperformed CSP (i.e. only data from other subjects after DTW alignments were used to obtain features).

Fig. 5.5 provides more insight into the results of the proposed DTW-RCSP-On algorithm compared to CSP. As shown in Fig. 5.5, although for a few cases the use of DTW-RCSP-On led to small deterioration in the accuracy, for the

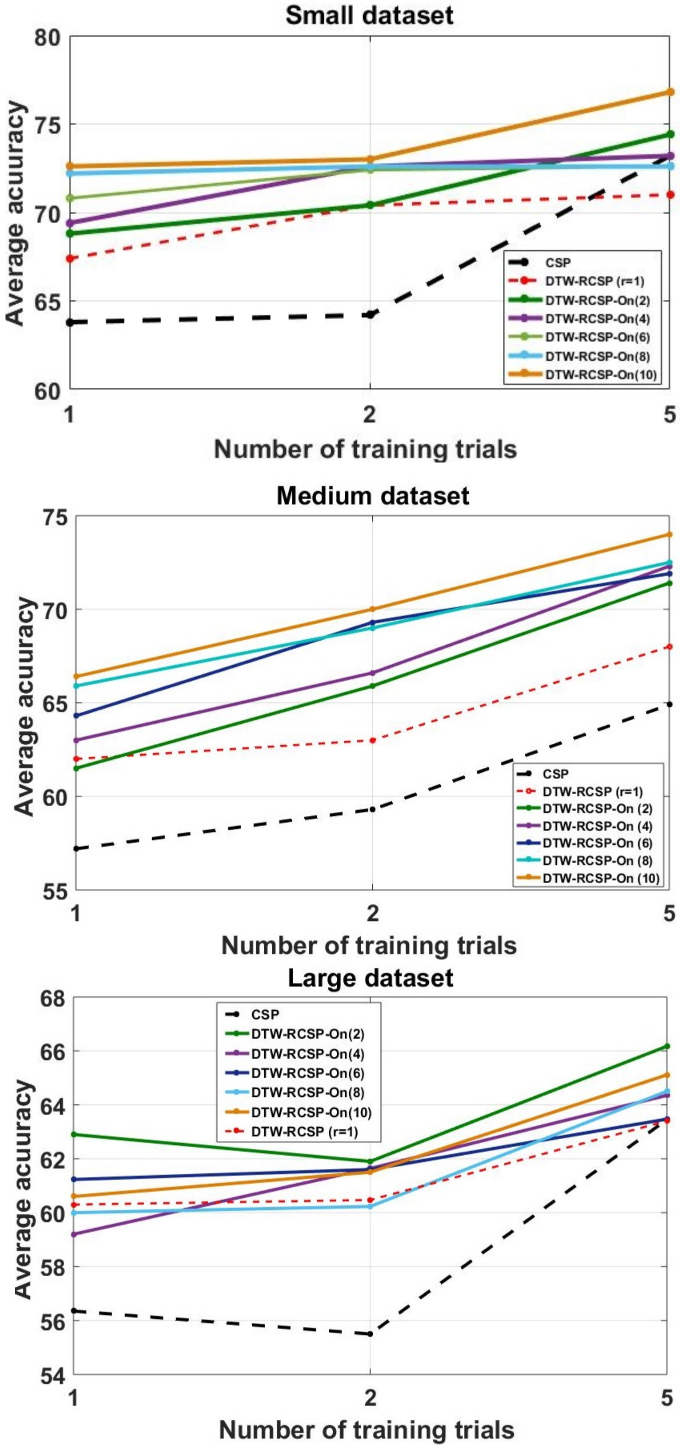
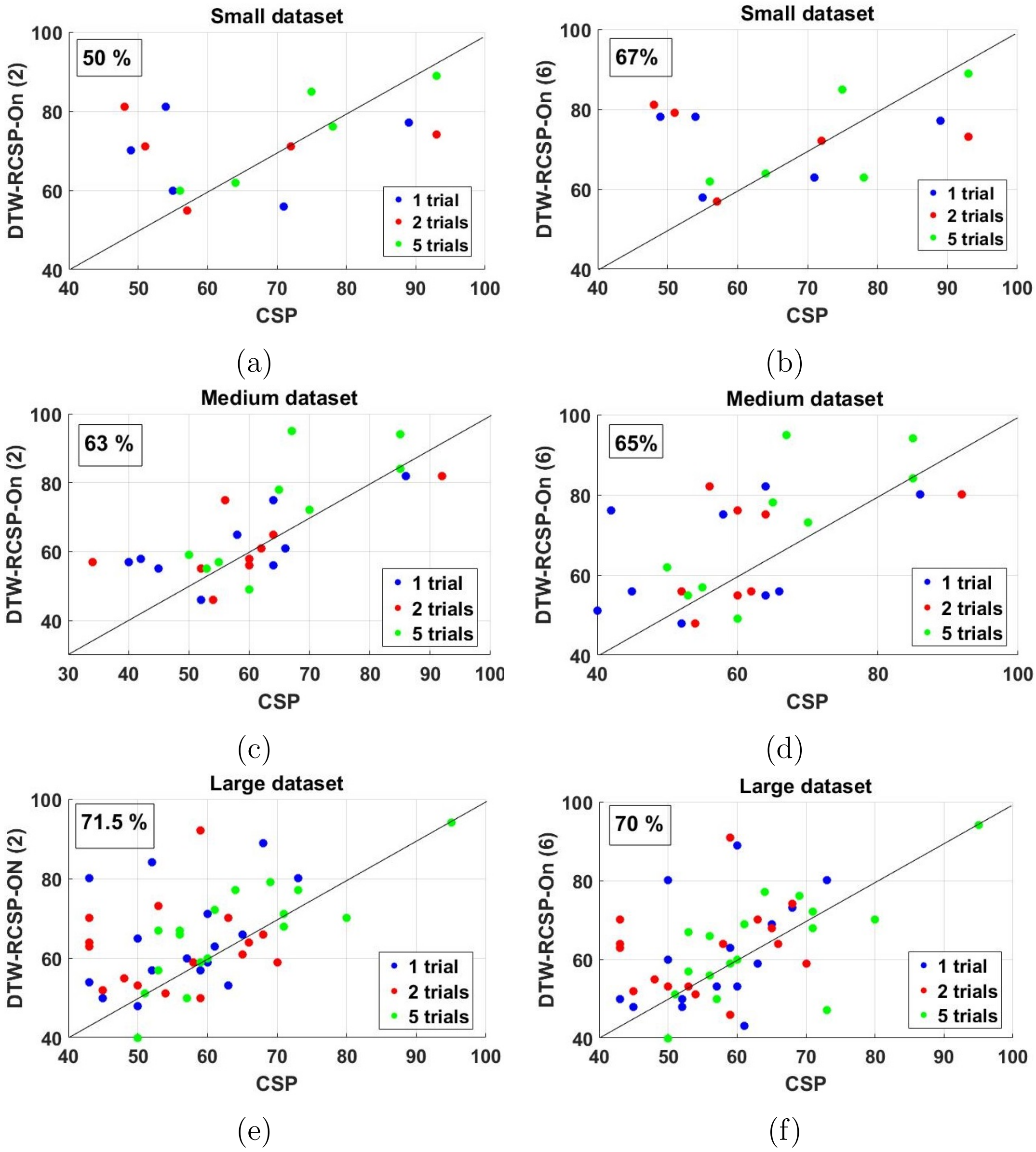


Figure 5.4: Comparing average classification results of the proposed DTW-RCSP-On using 2, 4, 6, 8, and 10 validation trials to select the regularization parameter, with those of DTW-RCSP with (r=1) and CSP when 1,2, and 5 trials per class were available for training from the new target subject.

majority of the subjects considerable improvements had been achieved. Indeed, in many cases the improvement was as large as 20% to 35%.

Concerning statistical significance, A 6 (Number of trials= 1, 2, 5, 10, 20, and 40 trials per class)× 6 (Algorithms= CSP, DTW-RCSP-On (2,4,6,8,10)) repeated measure ANOVA test was performed on the results of both datasets followed by post-hoc analyses. For the large dataset, statistical results revealed that using different algorithms had a main effect on the classification accuracy ($P = 0.003$). Based on the post-hoc analysis, DTW-RCSP-On with different number of validation trials significantly outperformed CSP with P values equal to 0.001, 0.017, 0.046, 0.035, and 0.027 respectively for 2, 4, 6, 8, and 10 validation trials used to select the regularization parameter. Interestingly, using the proposed DTW-RCSP-On(2) was significantly better than using any other number of testing trials (i.e. P values of 0.038, 0.05, 0.025, and 0.036 for 4, 6, 8, and 10 validation trials). Similarly, for the medium dataset, the statistical results revealed that using different algorithms had a main effect on the classification accuracy ($P = 0.012$). Based on the post-hoc analysis, DTW-RCSP-On with 2, 4, 6, 8, and 10 validation trials to select the regularization parameter significantly outperformed CSP with P values equal to 0.043, 0.043, 0.028, 0.022, and 0.023 respectively. However, using DTW-RCSP-On with 6, 8, or 10 testing trials to select the regularization parameter were not significantly different. Finally, for the small dataset, the statistical results revealed that using different number of trials had a main effect on the classification accuracy ($P = 0.005$). Based on the post-hoc analysis, DTW-RCSP-On with 10 validation trials was significantly better other DTW-RCSP-On with 2, 4, and 6 validation trials with P values equal to $0.016, 0.046$, and 0.046 respectively. Besides, using 40 trials per class for training were significantly better that using any other number of training trials with P values equal to $0.021, 0.017, 0.023, 0.024$, and 0.044 respectively. This outcome showed that when few subjects were available for transfer learning more training trials were required form the new subject to better improve the classification accuracy.

