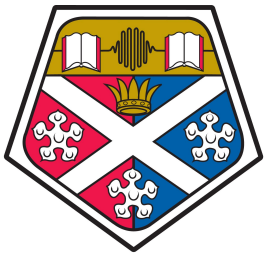

University of
Strathclyde

Application of Dynamic Mode Decomposition in MI-BCI

Lukasz Zapotoczny

September 28, 2022

Declaration

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

Signed: Lukasz Zapotoczny

Date: September 28, 2022

Dedication

In Memory of Brownisław Zapotoczny

To my beloved grandfather. Seeing you always tinker with something inspired me to follow your steps. We shared the joy together when I completed my undergraduate studies, and now, even though you are no longer here I know you would still be very proud of me.

Acknowledgement

I would like to express special thanks to my two PhD supervisors Prof Bernard Conway and Dr Gaetano Di Caterina who have stepped in and filled out the said positions amid some experienced difficulties. I am grateful for the numerous chats and discussions which challenged me and my perspective on research but in the end greatly helped me and allowed me to stay on track leading to finishing up my PhD studies.

I would also like to extend my gratitude to Dr Danial Kahani for helping me along my PhD journey. You were there from the very beginning, guiding me, discussing various topics (PhD and non-PhD related) and offering a friendly environment. Numerous brain-storming sessions and long conversations helped me formulate my research and pursue it in the right direction.

Additionally, I would like to offer thanks to Prof John Soraghan and Dr Lykourgos Petropoulakis for providing sound advice during some really challenging periods of my PhD and help me narrow down and correctly tackle my research. Your help during those times was very valuable and deeply appreciated.

Words cannot describe how grateful I am for my lovely girlfriend Liya Li who was there to encourage me in the lowest moments when I wanted to give up. Your presence and care helped me relax on many occasions and the loving atmosphere pushed me gently to finally finish this PhD journey. I know you will be grateful too for not having to listen to me complaining and having tenth ”PhD talk” this week.

I would not have been able to finish my PhD if it was not for my best friends Paul

Kirkland, Andreas Aßmann and Maria Urian who have listened to my almost non-stop moaning, groaning about my research and thesis writing. I thank you for your patience with me. I am happy that now I can finally join the cool doctorate group with Paul and Andy even though I am slightly late to the party.

Abstract

Brain-computer interfaces (BCIs) are a fruit of an impressive and long collaboration between fields of neuroscience and signal processing. The purpose of these complex systems is to interpret measured brain activity into useable commands and actions through implementation of different feature extraction, selection, and classification techniques. Depending on the application of a BCI and the exploited type of brain activity a specific set of methods would be implemented. In this thesis, electroencephalography (EEG) signals containing motor imagery (MI) information are analysed using a spatial-temporal technique called dynamic mode decomposition (DMD).

MI-EEG signals can be mainly described through two different types of brain activities: event-related de-/synchronisation (ERD/S) and event-related potentials (ERPs). The studies covered in this thesis focus on the former activity which exhibits strong characteristics in temporal, spectral and spatial domains. Despite being well described in the aforementioned domains, current state-of-the-art feature extraction techniques focus either on spectral (power spectral density (PSD) or bandpower) or spatial (common spatial patterns, CSP) side or on the combination of temporal and spectral domains (spectrograms and scalograms).

The introduction of DMD aims to address the lack of more spatial-oriented techniques and three different feature types were explored to extract features based on ERD/S phenomenon. Firstly, standard DMD modes are used to accomplish that task. The measured performance, while being relatively low, still provided valuable informa-

tion into the correct processing routes for DMD technique. With this knowledge, novel DMD spectrum features were extracted to cover a spatial-spectral domain combination. Despite the literature's suggestions and links of DMD spectrum to average Fast-Fourier transfroms (FFTs), the perceived performance clearly indicated that DMD spectrum is unfit to extract ERD/S features from MI-EEG signals. Lastly, novel implementation of DMD maps with convolutional neural network (CNN) aimed to fully exploit spatial characteristics of ERD/S phenomenon was not able to successfully do so. Even though all three proposed hypotheses were rejected based on the evidence seen from classification accuracy and kappa values, the author argues that DMD technique is still at the early stages of development and, given time and enough research, the performance of DMD modes and maps can be greatly improved.

Contents

Acronyms

BCI, brain-computer interface

BP, bereitschaftspotential

BSS, blind source separation

CAR, common average reference

CNN, convolutional neural network

CSP, common spatial pattern

CWT, continuous wavelet transform

DFT, discrete Fourier transform

DMD, dynamic mode decomposition

DWT, discrete wavelet transform

ECoG, electrocorticogram

EEG, electroencephalogram

EMD, empirical mode decomposition

EMG, electromyography

EOG, electrooculography

ERA, eigensystem realization algorithm

ERD/S, event-related de-/synchronisation

ERN, error-related negativity

ERP, event-related potentials

ErrP, error-related potential

FBCSP, filter-bank common spatial pattern

FFT, fast Fourier transform

FN, false negative

FP, false positive

ICA, independent component analysis

iEEG, intracranial electroencephalography

IMF, intrinsic mode function

JAD, joint-approximate diagonalization

JADE, join-approximation diagonalization of eigenmatrices

kNN, k-nearest neighbour

LDA, linear discriminant analysis

LFP, local field potential

LSTM, long short-term memory

MEG, magnetoencephalogram

MI, motor-imagery

MI-BCI, motor-imagery brain-computer interface

MIBFS, mutual information-based feature selection

MInf, mutual information

MI-EEG, motor-imagery electroencephalography

MLP, multi-layer perceptron

MP, motor potential

MRCP, motor-related cortical potentials

MRMR, minimal-redundancy maximum-relevance

NN, neural network

NS', negative slope

PC, principal component

PCA, principal component analysis

PMP, pre-motion positivity

PSD, power spectral density

QDA, quadratic discriminant analysis

RAP, reafferente Potentiale

RBF, radial basis function

RLDA, regularized linear discriminant analysis

ReLU, rectified linear unit

RNN, recurrent neural network

SBS, sequential backward selection

sEEG, surface electroencephalography

SFS, sequential forward selection

SL, surface Laplacian

SMR, sensorimotor rhythm

SNR, signal-to-noise ratio

SOBI, second order blind identification

SPD, symmetric positive definite

SSVEP, steady-state visual evoked potential

STFT, short-time Fourier transform

SVD, singular value decomposition

SVM, support vector machines

TN, true negative

TP, true positive

List of Figures

List of Tables

Chapter 1

Introduction

The concept of being able to move something without physically interacting with it, but by purely using the '*power of thought*', has been explored in both: fictional works as well as in the scientific research. While fiction adopted the idea of *psychokinesis* and attributed this skill to some 'superhero' powers or other supernatural events, it also explored more scientifically plausible ideas such as implementing some kind of interface which would be able to translate one's thoughts into some system-specific outputs or actions.

The most notable example of the latter is seen in a popular comic series "Spider-Man". There, one of the villains and enemies of Spider-Man, Dr Otto Octavius constructed a wearable harness with four tentacle-like arms which was controlled through a computer chip embedded into the brain of the user. The user would then be able to move the extra limbs through the same processes involved while moving 'regular' arms, as the embedded chip would be able to recognise the correct brain signals from the user and interpret them into correct actions. Interestingly, the character of Dr Octavius and his 'machine interface' was introduced in 1963, which preceded the most impactful introduction of the idea of motor imagery by several years (Richardson, 1967, 1969); and almost a decade before the first scientific publication coining the term of *brain-computer interface* (BCI) (Vidal, 1973), a system which is able to translate brain signals into usable and interpretable commands by a computer. Since then the idea of such sophisticated control of external robotic arms has been greatly explored in research, although at much smaller scale mostly looking at the control of a single robotic arm

(Hochberg et al., 2012; McMullen et al., 2014; Hortal et al., 2015; Meng et al., 2016). Besides the idea of extending brain control into additional external limbs, BCIs have been put to a great use in other applications, such as spellers and virtual keyboards (Farwell & Donchin, 1988; Pinegger et al., 2017; Riccio et al., 2013; Cecotti & Gräser, 2011), mobile robots and wheelchairs (Millán et al., 2004; Galán et al., 2008). There, they have created new means of communication and interaction for people who suffered from a range of injuries or ailments which impaired the functionality of muscles, spinal cord or brain.

A high-level representation of a typical BCI system reveals three main building blocks: a signal acquisition block, a signal processing block, and the output or application block. The signal acquisition block only concerns the method used on how and what signals are recorded from the brain. Those signals can be either acquired through invasive (needs physical access to the brain and thus requires surgical procedure) or non-invasive (no need for any surgical procedures) methods (Buzsáki et al., 2012). The two most prominent signals used in BCIs are *electroencephalogram* (EEG, non-invasive) and *electrocorticogram* (ECoG, invasive) signals; however it must be noted that these are not the only means through which brain activity can be measured. The appropriate modality from the aforementioned signals is chosen depending on the desired application of the BCI. This could cover investigating rhythmic phenomena found in continuous EEG, such as *event-related de-/synchronisation* (ERD/S) (Pfurtscheller & Lopes Da Silva, 1999), or specific signals appearing in brain when the user is subjected to an external stimulus, such as *P300* (Riccio et al., 2013) or *motor-related cortical potentials* (MRCPs) (Shibasaki & Hallett, 2006).

After the signal acquisition block, the recorded brain signals are transformed by various techniques employed in the signal processing block which comprises of few internal submodules. These can be a selection of the following components: preprocessing, feature extraction, feature selection and dimensionality reduction. Preprocessing steps are often necessary when working with EEG signals, as these are known to have low signal-to-noise ratio and therefore preprocessing requires a combination of frequency-domain and spatial filtering techniques to increase the quality of the recorded signal. Depending on what brain modality is investigated, raw EEG data often does not carry

enough class-separable information in itself and additional computing methods must be employed to extract more descriptive features from the original data. One will see here popular techniques such as Fourier transforms, short-time Fourier transforms, principal component analysis, common spatial patterns, as well as other standard statistical methods being used for that purpose. Some of these methods are also used in the next module, where the best performing features are extracted and used as the final feature set before training a classifier. Quite often the feature selection module integrates dimensionality reduction techniques which help lower the complexity of subsequent classification of the transformed signals. The author acknowledges that 'feature selection' and 'dimensionality reduction' terms might have separate meanings in other fields; however in this thesis these two terms are grouped together and are used interchangeably throughout the thesis but are always clarified when used.

Processed features are then ready to be utilised accordingly in the output block of BCI. The essential component of this block is a classifier which is trained on the supplied feature set, to learn how to separate between different classes present in the data, allowing it to later classify any new incoming signals appropriately. A number of different methods are used for signal classification: from classical discriminant techniques and support vector machines (Subasi & Gursoy, 2010; Bhattacharyya et al., 2010), to modern machine learning approaches utilising the power of neural networks (Sakhavi et al., 2015; Lawhern et al., 2018). Once the brain signals are processed accordingly, the outputs of the classifier can be used to drive the end application. As mentioned earlier that might be a mouse cursor, keyboard commands, control commands for a mobile robot, wheelchair or a robotic arm.

However, not every BCI seen in the research must necessarily end with a hardware-based application. Substantial amount of BCI research focuses on developing novel methods for discerning some of the more demanding brain modalities and improving the quality of the initial signal at the acquisition stage. In such cases the outputs of the classifier are not translated into specific commands controlling hardware, but instead focus on improving the classification accuracy of the investigated brain signals. In the case of the signal acquisition block, the improvements are split between hardware- and neurophysiology-based. Here, some of the research concerns development of new

electrodes and recording systems, which for example might include the inclusion of in-built amplification and filtering circuitry at individual electrode level (Müller-Putz, 2020). From the neurophysiological perspective, advancements can be made in gaining better understanding of the organisation and structure of neurons in the brain and the dynamics involved during various brain processes (Cohen, 2017). In the signal processing and output (classifier only) blocks the innovation is purely software based and revolves around the development of new algorithms, with the aim of discovering new methods which would improve the overall performance of the BCI system. This improvement can be measured in the accuracy of the classification or the rate, or speed, of response at which the BCI can operate.

The work presented in this thesis focuses on the signal processing block by introducing a novel application of a processing technique called *dynamic mode decomposition*, which is used to extract features from motor imagery EEG signals, by exploiting the spectral and spatial nature of *event-related de-/synchronisation* (ERD/S) phenomenon. Motor imagery is a profound type of activity found in the brain, which carries valuable information addressing the idea of moving objects with the 'power of thought'. In short, motor imagery manifests itself over the sensorimotor area of the brain when an individual voluntarily imagines motor movement of a limb. Most commonly, this activity can be observed both in continuous (rhythmic) EEG and event-related potentials (ERPs) appearing at specific individual electrodes. In the case of continuous EEG, motor imagery is well defined in the spectral domain, where it occupies α and β bands of brain activity at 8-13Hz and 13-30Hz respectively (Pfurtscheller et al., 1997; Blankertz et al., 2007), and is attributed to the ERD/S phenomenon which occurs on the contralateral side of the brain with respect to the imagined movement, i.e. if an individual imagined moving right hand, ERD/S would manifests on the left side of the sensorimotor cortex. ERPs and more specifically *motor-related cortical potentials* (MRCPs) follow similar spatial distribution of signals; however, they have much better defined temporal characteristics than spectral ones: a sharp decrease of electric potential during imagination of movement.

Since motor imagery can be observed in temporal, spectral and/or spatial domains it comes as no surprise that numerous methods addressing a single or a combination of

the aforementioned domains were transferred over from other disciplines, or were specifically developed to extract features used for later classification of different imaginations of movement. Temporal domain mostly relies on the use of classical statistical analysis, such as averaging or kurtosis analysis; while spectral methods utilise fast Fourier transforms (FFTs), power spectral density (PSD) analysis or, most frequently, average bandpower in α and β frequency bands to observe changes in frequency. However, only in the last 15 years spatial domain started receiving more attention from the research community, which produced one of the most widely used methods nowadays: *common spatial patterns* (CSP). Despite the excellent performance of features extracted with CSP-based methods and their good interpretability of ERD/S phenomenon (Blankertz et al., 2007), no new spatial methods were developed nor implemented after CSP, apart from some derivations of CSP. While it might be argued that the recent rise in the use of convolutional neural networks (CNNs) with spectro-temporal maps such as the ones obtained from spectrograms or scalograms eliminates the need for spatial techniques, since CNN looks for spatial relations between pixels in the supplied image data, the author postulates that such maps do not reflect original spatial relations between electrodes, since each spectrogram or scalogram is a time-frequency map for a single electrode. Therefore, the author identified a clear gap and an opportunity to introduce a novel application of a spatial technique used for extracting motor imagery features from EEG signals.

The proposed Dynamic Mode Decomposition (DMD) technique was originally introduced in the study of fluid flow and its non-linear dynamics, where it was used to successfully extract and investigate spatial-temporal patterns emerging from the incoming data (Schmid, 2010). One of the noteworthy features of DMD is its ability to describe the non-linear dynamics of the system without explicitly constructing sets of equations explaining different dynamics present. DMD achieves that by stating that the future state of the system can be approximated via linear transformation of present state and some linear operator. Conventionally, this linear operator would take the form of a tall matrix containing descriptions of infinitely long simultaneous equations, attempting to describe all possible non-linear dynamics present as it is seen in the case of Koopman analysis when computing Koopman operator. DMD mitigates this by di-

rectly approximating the Koopman operator based solely on the available data from the system, greatly decreasing computational time. The resultant low-rank matrix approximation is further exploited through eigendecomposition to obtain its eigenvalues and eigenvectors, which allow to calculate *DMD modes*: a combination of spatial information from PCA decomposition and temporal information extracted by DFT which, when combined, describes the relative influence of each channel of multi-variate data at specific characteristic frequencies dictated by relative eigenvalues. The temporal information obtained from DFT allows to asses modes' temporal evolution, i.e. the growth/decay rates of each mode over time. It is therefore clear that all the possible modalities of DMD should provide a substantial amount of valuable information compared to other spatial methods such as PCA, ICA and CSP, which lack any temporal information. Therefore, in the author's view, it is clear that DMD has the potential to fill the gap in providing valuable spatial-based features for motor imagery problems in EEG signals.

1.1 Contributions

DMD has been gathering momentum in fluid flow research, however it remained mostly unnoticed in other research areas, especially the ones concerning EEG and motor imagery. The author notes that there is only a handful of academic publications from the last five years which use DMD in the context of analysing brain signals, with majority of the publications using DMD for seizure detection (Brunton et al., 2016; Solaija et al., 2018; Seo et al., 2020; Shiraishi et al., 2020; Takeishi et al., 2021). Given that DMD produces spatial-temporal modes and previous research has shown that spatial-based features are well-suited for exploiting ERD/S phenomenon during motor imagery, to address those two matters the author proposes a BCI system based on novel spatial-based features extracted from different representations of DMD modes.

The author decided on exploiting the following three representations of DMD modes: set of DMD mode vectors, DMD spectrum and DMD maps. Each of the representations has been thoroughly tested on three publicly available datasets on motor imagery with an addition of dataset available from the previous studies at the local laboratory. The

contributions of this thesis are summarised as follows:

- First documented application of DMD technique to EEG recordings concerning motor imagery to extract novel spatial-temporal features and their subsequent classification, fully investigating and assessing the performance of DMD-based features.

- Complete investigation of the effect of spatial filtering and different scaling methods on the obtained DMD modes — original data used in the study was only filtered using band-pass filter without any spatial filters applied (Brunner et al., 2008; Cho et al., 2017; Ofner et al., 2017). As part of the research, common average reference method was implemented to investigate the effect of spatial filters on DMD modes. Following that, DMD modes were either scaled naturally or scaled by their SVD energy. Finally, DMD modes were translated into more appropriate format for RBF-SVM classifier with the help of two different projection methods, a recently developed Grassmanian projection kernel and PCA. The performance of every possible combination has been assessed providing metrics such as accuracy of the classification, specificity, sensitivity and kappa value.

- Investigation of DMD spectrum features — the power spectrum obtained from the extracted DMD modes was assessed to see if this method produced viable features for investigating ERD/S phenomenon in the frequency domain similarly to PSD or average bandpower. The usefulness of the DMD spectrum features was measured by the same accuracy, specificity, sensitivity and kappa metrics.

- A novel utilisation of DMD modes through DMD maps — an intensity map representing the absolute values and phase of DMD modes was used for the first time as an input to a convolutional neural network to create a novel image processing pipeline. The usefulness of the DMD maps was measured by the same accuracy, specificity, sensitivity and kappa metrics as the other two DMD approaches.

1.2 Structure of the thesis

This thesis contains six chapters in total. The current introduction chapter is succeeded by a background chapter explaining the processes and signals involved in the brain which provide modalities to drive BCIs. The third chapter provides a comprehensive overview of the current state-of-the-art signal processing tools involved in a regular BCI system i.e. preprocessing, feature extraction, selection and classification methods. Following this, the fourth chapter describes the proposed pipeline for the new MI-BCI and the experiments performed, which allowed assessing the performance of the proposed BCI. Furthermore all the datasets which have been used in the thesis are also described. Chapter five presents the results obtained from aforementioned experiments. These results are discussed and compared to the performance of the state-of-the-art systems. Lastly, chapter six recalls the initial statements and aims of the thesis and provides a conclusion to the thesis. Additionally, future work suggestions are provided with the intent of helping further research.

Chapter 2

Physiological background of brain signals

In the introduction chapter BCIs were described as complex systems constructed with several modules, whose purpose is to extract meaningful information from the measured brain signals and translate them into correct commands through some form of signal classification technique. This chapter will focus on expanding upon the background regarding type of brain signals generally used in BCIs and the chosen signal modality used in this thesis. The signals acquired from the brain relate to a wide range of activities and processes found in the brain; some reflect reaction to auditory or visual stimulus, while others measure ongoing concentration. The background presented in this section will be primarily looking at brain signals and their smaller subdivisions related to motor execution and imagination activity which is exploited in *motor-imagery* (MI)-BCIs. However, some other signals will be briefly covered too for the sake of completeness.