Another comparison was held to make sure that adding the validation trials used by DTW-RCSP-On for selecting the regularization parameter to the training trials of CSP would not achieve the same improvement as DTW-RCSP-On. Fig. 5.6 compares the average classification results of the proposed DTW-RCSP-On algorithm with the results of the CSP algorithm where the CSP was trained

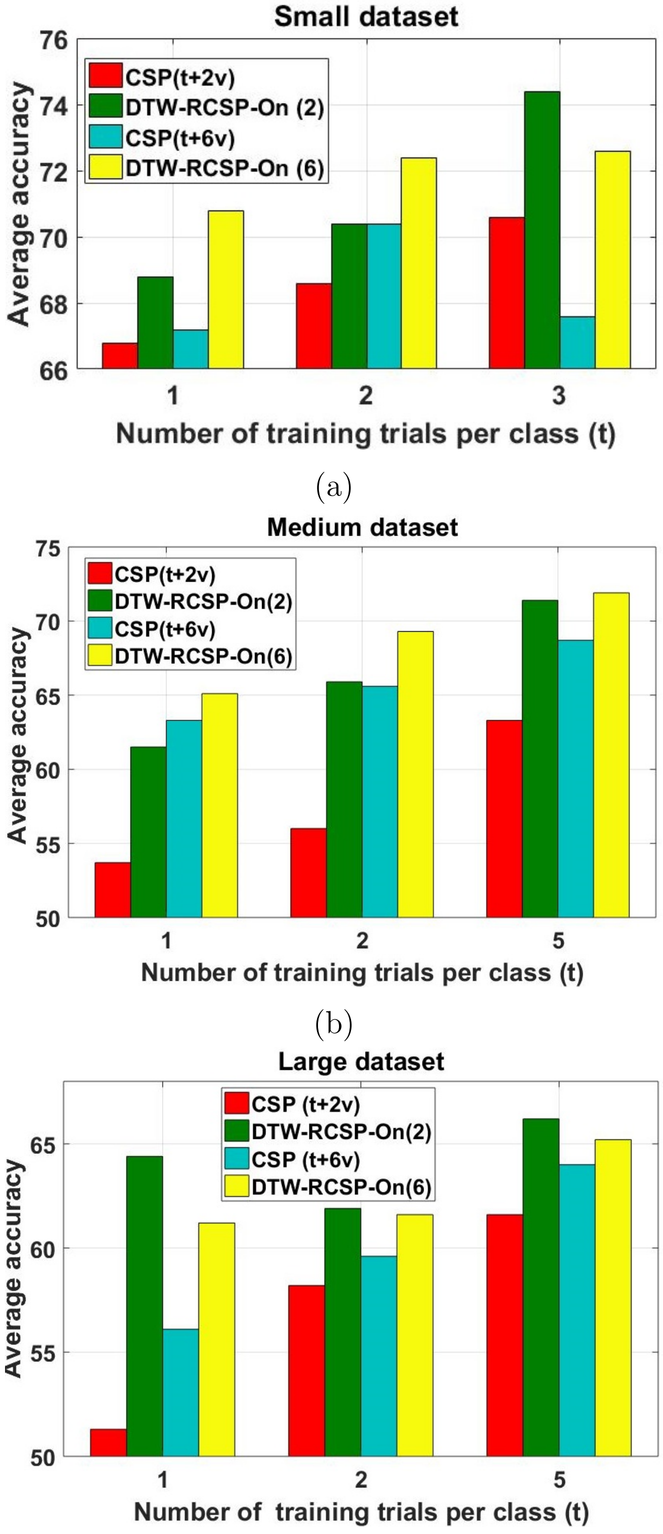


(a) (b)

(c) (d)

(e) (f)

Figure 5.5: Classification accuracy comparison for each individual subject in both datasets when 1, 2, and 5 trials were available for training from the new target subject. (a) CSP versus DTW-RCSP-On(2) for small dataset. (b) CSP versus DTW-RCSP-On(6) for small dataset. (c) CSP versus DTW-RCSP-On(2) for medium dataset. (d) CSP versus DTW-RCSP-On(6) for medium dataset. (e) CSP versus DTW-RCSP-On(2) for large dataset. (f) CSP versus DTW-RCSP-On(6) for large dataset. ”v” in DTW-RCSP-On(v) refers to the number of validation trials used for selecting the regularization parameter.



(a)

(b)

(c)

Figure 5.6: Comparison between DTW-RCSP-On(v) versus CSP trained with the available training trials(t) plus the used number of the validation trials (v) when 1, 2, and 5 trials were available for training from the target subject.

using the training trials plus the validation trials (i.e. CSP(t+v)). Interestingly, Fig. 5.6 shows that in all cases, considered in this comparison, DTW-RCSP-On outperformed the corresponding CSP(t+v).

A 2 (Algorithms= CSP(t+v), and DTW-RCSP-On) × 2 (Number of validation trials= 2, and 6)× 3 (Number of training trials per class= 1, 2, and 5)) repeated measure ANOVA tests were performed on the results of all datasets followed by post-hoc analyses. For the large dataset, there was a main effect of using different number of training trials with $P = 0.024$. Moreover, the ANOVA results showed that our proposed DTW-RCSP-On tended to be significantly better than CSP(t+v) with $P = 0.059$. Posthoc analyses revealed that using 5 training trials per class were significantly better than using 1, and 2 trials with P values equal to 0.025 and 0.043 respectively. For the medium dataset using different algorithms, different training trials and different validation trials had main effects on the results with P values $0.042, 0.034$, and 0.013 respectively. Thus, we can conclude that in the medium dataset our proposed DTW-RCSP-On was significantly better than CSP(t+v) with $p = 0.042$. Posthoc analyses showed that using 5 training trials per class were significantly better than 1, and 2 trials with P values equal to 0.016 and 0.023 respectively, and using 6 validation trials were significantly better than 2 with $P = 0.034$. Surprisingly, for the small dataset even only few subjects were available for transfer learning there was a main effect of using different number of validation trials with $P = 0.015$. Moreover, the ANOVA tests showed that our proposed DTW-RCSP-On tended to be significantly better than CSP(t+v) with $P = 0.08$.

In summary, our results showed that the proposed DTW-RCSP based transfer learning framework led to improved CSP features and hence improved BCI systems, particularly when a small subject-specific training data were available.

5.5 Conclusion

This chapter proposed a novel DTW-based transfer learning framework on raw EEG and feature domains to improve the CSP covariance matrix estimations and hence enhance MI-based BCI systems. The proposed framework minimizes the temporal variations between the EEG trials of other subjects and the few EEG trials of the target subject using DTW. Then the temporally aligned trials of other subjects are mixed with the few subject-specific trials in the CSP framework using a regularization parameter.

Our results suggested that applying the proposed DTW-based transfer learning framework reduced calibration of the MI-BCI systems. The results obtained showed that our proposed framework significantly outperformed the subject-specific CSP and CCSP algorithms in many different scenarios specially when data from a large number of subjects were available for transfer learning.

The proposed framework uses only one regularization parameter which is not computationally expensive compared to most of transfer learning-based regularized CSP algorithms that use 2 regularization parameters. Besides, the proposed online method required very slightly more computational time compared to CSP when the same number of trials are used. Thus, with these two benefits and with using only two validation trials the proposed DTW-RCSP-On could be potentially used for online applications.

Interestingly, the proposed DTW-based transfer learning framework yielded remarkable increase in the classification accuracy of majority of the participants specially when only few trials were available for training from the target subject. However, the observed improvement for some subjects with initially poor BCI performance was not pronounced. The possible reason might be having inseparable EEG signals between two classes either during training session and testing session or both. In the future, further investigation is needed to identify these participants and exclude their data from being used for transfer learning. In other case, if these subjects are the current users of the BCI system, some human-training strategies should be identified and provided for them to improve their BCI accuracy.

Chapter 6

An Ensemble Framework with Temporal Alignment for Improving BCI Performance in Small Sample Settings

6.1 Introduction

In previous chapters we proposed transfer learning approaches that use data from other subjects/sessions to improve the subject-specific BCI model when few trials are available for training from the current BCI user. However, if there are not enough data from other sessions/subjects, transfer learning is not possible using these previously proposed approaches. Thus, making BCI system robust using just data available from the current subject is highly desirable.

To address these drawbacks, different algorithms have been proposed to improve the CSP learning process. CSP improvement can be done either at CSP optimization objective function level [33, 151, 152], or at covariance matrix estimation level [104, 105, 153]. In this chapter, we are interested in improving CSP covariance matrix estimation, as the first part of our proposed framework is dealing with that problem. For example, previously, in [145], a modified version of the CSP has been proposed. Minimum covariance determinant estimator was used to obtain robust covariance estimates that replaced the classical covariance estimates of CSP. In [153], a regularized CSP algorithm has been proposed to improve the covariance matrix estimation using two regularization parameters. The first regularization parameter used the generic learning principle to improve the estimation stability by controlling the shrinkage of a subject-specific covariance matrix towards a generic covariance matrix. The second regularization parameter deals with with the limited availability of training trials by controlling the shrinkage towards a scaled identity matrix.

Despite outperforming the classical CSP to some extent, most of the existing improved CSP algorithms are computationally expensive due to calculating a number of regularization parameters. Besides, none of them considers the temporal variations between trials. Moreover, most of these algorithms focus only on training data without considering the inter-session non-stationarity when transferring to test session. Particularly, it was found that even great improvement for feature extraction using the available training data does not guarantee a perfect BCI system [154]. The problem might be related to the variation between the distribution of training and testing trials. Testing trials could have severe overlapping specially when the BCI user starts to feel fatigue or being distracted.

This chapter proposes a novel dynamic time warping (DTW)-based ensemble framework to deal with intra- and inter-session non-stationarity in motor imagery-based BCIs mainly when only a few trials are available for training and there is no available trials from previous sessions or other subjects. Our proposed framework is split into two parts. Firstly, DTW is used to improve CSP covariance matrix estimation, and hence feature extraction. We hypothesize that the alignment of EEG trials from the same class to their average might reduce within class temporal variations and non-stationarity. Following the previous assumption using DTW, the available trials from the same class get as close as possible to the mean of its class and also to each other. The new aligned trials are used to calculate the CSP covariance matrices. Secondly, DTW is used to reduce non-stationarities between the upcoming testing trials and the average of each class of the available few training trials. Aligning the upcoming test trial to the average of the two classes of the training trials results in two new aligned trials. These two trials are classified using the trained classifier individually. Then the ensemble decision is used to accept the trials if the two predicted labels of these new trials are the same, otherwise it will be rejected and the user will be asked to repeat the task performed.

The proposed framework was evaluated using one of the publicly available datasets with a moderate number of subjects. Performance of the proposed framework was also compared with two baseline algorithms to show its significance.

The remainder of this chapter is organized as follows. In Section 6.2, we will describe the proposed framework. Data description and results are discussed and analyzed in Section 6.3. Finally, conclusions are drawn in Section 6.4.

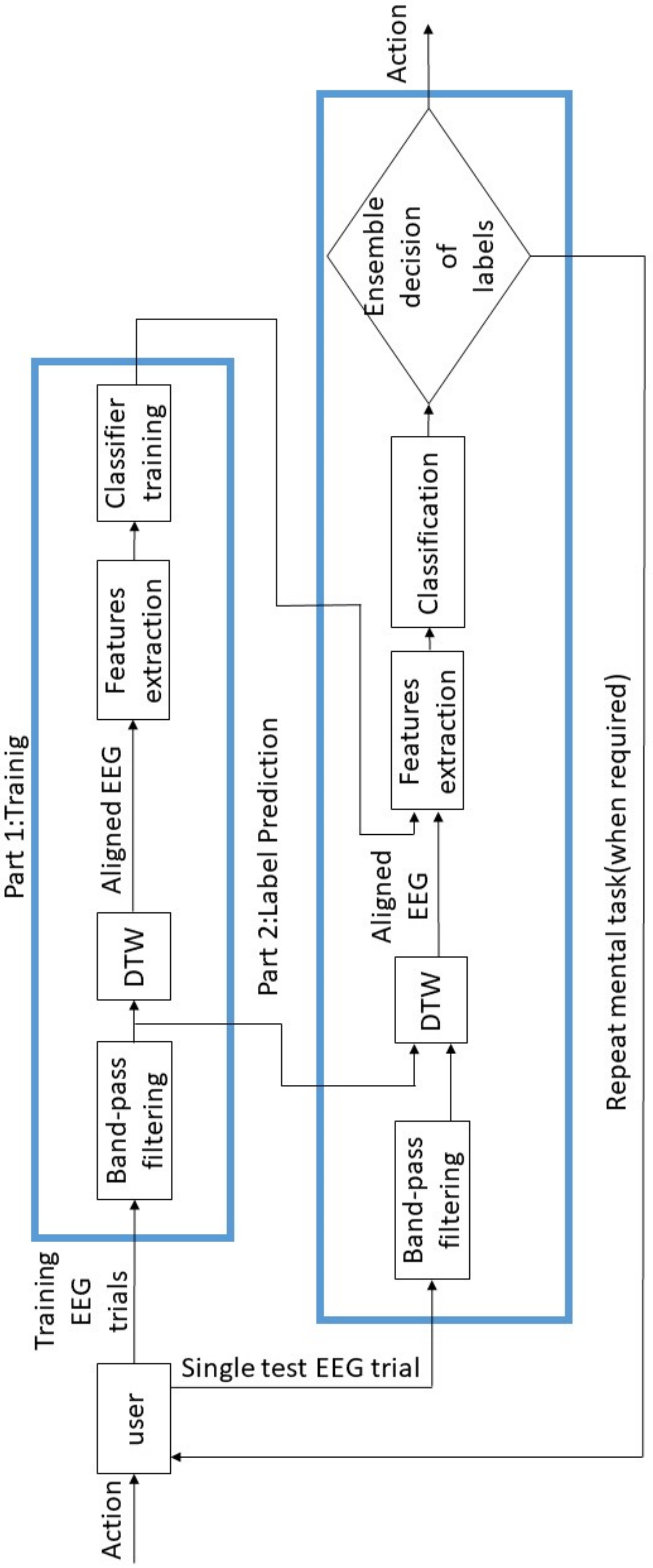


Figure 6.1: The proposed DTW-based Ensemble framework

6.2 Methodology

This section explains the proposed framework. Fig. 6.1 shows the two parts of the proposed framework. First part is the training of the proposed framework using the available training trials from the user. Second part is dealing with the testing trials where DTW and an ensemble decision criteria were used to evaluate the quality of the new coming trial.

In this chapter, we assume that a number of labelled EEG trials are available from each subject. The set of labelled EEG trials for each subject can be presented as $d = (\mathbf{X}_i, y_i)_{i=1}^{f}$, where f is the number of trials, and $\mathbf{X}_i$ and y_i respectively denote the instances matrix and the class label, $y_i \in \{-1, 1\}$, of the i^{th} trial. Each trial is a subset of $\mathbb{R}^{h \times V}$, where h is the number of EEG samples and V is the number of channels.

6.2.1 Part I: Robust DTW-based CSP training

Typically, the classifier is trained using the available labelled training features to predict the labels of the unlabelled trials. The commonly used BCI model uses CSP algorithm to extract features [33, 124]. Hence, in order to overcome the problem of non-robust CSP covariance matrices estimation, we use our proposed DTW-CSP algorithm. Thus, in the first part of the proposed framework, our novel DTW-CSP algorithm, explained in the previous chapter, aligns the available trials from each class to be as much similar to their average. Performing the proposed alignment leads to create new training trials that are less dissimilar in temporal domain and hence improve CSP covariance matrix estimation. This leads to improve CSP feature extraction.

6.2.2 Part II: Ensemble decision of the upcoming testing trial

At this point, CSP features were calculated and the classifier was trained using previously obtained features after using DTW-CSP algorithm. Now for any upcoming test trial $\mathbf{X}_{\mathrm{T}}$, DTW is used to find a similarity matrix between this test trial and the average signal of each class of the few available training trials computed using 5.6. Then two warping paths for these two new aligned trials are calculated in a way to minimize the cost function 5.7 under the constraints mentioned in discussed in 5.2.1.

Given the optimum warping path between the testing trial (T) and the average of each class of the the available training trials, $\mathbf{X}_{\mathrm{T}}$ is aligned to $\bar{\mathbf{X}}_{t_c}$ resulting in $\mathbf{X}_{\mathrm{T}}^c$. $\mathbf{X}_{\mathrm{T}}^c$ contains the time indices of $\mathbf{X}_T$ forming the minimum warping path between this testing trials and the average of the available training trials from class c. These time instances are the instances that will make $\mathbf{X}_{\mathrm{T}}$ to be as much similar to $\bar{\mathbf{X}}_{\mathbf{t_c}}$ as possible. Subsequently the covariance matrix of $\mathbf{X}_{\mathrm{T}_{aligned}}^c$ is calculated as follows:

$$\mathbf{\Sigma}_{\mathrm{T}_{\mathrm{aligned}}}^c = \frac{(\mathbf{X}_{\mathrm{T}_{aligned}}^c)^\top \mathbf{X}_{\mathrm{T}_{aligned}}^c}{\mathbf{tr}((\mathbf{X}_{\mathrm{T}_{aligned}}^c)^\top \mathbf{X}_{\mathrm{T}_{aligned}}^c)}. \tag{6.1}$$

These covariance matrices will be used to calculate CSP features. Then both of the resulting features are classified using the trained classifier. If the output label for the two new aligned trials is the same this testing trial is accepted and classified accordingly. If the labels are different this trial is rejected and the user is asked to repeat the action.

6.3 Experimental results

6.3.1 Data description

The proposed framework was compared to two algorithms. The first algorithms is the state of the art BCI algorithm using CSP features. This algorithms will be called (CSP) in the rest of this chapter. The second algorithm is the proposed DTW-CSP algorithm, explained in the previous chapter.

The proposed framework and the baseline algorithms were applied to data set 2a BCI Competition IV 2008 [130]. This data set consists of EEG data from 9 subjects performing 4 classes of motor imagery task. In this thesis, as mentioned before, only data from right and left hand motor imagery were used. Two sessions on different days were recorded for each subject. Each session is comprised of 6 runs, each run consists of 12 trials for each class.

EEG signal was recorded using 22 electrodes. EEG signals were sampled at 250 Hz, and were bandpass-filtered between 0.5 Hz and 100 Hz. Moreover, a 50 Hz notch filter was applied to remove power line noise. The proposed framework

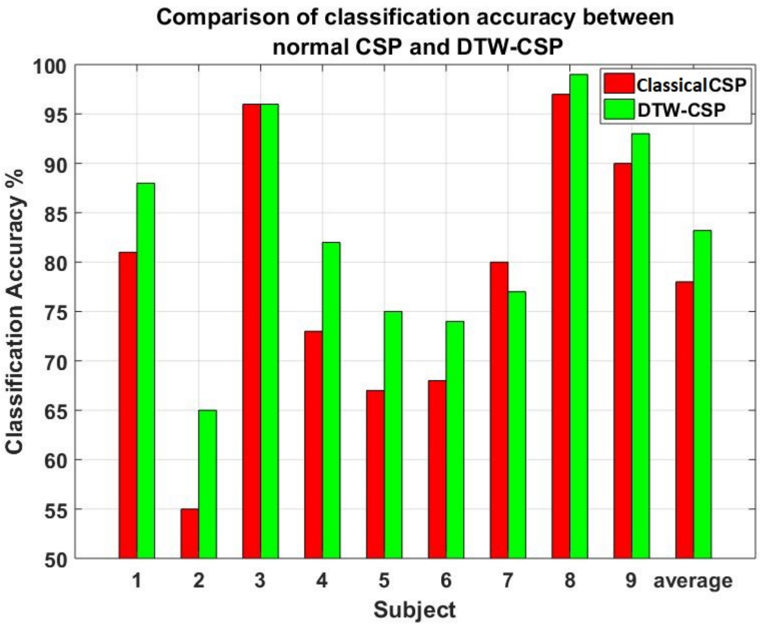


Figure 6.2: Comparison of classification accuracy between classical CSP and DTW-based CSP. Interestingly, it shows that the proposed DTW-CSP algorithm outperform classical CSP for all subject except subject 7. Moreover on average classification accuracy the proposed algorithm is better than normal CSP by almost 5%.

and the baseline algorithms were applied only on the trials recorded on the second day by dividing it to two sessions, one for training (consists of the first 42 trials recorded per class) and one for testing (consists of the last 30 trials recorded per class). This was done to establish a practical case that new subject data is coming from the same session. For the new subject, different training sizes were examined (i.e. 5, 10, 20 and 42 trials per class).

6.3.2 Evaluation and discussion

For each subject, the CSP and the DTW-CSP filters were learnt on the available training set. The log-variances of the spatially filtered EEG signal were then used as the input features of a Support vector machines (SVM) classifier. The classification accuracy was calculated based on how accurately the labels of testing sessions trials are estimated. Fig.6.2 shows that, except subject 7, the DTW-CSP algorithm outperformed classical CSP. The DTW-CSP algorithm outperformed

CSP by about 4% to 10% for each subject. On average, classification accuracy for all subject was increased from 78% to 83.3%. These results confirm that using DTW reduces temporal variations and non-stationarities between trials within the same class, and hence enhance the computed features. Particularly, with a closer look at the results, DTW-CSP algorithm was more valuable for subjects with poor and medium initial BCI performance (e.g.sub1, sub2, sub4, sub5, sub6) than subjects with initially high performance, whose performances were slightly changed. This finding makes sense as subjects with high initial accuracy already have their features well separated.

Fig.6.3 shows some examples of the spatial filters obtained with classical CSP and DTW-CSP algorithms for different subjects. Notably, these pictures show that classical CSP filters appear with large weights in several unexpected locations from a neurophysiological point of view. On the other hand, DTW-CSP filters were interestingly smoother and physiologically more relevant to the imagined hand. Contrary to classical CSP, DTW-CSP filters weights were more related to the expected motor cortex areas. This is another benefit of the DTW-CSP algorithm as it does not only make the trials of the same class get closer but also lead to filters that are neurophysiologically smoother and as such more illustratable. Moreover, our approach requires much less computational time as there is no need to calculate any regularization parameters either using cross-validation or by optimizing objective functions which are computationally expensive.

The classification accuracy was calculated based on how accurately the labels of testing sessions trials were estimated using three methods (CSP, DTW-CSP, and the proposed framework and this will be called DTW-Ensemble Framework). For each method we used different number of training trials.

Fig. 6.4 shows an example of one subject where great improvement in feature extraction using the available training data does not guarantee having a good BCI performance as testing trials might be very different from training trials due to intersession non-stationarity inherent in EEG signals. Training trials features from the two classes were completely separated, however, testing trials features from the two classes were some how overlapped. This figure presents the importance of using the second part of the proposed framework where ensemble decision criteria that predicts the test trials labels with the option of rejecting them to improve the BCI system accuracy was employed.