2.1 Brain activity signals

Measuring brain activity with the purpose of analysing it for motor execution or imagination is a complicated process. After all, it is estimated that human brain contains 100 billion minuscule neurons (0.004-0.1mm in size), where every neuron has approximately 10,000 connections to other neurons (Müller-Putz, 2020). Thus it is

evident that measuring activity of a single neuron or even a smaller population of neurons is an incredibly demanding task. Complicating matters further, the layers of dura, skull and scalp protecting the brain make it harder to measure the activity originating from the network of neurons non-invasively.

Nonetheless, throughout the decades of research concerning BCIs, numerous methods were devised to record brain activity. These methods can be divided into two distinctive groups: electrophysiological- and haemodynamic-based methods (Shih et al., 2012). Electrophysiological (electrical and magnetic) signals are directly related to the neuronal activity as they observe changes of electric potential in the brain during neuronal firing; while haemodynamic (metabolic) signals investigate blood flow changes in the veins which occur as the result of neuronal firing.

2.1.1 Electrophysiological signals

Electrophysiological signals are generally preferred as they are easier and cheaper to collect. The most popular types of such signals are *surface electroencephalogram* (sEEG or simply EEG), *electrocorticogram* (ECoG), *intracranial EEG* (iEEG) (Hermes & Miller, 2020) - also known as micro-depth EEG or *local field potential* (LFP) - and lastly *magnetoencephalogram*. It must be noted that all the aforementioned signals in fact refer to the same biophysical process taking place in the brain - LFPs (Buzsáki et al., 2012; Heldman & Moran, 2020), with the difference between being the physical depth of LFP acquisition.

In essence, an electrical current is generated as neurons exchange information between each other in a subnetwork, leading to extracellular potential changes V_e, which, if superimposed, will generate LFPs. Coincidentally, measuring electric potential with small-sized electrode in the brain is also called LFP, and this melapropism is highlighted in academic publications (Buzsáki et al., 2012). Hence to avoid confusion in this thesis, those type of measurements are referred to as iEEG. While iEEG measures LFPs in the cortex, ECoG measures LFPs on the cortical surface using a subdural grid electrodes, and EEG measures LFPs from the scalp with appropriate electrodes (disc, ring or pin electrodes), thus explaining the ”physical depth” or ”level” of signal acquisition. magnetoencephalogram recordings are a special case of measuring LFPs,

as they exploit the magnetic field produced by LFPs.

It can be seen from the above descriptions that neurophysiological signals can be further divided into *invasive* and *non-invasive* methods, with iEEG and ECoG falling into the former category, while EEG and magnetoencephalogram are part of the latter. Invasive methods require direct access to the brain, which can only be achieved through surgical means. The need to implant a micro-array of electrodes into the cortex, as it is the case with iEEG, or a grid of electrodes on the surface of the brain, as seen with ECoG, carries both short- and long-term risks. In the short-time scale, the most obvious risks involve the surgical procedure itself and recovery afterwards, while in long-term the biggest risks is the uncertainty of the functional stability of the implanted electrodes (Buzsáki et al., 2012). Despite the superior signal quality offered by iEEG and ECoG, they remain sparsely used in BCIs due to the risks just mentioned.

As a result, the non-invasive methods, with EEG in particular, are very common in BCIs. While the popularity of EEG can be attributed to its long presence in research (Berger, 1929), its safety, low cost and simplicity of implementation that makes it so popular. Even though EEG signals are heavily affected by both internal and external noise, researchers came up with various processing methods, allowing them to extract meaningful information. Over the years, the hardware used for recording EEG has greatly improved, allowing for much more compact and wireless systems to be developed, such as the ones made by the company g.tec medical engineering [1]. The same however cannot be said about magnetoencephalogram, which uses cumbersome equipment for recording, but offers a signal quality comparable to the one of iEEG.

2.1.2 Haemodynamic signals

While electrophysiological signals measure the electrical activity produced by single or combined group of neurons, haemodynamic (also called cerebrovascular) recordings measure brain activity through a principal known as *neurovascular coupling*, which states that blood supply in the brain is connected to the local metabolic demand at and near cortical tissue (Ramsey, 2020). Two most popular methods which exploit this principal are *functional magnetic resonance imaging* (Sorger & Goebel, 2020) and

[1] https://www.gtec.at/product/

functional near-infrared imaging (Villringer & Chance, 1997); however they achieve that through different means.

The principal of neurovascular coupling is based on the observations made on few processes taking place during increased neuronal activity. Changes in cerebral metabolic rate of oxygen, cerebral blood flow and cerebral blood volume lead to changes in the concentration of oxy- and deoxyganted haemoglobin. On its own, measuring the changes of concentration ratio does not provide meaningful information, however Ogawa et al. (1990) showed that both oxy- and deoxygenated haemoglobin posses dia- and paramagnetic magnetic properties respectively, which allow measuring brain activity through an effect called blood oxygenation level dependent. Due to those magnetic properties oxygenated haemoglobin is repelled by the magnetic field and deoxygenated haemoglobin is attracted by the magnetic field induced by the MRI scanner thus allowing to measure the changes of haemoglobin concentration. In other words, as the blood flow increases and vessels are drained of deoxygenated haemoglobin, the strength of the MRI signal increases and vice versa, as the oxygen demand increases in vessels and the levels of deoxygenated haemoglobin briefly increase causing the MRI signal to decrease. In comparison, functional near-infrared imaging measures the concentration of the two states of haemoglobin by using an infrared light source and passing it through the cortex. Since infrared light is capable of penetrating scalp and skull, functional near-infrared imaging can be implemented non-invasively. As the light reaches the cortical issue, oxy- and deoxygenated haemoglobin absorbs different infrared frequencies, which as a result cause a decrease in the intensity of those frequencies indicating a neural activity.

Both functional magnetic resonance imaging and functional near-infrared imaging are non-invasive techniques which makes them more attractive and easier to use compared to previously mentioned iEEG and ECoG. Despite that, functional magnetic resonance imaging and functional near-infrared imaging suffer from a major drawback which is rooted in the fundamental process governing them, the neurovascular coupling. The metabolic processes involved in neurovascular coupling have particularly negative effect on the temporal-based signals, such as the ones used in BCIs. As it is noted, the usual haemodynamic response takes nearly 30s to return to the baseline reading of

blood oxygenation level-dependent; however usually a 10s window is enough to detect neural activity. Nonetheless, such a long window is unfavourable while working with BCIs. Despite this, some efforts have been made in trying to adapt functional magnetic resonance imaging to be usable in BCIs either on its own (Lee et al., 2009), or through combining it with EEG (Goldman et al., 2002). Functional near-infrared imaging is subject to a similar situation, where using functional near-infrared imaging on its own has been shown to be able to recognise motor-related neural activity (Batula et al., 2017), while the combination of functional near-infrared imaging and EEG has been more widespread (Leamy et al., 2011; Blokland et al., 2014). Between the two, functional near-infrared imaging is generally preferred over functional magnetic resonance imaging due to its low cost and complexity of setup.

2.2 Electroencephalography

In this thesis, EEG signals are the chosen modality used to measure brain activity due to the low cost of implementation, along with the pre-existing knowledge at the research laboratory. While the general information on how EEG is produced has already been provided, some more detail will be presented in this section to fully show the complexity and obstacles faced while working with EEG.

Foremost, before the LFPs generated from neuronal activity can manifest at the scalp level as EEG signals, they have to travel through layers of brain tissue, cerebral fluids and skull. Such approximation of EEG as LFPs comes with two significant assumptions:

- the aforementioned layers attenuate the signal such that the electric field produced by LFP decays with the square of the distance from the original source; (Buzsáki et al., 2012)
- volume conductance of the aforementioned layers causes spatial smoothing over an approximate area of $10cm^2$ (Akhtari et al., 2002)

While these assumptions made it possible to understand, measure and interpret EEG in a reliable fashion, it further shows the fragility of EEG and highlights the need for

very careful choice of processing tools and techniques when working with EEG signals. More recently, those assumptions and resultant views came under heavy scrutiny by the research community (Cohen, 2017). For the most part of time LFPs were understood to be mostly constructed of postsynaptic potentials leading to the assumption that EEG reflect changes in extracellular currents, which themselves reflect changes in potential of millions of pyramidal cells. However research from the last ten years has shown that calcium and sodium spikes, glial cells, active and passive currents all contribute to LFPs (Buzsáki et al., 2012), and therefore it has been argued that the general statement that "EEG reflects integration of postsynaptic potentials across neural populations" is more of a very basic explanation of no explanatory power of what is a much more complex system (Cohen, 2017).

While gaining better understanding of the dynamics and interaction between cells within the same and other layers would be without a doubt of great benefit to the researcher community introduction of novel processing techniques to EEG-based BCIs is equivalently beneficial. The most practical challenges of EEG signal processing are its characteristics; the non-stationary, non-linear and non-gaussian properties of EEG and its poor signal-to-noise ratio (SNR) makes the analysis difficult and limits the use of conventional signal processing approaches. Despite those characteristics, there are signal processing techniques that can still extract relevant information from EEG signals. Such techniques will be discussed in the next chapter.

In terms of improving SNR, one of the most popular (physical) approaches is the use of high-density EEG electrodes, which refers to the system used for electrode placement during recording. The initial 10-20 system only allowed a handful of electrodes to be placed around the scalp (sites located at 10% and 20% from the nasion, inion, left and right preauricular points) (Jasper, 1958). Over the years more dense electrode systems have been introduced, most notably 10-10 and 10-5 (Oostenveld & Praamstra, 2001), with the latter offering a recording of up to 128 electrodes. The comparison of those placement systems is shown in Figure 2.1.

Electrodes used in recordings also have a big impact on the quality of the recorded EEG signal. Researchers have used different types of electrodes with hopes of improving SNR more. While standard silver-silver chloride (Ag-AgCl) ring electrodes remain

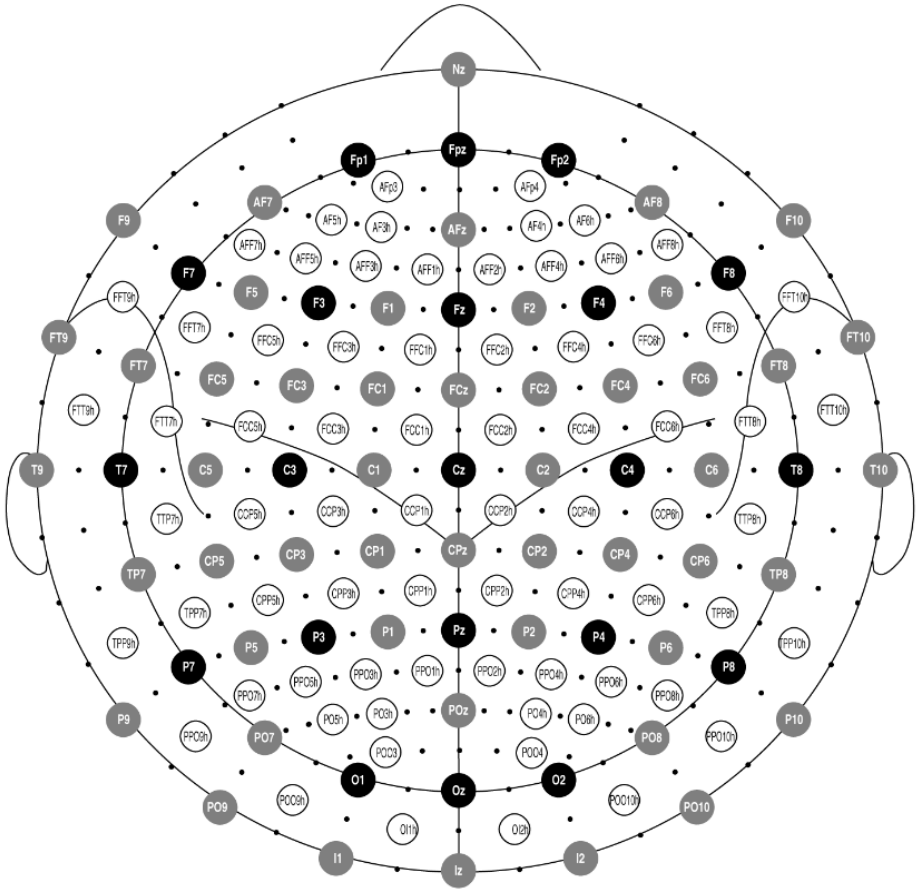

Figure 2.1: Comparison of 10-20 (black dots), 10-10 (grey dots) and 10-5 (white dots) electrode placement systems. Figures obtained from Oostenveld & Praamstra (2001)

popular in use due to their low cost, more advanced electrodes have been created to address some of the shortcomings or inconveniences present when using regular Ag-AgCl ring electrodes. These regular electrodes are referred to as *wet* because of the use of skin abrasive gel to firstly clean the scalp with subsequent application of electroconductive gel in order to maintain a contact between scalp and the disc electrode, and *passive* since they do not have any amplification circuitry at the individual electrodes. The use of abrasive and electroconductive gels is one of the most significant drawbacks of those traditional electrodes as the gels dry out between 30 and 45 minutes causing an impedance drop at the electrodes, negatively affecting the recorded signal. This becomes particularly problematic in long experiments as the gels would have to be reapplied several times to maintain good quality of signal. A recent example of newer electrodes which aim to tackle those problems are g.Sahara Hybrid electrodes offered by

g.tec [2], which are active and dry/wet electrodes. One of the most favourable advantages over regular Ag-AgCl electrodes is the lack of any gels as the needle-like pins on these electrodes make a good connection on the scalp. In addition to that, the signal recorded is being actively filtered and amplified at each electrode separately, cleaning up the signal before it reaches the main amplification unit, thus producing a much cleaner signal.

Research concerning EEG identified two major types of signals reflecting the brain activity, which lead to incorporating them into EEG-based BCIs. It has been observed that internally induced processes and mental tasks affect the ongoing EEG, consequently leading those type of signals to be named as *continuous* (or spontaneous) EEG. Additionally, exposure to external event or stimulus has been shown to lead to appearance of another distinct brain activity called *event-related potentials* (ERPs).

2.2.1 Continuous signals

It has been noted that the majority of continuous EEG, in a healthy functioning brain, take the form of rhythmic oscillations (rhythms or waves) which are found in very particular frequency bandwidths, while their amplitude varies between tenths of μV up to several μV . These bandwidths are referred to as follows: *delta* (δ, <1-4Hz), *theta* (θ,4-8Hz), *alpha* (α,8-13Hz), *beta* (β,13-30Hz) and *gamma* (γ,30-200Hz).

The lowest frequency band is occupied by δ-waves which correspond to brain activity observed at very low frequencies below 1Hz up to 4Hz, and are related to deep and unconscious sleep. However, those waves are most commonly found in infants and their strength diminishes with increasing age (Hobson & Pace-Schott, 2002), thus they are not frequently utilised in BCIs. θ-oscillations are found in the next frequency band covering 4-8Hz range. While θ-waves are mostly associated with different sleep states, it has been reported that those waves do carry some information related to mental effort (Cahn & Polich, 2006); however despite this θ-waves remain mostly unused in BCIs. The next frequency band grouping α-waves (found between 8 and 13Hz) have been generally used as an indicator of relaxed states in the brain, sometimes referred to as cognitive inactivity or cortical idling. However, α-waves recorded from the sensorimotor

[2] https://www.gtec.at/product/g-sahara-hybrid-eeg-electrodes/

areas of the brain (sometimes referred to as μ-rhythms and refined to be between 7 and 12Hz) have been found to contain information related to both movement execution and imagination (Schomer & Lopes da Silva, 2012). β-waves, which are located in the range of 13-30Hz, have been found to be related to active concentration, task engagement and attention. Similarly to α-waves, β-waves are an indicator of sensorimotor activity (Pfurtscheller & Lopes Da Silva, 1999) and have been widely used in BCIs. The last component of continuous EEG are γ-waves which occupy frequencies above 30Hz upto 200Hz, however non-invasive EEG only allows reliable detection of γ-waves up to 100Hz. In general, those waves are associated with integration of different stimuli into an overall coherent signal (Hughes, 2008). Due to closeness to power line (50Hz), γ-waves are not widely used in BCIs since they can get contaminated by power line noise and notch filtering might remove significant information from the signal.

Apart from the above frequency-dependent characteristics, continuous EEG has another important property which must be accounted for when using those type of signals in BCIs. While some components of continuous EEG are time-locked to some events, it is not *phase-locked*, meaning that simple averaging techniques commonly used to improve signal quality in brain activity analysis are not viable. Despite that, as different trials are being analysed and their amplitude might appear similar, since continuous EEG are rhythmic waves, the phases of EEG waves will be different. Thus averaging over trials, particularly in time domain, would in fact remove any meaningful information regarding continuous EEG from the analysed signal (Pfurtscheller, 2001). In the context of MI-BCIs a technique employed to mitigate shortcomings of simple averaging is called *event-related de-/synchronisation* (ERD/S) introduced firstly in 1970s and later defined more in-detail by Pfurtscheller (Pfurtscheller & Aranibar, 1977; Pfurtscheller & Lopes Da Silva, 1999). ERD/S exploits the previously described relation of α- and β-waves involved in brain activity around the sensorimotor areas. Firstly, a reference window during a resting period is established and band power values are calculated. Then, a sliding window is incorporated to move along the signal and obtain band power values for each step. Comparing the values from reference and activity windows yields a relative band power change expressed as percentages. Furthermore, if this ERD/S calculation is to be extended over several frequency bands,

then it is possible to represent ERD/S as time-frequency maps (Graimann et al., 2002).

It must be noted that in the literature, BCIs which utilise continuous EEG are often referred to as *sensorimotor-rhythms*(SMR)-BCIs; however since the main focus in this thesis is on motor imagery, *MI-BCI* naming will be used while providing reference to what exact brain modality is being used.

2.2.2 Event-related potentials

The naming of ERPs might suggest that ERPs only appear when a person is exposed to an external stimulus, however, in reality that is not fully accurate. While ERPs mainly originate from *evoked* potentials which are a result of stimulating the brain with either visual, auditory, somatosensory or olfactory stimuli, ERPs can also be generated internally through person's volition to perform a task. Analysing single-trial recordings, that is looking at each trial recording on its own, shows that ERPs are very hard to separate from continuous EEG as their potential is not high. However, since ERPs are both time- and phase-locked, time averaging over multiple trials allows to remove continuous EEG and emphasise any externally evoked brain activity.

2.2.2.1 Movement-related cortical potential

One of the most relevant ERPs which has found good application in MI-BCIs is the *Movement-Related Cortical Potential* (MRCP). By definition MRCP is a slow cortical potential preceding the onset of EMG signal by 500ms up to 2s during a voluntary action. This particular potential is composed of several components which are split into pre- and post-movement types. Initially, the former contained *bereitschaftspotential* (BP), *pre-motion positivity* (PMP) and *motor potential* (MP), while the latter contained *reafferente Potentiale* (RAP) (Kornhuber & Deecke, 1965). In later research, BP was split into early BP and *negative slope* (NS'), while PMP and MP (see Figure 2.2) were given alternative names as *P-50* and *N-10* respectively to reflect the polarity and time occurrence of those signals. The post-movement signals have also expanded to include *N+50*, *P+90* and *N+160*, while RAP was given a new name *P+300* (Shibasaki et al., 1980). Topologically MRCPs are mostly distributed around the sensorimotor cortex, however particular movements have their defined locations following

the cortical homunculus representation of the brain. For example feet movements are mainly concentrated on the midline precentral region with symmetrical distribution, while hand movements have been observed to emerge contralaterally in the precentral regions.