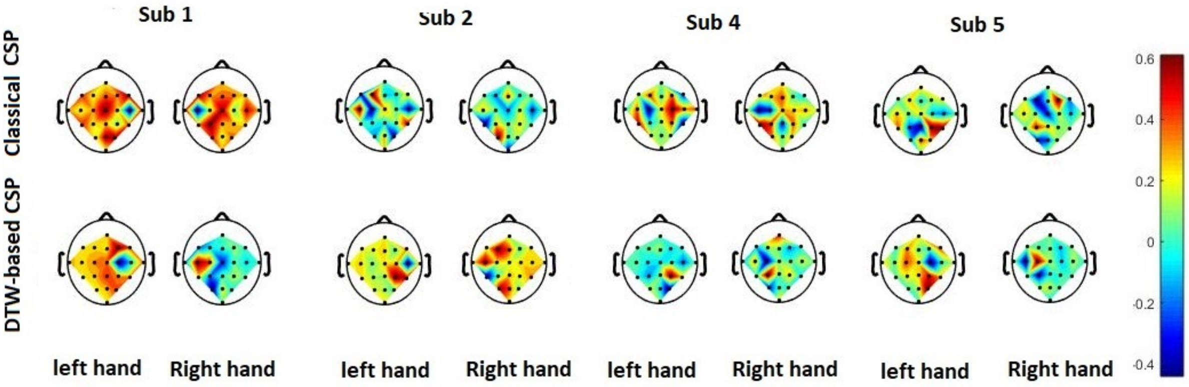


Figure 6.3: Some examples of spatial filters obtained with classical CSP and DTW-CSP algorithms for different subjects. Interestingly, DTW-CSP filters are smoother and physiologically more relevant to the imagined hand. Contrary to classical CSP, DTW-CSP filters weights are more related to the expected motor cortex areas.

Fig. 6.5 shows that the proposed DTW-Ensemble Framework outperformed the classical CSP and DTW-CSP algorithms using any number of training trials. On average, classification accuracy for all subject was increased from 65.5% to 67.77%, 70% to 74.7%, 73.5% to 77.55%, 75.4 % to 79.1% when using 5, 10, 20, and 40 trials for training, respectively.

Importantly, the proposed framework did not reject too many trials to improve the classification accuracy. For some subjects, only 1 or 2 trials were rejected with the maximum of 5 trials for the subjects with poor BCI performance. Interestingly, using the proposed DTW-Ensemble Framework with 20 trials per class for training outperformed the classical CSP with 40 trials per class for training. This means that proposed framework not only reduced the calibration time but also increased the classification accuracy. Moreover, this figure confirms that using DTW reduced the temporal variations and non-statinarties between the trials within the same class when only few trials are available to perform DTW, and hence enhanced the computed features.

Concerning statistical significance, a 4 (Number of trials)×3 (Algorithms) repeated measure ANOVA test with a Greenhouse-Geisser correction revealed that using different algorithms had a main effect on the classification accuracy with ($P = 0.021$), Greenhouse-Geisser correction is used as sphericity assumed condi-

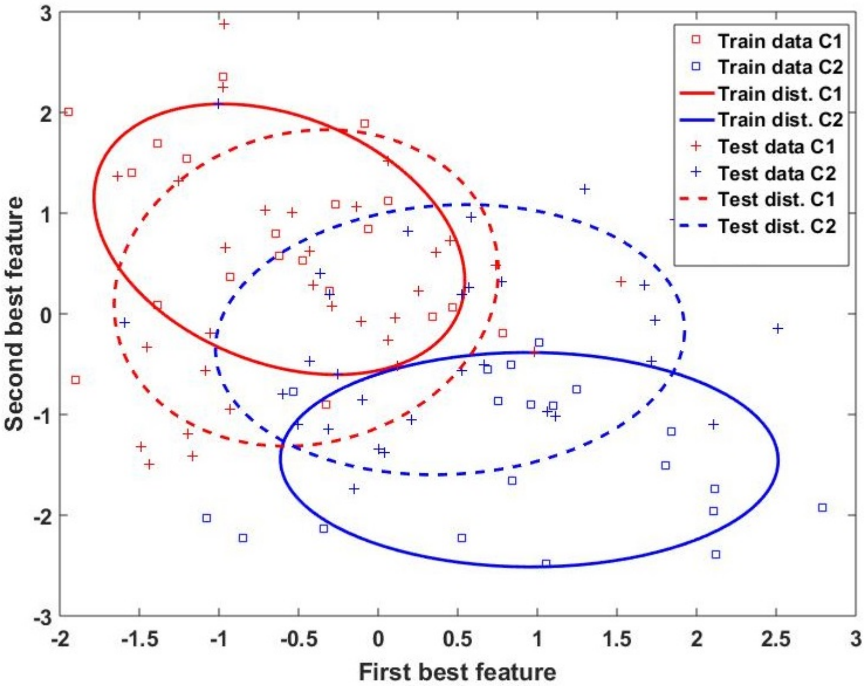


Figure 6.4: This figure shows an example of training and testing feature distribution from both classes for subject 7. It is shown that the distribution of training data from both classes are entirely separated. However, there is some overlapping between testing trials distribution from class1 and class2

tion was violated. However, using different numbers of trials did not affect the classification accuracy. Based on the post-hoc analysis, the proposed framework significantly outperformed CSP and DTW-CSP with the P values equal to 0.006 and 0.046 respectively.

Table 1 shows the individual classification accuracy for each subject calculated using the proposed DTW-Ensemble Framework, DTW-CSP, and CSP when 20 trials were used for training. It is shown that the DTW-Ensemble Framework outperformed or had the similar results regards to classical CSP. The proposed framework outperformed CSP by about 2% to 10% for different subjects. These results confirmed that using DTW not only with training data but also with testing data reduced the temporal variations and non-statinarties between training trials within the same class and the training and testing trials. Using DTW between the average of each class from the training data and the upcoming testing trials made this trial more similar to its related class and hence enhanced the

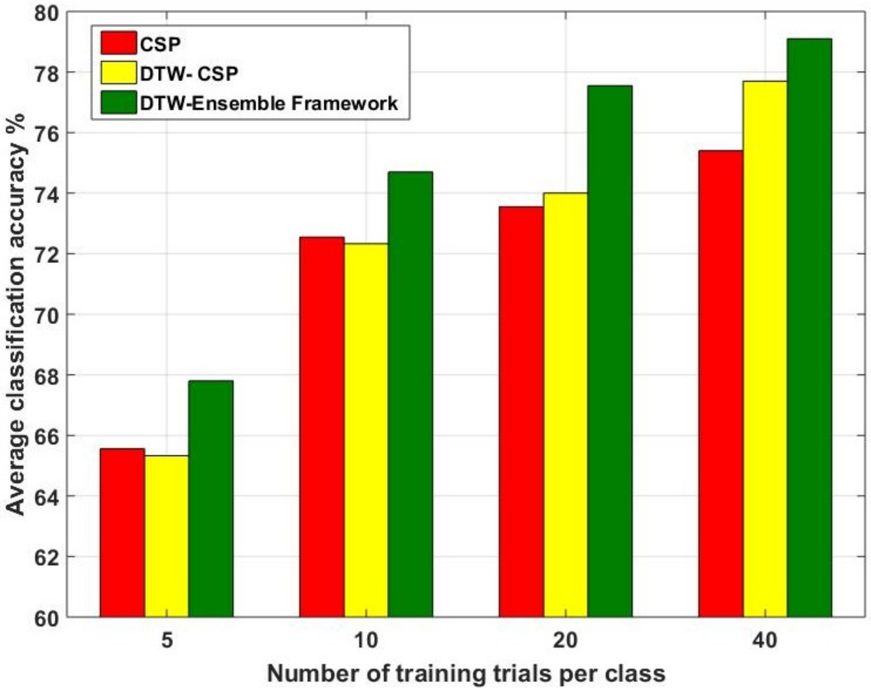


Figure 6.5: Comparison of the classification accuracy between CSP, DTW-CSP, and the proposed DTW-Ensemble Framework. Interestingly, it shows that the proposed framework outperformed the baseline algorithms using any number of training trials.

computed features. This led to improve the process of label estimation. Particularly, with a closer look at the results, suggests that the proposed framework was more valuable for subjects with poor and medium initial BCI performance (e.g., sub1, sub2, sub5, sub5, sub6, sub7, and sub8) than subjects with initially high performance, whose performances were not changed. This finding makes sense as subjects with high initial accuracy already had their features well separated.

Moreover, our framework doesn't require any regularization parameters calculations either using cross-validation or by optimizing objective functions which are computationally expensive.

In summary, our results showed that the novel proposed DTW-Ensemble Framework outperformed the baseline algorithms when only few trials were available for training. The proposed framework not only reduced the required calibration time but also enhanced the average classification accuracy. Moreover, the proposed framework could be applied to any temporal domain based feature

Table 6.1: Classification accuracies calculated using the baseline algorithms (CSP, and DTW-CSP) and the proposed framework for each individual subject when there were only 20 trials available for training per class from the user.

Algorithm	s1	s2	s3	s4	s5	s6	s7	s8	s9	Avg.
CSP	85	53	98	67	55	57	73	87	87	73.55
DTW-CSP	88	55	98	67	50	60	73	88	87	74
DTW-Ensemble Framework	**90**	**62**	97	67	**57**	**63**	**78**	**97**	87	**77.6**

space and could be used to accept or reject the upcoming testing trials using any ensemble learning algorithm.

6.4 Conclusions

This chapter proposed a novel DTW-Ensemble Framework to improve BCI systems. Our results suggest that using the proposed framework could lead to reducing the calibration time to 5 trials per class and at the same time enhancing the average accuracy of the MI-BCI systems. Interestingly, The proposed framework could be applicable not only to CSP but to any temporal domain based feature space and able to be used to accept or reject the upcoming testing trials using any ensemble learning algorithm. The results obtained show that the proposed DTW-Ensemble Framework significantly outperformed the state of the art BCI algorithm using CSP for any available number of the training trials.

Chapter 7

Conclusion and Future Work

The work presented in this thesis aimed to make BCI more reliable as a daily use system. Thus, this thesis focused on developing novel transfer learning approaches to reduce calibration time of BCI with minimum accuracy loss or even improve it. To achieve this objective two main challenges needed to be addressed based on the available training data from the current user and previous sessions or users. First, reducing inter-subjects/sessions non-satationarity. Second, reducing intra-session non stationarity.

7.1 Conclusions

Through this thesis, we proposed the following novel transfer learning approaches to address the previous issues and improve the usability of BCI:
1- A novel weighted multi-task algorithm on classification domain.
2- A novel weighted transfer learning algorithm on classification domain.
3- A novel DTW-RCSP based transfer learning framework on raw EEG and feature domains.
4- A novel subject-specific DTW based CSP algorithm.
5- A novel domain adaptation framework based on DTW for Improving BCI Performance in Small Sample Settings.

First, to achieve the objectives related to the first scenario, in chapter 3, we proposed novel weighted multi-task transfer learning algorithms in the classification domain to reduce the calibration time without sacrificing the classification accuracy of the MI-BCI systems. Previously recorded data were mined, processed and reused in a way that higher weights were given to the data that were more similar to the new data and less weights to data that were less similar. Two versions of weighted multitask learning were proposed, namely supervised and unsupervised. Results showed that the novel proposed unsupervised weighted logistic multi-task learning algorithm (UMLLog) outperformed all the other algorithms. The proposed UMLLog not only reduced the required calibration time but also enhanced the classification accuracy for most of the subjects.