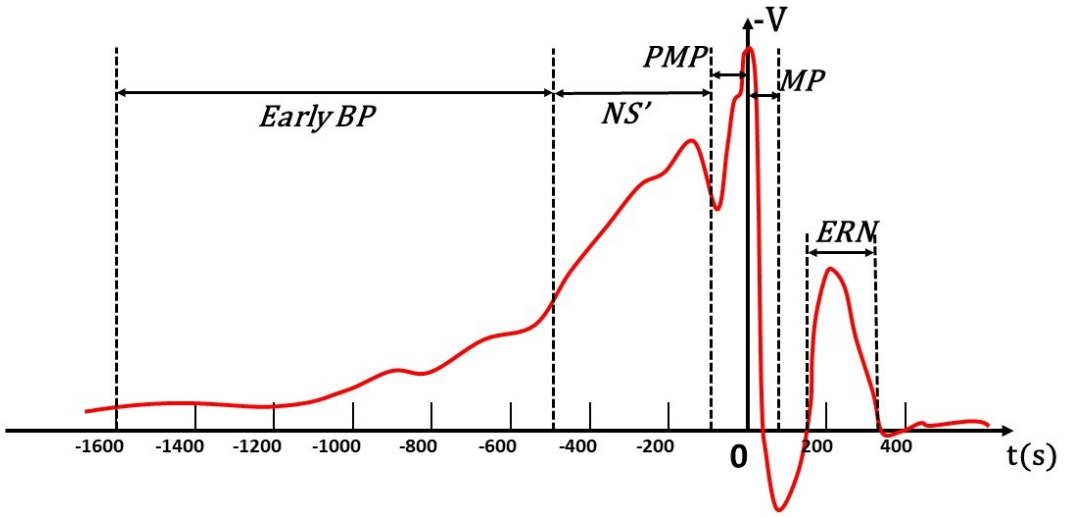

Figure 2.2: Diagram showing the main components of MRCPs. Note the reversed voltage axis direction which is a common practice when plotting MRCPs.

Despite covering such a wide spectrum of different signals, BP and NS' have been exploited the most, primarily due to their much earlier emergence in time (between 1.2 and 0.5s before EMG onset) and having clear features (sudden negative slope as seen in NS' reaching maximum negativity between 0.5 and 0s before the EMG onset) (Shibasaki & Hallett, 2006). Furthermore, BP and NS' also have been noticed during attempt to execute a movement or movement imagination, which led to incorporating BS and NS' as low δ-frequency signal into MI-BCIs, either on their own or in combination with ERD/S (Lew et al., 2012; López-Larraz et al., 2014; Ibáñez et al., 2014; Ofner et al., 2017), boosting the overall performance of BCIs.

2.2.2.2 Error potential

While on the topic of improving BCI performance, one distinct type of ERP has been of special interest. It has been postulated and later shown through brain recordings that a specific signal related to perceiving errors exists. The *error-related negativity* (ERN) or *error-related potential* (ErrP) is such signal which manifests itself on the frontal

midline, located just above the anterior cingulate cortex, a part of the brain which has also been shown to be involved in conflict monitoring and error processing (Carter et al., 1998). In general, ErrP has been observed to develop after one perceives an erroneous result or outcome, which varies between 100 and 500ms after response onset as shown in Figure 2.2. The generation of ErrP is accomplished through one of the four major actions: observation, feedback, response or interaction, where interaction ErrP has been shown to be of most use in BCIs (Ferrez & Del R. Millán, 2008). Observation ErrP manifests in the subject in the event when they observe an erroneous action or choice being committed by someone or something else. Feedback ErrP appears when the subject is informed by BCI that their action or choice was incorrect. Exposing a subject to a stimulus and requiring them to respond to it as fast as possible will yield response ErrP. Lastly, interaction ErrP is generated by the subject if they believe that BCI misinterpreted the issued command.

It can be seen from the above definitions the preference of interaction ErrP and its viability in MI-BCIs. In addition to being independent of external factors factors, as it has been shown in research, ErrP also appears in asynchronous recordings, i.e. experiments where subject decides for themselves when to perform an action and is not guided by any clues (Lopes Dias et al., 2018). In comparison, the other modalities of ErrP fail to appear while using such paradigms.

2.2.2.3 P300 component

Besides continuous EEG and MRCPs, P300 component (Figure 2.3) is the most researched brain signal which has been widely utilised in some of the first complete BCIs (Farwell & Donchin, 1988). It must be noted that P300 component described here is a different type of a signal than the previously mentioned P+300 or RAP signal. The P300 component is described as an indicator of processing information related to attentional and memory mechanisms (Sutton et al., 1965), and the ongoing research into P300 component has shown that it can be further split into two subcategories. The *novelty* P300 or *P3a* manifests as a positive potential with maximum amplitude appearing around 250-280ms post-stimulus over frontal/central electrode sites, and has been attributed to engaging attention and processing novelty. The *classic* P300 or *P3b*

is a positive potential with a maximum value appearing at 300ms over the midline parietal brain areas, and has been linked to observing likelihood of events; for example larger P3b is noticed when an event is less likely to occur. In practice, P300 component is best measured through a use of matrix of different flashing elements, focusing on the specific element (which would itself would be an intersection of a row and a column) and then counting its flashing occurrences. This paradigm has been mostly used for creating virtual keyboards enabling users to construct words and sentences with their minds (Riccio et al., 2013). However, since P300 component appears post-stimulus, it has not been as widely used in MI-BCIs as compared to other BCIs.

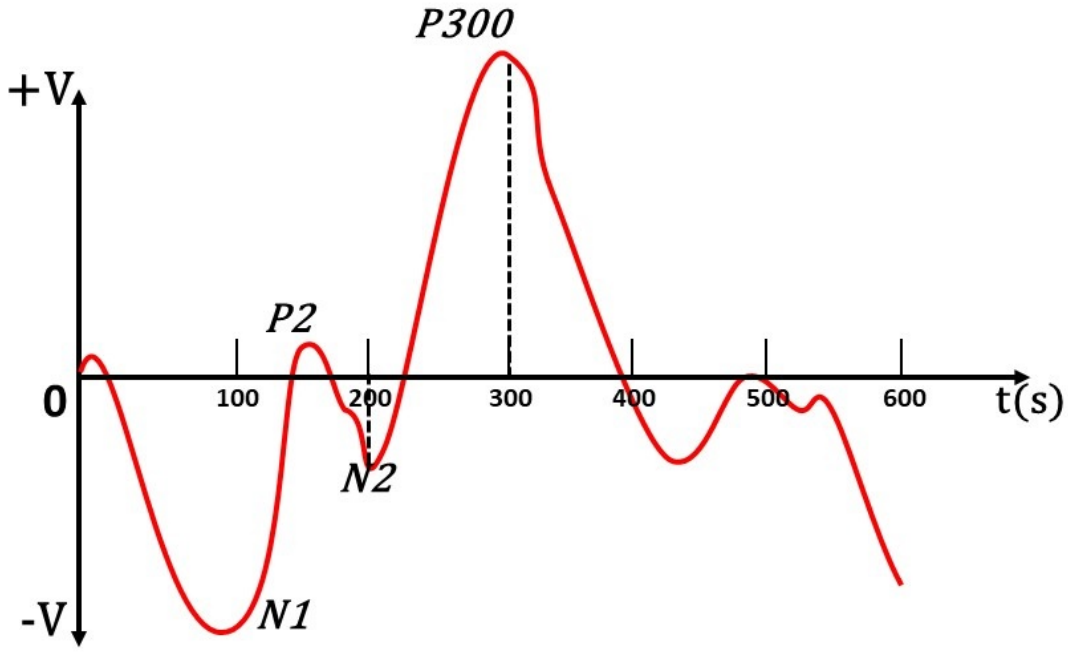

Figure 2.3: Example of brain signal showing different evoked potentials, including P300 response.

2.2.2.4 Steady-state evoked potential

The least used ERP in BCIs is *steady-state visual evoked potential* (SSVEP). Its most characteristic feature is the sinusoidal wave found in EEG recordings with the frequency reflecting the stimulation frequency. In order to evoke SSVEP in brain, subjects are exposed to multiple visual stimuli, which usually take form of a flashing LEDs or boxes on a screen. Each of the flashing elements has its own specific flicker

frequency which, when focused upon, will cause SSVEP to appear in the brain at the same frequency (Guger et al., 2012). As it is an ocular-driven method, SSVEPs are mainly recorded from the occipital lobe region of the brain, which is covered by electrodes *PO* and *O* in 10-20 system (refer to Figure 2.1 for positions).

2.3 Conclusion

Throughout this chapter the complexity of measurable brain activity has been unfolded and discussed. Two means of measuring brain signals were presented: electrophysiological signals and exploitation of haemodynamics processes. Discussion and presentation of current research regarding the two types of brain activity has revealed that electrophysiological signals are much more common in the literature due to their higher reliance and much greater ease of use. It has also been shown that electrophysiological signals are subdivided into more specific modalities out of which ECoG and EEG are the most popular choices. As a result, the author narrowed down further background to only concern EEG signals as they are the ones used in this thesis. The ease of preparation of the recording and the non-invasive nature of EEG are some of the most favourable characteristics of EEG despite its shortcomings observable in low SNR and low spatial resolution of EEG recordings. Brain activity observed through analysis of EEG signals in the context of motor imagery was split into two types: continuous signals and event-related potentials. The former was explained as continuous and rhythmic oscillations perceived in a living and healthy brain occurring at specific frequency bandwidths, while the latter grouped several sub-types of ERPs out of which MRCPs were of the most interest. Further discussion revealed that when studying motor imagery, ERD/S phenomenon and MRCPs are the most often used EEG modalities, and in the case of this thesis the author decided to exploit the ERD/S phenomenon in EEG recordings to extract information related to motor imagery. The means on how to prepare EEG signal for such analysis and the different methods used in the literature to extract relevant information is presented in the following chapter.

Chapter 3

Review of EEG analysis methods

The reader should now appreciate the complexity of EEG signals and understand the potential difficulties in working with such signals, especially when investigating motor imagery. In this chapter, the author expands upon the previous high-level representation of a typical BCI system (shown in Figure 3.1) and deconstructs said system into its individual submodules, providing a much more in-depth explanation of various techniques used in the literature. While the main focus of the thesis lies in the feature extraction module of BCI, the whole system needs to be described as every module plays an important part in the further performance assessment of the proposed feature extraction methods. The literature review starts off by presenting and discussing various preprocessing techniques used widely in the research community to precondition EEG signals. Preprocessing module is an essential part of a typical BCI as it helps removing artefacts and noise, such as power line noise, muscle movements or drifting, which are known to regularly contaminate EEG signals. Following this, the chapter concentrates on different feature extraction methods, where DMD technique is introduced. After that, feature selection and dimensionality reduction techniques currently adopted in the BCI systems are explored. Finally, the overview of a BCI is concluded with an outline of various classification techniques used in BCI systems nowadays.

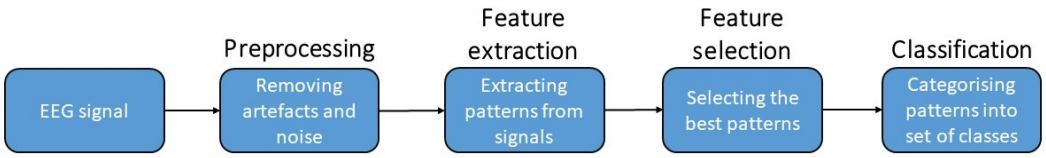

Figure 3.1: Diagram depicting the four main processing blocks in a typical BCI.

3.1 Preprocessing

The choice of methods in the preprocessing stage affects the quality of the recorded EEG signals to process and, if configured incorrectly, can cause adverse effects. Recalling the challenging nature of EEG signals and their susceptibility to different forms of interference, the correct steps and methodologies must be provided. This is reflected in the literature as researchers have proposed numerous methods and processes that enhance the quality of EEG signals. In this section, two methods of filtering will be discussed, namely frequency-domain filtering and spatial filtering, each of them having a very distinct purpose in the preprocessing stages.

3.1.1 Frequency-domain filtering

The most common method of preprocessing any signal is applying Fourier-based filtering on windowed EEG signals. This can take the form of either low- or high-pass filter, or the combination of both, as a band-pass filter (Bigdely-Shamlo et al., 2015). Low-pass filtering helps clean up the unwanted content found in the higher frequencies in the signals, which is beneficial for applications which focus on the analysis of slower brain dynamics e.g., δ-waves. High-pass filters have a similar function, where the low-frequency content is removed from the signal. However, the rationale behind this is to remove the DC offset and drifting, which is commonly found in the EEG signals recorded over longer periods of time. Sometimes a band-pass filter is applied which combines functionality of both aforementioned filters. Furthermore, band-pass filters are routinely used for isolating particular bands of brain activity. Notch filtering is the last method, used purely for removing the noise induced by the power lines (50 or 60 Hz depending on the world region).

The result of removing any unwanted components from signals helps improve SNR, however, it does produce a phase shift within the signal which is a frequently overlooked side effect of such filtering process. Therefore, one must be cautious when filtering EEG signals and the literature recommends the use a zero-phase filters in cases when one wants to preserve original phase information (Bigdely-Shamlo et al., 2015). A thorough review and discussion of different filtering methods is provided in de Cheveigné & Nelken (2019), where the authors question the implementation of filtering techniques by the research community while working with brain signals and urge to consider the exact setting of each study. Most notably, the authors encourage to look for alternative techniques for removing constant DC offset, drift and eye movement artefacts, such as: robust detrending or regression techniques.

The ERD/S phenomenon considered in this thesis is known to produce spatio-temporal patterns, therefore it is vital to preserve as much original temporal and spatial relations while applying filtering. Additionally, ERD/S phenomenon is not phase-locked to events (Section 2.2.1), thus if data processing is to be carried out offline, a non-causal digital filter can be used. This can be implemented by using `filtfilt()` function in MATLAB, which in turn can utilise Butterworth IIR digital filter, constructed with `butter()` function, as input.

3.1.2 Spatial filtering

One of the biggest challenges faced when recording and working with EEG signals is their low signal-to-noise ratio (SNR). Another characteristic of EEG signals which has been discussed in the previous section is their low spatial resolution. This is attributed to the problem of volume conduction, which becomes problematic in the study of faint signals such as motor intention and imagination (Blankertz et al., 2008).

Thus, in theory the application of spatial filtering should be beneficial to the overall performance of the system, however, while reviewing the ERD/S phenomenon the author did not find a conclusive answer specifying a need for applying spatial filters to spatial feature extraction methods such as DMD or CSP. Some publications concerning DMD did not implement any spatial filtering at all (Brunton et al., 2016; Seo et al., 2020; Solaija et al., 2018), while others applied common average reference (CAR)

(Shiraishi et al., 2020) or small Laplacian (Takeishi et al., 2021). Similarly with CSP, majority of research did not implement any spatial filters at all (Grosse-Wentrup & Buss, 2008; Ang et al., 2008, 2012) or only used small Laplacian (Müller-Gerking et al., 1999; Blankertz et al., 2007). This is perplexing as earlier literature, which both CAR and small Laplacian methods, showed that such filtering enhances the effects of ERD/S phenomena (McFarland et al., 1997).

3.1.2.1 Bipolar filter

The simplest and least computationally demanding filter is the bipolar filter (Lou et al., 2008). This method can be implemented in two ways. First approach requires placing a pair of electrodes, one anterior and one posterior in relation to the area of interest e.g., if considering C3 electrode, FC3 and CP3 electrodes will be used for recording (see Figure 3.2). The difference in their potential yields an improved SNR as the common noise is removed from the recording. The second method of applying bipolar filter is performed by recording a full multi-channel EEG recording and then iteratively testing all possible electrode combinations, which allows finding the most suitable filter. However, this method in general is more computationally demanding than the former method, and so the first method is preferable.

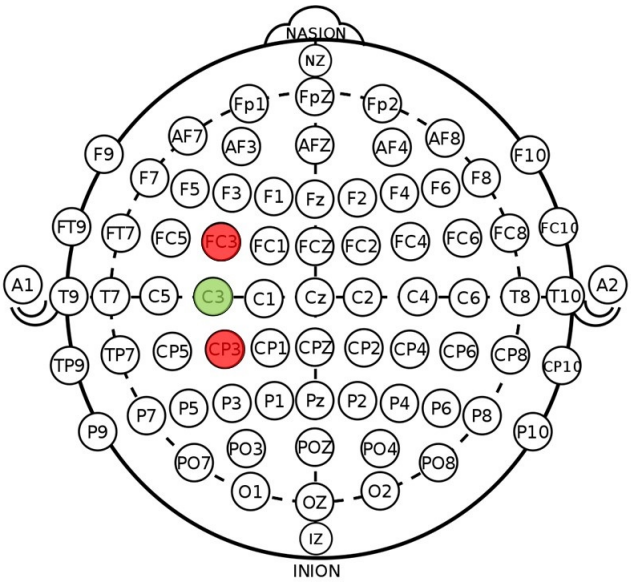

Figure 3.2: An example of a bipolar filtering. Signal at C3 electrode is derived from the difference between the anterior and posterior electrodes, FC3 and CP3.

3.1.2.2 Common average reference

The usual process of recording EEG signals includes an additional reference channel with the purpose of removing common noise present across scalp channels. However, as with any other electrodes, if the reference electrode makes a poor contact with the skin surface (usually a mastoid or ear lobe), it will introduce significant artefacts to the obtained recording (Bigdely-Shamlo et al., 2015). Additionally, those reference points can still be affected by other muscle (electromyography, EMG) or heart (electrocardiogram, ECG) activity or any movement artefacts, which contaminate EEG signals further. To help lower the impact of such artefacts a method called CAR was developed, and is frequently used as a post-recording step to remove the reference from the recorded signals. It must be noted that CAR is the preferred method in settings where high density EEG montages with equally spaced electrodes are used.

In the CAR algorithm, as an electrode of interest is chosen, the average contribution of all other electrodes is calculated and then subtracted from it. One can think of this process as high-passing EEG signals, which as a result amplifies local activity components present at each individual electrode (McFarland et al., 1997). As seen in (1), the average EEG activity of V_j is removed from an electrode of interest V_i, where n is the total number of electrodes (Yu et al., 2014).

$$V_i^{CAR} = V_i - \frac{1}{n}\sum_{j=1}^{n} V_j \tag{1}$$

3.1.2.3 Surface Laplacian

Another popular spatial filter commonly used in motor imagery problems is Surface Laplacian (SL), also known as current source density. Similarly to CAR, SL provides reference-free EEG readings which estimates radial current flow at the scalp level, by calculating the second derivatives, $\frac{\partial^2 V}{\partial x^2}$ and $\frac{\partial^2 V}{\partial y^2}$, of spatial voltage distribution. Over the years two methods emerged as the most prominent ones, one proposed by Hjorth (Hjorth, 1970, 1975) and the other one advocated by Perrin (Perrin et al., 1989).

Hjorth's approximation uses a finite difference method which looks at the difference in potential between the centre electrode V_i and the sum of weighted mean activity

of the neighbouring electrodes V_j. For this approach to be feasible, first it must be assumed that the scalp surface is locally flat and can be represented as a 3×3 grid system and, secondly, that all electrodes are equidistant. This allows SL to be defined in Cartesian coordinates as seen in (2) and subsequently approximated to the expression in (3).

$$V^{LAP} = \frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} \tag{2}$$

$$V_{ij}^{LAP} \approx \frac{1}{d^2} V_{(i-1,j)} + V_{(i+1,j)} + V_{(i,j-1)} + V_{(i,j+1)} - 4V_{(i,j)} \tag{3}$$

where, d is the distance between electrodes in cm. Alternatively, (3) can be expressed as follows:

$$V_i^{LAP} = V_i - \sum_{j \epsilon Si} g_{ij} V_j \tag{4}$$

where the weighting factor is defined as,

$$g_{ij} = \frac{1/d_{ij}}{\sum_{j \epsilon Si} 1/d_{ij}} \tag{5}$$

and d_{ij} is Euclidean distance between electrode of interest and particular neighbouring electrode, and S_i is the set of the neighbouring electrodes. These could be either nearest-neighbour (small Laplacian) or next nearest-neighbour (large Laplacian) (McFarland et al., 1997) as seen in Figure 3.3.