Despite success to some extent, the proposed algorithms in chapter 3 were computationally expensive as a big number of parameters needed to be optimized simultaneously. To further improve the classification accuracy and reduce the computational time of the BCI system, we proposed novel weighted transfer learning algorithms in chapter 4. In the proposed algorithms, the classification parameters of each of previous users with relatively large number of trials were calculated independently in a way to minimize the subject-specific classification error. Thereafter, the new user's classification parameters were calculated in a way that the classification error was minimized and at the same time got as close as possible to the classification parameters of other existing subjects. A regularization term was added into the classification objective function to make a trade-off between minimizing the classification error of the new user and dissimilarities with the classification parameters of previous users. The proposed weighted transfer learning algorithms yielded a significant reduction in the calibration time and a remarkable increase in the classification accuracy for most of the subjects that initially performed BCI with poor or medium accuracy.

Our results in chapter 4 showed that the observed improvement for a few subjects with initially low BCI performance was not pronounced. It seems that estimating better classification parameters for those subjects was not sufficient since their feature spaces for different classes were severely overlapped. Based on these outcomes transfer learning approaches should be applied in a different domain before the classification domain. Therefore, in chapter 5, we proposed a novel regularized covariance estimation framework for CSP (i.e. DTW-RCSP) based on dynamic time warping (DTW) and transfer learning. DTW-RCSP framework reduced temporal non-stationarity between the few available trails from the current user and the available trials from previous sessions or subjects and at the same time improved the covariance matrix estimation of CSP to extract more robust and relevant spatial filters. The proposed framework combined the subject-specific covariance matrix estimated using the few available trials from the new subject, with a novel DTW-based transferred covariance matrix estimated using previous subjects trials. In the proposed framework, the available labelled trials from the previous subjects were temporally aligned to the average of the few available trials of the new subject from the same class using DTW to reduce temporal variations and non-stationarities. The regularization parameter was selected using a novel method based on the confidence scores of the trained classifier on upcoming first few labelled testing trials. The proposed framework was evaluated on three datasets the classical CSP and CCSP. Results showed that DTW-RCSP significantly outperformed the classical CSP in various testing scenarios, particularly, when only a few trials were available for training. Impressively, our results showed that successful BCI interactions could be achieved with a calibration session as small as only one trial per class.

Finally, we observed that still the improvement for some users with initially poor BCI performance was not significantly enhanced. Improving the features extracted and hence improving the estimated classifier's parameters for these users were not effective, We found that unlike their training features their testing features for different classes were not separable. Therefore, to achieve the objective related to the second scenario, in chapter 6 a novel dynamic time warping (DTW)-based ensemble framework to deal with intra- and inter-session non-stationarity in motor imagery-based BCIs mainly when only a few trials are available for training was proposed. The proposed framework consists of two parts. First, DTW was used to make the CSP robust against intra-session variations when only a few trials are available for training. Second, as a domain adaptation method, DTW reduces the dissimilarity between testing and training trials. Finally, an ensemble decision making is used to predict the test trials labels with the option of rejecting them. The results showed that besides calibration time reduction, the average classification accuracy was enhanced compared to the classical CSP

In summary, using the proposed algorithms we addressed the mentioned challenging issues, and consequently we achieved our objective to make BCI systems more robust with less calibration time. These improvements along with improvements at different levels such as the neuroscience level and the human level learning will lead to a more robust and efficient BCI technology capable enough for daily use.

7.2 Limitations and directions of future Work

The work presented in thesis can be potentially extended to address limitations faced during this thesis, some other challenges in BCI or even other areas. Some of these future extensions are listed below.

- Transfer learning algorithms proposed in this thesis were compared to CSP as a subject specific algorithm and to the related transfer learning learning algorithms as well. The CSP algorithm has been chosen as it is the most commonly used subject specific algorithm in BCI. Although, filter-bank CSP algorithm has been used in several BCI applications, it tends to overfit when the available subject specific training trials are few. We have applied filter-bank CSP on the subjects of 2a BCI Competition IV 2008 when only 5 trials per class were used for training. We found that there was a huge loss in classification accuracy compared to CSP. However, in the future, a benchmark of different algorithms related to the proposed algorithms need to be addressed and compared with the proposed algorithms using the same datasets.

- Transfer learning algorithms proposed in chapter 3 and 4 used KL divergence for calculating similarity weights between the data from the new subject and previous subjects or sessions. The effectiveness of the proposed weighted transfer learning can be further improved by exploring new methods,e.g. Riemannian geometry, to measure the similarity weights between the previous subjects/sessions data and the few trials from the new subject [155].

- Concerning DTW-R-CSP framework, our offline analyses showed that some subjects could have achieved much more improvement in accuracy if a different regularization parameter could have been selected. Thus, it would be interesting if another method could be explored to calculate the optimum regularization parameter that can maximize the classification accuracy with less number of trials and less computational time [156, 157].

- The results presented in this thesis were obtained using offline analysis. Although, these results were very promising, it is recommended to conduct online experiments to check the liability of proposed algorithms in real time scenarios. For example, it would be interesting to see if rejecting some of testing trials using DTW-Ensemble framework during an online experiment and asking the participant to repeat performing the mental tasks would improve the human learning process in BCI [158, 159]. Moreover, other methods for how to reject the bad testing trials need to be explored and compared to the current method until find the best optimum method.

- In this thesis, we proposed 4 novel transfer learning frameworks and algorithms to reduce the calibration time of MI-based BCI systems. The proposed algorithms rely on different processing methods. Thus, the classification results might be further improved, especially for subjects with poor BCI performance, by combining these algorithms and frameworks in a complementary way. Accordingly, different complementary combinations need to be explored in order to design a more accurate and effective BCI with minimum calibration time.

In addition to the above-mentioned future works, the following long-term extensions might be of interest.

- Basically, zero calibration is the optimum case for a real time BCI system that can be used in daily life tasks. Too frequent recalibrations might negatively affect the learning process of the BCI user and make the user confused regarding the received feedback. Thus, recalibration should be done very carefully to prevent the BCI user from being confused. Moreover, some research should be conducted at the human level by developing more advanced and successful user training techniques to better improve the BCI user learning process. Modeling human learning using signal processing and machine learning, and using those models to identify when recalibration is needed would be a very interesting and important future study in BCI community.

- It would be a promising idea to extend our proposed algorithms on other neurophysiological signals, not only motor imagery. Beside reducing the calibration time, the accuracy and robustness of BCI needs to be improved, it is highly desirable to apply and generalize the proposed algorithms to other areas that are affected by noise and non-stationary data. Future research should focus on developing a novel general robust subject-independent BCI framework that can be used for any subject with minimum calibration time [160, 161].

- Developing new transfer learning approaches to improve the performance of BCI users especially the users doing poor BCI. As, discovered in the work done through this research, after some time, for some subjects, brain signals from different classes became overlapped. Thus, new training techniques and new feedback types should be explored to achieve more accurate and robust BCI performance with minimum training time [120, 162].

- The discussions and analysis of the proposed algorithms and frameworks presented in this study focused on EEG patterns from only two classes of motor imagery tasks. It is highly desirable to develop algorithms that can accurately classify a larger number of mental tasks (classes). More mental tasks to be identified accurately means more commands for controlling a device for communication by the BCI user. In future, EEG signals from multi-class mental tasks have to be analyzed either by extending our proposed algorithms to multi-class paradigms or developing new algorithms to improve the accuracy and robustness of multi-class BCIs with minimum calibration time [163–165].

References

[1] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, “Brain–computer interfaces for communication and control,” *Clinical Neurophysiology*, vol. 113, no. 6, pp. 767–791, 2002.

[2] E. A. Curran and M. J. Stokes, “Learning to control brain activity: a review of the production and control of EEG components for driving brain–computer interface (BCI) systems,” *Brain and Cognition*, vol. 51, no. 3, pp. 326–336, 2003.

[3] J. J. Vidal, “Toward direct brain-computer communication,” *Annual review of Biophysics and Bioengineering*, vol. 2, no. 1, pp. 157–180, 1973.

[4] C.-S. Lin, C.-W. Ho, W.-C. Chen, C.-C. Chiu, and M.-S. Yeh, “Powered wheelchair controlled by eye-tracking system.” *Optica Applicata*, vol. 36, 2006.

[5] M. K. Holden, “Virtual environments for motor rehabilitation,” *Cyberpsychology & behavior*, vol. 8, no. 3, pp. 187–211, 2005.

[6] H. I. Krebs, J. J. Palazzolo, L. Dipietro, M. Ferraro, J. Krol, K. Rannekleiv, B. T. Volpe, and N. Hogan, “Rehabilitation robotics: Performance-based progressive robot-assisted therapy,” *Autonomous robots*, vol. 15, no. 1, pp. 7–20, 2003.

[7] J. P. Hansen, K. Tørning, A. S. Johansen, K. Itoh, and H. Aoki, “Gaze typing compared with input by head and hand,” in *Proceedings of the 2004 symposium on Eye tracking research & applications.* ACM, 2004, pp. 131–138.

[8] U. Chaudhary, B. Xia, S. Silvoni, L. G. Cohen, and N. Birbaumer, “Brain–computer interface–based communication in the completely locked-in state,” *PLoS biology*, vol. 15, no. 1, p. e1002593, 2017.

[9] S. N. Abdulkader, A. Atia, and M.-S. M. Mostafa, “Brain computer interfacing: Applications and challenges,” *Egyptian Informatics Journal*, vol. 16, no. 2, pp. 213–230, 2015.

[10] E. Niedermeyer and F. L. da Silva, *Electroencephalography: basic principles, clinical applications, and related fields.* Lippincott Williams & Wilkins, 2005.

[11] F. Lotte, L. Bougrain, and M. Clerc, “Electroencephalography (EEG)-based brain–computer interfaces,” *Wiley Encyclopedia of Electrical and Electronics Engineering*, pp. 1–20, 2015.

[12] G. Pfurtscheller and F. L. Da Silva, “Event-related eeg/meg synchronization and desynchronization: basic principles,” *Clinical neurophysiology*, vol. 110, no. 11, pp. 1842–1857, 1999.

[13] M. Krauledat, M. Schröder, B. Blankertz, and K.-R. Müller, “Reducing calibration time for brain-computer interfaces: A clustering approach,” in *Advances in Neural Information Processing Systems*, 2007, pp. 753–760.

[14] J. d. R. Millán, R. Rupp, G. R. Müller-Putz, R. Murray-Smith, C. Giugliemma, M. Tangermann, C. Vidaurre, F. Cincotti, A. Kübler, R. Leeb *et al.*, “Combining brain–computer interfaces and assistive technologies: state-of-the-art and challenges,” *Frontiers in neuroscience*, vol. 4, 2010.

[15] P. Wang, J. Lu, B. Zhang, and Z. Tang, “A review on transfer learning for brain-computer interface classification,” *2015 5th International Conference on Information Science and Technology, ICIST 2015*, pp. 315–322, 2015.