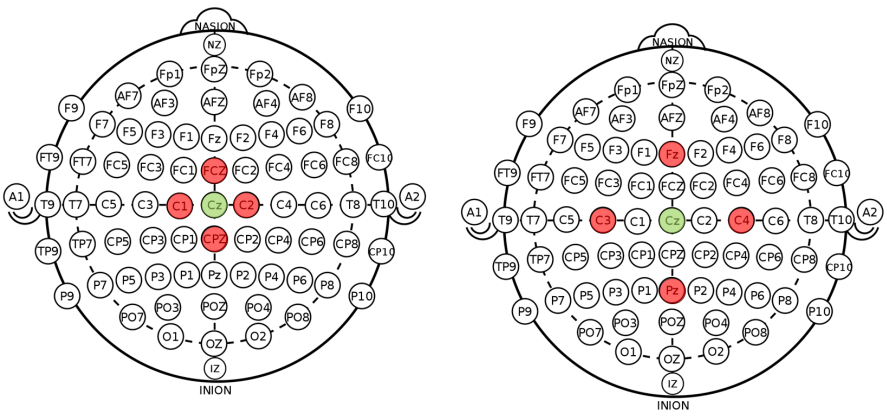

Figure 3.3: Comparison of small Laplacian (left) and large Laplacian (right) configuration

One of the biggest drawbacks of Hjorth's approximation is its inability to estimate activity of corner and edge electrodes. Therefore, it has been proposed to expand the 3×3 grid to a larger 9×9 grid, and introduce a SL matrix L (Equation (6)) which is a sparse matrix with weights related to the contributions of each electrode in the said 9×9 grid setting. Each row shows the contributions of electrodes at a specific location, i.e. first row represents weights for electrodes while considering top-left electrode in the 9×9 grid, second row represents top-middle electrode and so on (Carvalhaes & De Barros, 2015).

$$L = \frac{1}{d^2}\begin{pmatrix} 2 & -2 & 1 & -2 & 0 & 0 & 1 & 0 & 0 \\ 1 & -1 & 1 & 0 & -2 & 0 & 0 & 1 & 0 \\ 1 & -2 & 2 & 0 & 0 & -2 & 0 & 0 & 1 \\ 1 & 0 & 0 & -1 & -2 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & -4 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & -2 & -1 & 0 & 0 & 1 \\ 1 & 0 & 0 & -2 & 0 & 0 & 2 & -2 & 1 \\ 0 & 1 & 0 & 0 & -2 & 0 & 1 & -1 & 1 \\ 0 & 0 & 1 & 0 & 0 & -2 & 1 & -2 & 2 \end{pmatrix} \tag{6}$$

Another shortcoming of Hjorth's approximation is the assumption of local flat surfaces on the scalp. This led to the development of mesh-free methods, out of which spherical splines implementation (Perrin et al., 1989) became the most popular. To start off, scalp electrode positions must be firstly projected onto a sphere model. This is done by converting 3-D positions of the electrodes from Cartesian into unit sphere space, followed by calculating angles between electrodes with

$$cos(r_i, r_j) = 1 - \frac{(x_i - x_j)^2 + (z_i - z_j)^2 + (z_i - z_j)^2}{2} \tag{7}$$

where r_i and r_j are electrode coordinates. With the appropriate projection of the electrodes one may use the following smoothing/interpolating function, which utilizes spherical splines:

$$f_{sph}(r) = \sum_{i=1}^{n} c_i g(cos(r, r_i)) + d \tag{8}$$

where,

$$g(x) = \frac{1}{4\pi} \sum_{n=1}^{\infty} \frac{2n+1}{n^m (n+1)^m} Pn(x) \tag{9}$$

Function $g(x)$ allows calculating the geodesic distance between two electrodes based on their previously calculated angle. Parameter m is typically specified between 2 and 6, and Pn is the nth degree Legendre polynomial. It is worth noting that summation over Legendre polynomials as seen in (9) act as a Butterworth filter, which downweights high-frequency spatial components. A complete solution for SL with spherical splines is shown in (10)

$$Lap_s(f(r)) = -\frac{1}{r^2} \sum_{i=1}^{N} c_i g_{m-1}(r, r_i) \tag{10}$$

3.2 Feature extraction methods

The core module of any BCI system is the method used for extracting meaningful patterns from EEG signals. Those patterns, called features, can be expressed in different forms and the methodologies behind feature extraction cover a variety of signal processing topics. Depending on the choice of EEG characteristic being analysed it can be necessary to transform the original signals into a different domain. While ERPs can contain useful temporal features (majority of which use statistical approaches), spectral domain is the preferred choice whilst employing ERD/S phenomenon. That said however, features from different domains (temporal, spectral or spatial) can be combined and used together and in some cases have been shown to improve performance of BCIs (López-Larraz et al., 2014; Ibáñez et al., 2014; Kevric & Subasi, 2017). In this section the most popular feature extraction techniques will be introduced and concisely discussed in order to provide general understanding of the current state-of-the-art in BCI field.

3.2.1 Bandpower

One of the first introduced techniques when working with ERD/S involves calculating bandpower of the signal. In some earlier papers this process refers to calculating power values at α and β bands (Pfurtscheller et al., 1997) and looking for decreases or increases in power, while later papers are more specific about the process. Usually, the first step requires filtering the raw signal with a band-pass filter so a band specific (α or β) signal is obtained. This is followed by squaring each amplitude sample which, as a result, produces power samples. The final step of averaging power samples over all trials produces bandpower values along the original time window from which the relative power can be then calculated as

$$ERD\% = \frac{A - R}{R \times 100} \tag{11}$$

where A is the bandpower value at a specific sample and R is the average bandpower value from a reference period before activity, and the relative power $ERD\%$ is expressed in percentages. Alternatively, averaging over trials can be omitted and replaced by taking the log of the power samples. This would result in the instantaneous bandpower values. In addition to that, another method of calculating bandpower based on power spectral density (PSD) is available (Bhattacharyya et al., 2010). Firstly, a smaller window of bandpass filtered signal is transformed to spectral domain, where PSD can be calculated, and average bandpower is subsequently computed from the integration of the provided PSD estimate.

3.2.2 Fourier-based

Fourier transforms (FT) and additional extensions to FT have been particularly well-suited for motor imagery problems as they are the simplest techniques for transforming signals from temporal to spectral domain. FTs show frequency components of a windowed signal and their amplitude, and thus showing their contribution to the original signal. Two popular techniques used in BCIs are Fast Fourier Transforms (FFTs) and power spectral density (PSD). The former is an optimised algorithm for calculating discrete Fourier transforms (DFT) and the latter is the FT of the windows

signal's autocorrelation function. One must note that this is only applicable if the signal is stationary. In case of EEG, which is non-stationary in its nature, a sliding window approach must be used to create a pseudo-stationary scenario. PSD, which usually measures signal's power against frequency, in particular has been a useful feature as its estimation can be integrated and used as an average bandpower feature (as mentioned in the previous subsection), forming feature vectors containing such average band powers at specific electrodes as the entries (Bhattacharyya et al., 2010).

The main and biggest drawback of FFT and PSD approaches is their inability to provide temporal information regarding the frequency components. Time-frequency analysis overcomes this issue by implementing one of the following techniques, namely Short-time Fourier Transform (STFT) and wavelets. While the STFT is explained below, the latter will be discussed in the next section. STFT provides time-localized frequency information by firstly dividing a longer signal into shorter, often overlapping, windows of equal length followed by calculating FT for each window. However, it must be noted that STFT is subject to Heisenberg's uncertainty principle or Gabor limit, which states that the transformed signal cannot have good resolution in time and frequency simultaneously. Therefore, one must make a decision between either sharper temporal and wider spectral resolution, or vice versa. STFTs are usually presented as time-frequency maps called spectrograms, which in recent years have been exploited in neural networks following the rise of image processing techniques (Mammone et al., 2020; Xu et al., 2019; Wang et al., 2018b).

3.2.3 Wavelets

Despite STFT being able to deal with time-localisation issues, the aforementioned Gabor limit still holds back STFT from fully realising the potential of time-frequency analysis. Wavelets, the alternative method for implementing time-frequency analysis, are able to overcome this limitation and are particularly effective in dealing with nonstationary signals. They can be classified as either continuous or discrete wavelet transforms. The core of any wavelet implementation is the choice of the mother wavelet, which acts as a band-pass function. Over the years several popular mother wavelets have been developed and adopted in the field of signal processing; Haar, Morlet, Symlet

and Daubechies are some of the most commonly used (Li & Chen, 2014). Each of these mother wavelets have their own properties making them suitable for more particular scenarios. A continuous wavelet transform (CWT) can be described as a convolution of the original signal $x(t)$ with dilated and shifted versions of wavelet function $\psi(t)$

$$X(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t)\psi \frac{(t-b)}{a} dt \tag{12}$$

where a is the scaling factor of the wavelet (dilation or compression of the signal) and b is time shift factor of the wavelet. However, the constant change of scaling and time shifting factors combined with integration carries a heavy computational cost. Therefore, the discrete wavelet transform (DWT) was introduced. Instead of scaling and shifting the mother wavelet, DWT decomposes the signal into two new representations using two filters: a low-pass $h[n]$ and a high-pass $g[n]$, which in turn are downsampled by a factor of two. The downsampled result of $g[n]$, called the detail D and formulated as in (13), is kept as is, and its values represent the coefficient of the wavelet for this level; while result of $h[n]$, called the approximation A (14), is fed to another decomposition level and the process is repeated until the desired level of decomposition is reached. The resulting wavelet coefficients found in the detail signals D are then used as features for classification.

$$D_j[i] = \sum_k x[k] \cdot g[2 \cdot i - k] \tag{13}$$

$$A_j[i] = \sum_k x[k] \cdot h[2 \cdot i - k] \tag{14}$$

A further extension to DWT called wavelet packet decomposition is a proposed alternative allowing a better frequency resolution, as wavelet packet decomposition decomposes the detail coefficients as well and as a result increasing the total number of coefficient sets from $j + 1$ to 2^j.

3.2.4 Common spatial patterns

Despite spectral domain carrying valuable information, as evident from the literature so far, it is the spatial domain methods that have been shown to be particularly useful in MI-BCI systems which utilise ERD/S signals. One of the most renown and frequently used such methods are common spatial patterns (CSP). Since its introduction (Koles et al., 1990) and further popularisation by Pfurtscheller (Müller-Gerking et al., 1999; Ramoser et al., 2000; Brunner et al., 2007), CSP became a fundamental approach used in many BCI systems and a solid benchmark for comparisons with newer techniques in the following years (Ang et al., 2008, 2012).

At the core, CSP is a spatial filter which allows projecting raw EEG signals to new time series through a linear transform as shown in (15). The variances of those new time series components provide optimal features for discriminating between two different conditions (classes).

$$X_{CSP} = W^T X \tag{15}$$

$$C_{X_1} W = (C_{X_1} + C_{X_2}) W \Lambda \tag{16}$$

The spatial filter (or projection matrix) W is obtained through solving an eigenvalue decomposition problem (16) based on the covariance matrices of the two classes C_{X_1} and C_{X_2}. The subsequent projection of the original data transforms it in such a way that the first new 'channel'[1] maximises the variance of the first class, while minimising it for the second class. A common practice while using CSP is to choose first and last two 'channels' and calculate their log-variance, which are then used to construct a 4-by-1 feature vector X_P (17).

$$X_P = log\left(\frac{var(X^i_{CSP})}{\sum_{i=1}^{2m} var(X^i_{CSP})}\right) \tag{17}$$

While being a highly effective technique, the basic CSP algorithm is only capable of

[1] "Channel" is used here as a reference to the original orientation of data. If in the original data channels were rows then CSP filtered data will have rows as channels too. However, it must be noted that channels of CSP are not the same as the original channels in terms of location e.g., if channel 1 of X_{orig} represents a particular electrode, channel 1 of X_{CSP} will not.

working with two-class problems. Alternative approaches have been proposed to deal with this shortcoming; one-versus-rest (OVR) and joint-approximate diagonalization (JAD) are two of the most prevalent techniques in the literature (Grosse-Wentrup & Buss, 2008). The former calculates the eigendecomposition for every possible combination of classes using (16) and creates a new projection matrix W_{OVR}, which stores first and last filter values (which hold the maximum variance for each class combination). Equation (15) is then used as in the case of basic CSP to filter all original signals, followed by (17), which generates features for multi-class CSP.

JAD approach is based on the fundamental understanding of CSP as a diagonalization of two covariance matrices and expands it so that an approximate of the diagonalization of multiple covariance matrices, W_{JAD}, can be obtained such that

$$W_{JAD}^T C_{X_i} W_{JAD} = D_{c_i} \tag{18}$$

where D_{c_i} is a diagonal matrix for the i^{th} class. However, this method, as introduced by Ziehe et al. (2004), relies on a heuristic approach and does not provide meaningful information in regards to which spatial filters are the most optimal for further processing. Grosse-Wentrup & Buss (2008) complements JAD method by incorporating calculation allowing to approximate mutual information (19) and then extract L spatial filters, which have shown to contain maximum mutual information. This allows to create W_{JAD} with optimal spatial filters.

$$I(c, W_j^T x) \approx -\sum_{i=1}^{M} P(c_i) log\sqrt{w_j^T C_{x|c_i} w_j} - \frac{3}{16}\left(\sum_{i=1}^{M} P(c_i)((w_j^T C_{x|c_i} w_j)^2 - 1)\right)^2 \tag{19}$$

In the above final equation from Grosse-Wentrup & Buss (2008), the mutual information $I()$, which is based on the class information c and the spatially filtered signal $W_j^T x$, is the approximation of the difference between entropy of recorded data x (the first term) and the sum of the product between the probability score of a certain class $P(c_i)$ and the entropy of the recorded data x given class c_i (second term). However, since there is no closed-form solution of directly calculating entropy ox x, it can be defined and approximated through negentropy of x. For the complete steps and expla-

nation please refer to the original work by Grosse-Wentrup & Buss (2008).

In addition to the multi-class CSP extensions discussed above, a popular practice is to include a filter bank prior to executing CSP algorithm, thus creating a Filter-Bank CSP (FBCSP) (Ang et al., 2008). Instead of band-pass filtering the signal between 7-30Hz, a filter bank is set up such that multiple signals are obtained. This allows a more neurophysiologically-based analysis of the signals, as they can separated into their own respective bands e.g., α or β bands. Implementing this procedure has been shown to provide significant improvements to the accuracy of the BCI (Ang et al., 2012). In recent years FBCSP methods have surged in popularity, as they have been paired with neural networks as a method for further feature selection and dimensionality reduction (Wu et al., 2019; Olivas-Padilla & Chacon-Murguia, 2019; Wang et al., 2020).

3.2.5 Principal component analysis

Continuing the trend of reliable spatial techniques with applications in EEG, principal component analysis (PCA) was found to be a very powerful and flexible statistical method with many different uses in BCIs. While in BCI field, PCA is predominantly used as a feature selection method (or a dimensionality reduction technique), there are instances where PCA is used as a feature extraction method. Nonetheless, the processes involved in both feature extraction and selection are the same (see 3.3.3.1).

PCA is a linear transformation technique which converts a set of measurements that might be correlated to some degree, to a new set of orthogonal values, which are linearly uncorrelated and they are called principal components. In literature, PCA is often referred to as a Karhunen-Loéve transform, Hotelling transform, or proper orthogonal decomposition (which itself is also an alternative name for singular value decomposition, SVD). It is also referred to be a blind source separation technique (BSS).

In this thesis to keep consistent naming, PCA will be simply called 'PCA' and if PCA is calculated with the SVD method, it will be referred to as 'PCA through SVD'. Furthermore, there are two methods through which PCA can be calculated: through eigendecomposition of the covariance matrix (Yu et al., 2014) or through SVD (Lee & Choi, 2002). The prerequisite for either of the methods is to have the original data matrix X centred, i.e. with removed mean. Following this, one of the aforementioned

approach can be used.

Algorithm 1 PCA through eigendecomposition of covariance matrix C_X

Require: $X \in \mathbb{R}^{m \times n}$ ▷ m-channels, n-samples
Ensure: $X_{new} = X - \bar{X}$ ▷ $\bar{X}$ is mean of X
$C_X W = \Lambda W$
$PC = X_{new} \times inv(diag(\Lambda)^T)$

Algorithm 1 provides an overview of performing PCA using the eigendecomposition method. Starting with the original data matrix X and subsequent centring of data, the covariance of the zero-mean data matrix X_{new} is calculated. It must be stressed that for PCA to work well, the mean must be removed. This is because the aim of PCA is to obtain eigenvalues which maximise the variance of different sources present in the data. That ensured, eigenvalues Λ can be obtained from the eigendecomposition as shown in Algorithm 1. Lastly, principal components (PCs) are the projections of the calculated eigenvalues onto X_{new}. A common practice at this stage is to select a specific number of eigenvalues which describe the most important features and thus further reduce dimensionality of data. This number is equivalent to the number of eigenvalues which contribute to either 95% or 99% of total explained variance of PCs.

Algorithm 2 PCA through SVD

Require: $X \in \mathbb{R}^{m \times n}$ ▷ m-channels, n-samples
Ensure: $X_{new} = X - \bar{X}$ ▷ $\bar{X}$ is mean of X
$X_{new} = U \Sigma V^T$
$PC = U \times \Sigma^T$

To calculate PCs with the help of SVD, the data centring operation is still required. Then SVD is performed on the centred data matrix X_{new} as outlined in Algorithm 2, where the resultant U, Σ and V are unitary matrix, diagonal matrix of singular values and right singular values respectively. Finally, PCs are calculated by projecting singular values Σ onto the unitary matrix U. Here, the same dimensionality reduction technique can be used as in the previous approach for PCA. In general, calculating PCA with SVD is favoured as it provides a more stable and reliable numerical method for obtaining PCs.

3.2.6 Independent component analysis

While PCA focused on exploiting the second moment of statistics (variance), independent component analysis (ICA) looks at higher order statistical moments and provides source separation of equally important components called independent components (ICs). That said, PCA is often used as a preprocessing tool for ICA (Bugli & Lambert, 2007), however that became an object of scrutiny in recent years with some researchers suggesting that this process adversely affects the performance of ICA (Artoni et al., 2018).

ICA problem is synonymous to the "*cocktail-party problem*", where the aim is to separate a mixture of all signals into their own respective sources. This is achieved by assuming a linear matrix model $x = As$, where the only known is the matrix of observables x which is used to then estimate the mixing matrix A and statistically independent components s. Once A is estimated, ICs can be expressed as $s = Wx$, where W is the inverse of the mixing matrix A. The main measurement used for discerning between the sources is their non-Gaussian nature, meaning that Gaussian signals would not be suitable for any ICA methods (Hyvärinen & Oja, 2000). The resultant ICs have several uses; sometimes ICs are used for removing noisy sources and therefore help clean up signal from artefacts (mainly electrooculographic signals, EOG), while in other cases ICs can be selected as features and then used for subsequent classification.

An important part of ICA is the choice of the solving algorithm used for estimating the unmixing matrix and infomax. FastICA, Joint Approximation Diagonalization of Eigenmatrices (JADE) and Second Order Blind Identification (SOBI) approaches are two of the most well-known ones; however the focus here will be on Infomax and FastICA methods.

Infomax algorithm implements a function derived from neural networks which maximises the output entropy resulting in minimisation of the mutual information of the outputs thus yielding ICs. Those outputs take the form of $\phi_i(w_i^T x)$, where the input x and the weight vectors of the neurons w_i are used in a non-linear scalar function ϕ_i. Using cumulative distribution function as ϕ_i allows to obtain ICs, as shown in (20),

through a method that is equivalent to maximum likelihood estimation (Hyvärinen & Oja, 2000).

$$L_2 = H(\phi_1(w_1^T x), ..., \phi_n(w_n^T x)) \tag{20}$$

Algorithm 3 FastICA algorithm

Require: Centred and whitened matrix of mixed signals $\mathbf{x}$
  Choose initial random weight vector $\mathbf{w_i}$
  repeat
    while w_i not converged **do**
      $w_i^+ = E\left\{xg(w_i^T x)\right\} - E\left\{xg'(w_i^T x)\right\} w_i$
      $w_i = w_i^+ / ||w_i^+||$
      if $i = 1$ **then**
        if w_i converged **then** *break*
        end if
      else
        $w_i^+ = w_i - \Sigma_{j=1}^{i-1} w_i^T w_j w_j$
        $w_i = w_i^+ / ||w_i^+||$
      end if
    end while
    $i = i + 1$
    return $\mathbf{w_i}$
  until all w_i are obtained

Based on the idea of utilising neural network learning rules as seen in Infomax approach, FastICA method introduced by Hyvärinen is a fixed-point iteration of such process (Hyvärinen & Oja, 1997). In comparison to Infomax, FastICA does not require any user-defined parameters and boasts a great performance boost (Sahonero-Alvarez & Calderon, 2017). As stated before, the aim of ICA is to look at the nongaussianity of the signals which then allows for extracting ICs. As such, FastICA uses the approximation of negentropy $J(w^T x)$ as means of measuring the nongaussianity.

FastICA offers two methods of calculating ICs: either through one-by-one estimation (equivalent to projection pursuit method) or through symmetric decorrelation allowing for parallel estimation of the weight matrix w. Before application of FastICA, the matrix of observables x must be firstly centred, which simplifies running of the algorithm, and then whitened, so that the calculated ICs will be uncorrelated and have unit variance. Following that, a random weight vector w_i is initialised and new w_i^+ is obtained. The negentropy is approximated by using first and second derivatives (g

and g') of a nonquadratic function G chosen before executing the ICA algorithm. The choice of G mostly affects the robustness of the obtained ICs as it controls the speed of the convergence of the weight vector w_i. This vector is then normalised and checked if it has converged: if it did, the weight vector w_i is returned as ICs and the algorithm is repeated again as shown in Algorithm 3.

3.2.7 Empirical mode decomposition

Empirical mode decomposition (EMD) is a data-driven method highly suitable for working with non-linear and nonstationary signals, something that Fourier based methods are known to struggle with (due to windowing constraints). Compared to other methods discussed so far, EMD is a relatively new technique in signal processing only introduced in 1998 (Huang et al., 1998) with its first uses in EEG analysis presented just in 2004 to assess synchronisation of neuronal activity (Sweeney-Reed et al., 2004).

$$x(t) = \sum_{i=1}^{n} IMF_i(t) + r(t) \tag{21}$$

EMD is a decomposition technique which is described as a sum of finite number of intrinsic mode functions (IMFs) with the addition of residual signal as seen in (21). Those IMFs are functions extracted from the data through an iterative sifting process during which the candidate IMFs have to satisfy the following two conditions in order for them to be valid:

(1) *number of extrema in the whole data set must be either equal or differ at most by one to the number of zero crossings*

(2) *mean value of the envelope defined by both local maxima and minima is equal to zero at any data point*

In practice those conditions are fulfilled by firstly identifying extrema of the signal and then applying a cubic spline interpolation between local maxima and minima producing the upper and lower envelopes as a result. A valid IMF is acquired if the difference between the signal and the mean of the envelopes is close or equal to zero. Otherwise the sifting process is repeated until this condition is satisfied. Once the

desired number of IMFs has been extracted the original signal can be expressed with equation (21), where $r(t)$ is the residual signal left after extracting the last IMF. It must be noted that the first IMF corresponds to the highest frequency component of $x(t)$.

Algorithm 4 General EMD

```
Provide a desired number i of IMFs to be extracted
for i times do
    let h = x
    identify extrema in h
    while h − μ ≠ 0 do
        identify local maxima and minima
        fit cubic spline to create envelopes e_U and e_L
        μ = (e_U + e_L)/2
        h_1 = h − μ
    end while
    IMF_i = h_1
    if last i then
        r = h − h_1                    ▷ where r is the residual signal left
    else
        x = h − h_1
    end if
end for
```

A common method for extracting features from IMFs is to obtain the instantaneous frequency of each respective IMF using Hilbert Transform which in turn is calculated according to (22). From there a new analytic signal Z_t is formed from input signal X_t and its Hilbert Transform Y_t. This new analytic signal can also be expressed in polar coordinates, as seen in (23), where a_t is the series of instantaneous amplitudes and the instantaneous phase is θ_t. Finally, instantaneous frequency f_t is described as the rate of change of θ_t and is commonly used to construct a feature vector for further classification.

$$Y_t = \frac{1}{\pi} P \int_{-\infty}^{\infty} \frac{X_{t'}}{t - t'} dt' \tag{22}$$

$$Z_t = X_t + iY_t = a_t e^{i\theta_t} \tag{23}$$

However, the basic formulation of EMD has some significant drawbacks, most no-

tably its reliance on the existence of extrema in data (either in amplitude or in curvature) and the inability to appropriately deal with multivariate data. Two methods were developed to tackle those shortcomings: ensemble EMD its extension multi-dimensional ensemble EMD and multivariate EMD.

Soon after the introduction of EMD some researchers have noticed that a phenomenon named "mode mixing" was occurring during the extraction of IMFs (Wu & Huang, 2009). It was found that sometimes the extracted IMFs contained hugely different oscillations leading to IMFs forfeiting useful physical meaning as they would suggest a presence of false modes in data. Thus, it has been proposed to obtain an ensemble of IMFs and use the mean of such ensemble as the final IMF for a particular level of decomposition. The ensemble is calculated through an iterative process of adding different white noise to the original signal and then decomposing according to normal EMD procedure (as seen in Algorithm 4). Since white noise is used in ensemble EMD, the final IMF will not contain any white noise as it will cancel itself out during calculating the final mean.

Development of multivariate EMD (Rehman & Mandic, 2010) followed some earlier attempts of dealing with multivariate signals, such as bivariate EMD and trivariate EMD. As the direct definition of local maxima and minima is not possible for multivariate signals, multivariate EMD approach solves this problem by projecting the p-variate signal $x_i(t)$ onto a $(p-1)$ hypersphere. This process generates p-dimensional envelopes $e^{\theta_k}(t)$ of the projections $q^{\theta_k}(t)$ through the interpolation of the local extrema. Projections are defined here as $q^{\theta_k}(t) = x_i(t) \cdot (v^{\theta_k})^T$, where v^{θ_k} is the projection vector along the direction given by angle θ_k and $k = 1, 2, \ldots, K$ with K being the number of uniformly distributed θ_k. As the projections are calculated, the time instants of maxima found in q^{θ_k} are obtained and used for later interpolation producing $e^{\theta_k}(t)$ as a result. Averaging those envelopes with (24) yields the mean of envelope curves $m(t)$ which is used in a similar way as in general EMD method. A candidate IMF can be obtained as $s(t) = x_i(t) - m(t)$ and if it satisfies the sifting stopping criterion, $s(t)$ is set as one of the multivariate IMFs (MIMFs) and the whole process is repeated until exhaustion of meaningful multivariate IMFs.

$$m(t) = \frac{1}{K} \sum_{k=1}^{K} e^{\theta_k}(t) \tag{24}$$

3.2.8 Dynamic mode decomposition

A new decomposition technique recently emerged from the field of fluid dynamics analysis (Schmid, 2010) called Dynamic Mode Decomposition (DMD). Similarly to EMD technique, DMD is a purely data-driven method well suited for non-linear and dynamic systems with multivariate signals. DMD decomposes data into dynamic, spatio-temporal modes which reflect low-rank dynamics present in the data. Those low-rank dynamics are approximated without using any equations to directly describe the dynamics of the systems. Over the last five years, DMD has been gradually gaining momentum in the field of analysing brain signals thanks to the initial paper presenting the application of DMD to ECoG recordings (Brunton et al., 2016) in the context of analysis of movement tasks. Since that first publication, DMD has seen few more notable academic works relevant to these topics: application of DMD in seizure detection (Solaija et al., 2018), studying epilepsy (Seo et al., 2020), decoding movement from ECoG signals (Shiraishi et al., 2020). Additionally, work done by Bito et al. (2019) lays down a foundation in how DMD can be used to separate and identify clusters of human activities from movement data. In the context of applying DMD to MI signals found in EEG only one paper has been found (Takeishi et al., 2021).

DMD problem is formulated on the basis of a dynamical system being described by a set of differential equations, however, describing non-linear signals such as EEG with differential equations is a near impossibility. Therefore, a proxy is introduced which attempts to instead approximate a locally linear dynamic system as shown:

$$\frac{dx}{dt} = f(x,t) \approx \mathcal{A}x \tag{25}$$

which can be also represented in a discrete-time system as:

$$X_2 \approx AX_1 \tag{26}$$

Here, the pair of state matrices X_1 and X_2 are single-sample shifted matrices of the original matrix $X \in \mathbb{R}^{c \times m}$, which in case of usual EEG recordings is a matrix of column vectors $x_k \in \mathbb{R}^c$, where c is the number of recorded channels and m is the number of samples in the matrix X.

$$\begin{aligned} X &= [x_0, x_1, x_2, \ldots, x_m] \\ X_1 &= [x_0, x_1, x_2, \ldots, x_{m-1}] \\ X_2 &= [x_1, x_2, x_3, \ldots, x_m] \end{aligned} \tag{27}$$

While the above arrangement of data matrices X_1 and X_2 would be acceptable for the originally intended fluid flows as the measured data has $c \gg m$ (the preferred combination for DMD algorithm) in the case of windowed EEG signals the opposite $c \ll m$ holds. Therefore, in order to satisfy this requirement, X_1 and X_2 are augmented by shift-stacking column vectors and producing respective Hankel matrices (28) as a result. These new matrices are constant in skew diagonal and their new "height" is controlled by a stacking factor h, which is the smallest integer satisfying the following $hc > 2m$ inequality. This is based on the findings from Tu et al. (2014) where it has been shown that dynamics of standing wave can be determined by DMD algorithm if such time shift-stacking method is applied and then was expanded to cover data which does not meet $c \gg m$ criteria (such as EEG data).

$$X_{1,aug} = \begin{bmatrix} x_0 & x_1 & \cdots & x_{m-h} \\ x_1 & x_2 & \cdots & x_{m-h+1} \\ & & \vdots & \\ x_h & x_{h+1} & \cdots & x_{m-1} \end{bmatrix}, \quad X_{2,aug} = \begin{bmatrix} x_1 & x_2 & \cdots & x_{m-h+1} \\ x_2 & x_3 & \cdots & x_{m-h+2} \\ & & \vdots & \\ x_{h+1} & x_{h+2} & \cdots & x_m \end{bmatrix} \tag{28}$$

The last undefined left from (26) is the linear operator A. As stated earlier, this operator attempts to describe non-linear dynamics which relate the two state matrices. In the literature this method of analysing signal is often called Koopman spectral analysis. There, a linear operator named Koopman operator is an infinite-dimensional linear operator which represents finite-dimensional, non-linear dynamics of the system. DMD

heavily relies on this approach and the aim of DMD algorithm is to approximate this Koopman operator into a matrix representation and exploit its eigenvalues and eigenvectors to then calculate DMD modes. Therefore, Equation (26) must be transformed such that A can be found. As seen in Equation (29) the Moore-Penrose pseudoinverse of X_1 is obtained by calculating its SVD, through the built-in MATLAB function `svd()`, thus allowing A to be determined. However, the size of A must be noted here and taken into the account of computing time. Given that A will be $hc \times hc$, it can yield very big matrices which then will lead to longer computation time of the eigendecomposition, therefore, it has been proposed to find a low-rank approximation $\tilde{A}$ instead. This is achieved by r-rank truncation of $X_{1,SVD}$ allowing to obtain $\tilde{A}$ of size $hc \times r$ which is much more suitable for the subsequent eigendecomposition.

$$A \approx X_2 X_1^{\dagger} \triangleq X_2 V \Sigma^{-1} U^*, \qquad \tilde{A} = U_r^* X_2 V_r \Sigma_r^{-1} \tag{29}$$

In the case of this thesis, r-rank number was set to 100 ranks. Although approaches that can calculate optimal number of r-ranks exist (Gavish & Donoho, 2014), the calculated optimal r-rank tends to vary across different data windows. Therefore, to keep all decomposed trials equal, a single set value has been provided instead. While selecting 100 ranks might seem arbitrary, the reasoning behind the chosen value is that such number of ranks should be more than enough to capture approximately 95% of the total SVD energy in the analysed window. Furthermore, preserving too many ranks has corruptive effect on the quality of the computed DMD modes as noted in Chapter 8 of Kutz et al. (2016).

Algorithm 5 DMD algorithm, based on Brunton et al. (2016)

Require: $X \in \mathbb{R}^{c \times m}$
Build X_1 and X_2 based on Hankel shift-stacking method
$X_{1,SVD} = U \Sigma V^*$
$\tilde{A} = U_r^* A U_r = U_r^* X_2 V_r \Sigma_r^{-1}$
$\tilde{A} W = W \Lambda$
$\Phi = X_2 V_r \Sigma_r^{-1} \mathbf{W}$

The eigendecomposition problem is formulated as $\tilde{A}\mathbf{W} = \mathbf{W}\mathbf{\Lambda}$, where $\mathbf{W}$ are eigenvectors and $\mathbf{\Lambda}$ is a diagonal matrix of eigenvalues λ. Lastly, DMD modes can be computed. Two methods for computation are prevalent in the literature: *projected* DMD

modes and *exact* DMD modes (Tu et al., 2014), where the former projects the modes onto the initial state matrix X_1 through use of U_r matrix meaning that DMD modes φ_{proj} are not the direct eigenvectors of $\tilde{A}$ (Equation (30)). Exact DMD modes φ_{exc} take into consideration projection onto the next state matrix X_2, meaning that φ_{exc} are direct eigenvectors of $\tilde{A}$ and have been proven as such (Tu et al., 2014). Because of the exact modes being calculated in the image of the future state matrix X_2 only exact DMD modes (Equation (31)) are used in this thesis, and will be represented as Φ when referring to the full matrix of modes or as ϕ_i when referring to an individual mode.

$$\varphi_{proj} \triangleq U_r \mathbf{W} \tag{30}$$

$$\varphi_{exc} \triangleq X_2 V_r \Sigma_r^{-1} \mathbf{W} \tag{31}$$

Calculated DMD modes Φ are non-orthogonal, will have the same size as $\tilde{A}$ i.e., $hc \times m$ however only first c rows of Φ are considered for the later analysis as the rest of the modes are just shift-stacked copies (Brunton et al., 2016). Each column is a mode corresponding to its i-th eigenvalue and it comes in conjugate pairs, therefore the matrix Φ can be further pruned by selecting every second mode giving the final size of Φ to be $c \times \frac{r}{2}$. Since a single mode φ_i is a complex number its magnitude and phase provides valuable information which can be exploited for feature extraction. The magnitude provides information regarding the relative influence of all channels on the associated mode frequency, which in turn is obtained from the relative eigenvalue as shown below

$$f_i = abs\left(\frac{log(\lambda_i)}{2\pi\Delta t}\right) \tag{32}$$

where, Δt is the sampling period of the signal. It must be noted that the mode frequency is characteristic to the specific window and not fixed across trials and therefore appropriate methods must be incorporated to find the similarities between modes of different trials. These methods will be discussed later in the "Methodology" chapter of

this thesis. One way of visualising the effect of modes on their characteristic frequencies is by producing a DMD spectrum which is a plot of $|\Phi|$ against f.

Following the described method from Algorithm 5 will produce DMD modes in unit norm. This normalized state shows which modes contribute the most dynamically at the specific characteristic frequencies and modes with greater magnitudes can be seen as more dynamically important. Findings by Tu et al. (2014) have shown that DMD modes can be scaled through alternative means. The authors of the referenced paper presented that under certain conditions modes obtained by DMD are related to modes obtained by eigensystem realization algorithm (**ERA**). In particular, when system under investigation has been subjected to shift-stacking procedure outlined earlier a similarity is found between approaches used for calculating the low-dimensional approximation $\tilde{A}$ of DMD and low-dimensional approximation A_{ERA} of ERA. Let H and H' be Hankel matrices of some data similar to matrices $X_{1,aug}$ and $X_{2,aug}$. In ERA, the low-rank approximation A_{ERA} can be then calculated as follows:

$$A_{ERA} = \Sigma_r^{-1/2} U_r^* H' V_r \Sigma_r^{-1/2} \tag{33}$$

where U_r, Σ_r and V_r are r-rank truncated SVD of $H = U\Sigma V^*$. As Tu et al. (2014) points out, if the same H' matrix was to be used to compute DMD modes by substituting $X_{2,aug}$ in (29) such that $\tilde{A} = U^* H' V \Sigma^{-1}$ then it becomes apparent that A_{ERA} and $\tilde{A}$ can be related by a similarity transform

$$A_{ERA} = \Sigma^{-1/2} \tilde{A} \Sigma^{1/2} \tag{34}$$

It must be noted here that eigenvalues of A_{ERA} and $\tilde{A}$ are equal and because of that the eigenvectors of those low-rank approximations have a special relation too. If

$$A_{ERA} W_{ERA} = \lambda_{ERA} W_{ERA} \tag{35}$$

and $\lambda_{ERA} = \Lambda$ then by looking at the relation below it can be clearly seen that eigenvectors of $\tilde{A}$ are $W = \Sigma^{1/2} W_{ERA}$

$$\tilde{A}W = \tilde{A}\Sigma^{1/2}W_{ERA} = \Sigma^{1/2}A_{ERA}W_{ERA} = \lambda\Sigma^{1/2}W_{ERA} = \lambda W \tag{36}$$

This method is incorporated by Brunton et al. (2016) where the authors call it *scaling modes by SVD energy* and in this thesis the author refers to such scaled modes as *energy-* or *SVD*-scaled DMD modes. This results in mode's magnitude now displaying their energy content rather than relative influence over channels as with modes in unit norm. Additionally, the authors showed that by using the relations of A_{ERA} and $\tilde{A}$ the subsequent calculation of the mode amplitude P which is the square of modes' magnitude can be plotted against the characteristic frequencies f with the result closely resembling the shape of average FFT of the exact same windowed signal. Notably, DMD spectrum corresponds only to the energy of a single specific mode across all channels while FFT power spectrum is calculated for each channel individually.

$$P_i = |\phi_i|_2^2 \tag{37}$$

Algorithm 6 modified DMD algorithm

Require: $X \in \mathbb{R}^{c\times m}$
Build X_1 and X_2 based on Hankel shift-stacking method
$X_{1,SVD} = U\Sigma V^*$
$\tilde{A} = U_r^* A U_r = U_r^* X_2 V_r \Sigma_r^{-1}$
$\hat{A} = \Sigma_r^{-1/2}\tilde{A}\Sigma_r^{1/2}$
$\hat{A}\hat{W} = \hat{W}\Lambda$
$\mathbf{W} = \Sigma_r^{1/2}\hat{W}$
$\Phi = X_2 V_r \Sigma_r^{-1}\mathbf{W}$

3.3 Feature selection methods

After extracting features from the signals under investigation, it is often desired to further reduce the size of the feature space to improve the speed of subsequent training and classification. Depending on what the initial features are, feature selection methods might look for a set of particularly influential channels or frequencies that according to some algorithm carry the most significant data which can be used for classification. These algorithms could be either purely statistical in their nature or rely on some

alternative methods (mostly projection-based methods). This section aims to provide an overview of the most commonly used methods for feature selection found in the literature.