[16] S. J. Pan, Q. Yang *et al.*, “A survey on transfer learning,” *IEEE Transactions on knowledge and data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[17] C. Jeunet, B. NKaoua, S. Subramanian, M. Hachet, and F. Lotte, “Predicting mental imagery-based BCI performance from personality, cognitive profile and neurophysiological patterns,” *Plos one*, vol. 10, no. 12, p. e0143962, 2015.

[18] L. da Silva-Sauer, A. d. l. Torre-Luque, J. S. Silva, and B. Fernández-Calvo, "New perspectives for cognitive rehabilitation: Could brain-computer interface systems benefit people with dementia?" *Psychology & Neuroscience*, vol. 12, no. 1, p. 25, 2019.

[19] F. Lotte and C. Jeunet, "Towards improved BCI based on human learning principles," in *The 3rd International Winter Conference on Brain-Computer Interface*. IEEE, 2015, pp. 1–4.

[20] C. Jeunet, B. NKaoua, and F. Lotte, "Advances in user-training for mental-imagery-based BCI control: Psychological and cognitive factors and their neural correlates," in *Progress in brain research*. Elsevier, 2016, vol. 228, pp. 3–35.

[21] S. K. Card, T. P. Moran, and A. Newell, "The keystroke-level model for user performance time with interactive systems," *Communications of the ACM*, vol. 23, no. 7, pp. 396–410, 1980.

[22] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain-computer interfaces for communication and control." *Clinical Neurophysiology : official journal of the International Federation of Clinical Neurophysiology*, vol. 113, no. 6, pp. 767–91, 2002. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/12048038

[23] M. Ahn, M. Lee, J. Choi, and S. Jun, "A review of brain-computer interface games and an opinion survey from researchers, developers and users," *Sensors*, vol. 14, no. 8, pp. 14 601–14 633, 2014.

[24] A. Kübler, F. Nijboer, J. Mellinger, T. M. Vaughan, H. Pawelzik, G. Schalk, D. J. McFarland, N. Birbaumer, and J. R. Wolpaw, "Patients with als can use sensorimotor rhythms to operate a brain-computer interface," *Neurology*, vol. 64, no. 10, pp. 1775–1777, 2005.

[25] N. Birbaumer, L. G. Cohen, and Kübler, "Brain-computer interfaces: communication and restoration of movement in paralysis," *The Journal of Physiology*, vol. 579, no. 3, pp. 621–636, 2007. [Online]. Available: http://doi.wiley.com/10.1113/jphysiol.2006.125633

[26] M. Kaper, P. Meinicke, U. Grossekathoefer, T. Lingner, and H. Ritter, "Bci competition 2003-data set iib: support vector machines for the p300 speller paradigm," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 6, pp. 1073–1076, 2004.

[27] C. Guger, "Using a brain / computer interface for smart-home control," *PerAdaMagazine*, pp. 1–2, 2009.

[28] C.-T. Lin, B.-S. Lin, F.-C. Lin, and C.-J. Chang, "Brain computer interface-based smart living environmental auto-adjustment control system in upnp home networking," *IEEE Systems Journal*, vol. 8, no. 2, pp. 363–370, 2014.

[29] K. K. Ang and C. Guan, "Brain-computer interface in stroke rehabilitation," *Journal of Computing Science and Engineering*, vol. 7, no. 2, pp. 139–146, 2013.

[30] M. D. Matthews and D. M. Schnyer, *Human Performance Optimization: The Science and Ethics of Enhancing Human Capabilities.* Oxford University Press, 2018.

[31] H. Huang, Q. Xie, J. Pan, Y. He, Z. Wen, R. Yu, and Y. Li, "An eeg-based brain computer interface for emotion recognition and its application in patients with disorder of consciousness," *IEEE Transactions on Affective Computing*, 2019.

[32] C. Pandarinath, P. Nuyujukian, C. H. Blabe, B. L. Sorice, J. Saab, F. R. Willett, L. R. Hochberg, K. V. Shenoy, and J. M. Henderson, "High performance communication by people with paralysis using an intracortical brain-computer interface," *eLife*, vol. 6, pp. 1–27, 2017.

[33] M. Arvaneh, C. Guan, K. K. Ang, and C. Quek, "Optimizing spatial filters by minimizing within-class dissimilarities in electroencephalogram-based brain–computer interface," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 4, pp. 610–619, 2013.

[34] F. Lotte, "Signal processing approaches to minimize or suppress calibration time in oscillatory activity-based brain–computer interfaces," *Proceedings of the IEEE*, vol. 103, no. 6, pp. 871–890, 2015.

[35] S. Saha, K. I. U. Ahmed, R. Mostafa, L. Hadjileontiadis, and A. Khandoker, "Evidence of variabilities in EEG dynamics during motor imagery-based multiclass brain–computer interface," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 2, pp. 371–382, 2018.

[36] W. Tu and S. Sun, "A subject transfer framework for EEG classification," *Neurocomputing*, vol. 82, pp. 109–116, 2012.

[37] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[38] A. M. Azab, J. Toth, L. S. Mihaylova, and M. Arvaneh, "A review on transfer learning approaches in braincomputer interface," in *Signal Processing and Machine Learning for Brain-Machine Interfaces.* The Institution of Engineering and Technology (IET), 2018, ch. 5.

[39] G. Pfurtscheller, C. Neuper, C. Guger, W. Harkam, H. Ramoser, A. Schlogl, B. Obermaier, and M. Pregenzer, "Current trends in Graz brain-computer interface (bci) research," *IEEE Transactions on Rehabilitation Engineering*, vol. 8, no. 2, pp. 216–219, 2000.

[40] C. Wang, K. S. Phua, K. K. Ang, C. Guan, H. Zhang, R. Lin, K. S. G. Chua, B. T. Ang, and C. W. K. Kuah, "A feasibility study of non-invasive motor-imagery bci-based robotic rehabilitation for stroke patients," in *2009 4th International IEEE/EMBS Conference on Neural Engineering.* IEEE, 2009, pp. 271–274.

[41] S. M. Grigorescu, T. Lüth, C. Fragkopoulos, M. Cyriacks, and A. Gräser, "A bci-controlled robotic assistant for quadriplegic people in domestic and professional life," *Robotica*, vol. 30, no. 3, pp. 419–431, 2012.

[42] B. Dal Seno, M. Matteucci, and L. Mainardi, "Online detection of p300 and error potentials in a bci speller," *Computational intelligence and neuroscience*, vol. 2010, p. 11, 2010.

[43] L. F. Nicolas-Alonso and J. Gomez-Gil, "Brain computer interfaces, a review," *Sensors*, vol. 12, no. 2, pp. 1211–1279, 2012.

[44] M. Teplan *et al.*, “Fundamentals of eeg measurement,” *Measurement science review*, vol. 2, no. 2, pp. 1–11, 2002.

[45] J.-D. Haynes and G. Rees, “Neuroimaging: decoding mental states from brain activity in humans,” *Nature Reviews Neuroscience*, vol. 7, no. 7, p. 523, 2006.

[46] E. C. Leuthardt, G. Schalk, J. R. Wolpaw, J. G. Ojemann, and D. W. Moran, “A brain–computer interface using electrocorticographic signals in humans,” *Journal of neural engineering*, vol. 1, no. 2, p. 63, 2004.

[47] D. Millett, “Hans berger: From psychic energy to the EEG,” *Perspectives in biology and medicine*, vol. 44, no. 4, pp. 522–542, 2001.

[48] S. Baillet, J. C. Mosher, and R. M. Leahy, “Electromagnetic brain mapping,” *IEEE Signal processing magazine*, vol. 18, no. 6, pp. 14–30, 2001.

[49] A. Searle and L. Kirkup, “A direct comparison of wet, dry and insulating bioelectric recording electrodes,” *Physiological measurement*, vol. 21, no. 2, p. 271, 2000.

[50] G. H. Klem, H. O. Lüders, H. Jasper, C. Elger *et al.*, “The ten-twenty electrode system of the international federation,” *Electroencephalogr Clin Neurophysiol*, vol. 52, no. 3, pp. 3–6, 1999.

[51] M. Abo-Zahhad, S. M. Ahmed, and S. N. Abbas, “A new EEG acquisition protocol for biometric identification using eye blinking signals,” *International Journal of Intelligent Systems and Applications*, vol. 7, no. 6, p. 48, 2015.

[52] B. Burle, L. Spieser, C. Roger, L. Casini, T. Hasbroucq, and F. Vidal, “Spatial and temporal resolutions of eeg: Is it really black and white? a scalp current density view,” *International Journal of Psychophysiology*, vol. 97, no. 3, pp. 210–220, 2015.

[53] J. R. Wolpaw, G. E. Loeb, B. Z. Allison, E. Donchin, O. F. do Nascimento, W. J. Heetderks, F. Nijboer, W. G. Shain, and J. N. Turner, “Bci meeting 2005-workshop on signals and recording methods,” *IEEE Transactions on*

neural systems and rehabilitation engineering, vol. 14, no. 2, pp. 138–141, 2006.

[54] V. Jayaram, M. Alamgir, Y. Altun, B. Schölkopf, and M. Grosse-Wentrup, "Transfer Learning in Brain-Computer Interfaces," pp. 1–20, 2015. [Online]. Available: http://arxiv.org/abs/1512.00296{%}0Ahttp://dx.doi.org/10.1109/MCI.2015.2501545

[55] D. Wu, B. Lance, and V. Lawhern, "Transfer learning and active transfer learning for reducing calibration data in single-trial classification of visually-evoked potentials," in *IEEE International Conference on Systems, Man and Cybernetics (SMC), 2014*. IEEE, 2014, pp. 2801–2807.

[56] T. Picton, "Human brain electrophysiology. evoked potentials and evoked magnetic fields in science and medicine." *Journal of Clinical Neurophysiology*, vol. 7, no. 3, pp. 450–451, 1990.

[57] G. Garcia-Molina and D. Zhu, "Optimal spatial filtering for the steady state visual evoked potential: Bci application," in *Neural Engineering (NER), 2011 5th International IEEE/EMBS Conference on*. IEEE, 2011, pp. 156–160.

[58] G. R. Müller-Putz, R. Scherer, C. Brauneis, and G. Pfurtscheller, "Steady-state visual evoked potential (ssvep)-based communication: impact of harmonic frequency components," *Journal of neural engineering*, vol. 2, no. 4, p. 123, 2005.

[59] W. Yijun, W. Ruiping, G. Xiaorong, and G. Shangkai, "Brain-computer interface based on the high-frequency steady-state visual evoked potential," in *Neural Interface and Control, 2005. Proceedings. 2005 First International Conference on*. IEEE, 2005, pp. 37–39.

[60] M. Middendorf, G. McMillan, G. Calhoun, and K. S. Jones, "Brain-computer interfaces based on the steady-state visual-evoked response," *IEEE transactions on rehabilitation engineering*, vol. 8, no. 2, pp. 211–214, 2000.

[61] J. D. Bayliss, “Use of the evoked potential p3 component for control in a virtual apartment,” *IEEE transactions on neural systems and rehabilitation engineering*, vol. 11, no. 2, pp. 113–116, 2003.