As highlighted in the introduction, the author decided to group feature selection and dimensionality reduction techniques together as they achieve similar goals: feature selection selects a subset of most useful features from the initial feature vector thus reducing the initial dimensions in the process.

3.3.1 Mutual information-based

Mutual information-based feature selection (MIBFS) methods are one of the biggest accumulations of feature selection methods and they can be separated into two main approaches: feature scoring and feature subset selection algorithms (Pohjalainen et al., 2015). The former techniques evaluate the usefulness of features by calculating score values of each individual feature and returning their ranking based on different criteria, however, they do not provide information regarding how many of the features should be selected for the most optimal performance. On the other hand, feature subset selection techniques aim to rectify that shortcoming by combining information from the previously mentioned approach with an additional intrinsic determination of feature set size.

The basis for MIBFS is the fundamental formulation of obtaining mutual information (MInf), which is a measure of mutual dependence between two variables and is part of feature scoring approaches. When using MInf in the context of EEG features, feature values are expressed as x and class labels corresponding to those features are y. The value for MInf is the product of pair joint probability density function of feature x and its class label y and the log of the ratio between pair joint probability density function and the product of individual probability density functions as seen in Equation (38). Since MInf is a feature scoring method the returned results contain scoring and relative ranking of each feature leaving the number of chosen features up to personal judgement.

$$MInf = \sum_{x \in X} \sum_{y \in Y} p(x,y) log \left(\frac{p(x,y)}{p(x)p(y)} \right) \tag{38}$$

Moving onto feature subset selection algorithms, a group of closely related and well known methods called sequential feature selection algorithms are found. Every method from this family-tree contains two essential building blocks: objective function and a sequential search algorithm. One of the earliest and still widely used nowadays sequential feature selection techniques is the sequential forward selection (SFS) method (Whitney, 1971). Here, the best features X are obtained through iteratively populating an initially empty feature subset vector X_0 with best feature candidates x^+ which maximise the specified criterion function $argmaxJ()$. The procedure is repeated until the desired number of features p is obtained as seen in Algorithm 7. Another popular method is sequential backward selection (SBS) which is SFS but in reverse i.e., features are iteratively removed from the initial feature set so that eventually only the best features remain in the new subset.

Algorithm 7 SFS algorithm

repeat
  $X_0 = \emptyset, k = 0$
  $x^+ = argmaxJ(X_k + x)$ ▷ $x \in Y - X_k$
  $X_{k+1} = X_k + x^+$
  $k = k + 1$
until $k = p$

A feature subset selection algorithm which addresses this drawback is a filter-based technique called minimal-redundancy maximum-relevance (MRMR) (Peng et al., 2005). Through analysis of the correlation and mutual information, MRMR simultaneously minimises redundancy and maximises relevance between features and the provided class information. This allows MRMR to select a subset of features which have the most correlation with a class (maximising relevance of the features V_s with respect to the class labels) and the least correlation between the features themselves (addressing the redundancy W_s). The criterion controlling determination of the best feature is mutual information quotient and it is calculated as seen in Equation (41).

$$V_S = \frac{1}{|S|} \sum_{x \in S} I(x, y) \tag{39}$$

$$W_S = \frac{1}{|S|^2} \sum_{x,z \in S} I(x,z) \tag{40}$$

$$maxMIQ_x = max \frac{V_x}{W_x} = max \frac{I(x,y)}{\frac{1}{|S|} \sum_{z \in S} I(x,z)} \tag{41}$$

3.3.2 Linear discriminant analysis

Linear discriminant analysis (LDA), while mostly used as a classification technique (see section 3.4.1), has been shown that with slight modifications LDA can be turned into a simple and efficient feature selection technique (Song et al., 2010). The approach exploits a particular step of LDA calculation where eigenvalues and eigenvectors are calculated. Assuming that the LDA-based feature can be expressed as the product of i-th eigenvector W and sample x such that:

$$z = x^T W = \sum_{i=1}^{N} x_i W_i \tag{42}$$

it has been noted that the magnitude of W_i statistically reflects to the contribution of the i-th sample, thus Song et al. (2010) postulated that removing $x_i W_i$ with small $|W_i|$ will have negligible effect on the accuracy of the classification and therefore such features are safe to be removed from the initially calculated features z. Algorithmically it has has been proposed to calculate individual contributions of eigenvectors based on the selection of m largest corresponding eigenvalues, denoting the selected eigenvectors as $V_1, \ldots, V_m$. The newly introduced term for contribution is denoted as c_j as seen in Equation (43) and it reflects the contribution of the j-th sample based on the j-th element of p-th eigenvector V_{pj}, where $j = 1, 2, \ldots, N$ and $p = 1, 2, \ldots, m$.

$$c_j = \sum_{p=1}^{m} |V_{pj}| \tag{43}$$

3.3.3 Projection methods

An alternative method of feature selection involves projecting the obtained features onto another space representation. This is the preferred method of selecting features

in the cases where for single trial a matrix of features is produced instead of a vector and where vectorisation of matrices could remove important data from the extracted features.

3.3.3.1 Principal Component Analysis

Feature selection using PCA is accomplished by projecting feature matrix X onto its left singular values U to obtain a projected matrix of new features corresponding. This is accomplished by first obtaining SVD of feature matrix X and then selecting d-elements of U which correspond to d number of biggest energy found in singular values Σ. The most notable examples of such successful projections can be seen in Brunton et al. (2016) and Seo et al. (2020).

$$SVD(X) = U\Sigma V^*, \quad a = U_d^T X \tag{44}$$

In practice, individual DMD modes φ_i are vectorised and stacked vertically to create tall matrices M_{PCA}, (45). Each vectorised trial Φ_t is stacked column-wise to create a new feature set $\mathcal{F}_{PCA}$, (46). This newly created feature matrix replaces X matrix to form (47), allowing to select best features from DMD modes using PCA method.

$$M_{PCA} = vec(\Phi_t) = \begin{bmatrix} \varphi_1 \\ \varphi_2 \\ \vdots \\ \varphi_{r/2} \end{bmatrix} \in \mathbb{C}^{cr/2 \times 1} \tag{45}$$

$$\mathcal{F}_{PCA} = [M_{1,PCA} M_{2,PCA} \cdots M_{t,PCA}] \in \mathbb{C}^{cr/2 \times t} \tag{46}$$

$$|\mathcal{F}_{PCA}| = U\Sigma V^*, \qquad a_d = U_d^T |\mathcal{F}_{PCA}| \tag{47}$$

3.3.3.2 Riemannian manifold

Riemannian manifolds are a part of studies concerning Riemannian geometry which investigates smoothly curved spaces (manifolds) that locally exhibit behaviour to the one seen in Euclidean spaces. At each point of the manifold a linear approximation can be calculated creating a tangent space. This tangent space can be equipped with some metric, which varies from point to point and can be then exploited in feature selection process. In the context of application of Riemannian manifolds in BCI, the introduction and popularisation of the method can be traced to A. Barachant who has extensively studied Riemannian manifolds and paved a way for their implementation in MI-BCIs (Barachant et al., 2010, 2012, 2013; Congedo et al., 2017). In those works a connection between CSP filters, covariance matrices and Symmetric Positive Definite (SPD) matrices has been made while proposing the use of Riemannian distance δ_r and mean $\bar{P}$ as the metrics used on the tangential space for feature separation and selection.

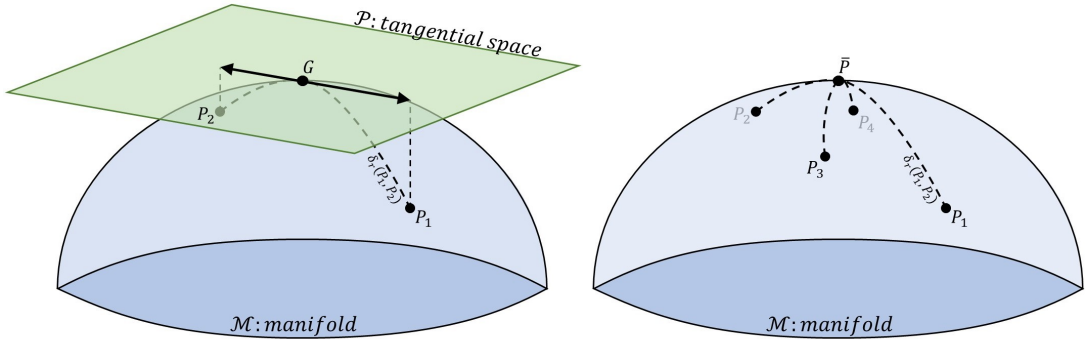

Figure 3.4: Representations of Riemannian manifold $\mathcal{M}$. On the top, a tangential space $\mathcal{P}$ seen at point G shows two points P_1 and P_2 being connected by their geodesic distance δ_r. These points can be mapped on $\mathcal{P}$ as straight lines instead. On the bottom, the Riemannian mean $\bar{P}$ of set of P_i points is shown

Recalling Equation (16), the covariance matrices found there have been shown to be SPD matrices (Barachant et al., 2010) and therefore can be used to populate the tangential space $\mathcal{P}_n$ at points P. Since points P would be spread on the Riemannian manifold $\mathcal{M}$, two points P_1 and P_2 can be related by their shortest path i.e., the geodesic distance δ_r as seen below.

$$\delta_r(P_1, P_2) = ||log(P_1^{-\frac{1}{2}} P_2)||_F \tag{48}$$

While tangential space $\mathcal{P}$ is regarded to be locally Euclidean, using Euclidean distance to measure separation of two points P_1 and P_2 does not produce the correct representation of the distance since Euclidean distance would ignore the shape of the manifold. Instead, δ_r accounts for the geometry of the manifold, producing the correct distance of the two aforementioned points. Possession of δ_r also allows to calculate the Riemannian mean (also known as geometric mean) through an optimisation problem as shown in Equation (49). The recommended optimisation method is the gradient-descent method (Barachant et al., 2010).

$$\bar{P} = \underset{P \in \mathcal{P}_n}{\arg\min} \sum_{i=1}^{N} \delta_r^2(P_i, P) \tag{49}$$

3.3.3.3 Grassmannian manifold

It is evident from the literature that Riemannian manifolds are only useful in EEG analysis and BCI applications if the features extracted are SPD or covariance matrices. This greatly limits the type of features that can be utilised leading researchers to explore and investigate alternatives. While Lotte et al. (2018) observed that Stiefel and Grassmann manifolds have been shown to be well suited for subspace projections and orthogonal matrices, the latter was seen particularly attractive. Although Grassmann manifolds have been well defined and offer a variety of different metrics (Hamm, 2008; Hamm & Lee, 2008, 2009) they remain mostly unutilised. This was clearly reflected in the quick diminish of the publications following the initial ones. Despite Grassmann being explored with the established metrics (Chevallier et al., 2014) or incorporating additional metrics based on Mahalanobis distance (Washizawa & Hotta, 2012) or geodesic distance (Li et al., 2014) the field has seen lack of interest until recently. With the rising popularity of DMD methods, it would seem that Grassmann manifolds might have found much better suited features than the ones offered by CSP (Bito et al., 2019; Shiraishi et al., 2020).

Grassmann manifold works particularly well when it is supplied with orthogonal matrices or if it is tasked with looking for similarity between different subspaces. By definition, a Grassmann manifold $\mathcal{G}(m, D)$ is a $m(D-m)$-dimensional compact Rieman-

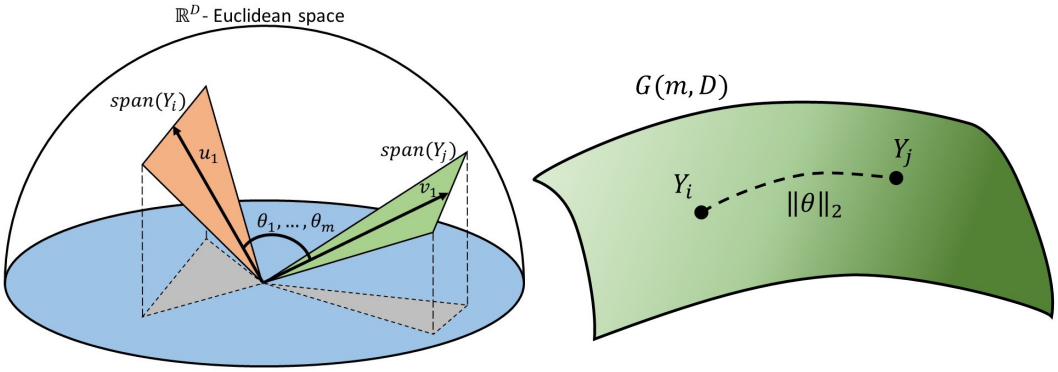

Figure 3.5: Span of subspaces Y_i and Y_j in a Euclidean space (left) and their representation on a Grassmaninan manifold (right).

nian manifold which consists of a set of m-dimensional linear subspaces of a Euclidean space $\mathbb{R}^D$. Considering two different subspaces which are represented in $\mathcal{G}(m, D)$ as orthonormal matrices Y_i and Y_j, their span in Euclidean space can be seen on the left of Figure 3.5, while their location on the Grassman manifold is shown on the right of Figure 3.5.

Previously, when investigating Riemannian manifolds the geodesic distance (Equation (48)) was used to relate two points on the manifold. Since the Euclidean space here was defined as a collection of linear subspaces, they can be related in $\mathbb{R}^D$ by the principal angle, θ, between the span of two subspaces Y_i and Y_j and then can be directly used on Grassmann manifold $\mathcal{G}(m, D)$ as simply:

$$d(Y_i, Y_j) = ||\theta||_2 \tag{50}$$

In order to obtain the principal angles, SVD of the two subspaces can be performed in the following manner:

$$Y_i' Y_j = U(cos\Theta)V' \tag{51}$$

where $U = [u_1 \ldots u_m]$, $V = [v_1 \ldots v_m]$ and $cos\Theta$ is a diagonal matrix containing the cosines of the principal angles $cos\theta_m \ldots cos\theta_m$, which are known in the literature as *principal correlations* or *canonical correlations*. Based on these principal angles, a set of different distance metrics and kernels has been developed which are presented in

Table 3.1 and 3.2.

Table 3.1: Grassmannian subspace distances as defined by Hamm & Lee (2008)

	Distance equation
Projection distance	$d_P(Y_1, Y_2) = \left(\sum_{i=1}^{m} sin^2\theta_i\right)^{1/2} = \left(m - \sum_{i=1}^{m} cos^2\theta_i\right)^{1/2}$
Binet-Cauchy	$d_{BC}(Y_1, Y_2) = \left(1 - \Pi_i cos^2\theta_i\right)^{1/2}$
Max Correlation	$d_{Max}(Y_1, Y_2) = \left(1 - cos^2\theta_1\right)^{1/2}$
Min Correlation	$d_{Min}(Y_1, Y_2) = \left(1 - cos^2\theta_m\right)^{1/2}$

Table 3.2: Grassmannian subspace kernels as defined by Hamm & Lee (2008)

	Distance equation
Projection metric	$k_P(Y_i, Y_j) = \|\|Y_i' Y_j\|\|_F^2$
Binet-Cauchy	$k_{BC}(Y_i, Y_j) = (detY_i' Y_j)^2$

Following the example of Bito et al. (2019), collection of dynamic modes can be regarded as a set of feature vectors representing bases for a subspace. Firstly, matrices of DMD modes can be vectorised and stacked horizontally, as shown in (52), where Φ_T is a vectorised DMD matrix for a trial t, and φ_i are individual modes. From a complete feature set $\mathcal{F}_{pk}$, two trials Φ_i and Φ_j can be selected and represent two linear subspaces Y_i and Y_j.

$$M_{pk} = vec(\Phi_t)' = [\varphi_1' \varphi_2' \cdots \varphi_{r/2}'] \in \mathbb{C}^{1 \times cr/2}, \qquad \mathcal{F}_{pk} = \begin{bmatrix} M_{1,pk} \\ M_{2,pk} \\ \vdots \\ M_{t,pk} \end{bmatrix} \in \mathbb{C}^{t \times cr/2} \tag{52}$$

However, before any distance metrics explaining separation between the two subspaces can be calculated, DMD modes have to be orthogonalised. Method presented in Bito et al. (2019) specifies the use of QR decomposition to accomplish that, and is shown as:

$$Y_i = Q_i R_i, \qquad Y_j = Q_j R_j \tag{53}$$

where Q is $c \times c$ orthonormal basis for the respective subspace and R is $c \times r/2$ upper-triangular matrix containing Gram-Schmidt coefficients, relating original Y matrix to the new Q orthonormal basis. A critical note must be made here: for the above QR decomposition to be valid, Y must satisfy $c \geq r/2$ condition, otherwise QR decomposition will not produce valid orthonormal matrix Q. From there, the new orthonormal matrices Q_i and Q_j can be used in the calculation of the projection kernel to find the distance between two subspaces using the equation below, extracted from Table 3.2. The result of the iterative calculation process between every trial modes yields a $t \times t$ symmetric Gram matrix, which then can be used as features for training a classifier.

$$k_P(Q_i, Q_j) = ||Q_i' Q_j||_F^2 \tag{54}$$

3.4 Classification

A BCI system is completed with a classification module which is responsible for assigning features into respective classes through some discriminative process. However, for a classifier to be able to do that it must be firstly trained which is usually accomplished by supplying training features and correct class labels. Once the training process is complete, the whole system can be tested with testing samples to assess its performance. Here, some of the most popular classification methods used will be reviewed.

3.4.1 Discriminant analysis

One of the oldest and most popular methods used for classification is a generalisation of Fisher's linear discriminant called linear discriminant analysis (LDA) (Fisher, 1936; Lotte et al., 2007). The fundamental idea behind this approach is to approximate special boundaries using hyperplanes which allow the best separation between classes and are characterised as:

$$[w_1, \ldots, w_p]^T [x_1, \ldots, x_p] + w_0 = \mathbf{w}^T \mathbf{x} + w_0 = 0 \tag{55}$$

where, $\mathbf{w}$ is the normal vector of the hyperplane, w_0 is the threshold and $\mathbf{x}$ is the

input p-dimensional feature vector. The two main assumptions in 2-class LDA are that class-conditional distributions are normal distributions with some mean μ_c and covariance Σ_c for two classes $c \in \{1, 2\}$, and where the class covariances are set to be equal i.e. $\Sigma_1 = \Sigma_2 = \Sigma$. A class label $y = +1$ or $y = -1$ is assigned to a new feature vector $\mathbf{x}$ according to Equation (56) if the linear projection $\mathbf{w}^T x$ is above or below a threshold c as shown in Equation (57).

$$y = sign(\mathbf{w}^T\mathbf{x} + w_0) \tag{56}$$

$$\mathbf{w}^T\mathbf{x} > c \tag{57}$$

$$\mathbf{w} = \Sigma^{-1}(\mu_1 - \mu_2) \tag{58}$$

$$c = \mathbf{w}^T(\mu_1 + \mu_2)/2 \tag{59}$$

In addition to the standard formulation of LDA there are two other popular alternatives: regularized linear discriminant analysis (RLDA) and quadratic discriminant analysis (QDA). In the former, the only alteration concerns the common covariance which now becomes regulated by an additional parameter λ such that new covariance becomes as

$$\Sigma_\lambda = (1 - \lambda)\Sigma + \lambda I \tag{60}$$

In the case of QDA, the biggest difference is the assumption that the class covariance varies between the classes thus resulting in a quadratic decision boundary and not a simple linear one. Thus the boundary expression from Equation (55) is now different and is defined as the square of the Mahalanobis distance $m_c$as:

$$m_c(\mathbf{x}) = (\mathbf{x} - \mu_c)^T\Sigma_c^{-1}(\mathbf{x} - \mu_c) \tag{61}$$

which subsequently leads to an alternative way of classification changing from Equation (56) to the one shown below

$$y = sign(m_1(\mathbf{x}) - m_2(\mathbf{x}) - T) \tag{62}$$

where T is a predetermined threshold.