[62] A. Rakotomamonjy and V. Guigue, “Bci competition iii: dataset ii-ensemble of svms for bci p300 speller,” *IEEE transactions on biomedical engineering*, vol. 55, no. 3, pp. 1147–1154, 2008.

[63] N. Birbaumer, A. Kubler, N. Ghanayim, T. Hinterberger, J. Perelmouter, J. Kaiser, I. Iversen, B. Kotchoubey, N. Neumann, and H. Flor, “The thought translation device (ttd) for completely paralyzed patients,” *IEEE Transactions on rehabilitation Engineering*, vol. 8, no. 2, pp. 190–193, 2000.

[64] G. Pfurtscheller, C. Brunner, A. Schlögl, and F. L. Da Silva, “Mu rhythm (de) synchronization and eeg single-trial classification of different motor imagery tasks,” *NeuroImage*, vol. 31, no. 1, pp. 153–159, 2006.

[65] N. Proesmans, “Reducing calibration time in Motor Imagery Brain-Computer Interfaces using Machine Learning,” 2015. [Online]. Available: http://lib.ugent.be/fulltxt/RUG01/002/300/488/RUG01-002300488{_}2016{_}0001{_}AC.pdf

[66] J. R. Wolpaw, D. J. McFarland, and T. M. Vaughan, “Brain-computer interface research at the wadsworth center,” *IEEE Transactions on Rehabilitation Engineering*, vol. 8, no. 2, pp. 222–226, 2000.

[67] H.-J. Hwang, K. Kwon, and C.-H. Im, “Neurofeedback-based motor imagery training for brain–computer interface (bci),” *Journal of neuroscience methods*, vol. 179, no. 1, pp. 150–156, 2009.

[68] G. Pfurtscheller, C. Neuper, G. Muller, B. Obermaier, G. Krausz, A. Schlogl, R. Scherer, B. Graimann, C. Keinrath, D. Skliris *et al.*, “Graz-bci: state of the art and clinical applications,” *IEEE Transactions on neural systems and rehabilitation engineering*, vol. 11, no. 2, pp. 1–4, 2003.

[69] A. Bashashati, M. Fatourechi, R. K. Ward, and G. E. Birch, “A survey of signal processing algorithms in brain–computer interfaces based on electrical brain signals,” *Journal of Neural engineering*, vol. 4, no. 2, p. R32, 2007.

[70] F. Lotte and C. Guan, "Learning from other subjects helps reducing brain-computer interface calibration time," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pp. 614–617, 2010.

[71] M. Arvaneh, C. Guan, K. K. Ang, and C. Quek, "EEG data space adaptation to reduce intersession nonstationarity in brain-computer interface," *Neural Computation*, vol. 25, no. 8, pp. 2146–2171, 2013.

[72] A. Subasi and M. I. Gursoy, "Eeg signal classification using pca, ica, lda and support vector machines," *Expert systems with applications*, vol. 37, no. 12, pp. 8659–8666, 2010.

[73] J. Pardey, S. Roberts, and L. Tarassenko, "A review of parametric modelling techniques for eeg analysis," *Medical engineering & physics*, vol. 18, no. 1, pp. 2–11, 1996.

[74] Y. M. Chi, Y.-T. Wang, Y. Wang, C. Maier, T.-P. Jung, and G. Cauwenberghs, "Dry and noncontact eeg sensors for mobile brain–computer interfaces," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 20, no. 2, pp. 228–235, 2012.

[75] W. Tu and S. Sun, "A subject transfer framework for EEG classification," *Neurocomputing*, vol. 82, pp. 109–116, 2012. [Online]. Available: http://dx.doi.org/10.1016/j.neucom.2011.10.024

[76] S. Fazli, F. Popescu, M. Danóczy, B. Blankertz, K. R. Müller, and C. Grozea, "Subject-independent mental state classification in single trials," *Neural Networks*, vol. 22, no. 9, pp. 1305–1312, 2009. [Online]. Available: http://dx.doi.org/10.1016/j.neunet.2009.06.003

[77] E. Baralis, S. Chiusano, and P. Garza, "A lazy approach to associative classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 2, pp. 156–171, 2008.

[78] X. Zhu, "Semi-supervised learning literature survey," *Computer Science, University of Wisconsin-Madison*, vol. 2, no. 3, p. 4, 2006.

[79] S. Thrun and L. Pratt, "Learning to learn: Introduction and overview," in *Learning to learn.* Springer, 1998, pp. 3–17.

[80] C. C. Aggarwal and C. Zhai, *Mining text data.* Springer Science & Business Media, 2012.

[81] A. Quattoni, M. Collins, and T. Darrell, "Transfer learning for image classification with sparse prototype representations," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on.* IEEE, 2008, pp. 1–8.

[82] S. J. Pan, V. W. Zheng, Q. Yang, and D. H. Hu, "Transfer learning for wifi-based indoor localization," in *Association for the advancement of artificial intelligence (AAAI) workshop*, 2008, p. 6.

[83] J. Lu, V. Behbood, P. Hao, H. Zuo, S. Xue, and G. Zhang, "Transfer learning using computational intelligence: a survey," *Knowledge-Based Systems*, vol. 80, pp. 14–23, 2015.

[84] K. Weiss, T. M. Khoshgoftaar, and D. Wang, *A survey of transfer learning.* Springer International Publishing, 2016, vol. 3, no. 1. [Online]. Available: http://journalofbigdata.springeropen.com/articles/10.1186/s40537-016-0043-6

[85] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu, "Boosting for transfer learning," in *Proceedings of the 24th international conference on Machine learning.* ACM, 2007, pp. 193–200.

[86] M. Alamgir, M. Grosse-Wentrup, and Y. Altun, "Multitask learning for brain-computer interfaces," *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS-10)*, vol. 9, pp. 17–24, 2010.

[87] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning: transfer learning from unlabeled data," in *Proceedings of the 24th international conference on Machine learning.* ACM, 2007, pp. 759–766.

[88] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big Data*, vol. 3, no. 1, p. 9, 2016.

[89] H. Shimodaira, "Improving predictive inference under covariate shift by weighting the log-likelihood function," *Journal of statistical planning and inference*, vol. 90, no. 2, pp. 227–244, 2000.

[90] X. Liao, Y. Xue, and L. Carin, "Logistic regression with an auxiliary data source," in *Proceedings of the 22nd international conference on Machine learning.* ACM, 2005, pp. 505–512.

[91] T. Evgeniou and M. Pontil, "Regularized multi–task learning," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM, 2004, pp. 109–117.

[92] L. Mihalkova and R. J. Mooney, "Transfer learning by mapping with minimal target data," 2008.

[93] W. Samek, F. C. Meinecke, and K.-R. Müller, "Transferring subspaces between subjects in brain–computer interfacing," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 8, pp. 2289–2298, 2013.

[94] P. Wang, J. Lu, B. Zhang, and Z. Tang, "A review on transfer learning for brain-computer interface classification," in *Information Science and Technology (ICIST), 2015 5th International Conference on.* IEEE, 2015, pp. 315–322.

[95] Y. Li, H. Kambara, Y. Koike, and M. Sugiyama, "Application of covariate shift adaptation techniques in brain–computer interfaces," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 6, pp. 1318–1324, 2010.

[96] M. Sugiyama, S. Nakajima, H. Kashima, P. V. Buenau, and M. Kawanabe, "Direct importance estimation with model selection and its application to covariate shift adaptation," in *Advances in neural information processing systems*, 2008, pp. 1433–1440.

[97] F. Abid, A. Hassan, A. Abid, M. Jochumsen, M. S. Navid, R. W. Nedergaard, and I. K. Niazi, "Transfer learning for electroencephalogram signals," in *International Conference on Computer and Electrical Engineering, ICCEE*, vol. 7, no. 3, 2017, pp. 143–152.

[98] T. Kanamori, S. Hido, and M. Sugiyama, “A least-squares approach to direct importance estimation,” *Journal of Machine Learning Research*, vol. 10, no. Jul, pp. 1391–1445, 2009.

[99] I. Hossain, A. Khosravi, and S. Nahavandhi, “Active transfer learning and selective instance transfer with active learning for motor imagery based BCI,” in *Proceedings of International Joint Conference on Neural Networks (IJCNN).* IEEE, 2016, pp. 4048–4055.

[100] B. Settles, “Active learning literature survey,” *University of Wisconsin, Madison*, vol. 52, no. 55-66, p. 11, 2010.

[101] E. Jeon, W. Ko, and H.-I. Suk, “Domain adaptation with source selection for motor-imagery based bci,” in *2019 7th International Winter Conference on Brain-Computer Interface (BCI).* IEEE, 2019, pp. 1–4.

[102] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K.-R. Muller, “Optimizing spatial filters for robust eeg single-trial analysis,” *IEEE Signal processing magazine*, vol. 25, no. 1, pp. 41–56, 2008.

[103] G. Schalk, D. J. McFarland, T. Hinterberger, N. Birbaumer, and J. R. Wolpaw, “Bci2000: a general-purpose brain-computer interface (bci) system,” *IEEE Transactions on biomedical engineering*, vol. 51, no. 6, pp. 1034–1043, 2004.

[104] F. Lotte and C. Guan, “Learning from other subjects helps reducing brain-computer interface calibration time,” in *In proceedings of IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), 2010.* IEEE, 2010, pp. 614–617.

[105] H. Kang, Y. Nam, and S. Choi, “Composite common spatial pattern for subject-to-subject transfer,” *IEEE Signal Processing Letters*, vol. 16, no. 8, pp. 683–686, 2009.

[106] D. Devlaminck, B. Wyns, M. Grosse-Wentrup, G. Otte, and P. Santens, “Multisubject learning for common spatial patterns in motor-imagery BCI,” *Computational Intelligence and Neuroscience*, vol. 2011, p. 8, 2011.

[107] W. Samek, M. Kawanabe, and K.-R. Müller, “Divergence-based framework for common spatial patterns algorithms,” *IEEE Reviews in Biomedical Engineering*, vol. 7, pp. 50–72, 2014.

[108] M. Spüler, W. Rosenstiel, and M. Bogdan, “Principal component based covariate shift adaption to reduce non-stationarity in a meg-based brain-computer interface,” *EURASIP Journal on Advances in Signal Processing*, vol. 2012, no. 1, p. 129, 2012.

[109] M. Sugiyama, M. Krauledat, and K.-R. MÃžller, “Covariate shift adaptation by importance weighted cross validation,” *Journal of Machine Learning Research*, vol. 8, no. May, pp. 985–1005, 2007.

[110] H. Raza, H. Cecotti, Y. Li, and G. Prasad, “Adaptive learning with covariate shift-detection for motor imagery-based brain–computer interface,” *Soft Computing*, vol. 20, no. 8, pp. 3085–3096, 2016.

[111] H. Lee, A. Cichocki, and S. Choi, “Nonnegative matrix factorization for motor imagery eeg classification,” *Artificial Neural Networks–ICANN 2006*, pp. 250–259, 2006.

[112] H. Lee and S. Choi, “Group nonnegative matrix factorization for eeg classification,” in *Artificial Intelligence and Statistics*, 2009, pp. 320–327.