3.4.2 Support vector machines

The second very popular method used for classification are Support Vector Machines (SVMs) which have been exceptionally useful in both offline and online BCIs and, to this day, are considered as one of the best classifier types for BCIs (Lotte et al., 2018). SVMs have been shown to be a remarkably versatile type of classifier, working very well with both linearly and non-linearly separable data. The basic implementation of SVM is similar to the idea shown in the approach used for LDA (Equation (55)), where a hyperplane was chosen to separate two classes. It has been observed that such hyperplane could be one of potentially infinite possibilities and therefore SVM addresses that issue by looking for a separation hyperplane for which the margin (separation) between two classes, or the gap, is maximised. This concept was developed further by introducing the idea of a *soft margin* (Sain & Vapnik, 1996; Vapnik, 1999) which allowed SVM to work on data which is not linearly separable. The soft margin is an optimisation problem which incorporates the use of a slack variable ξ_i to measure the misclassification distance of the i-th input features described as follows:

$$\mathbf{w}, \overset{min}{\xi}, w_0 \left\{ \frac{1}{2}||\mathbf{w}||_2^2 + \frac{C}{K}||\xi||_1 \right\} \tag{63}$$

$$y_i(\mathbf{w}^T\mathbf{x}_i + w_0) \geq 1 - \xi_i \tag{64}$$

where C is the softening, regularisation or strength of penalty parameter, K is the number of input features for $\xi_i \geq 0, i = 1, \ldots, K$ (Rao & Scherer, 2010). While this is a useful modification to SVM, the biggest strength of SVMs lies in the use of *kernel trick*. The use of kernel functions in SVM allows mapping feature samples from one space to another without the need of transforming the whole data set into higher dimension, which is an often occurrence when data is not separable at a specific dimension. In the case of BCI, the most used and useful kernel K is radial basis function (RBF) kernel which forms RBF-SVM classifier (Lotte et al., 2007) and uses a radial width σ. In cases where the analysed data contains multiple classes, `fitcecoc()` MATLAB

function can be used to construct an ensemble of SVM classifiers (since they lack ability to classify more than two classes at time), where each SVM classifier is created using `templateSVM()` function, with specifically set kernel parameters.

$$K(x, y) = exp\left(\frac{-||x - y||^2}{2\sigma^2}\right) \tag{65}$$

3.4.3 k-Nearest Neighbour

An example of a discriminative non-linear classifier would be a family of nearest neighbour classifiers, from which k-nearest neighbours (kNN) is the most popular. As the name suggest, the idea behind those type of classifiers is to simply assign feature vectors based on their nearest neighbours, synonymous to clustering methods. Here, Euclidean distance is used to find the nearest neighbours for the corresponding feature vectors. While kNN might appear to be a very tempting option for BCI applications, it is well documented that kNN suffers from "*curse of dimensionality*" (Blankertz et al., 2002; Müller et al., 2004; Lotte et al., 2007). Research concerning machine learning often faces the aforementioned "curse", which states that as the number of dimensions grows the feature space grows exponentially. This directly affects data points as with every dimension the data points get farther apart which becomes problematic for approaches such as kNN as it loses its predictive powers if the datapoints are far away from each other. The only counter-measure for such a problem is for the original dataset to follow the same exponential growth in size, however, it quickly becomes apparent that it is a non-desirable solution for BCIs as very often it is impossible to provide more data to meet this exponential growth requirement.

3.4.4 Neural networks

As opposed to the other approaches presented in this section, neural networks (NNs) are by nature a non-linear technique which produce non-linear decision boundaries. Those non-linear characteristics of NNs have been sought after by the BCI research field for a very long time, with some first papers implementing NNs in BCI as early as 1990s (Hiraiwa et al., 1990; Anderson & Sijercic, 1996). However, the limitations of the

computing power in the early days made it hard to realise the potential of NNs, as it was also observed in other fields such as speech processing or image recognition (Lotte et al., 2018). With the rapid technological advancement and significant increase of available computational power in the recent years, variations of NNs have seen a similar increase in the implementation in the aforementioned fields and BCI systems were not excluded. Three particular variants of NNs have been widely used in the research: multi-layer perceptron (MLP, sometimes referred to as artificial neural networks), convolutional neural networks (CNNs) and recurrent neural networks (RNNs) (Craik et al., 2019).

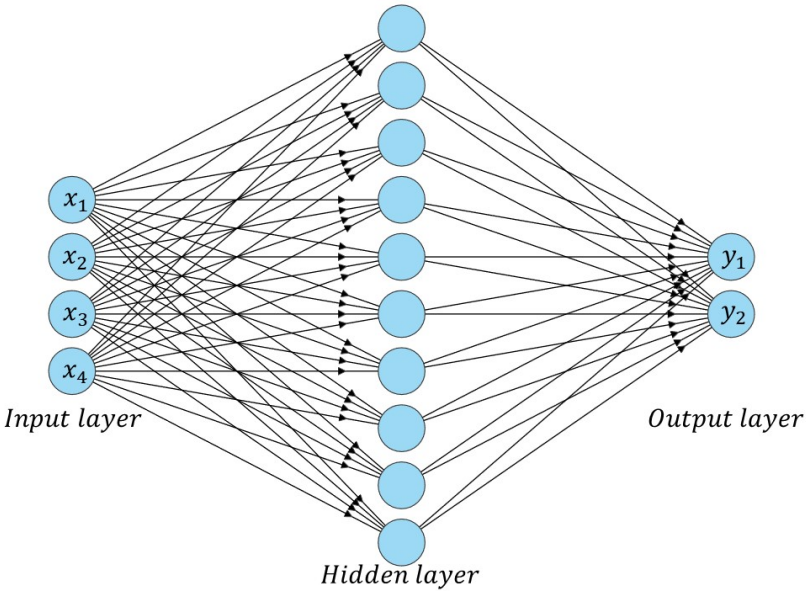

Figure 3.6: An example of an MLP network with 4 input neurons, 10 neurons in the hidden layer and 2 output neurons

One of the oldest and simplest networks used is MLP which is constructed from different layers of neurons: an input layer, several hidden layers and an output layer, where the outputs of each layer are connected directly to the inputs of the next layer, a classic example of a feed-forward network, as seen in Figure 3.6. On the interesting note, if the hidden layers were to be removed from an MLP, the resulting network simply called a *perceptron* is equivalent to LDA. MLP are classified as *universal approximators* meaning they can approximate any continuous function given a sufficient number of neurons is provided. However, this also has its drawbacks as such universal approximators are very sensitive to overtraining. Therefore a special care must be taken when implementing MLP with EEG data as its noisy and non-stationary nature

can have adverse effects on the performance of the classifier (Lotte et al., 2007).

A CNN is a slight modification of an MLP where in addition to a hidden layer a convolution and pooling layers are included in the architecture (see Figure 3.7 and 3.8). Due to the addition of those convolution layers, CNNs are particularly well-suited for image-based problems and thus have been particularly popular in BCIs which use spectrograms or scalograms as features (Sakhavi et al., 2015; Mammone et al., 2020; Bassi & Attux, 2021), however, raw EEG in combination with CNNs has also been observed in the last few years (Schirrmeister et al., 2017; Lawhern et al., 2018; Sakhavi et al., 2018; Zhao et al., 2019; Ingolfsson et al., 2020; Lashgari et al., 2021; Ko et al., 2021).

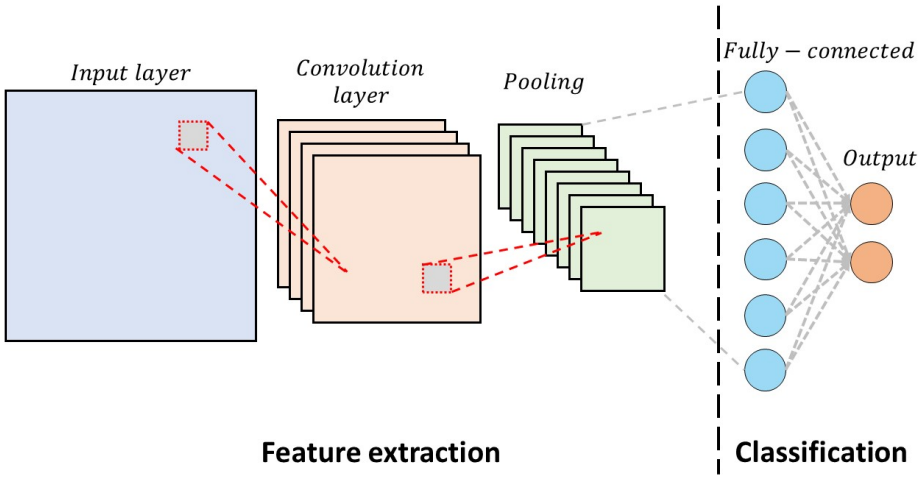

Figure 3.7: A diagram showing a simple CNN. Blocks to the left of the black dashed line are part of the feature extraction part of CNN and blocks on the right of the dashed line are part of the classification part of CNN

The convolutional layer employs a $n \times n$ kernel filter which convolves the input matrix and comes with three distinctive parameters: number of filters, stride and padding. Number of kernels dictates the depth of the output convolution layer e.g., if three kernels were chosen, the convolution layer would yield three different $n \times n$ feature maps. Stride controls the distance that kernel moves over the input matrix e.g., a value of *one* would cause the kernel to move pixel by pixel over the input matrix while a value of *two* would mean that kernel moves every *two* pixels. Lastly, padding controls the size of the output matrix. If valid or zero-padding is used then the last

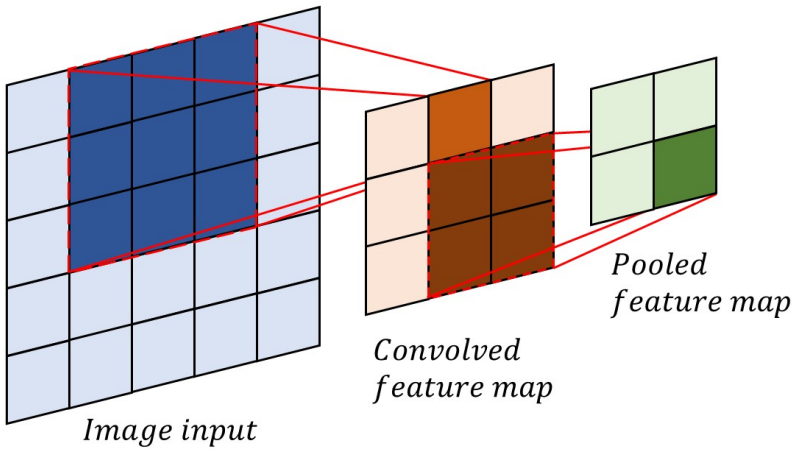

Figure 3.8: Example of convolution and pooling layers with their kernels at work. A 3×3 convolution kernel (red boundary box on blue image input) maps the convolution feature (orange). Subsequently a pooling layer reduces the dimension further by using a 2×2 kernel and maps the result on the pooling feature map (green)

convolution operation is dropped in the case that dimensions of the kernel and input do not align. Same padding ensures that the size of the output matrix is equal to the size of the input matrix. Full padding augments the input matrix by adding a border around it comprised of zero values and thus increases the size of the output matrix (Lotte et al., 2018).

A common practice is to include a non-linearity (activation function) in the convolution layer which is usually placed after convolution operations and transforms convolved matrices into their feature maps. Some of the examples of such functions would be sigmoid functions, tanh function or a rectified linear unit (ReLU) which is the most popular transformation used in CNNs. Following this transformation a further dimensionality reduction is possible by employing pooling layers. These work in the similar fashion as the convolution layers i.e. they use a kernel filter which creates a smaller feature map based either on the maximum value in a given kernel (max-pooling) or the mean of all the values (average pooling). The reduced features are then ready to be classified (usually through softmax algorithm), which in the case of a traditional CNN involves feeding features into a fully-connected layer, which in fact is just an MLP network. The more fully-connected layers the CNN contains the ”deeper” the network is, leading to the idea of deep CNNs.

The last popular architecture used for NNs is an RNN (Figure 3.9) (Dutta, 2019) and its variant long short-term memory (LSTM) (Wang et al., 2018a; Tayeb et al., 2019; Freer & Yang, 2020) which excel at working with sequential and time series data such as speech and language processing. In the contrast to other network types, RNNs rely on the prior outputs of the network from the supplied sequence. Since EEG is time series data and motor-related EEG signals have been shown to have a specific temporal structure (particularly ERPs), the motivation for using RNNs for classification of such signals becomes very clear.

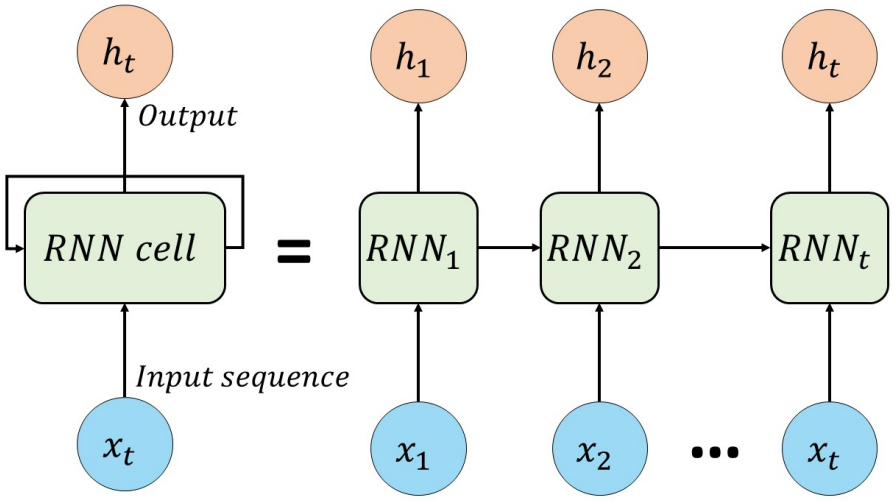

Figure 3.9: An example of a one-to-one RNN. The general diagram for RNN can be seen on the left, while on the right an unrolled RNN is presented

The architecture of RNN is similar to that of an MLP where the neurons in the hidden layer are replaced with "recurrent" cells containing neurons with addition of hidden states and loops, which allows storing past information of the input sequence. Moreover, RNN contains two sets of weights between neurons to fully exploit the past information, one for the inputs (as in a standard NN) and the second one for the hidden state. Thus, the output of the network is then based on the combination of the current input and the hidden state. As appealing as RNN might sound, it suffers from a lack of long-term dependency as well as due to being trained by a back-propagation RNN can experience a vanishing or exploding gradient problem where the network weights become either very small or large, decreasing the effectiveness of the classification. The previously introduced variant of an RNN, *LSTM*, overcomes the issue

of vanishing gradients by modifying the recurrent cell and employing additional *memory cells* which allow control of what sequence information is remembered or forgotten as seen in Figure 3.10 (Hochreiter & Schmidhuber, 1997). Here, the hidden state and input sequence are combined together before being processed by few sigmoid function layers. Memory cell from the previous step of the sequence is compared along in the cell. The memory cell and hidden state are updated and produced as outputs of LSTM cell.

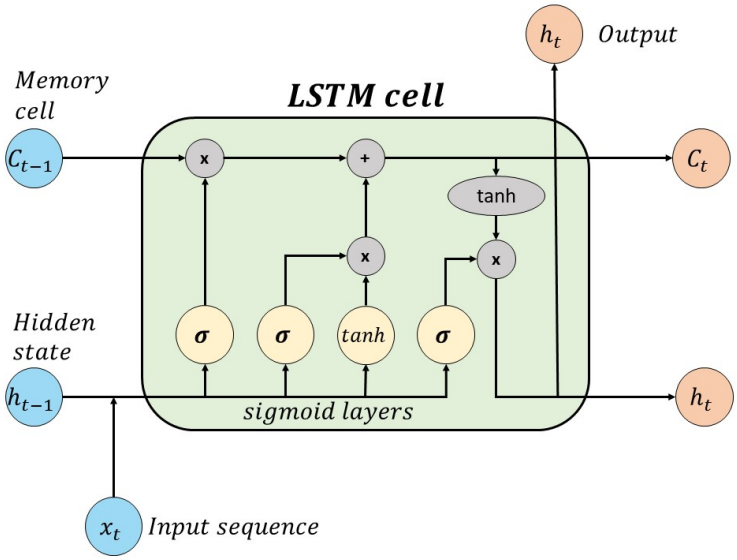

Figure 3.10: An example of an LSTM cell

As it is evident from the literature review, utilising NNs for BCIs is highly appealing as they are able to combine feature extraction, selection, dimensionality reduction and classification in a single module and thus simplify the overall system. However one must be aware of caveats found in NNs (especially CNNs), in particular the extensive use of back-propagation can lead to false sense of robustness of the system. Supplying labelled data with pre-extracted features can lead to a bias while using back-propagation meaning that neural network under training might then omit some other valuable features present in the data or completely misinterpret data. This situation would be synonymous to stating that all cars are of single specific colour, where in terms of NNs functionality the discriminant feature of a car would be its colour therefore leading to misclassification in cases where a car of a different colour was supplied to the network. It is obvious that this is not the case in the real world, as cars have other, much bet-

ter and more apparent features which differentiate them from other objects or vehicles. Therefore, while working with CNNs or RNNs, it is important to be aware of what kind of EEG data is being fed into the training thus raw EEG data is often recommended (Chiarelli et al., 2018; Amin et al., 2019a,b).

3.4.5 Performance metrics

In order to measure how well the classifier performs the author suggests to incorporate frequently used metrics in machine learning problems, namely: accuracy, sensitivity and specificity (Hudson & Cohen, 1999; Seliya et al., 2009) which can be conveniently extracted from a confusion matrix (example seen in Figure 3.11). Another metric that is very often used in MI-EEG analysis is Cohen's kappa (κ) value (Brunner et al., 2008; McHugh, 2012). When investigating a confusion matrix four terms are used to describe possible outcomes of the classifier: a True Positive (TP), a True Negative (TN), a False Positive (FP) and a False Negative (FN). These refer to an outcome correctly indicating a presence of a certain condition, a correctly observed absence of a certain condition, incorrectly indicating presence of a certain condition and incorrectly indicating absence of a certain condition respectively.

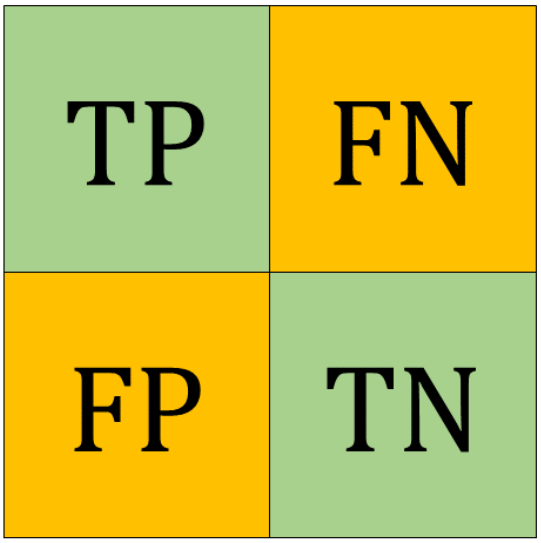

Figure 3.11: Example of a simple confusion matrix.