[113] A. Bamdadian, C. Guan, K. K. Ang, and J. Xu, “Improving session-to-session transfer performance of motor imagery-based bci using adaptive extreme learning machine,” in *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE.* IEEE, 2013, pp. 2188–2191.

[114] L. Bruzzone and M. Marconcini, “Domain adaptation problems: A dasvm classification technique and a circular validation strategy,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 5, pp. 770–787, 2010.

[115] D. Wu, V. J. Lawhern, and B. J. Lance, “Reducing offline bci calibration effort using weighted adaptation regularization with source domain selection,” in *Systems, Man, and Cybernetics (SMC), 2015 IEEE International Conference on.* IEEE, 2015, pp. 3209–3216.

[116] S. Dalhoumi, G. Dray, J. Montmain, G. Derosière, and S. Perrey, "An adaptive accuracy-weighted ensemble for inter-subjects classification in brain-computer interfacing," in *Neural Engineering (NER), 2015 7th International IEEE/EMBS Conference on.* IEEE, 2015, pp. 126–129.

[117] N. R. Waytowich, V. J. Lawhern, A. W. Bohannon, K. R. Ball, and B. J. Lance, "Spectral transfer learning using information geometry for a user-independent brain-computer interface," *Frontiers in neuroscience*, vol. 10, 2016.

[118] P. Zanini, M. Congedo, C. Jutten, S. Said, and Y. Berthoumieu, "Transfer learning: a riemannian geometry framework with applications to brain-computer interfaces," *IEEE Transactions on Biomedical Engineering*, 2017.

[119] N. R. Waytowich, J. Faller, J. O. Garcia, J. M. Vettel, and P. Sajda, "Unsupervised adaptive transfer learning for steady-state visual evoked potential brain-computer interfaces," in *IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2016.*

[120] H. He and D. Wu, "Transfer learning for brain-computer interfaces: A euclidean space data alignment approach," *IEEE Transactions on Biomedical Engineering*, 2019.

[121] M. Arvaneh, I. Robertson, and T. E. Ward, "Subject-to-subject adaptation to reduce calibration time in motor imagery-based brain-computer interface," in *36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Aug 2014, pp. 6501–6504.

[122] V. Jayaram, M. Alamgir, Y. Altun, B. Scholkopf, and M. Grosse-Wentrup, "Transfer learning in brain-computer interfaces," *IEEE Computational Intelligence Magazine*, vol. 11, no. 1, pp. 20–31, 2016.

[123] K.-H. Fiebig, V. Jayaram, J. Peters, and M. Grosse-Wentrup, "Multi-task logistic regression in brain-computer interfaces," in *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics (SMC).* IEEE, 2016, pp. 002 307–002 312.

[124] F. Lotte and C. Guan, “Regularizing common spatial patterns to improve bci designs: Unified theory and new algorithms,” *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 2, pp. 355–362, 2011.

[125] H. Ramoser, J. Muller-Gerking, and G. Pfurtscheller, “Optimal spatial filtering of single trial EEG during imagined hand movement,” *IEEE Transactions on Rehabilitation Engineering*, vol. 8, no. 4, pp. 441–446, 2000.

[126] R. Tomioka and S. Lemm, “Filters for Robust EEG,” *IEEE Signal Processing Magazine*, no. January 2008, pp. 41–56, 2008.

[127] J. M. Hilbe, *Logistic regression models.* Chapman and hall/CRC, 2009.

[128] M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, 1952, vol. 49, no. 1, PP 409–436.

[129] I. Iturrate, L. Montesano, and J. Minguez, “Task-dependent signal variations in EEG error-related potentials for brain–computer interfaces,” *Journal of Neural Engineering*, vol. 10, no. 2, p. 026024, 2013.

[130] C. Brunner, R. Leeb, G. Müller-Putz, A. Schlögl, and G. Pfurtscheller, “BCI competition 2008–Graz data set A,” *Institute for Knowledge Discovery (Laboratory of Brain-Computer Interfaces), Graz University of Technology*, 2008.

[131] C. Vidaurre, M. Kawanabe, B. Blankertz, K. Müller *et al.*, “Toward unsupervised adaptation of LDA for brain-computer interfaces.” *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 3, pp. 587–597, 2011.

[132] C. Vidaurre, A. Schlogl, R. Cabeza, R. Scherer, and G. Pfurtscheller, “Study of on-line adaptive discriminant analysis for EEG-based brain computer interfaces,” *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 3, pp. 550–556, 2007.

[133] P. Shenoy, M. Krauledat, B. Blankertz, R. P. Rao, and K.-R. Müller, “Towards adaptive classification for BCI,” *Journal of Neural Engineering*, vol. 3, no. 1, p. R13, 2006.

[134] S. Fazli, F. Popescu, M. Danóczy, B. Blankertz, K.-R. Müller, and C. Grozea, “Subject-independent mental state classification in single trials,” *Neural networks*, vol. 22, no. 9, pp. 1305–1312, 2009.

[135] C. M. Bishop, *Pattern recognition and machine learning.* springer, 2006.

[136] S. Shalev-Shwartz and A. Tewari, “Stochastic methods for l1-regularized loss minimization,” *Journal of Machine Learning Research*, vol. 12, pp. 1865–1892, 2011.

[137] C. Robert, “Machine learning, a probabilistic perspective,” 2014, Taylor & Francis.

[138] J. Shore and R. Johnson, “Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy,” *IEEE Transactions on Information Theory*, vol. 26, no. 1, pp. 26–37, 1980.

[139] B.-C. Kuo and D. A. Landgrebe, “Nonparametric weighted feature extraction for classification,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 5, pp. 1096–1105, 2004.

[140] M. Arvaneh, C. Guan, K. K. Ang, T. E. Ward, K. S. Chua, C. W. K. Kuah, G. J. E. Joseph, K. S. Phua, and C. Wang, “Facilitating motor imagery-based brain–computer interface for stroke patients using passive movement,” *Neural Computing and Applications*, vol. 28, no. 11, pp. 3259–3272, 2017.

[141] M. Naeem, C. Brunner, R. Leeb, B. Graimann, and G. Pfurtscheller, “Seperability of four-class motor imagery data using independent components analysis,” *Journal of neural engineering*, vol. 3, no. 3, p. 208, 2006.

[142] G. Dornhege, B. Blankertz, G. Curio, and K.-R. Muller, “Boosting bit rates in noninvasive EEG single-trial classifications by feature combination and multiclass paradigms,” *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 6, pp. 993–1002, 2004.

[143] C. S. Nam, A. Nijholt, and F. Lotte, *Brain–computer interfaces handbook: technological and theoretical advances.* CRC Press, 2018.

[144] Y. Guo, “Regularized discriminant analysis and its application in microarrays,” *Biostatistics*, vol. 1, no. 1, pp. 1–18, 2005.

[145] X. Yong, R. K. Ward, and G. E. Birch, “Robust common spatial patterns for eeg signal preprocessing,” in *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society.* IEEE, 2008, pp. 2087–2090.

[146] W. Holmes, *Speech synthesis and recognition.* CRC press, 2001.

[147] W. A. Chaovalitwongse, Y.-J. Fan, and R. C. Sachdeo, “On the time series k-nearest neighbor classification of abnormal brain activity,” *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 37, no. 6, pp. 1005–1016, 2007.

[148] P. Senin, “Dynamic time warping algorithm review,” in *Progress in brain research.* Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA., 2008, vol. 885, no. 1-23, p. 40.

[149] A. M. Azab, L. Mihaylova, H. Ahmadi, and M. Arvaneh, “Robust common spatial patterns estimation using dynamic time warping to improve bci systems,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2019, pp. 3897–3901.

[150] J. H. Friedman, “Regularized discriminant analysis,” *Journal of the American statistical association*, vol. 84, no. 405, pp. 165–175, 1989.

[151] F. Lotte and C. Guan, “Spatially regularized common spatial patterns for EEG classification,” in *20th International Conference on Pattern Recognition.* IEEE, 2010, pp. 3712–3715.

[152] W. Samek, C. Vidaurre, K.-R. Müller, and M. Kawanabe, “Stationary common spatial patterns for brain–computer interfacing,” *Journal of Neural Engineering*, vol. 9, no. 2, p. 026013, 2012.

[153] H. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, “Regularized common spatial patterns with generic learning for EEG signal classification,” in *Annual International Conference of the IEEE,Engineering in Medicine and Biology Society, EMBC 2009.* IEEE, 2009, pp. 6599–6602.

[154] A. M. Azab, L. Mihaylova, H. Ahmadi, and M. Arvaneh, "Robust common spatial patterns estimation using dynamic time warping to improve bci systems," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 3897–3901.

[155] F. Yger, M. Berar, and F. Lotte, "Riemannian approaches in brain-computer interfaces: a review," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 10, pp. 1753–1762, 2016.

[156] A. Meinel, F. Lotte, and M. Tangermann, "Tikhonov regularization enhances eeg-based spatial filtering for single trial regression," 2017.

[157] A. Meinel, S. Castaño-Candamil, B. Blankertz, F. Lotte, and M. Tangermann, "Characterizing regularization techniques for spatial filter optimization in oscillatory eeg regression problems," *Neuroinformatics*, vol. 17, no. 2, pp. 235–251, 2019.

[158] C. Jeunet, F. Lotte, J.-M. Batail, P. Philip, and J.-A. M. Franchi, "Using recent bci literature to deepen our understanding of clinical neurofeedback: A short review," *Neuroscience*, vol. 378, pp. 225–233, 2018.

[159] F. Lotte and C. Jeunet, "Defining and quantifying users mental imagery-based bci skills: a first step," *Journal of neural engineering*, vol. 15, no. 4, p. 046030, 2018.

[160] A. Rezeika, M. Benda, P. Stawicki, F. Gembler, A. Saboor, and I. Volosyak, "Brain–computer interface spellers: A review," *Brain sciences*, vol. 8, no. 4, p. 57, 2018.

[161] H. K. AlJobouri and F. E. Ali, "Brain-computer interface based on vep and fmri package." *American Journal of Biomedical Sciences*, vol. 11, no. 1, 2019.

[162] P. L. C. Rodrigues, C. Jutten, and M. Congedo, "Riemannian procrustes analysis: Transfer learning for brain-computer interfaces," *IEEE Transactions on Biomedical Engineering*, 2018.

[163] P. Gaur, R. B. Pachori, H. Wang, and G. Prasad, “A multi-class eeg-based bci classification using multivariate empirical mode decomposition based filtering and riemannian geometry,” *Expert Systems with Applications*, vol. 95, pp. 201–211, 2018.

[164] I. Hossain, A. Khosravi, I. Hettiarachchi, and S. Nahavandi, “Multiclass informative instance transfer learning framework for motor imagery-based brain-computer interface,” *Computational intelligence and neuroscience*, vol. 2018, 2018.

[165] F. Lotte, L. Bougrain, A. Cichocki, M. Clerc, M. Congedo, A. Rakotomamonjy, and F. Yger, “A review of classification algorithms for eeg-based brain–computer interfaces: a 10 year update,” *Journal of neural engineering*, vol. 15, no. 3, p. 031005, 2018.