The total number of all positive outcomes is denoted as P and, similarly, the total number of all negative outcomes is denoted as N. In binary classification problems accuracy is defined as the ratio between sum of TP and TN and sum of all P and N and it depicts how well the classifier correctly identifies conditions. Sensitivity, also referred to as recall or true positive rate, shows the ratio between TP and the sum of TP and FN. It indicates the probability of the observed true outcome to be definitely true, an important metric to be considered when implementing a BCI since a higher sensitivity would mean that the BCI definitely issued to correct command leading to less frustration from the user. Specificity, also referred to as true negative rate, can be seen as the opposite of sensitivity where it indicates the probability of the observed negative outcome to be definitely negative. Lastly, kappa value is used to measure the reliability of the used data. Low scores indicate that there is no obvious connection in the provided data which would allow robust classification. It is described as the ratio between the difference of observed and expected accuracy (p_o and p_e respectively) and the probability of a random guess ($1 - p_e$). The advantage of using kappa value to measure performance is that it also indicates how much better the classifier performs than a simply guess classifier. For the convenience and future reference all the metrics discussed above are summarised in Table 3.3.

Table 3.3: Metrics used for evaluating performance of the classifier.

	Equation
Accuracy, %	$\frac{TP+TN}{P+N}$
Sensitivity	$\frac{TP}{TP+FN}$
Specificity	$\frac{TN}{TN+FP}$
Cohen's kappa	$\frac{p_o-p_e}{1-p_e}$

Calculating accuracy, sensitivity and specificity becomes a bit more challenging when tackling multiclass problems. Figure 3.12 shows how a multiclass confusion matrix can be interpreted allowing to still implement the equations from Table 3.3 in order to calculate the aforementioned performance metrics. As seen in Figure 3.12 TP can be any diagonal entry in the confusion matrix. FPs are now regarded as all entries in the same column as TP and FNs are all entries in the same row as TP. Lastly, TNs

Table 3.4: Interpretation of kappa values, κ, per McHugh (2012)

κ value	Level of agreement	% of reliable data (κ^2)
0 - 0.20	None	0-4%
0.21 - 0.39	Minimal	4-15%
0.40 - 0.59	Weak	15-35%
0.60 - 0.79	Moderate	35-63%
0.80 - 0.90	Strong	64-81%
> 0.90	Almost perfect	82-100%

are all other entries in the confusion matrix that do not fall into any of the mentioned places. While accuracy can still be easily extracted from a multiclass confusion matrix as it is just a sum of all the diagonal outcomes divided by the number of all outcomes, specificity can be a subject to calculating — not necessarily incorrect, but slightly misleading probabilities. This is due to an inflated number of TNs as seen in Figure 3.12. Sensitivity does not suffer from such problems and can be calculated as previously.

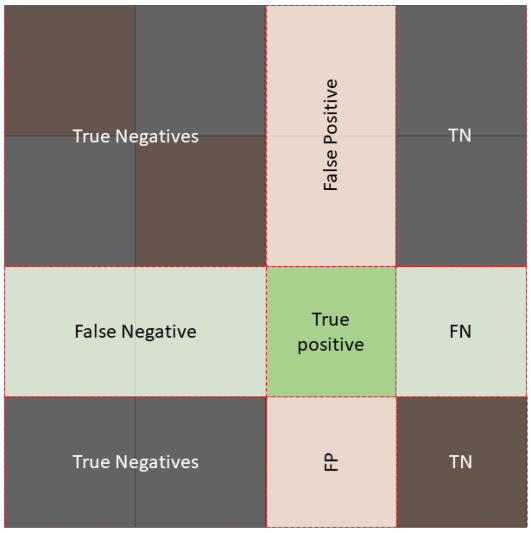

Figure 3.12: Example of a confusion matrix in a multiclass problem: green square represents a certain true class (TP), other values at that row are the false negatives (FN), while the column values represent false positives (FP). All dark squares are treated as true negatives (TN)

3.5 Conclusion

The presented literature in this chapter shows that analysing EEG signals is not a trivial task and requires a sophisticated and complex systems to accurately translate the incoming data into comprehensible commands or outputs. Due to the complexity of a BCI and many interconnected factors within the system dictating its performance, focusing on a single submodule would most likely not provide a substantial increase in the performance. Therefore, a BCI needs to be worked on as a whole. The first module touched upon preprocessing methods: temporal and spatial filtering. In general, the reviewed literature agreed on the need of filtering; however there were instances where this topic came under great scrutiny arguing potential loss of important data in the said process. However, the author notes that in some sense this dilemma cannot be fully solved, as extracting spatial-based features from motor imagery works best when the initial EEG signal is band-pass filtered into a meaningful bandwidth i.e., 7-30Hz, which enhances the ERD/S phenomenon. Following the preprocessing module, different feature extraction techniques were discussed. There, the author fully introduced DMD method fully explaining its functionality and possible means of extracting valuable features. The next module concerned different techniques used for choosing the best features or transforming features into more suitable representations. Along with statistical methods, the author also reviewed feature selection using PCA method as well as introduced a recent technique which involved projecting initial features onto a Grassmanian manifold which in turn allowed to calculate a new matrix based on distance between DMD modes on the said manifold. This was followed by a discussion concerning different classifiers, from classic methods such as LDA and SVM to newer and more advanced techniques of neural networks. Lastly, the author introduced four performance metrics: accuracy, sensitivity, specificity and kappa value, which are used to measure how well the implemented systems perform. Next chapter utilises the techniques discussed in this chapter to propose new processing pipelines for a BCI system based on the DMD approach.

Chapter 4

Methodology

The literature discussed in the background chapter presented a number of different methods for measuring and recording brain signals. That knowledge combined with the remarks concerning EEG signal in the introductory chapter of this thesis makes a valid case for using EEG signals. Furthermore, the discussion about the use of ERPs and continuous EEG for investigating imagination of movement showed that continuous EEG, ERD/S phenomenon specifically, is a better suited modality for this purpose. This is due to its neurophysiological processes and characteristics, especially the ability to appear in recordings which do not employ a cue-based system, i.e. the user executes or imagines actions at their own pace without any external cues.

Further investigation into the methods commonly used in analysing and extracting features from such signals revealed that the most popular method is a spatial-based CSP approach. Thus, the proposed DMD method, which produces spatio-temporal patterns from the analysed signals, is a justifiable approach as it builds on the previously successful spatial-based methods. In addition, the literature review showed only few academic publications which considered DMD in some brain signal analysis tasks, strengthening the novelity of DMD in the analysis of MI signals.

An essential part of the work required during the doctoral studies covered in this thesis was data collection through experimental means. Initially, the author intended to carry out a medium scale experiment recording self-paced (asynchronous) EEG signals from 10 subjects. Such experiments allow the volunteers to execute an action at their own pace without the need for waiting for a cue to appear, making the experiment

much closer to how BCI would operate in real-life situations. During the recordings the subjects would have been asked to perform and imagine (in separate trials) any of the 4 movements: elbow flexion, shoulder flexion, extension or abduction. Throughout the duration of the trials, EEG signals would have been recorded from the sensorimotor cortex area with Ag-AgCl disc electrodes placed according to 10-10 system. Such experimental paradigm was devised as the experiment was intended to be a collaborative effort with a project focusing on the development of upper-limb exoskeleton for rehabilitation purposes.

However, a number of factors made it impossible to perform the experiments described. At the start of the doctoral studies in October 2017, the entire Biomedical Engineering Department had to vacate the main building as renovation works began. These were meant to take only 3 months, however, this period overextended to almost 3 years (until November 2020). In theory, this should not have had such an impact on the experiments, however the temporary substitute facilities, which were assigned to the research group, suffered from great power line noise rendering any recorded EEG signals unreadable and unusable.

A suitable room was found only in November 2019, however it required to be appropriately converted to facilitate the requirements for recordings, thus it only became fully functional in January 2020. By the time all the necessary arrangements and ethics[1] were amended appropriately to accommodate the new recording room, entire UK went into the first lockdown in March 2020 due to COVID-19 pandemic and all the work had to be halted. The prolonged lockdown and restriction measures made it impossible to recruit any participants for the outlined study. By the time the university campus reopened in August 2021 it was too late to conduct any experiments due to the time constraints of the PhD.

The uncertainty associated with the development of the pandemic forced the author to explore alternative data collection methods. This led the author to decide to use 3 publicly available datasets: BCI Competition IV Dataset 2a (Brunner et al., 2008), BNCI Horizon 2020 (Ofner et al., 2017) and GIST-MI (Cho et al., 2017), with the

[1]Initial ethics application was granted on the 23rd of January 2020 (ID:1394, "Study of free-will movement intention using brainwave analysis".

addition of data recorded at the local laboratory from a previous study (Syam, 2017). All these datasets are summarised in the section 4.1, providing the general information regarding: the electrode setup, recording parameters and the paradigm used for experiments. In the next parts of this chapter, the author introduces the proposed BCI, outlining the complete structure of the processing pipelines. All preprocessing steps are explained in detail, following the recommendations from the previous chapter. After that, the author presents three different pipelines based on three different DMD features: DMD modes, DMD spectrum and DMD maps. In each case, the full process is thoroughly explained in terms of how each feature is extracted and transformed, before appropriate feature selection technique is shown. Each processing pipeline is then concluded with a choice of a classifier, outlining the parameters used during the training.

4.1 Dataset description

Thorough and rigorous testing is a common practice in literature, and demonstrates sufficient robustness of the presented approaches to different types of data and lack of bias. The first dataset, BCI Competition IV Dataset 2a (see 4.1.1), can be considered to be the MI-EEG equivalent of 'MNIST' database[2]. Just as 'MNIST' is regarded to be the fundamental dataset used for training various image processing systems and machine learning, BCI Competition IV Dataset 2a holds the same level of importance in MI-EEG field. Second dataset used (BNCI Horizon 2020, see 4.1.2) comes from an in-depth study exploring multiple classes of movement and imagination. It is worth noting that both of those datasets come from world renown BCI laboratories and have been widely used in the research community, thus showing their credibility and reliance. The third dataset (GIST-MI, see 4.1.3) is from an extensive study on motor-imagery containing additional recordings on non-task related EEG such as rest or facial movement. The last dataset (see 4.1.4) comes from past experiments carried out at the Neurophysiology Laboratory in the Department of Biomedical Engineering at the University of Strathclyde.

[2] http://yann.lecun.com/exdb/mnist/

4.1.1 BCI Competition IV Dataset 2a

This dataset recorded by the laboratory in Graz (Brunner et al., 2008) consists of EEG data obtained from 9 healthy subjects. A cue-based paradigm was employed to record four different motor-imagery tasks which involved left hand (class 1), right hand (class 2), both feet (class 3) and tongue (class 4). EEG signals were recorded with 250 Hz sampling frequency, bandpass filtering between 0.5 Hz and 100 Hz, with twenty-two Ag/AgCl electrodes placed according to the international 10-20 system (Figure 4.1). In addition to the mentioned EEG, three EOG channels were recorded as well to be used for artefact rejection. These signals were sampled at the same 250 Hz frequency as EEG signals.

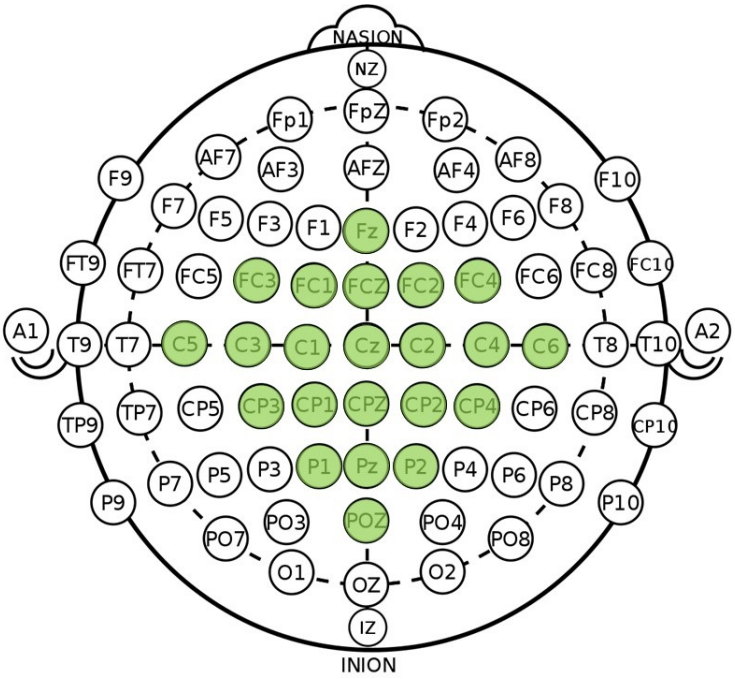

Figure 4.1: Electrode montage used in BCI Competition IV Dataset 2a. Green-coloured electrodes are the electrodes which were used for recording EEG signals.

Each trial started with a fixation cross on a black screen and an audible beep ($t = 0s$). A cue corresponding to one of the four motor-imagery tasks appeared on the screen after 2 seconds ($t = 2s$) and stayed on for the next 1.25 seconds. This in turn prompted the subjects to perform the displayed task and sustain it until the disappearance of the cue at $t = 6s$. The subjects then were given few seconds break before the onset of the next trial. Subjects had 6 experimental runs, where each run

contained 48 trials (12 for each class), meaning that each subject performed 288 trials in total. The diagram for this paradigm can be seen in Figure 4.2.

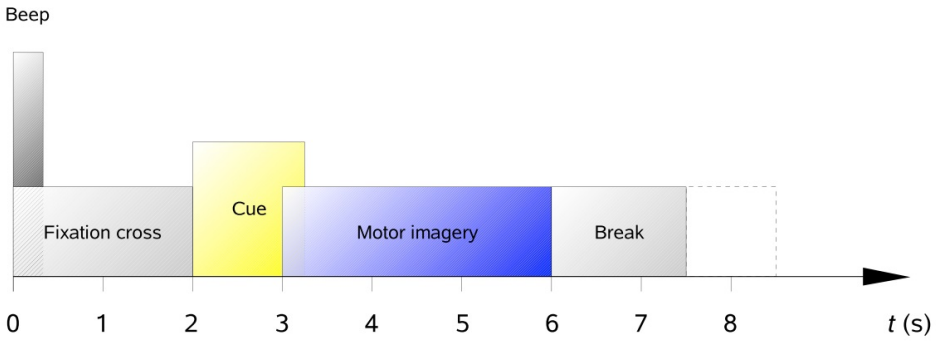

Figure 4.2: Paradigm used in BCI Competition IV Dataset 2a recordings. Sourced from Brunner et al. (2008).

4.1.2 BNCI Horizon 2020

A more recent dataset from the Graz Laboratory contains recordings from 15 healthy subjects performing both motor execution and motor imagery tasks. The number of classes has been extended to six and now accommodate the following tasks: elbow flexion, elbow extension, forearm supination, forearm pronation, hand close and hand open with additional 'rest' (non-task activity) class. The EEG signals were recorded with 61 electrodes covering frontal, central, parietal and temporal scalp areas with surface active electrodes from g.tec medical engineering, with referenced placed on the right mastoid and the ground placed on AFz location. Sampling frequency was set to 512 Hz and an 8^{th} order Chebyshev bandpass filter between 0.01 Hz and 200 Hz with addition of a notch filter at 50 Hz was applied. Arm joint angles and finger positional data were also recorded with the help of an exoskeleton with anti-gravity support and 5DT Data Glove. The location of EEG channels used in this dataset is seen Figure 4.3 and their description can be found in the supportive documents in Ofner et al. (2017).

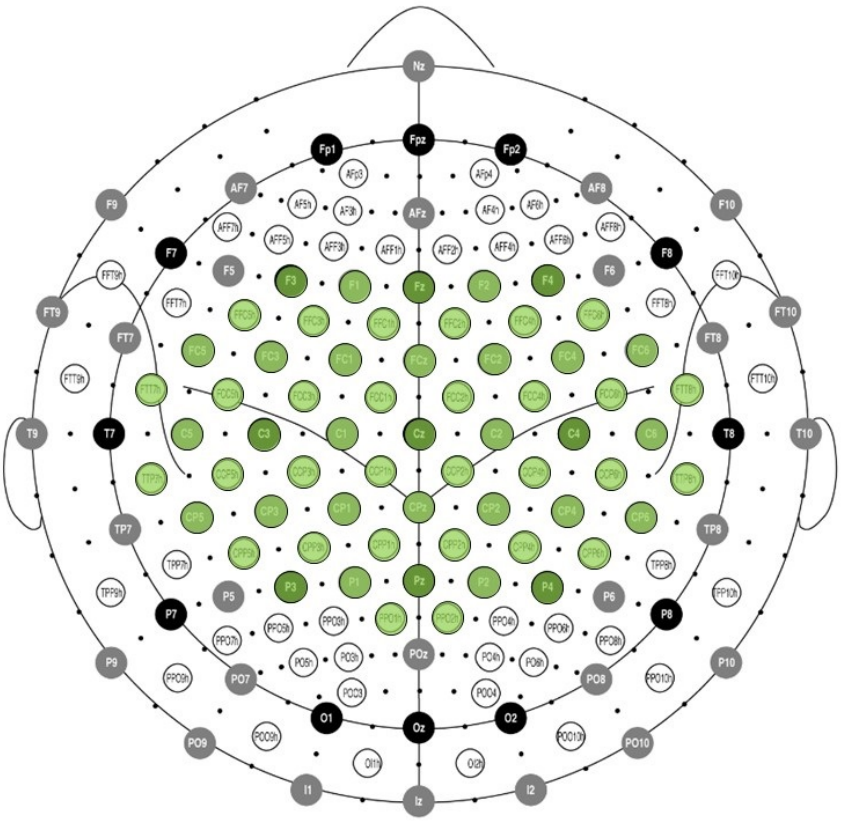

Figure 4.3: Electrode montage used in BNCI Horizon 2020. Green-coloured electrodes are the electrodes which were used for recording EEG signals.

A cue-based paradigm similar to the one from Brunner et al. (2008) was used to record trial runs in this experiment. The start of the trial was marked by a beep sound and an appearance of a fixation cross on the screen in front of the seated subject at $t = 0s$. At $t = 2s$ the subject was presented with a pictorial cue showing the required task and was asked to sustain it for the next 3 seconds until $t = 5s$. In the motor execution trials the subject would move from neutral position to the required position, sustaining it until moving back to neutral position after the disappearance of the cue at $t = 5s$. In the motor imagery trials the subjects would similarly sustain the thought of the movement until the disappearance of the cue. Subjects were given a few second break before starting the next trial. Each subject had 10 experimental runs with 42 trials per run. This resulted in recording 60 trials for each class (including the resting class) for each subject. The diagram for this paradigm is shown in Figure 4.4.

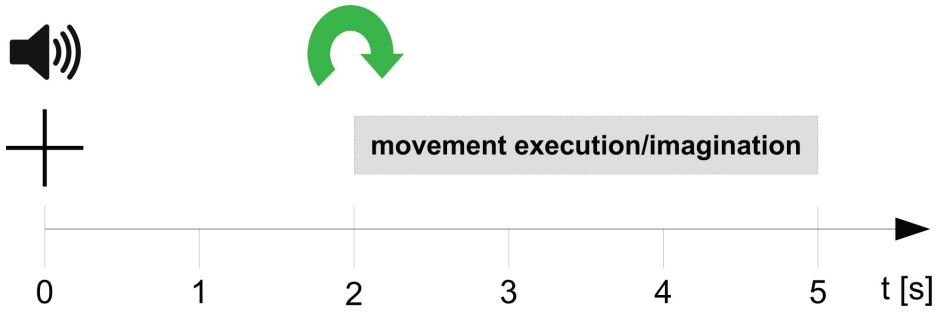

Figure 4.4: Cue-based paradigm used in BNCI Horizon 2020 for the EEG recordings. Sourced from Ofner et al. (2017).

4.1.3 GIST-MI

A rich dataset focusing on recording motor imagery movement signals is presented by Cho et al. (2017). A total of 52 healthy subjects took part in this substantial study. 64 Ag/AgCl active electrodes placed according to the international 10-10 system were used to record EEG signals with sampling rate of 512 Hz, with addition of four EMG electrodes placed on the subjects' forearms to monitor muscular activity. The placement of the EEG electrodes and their relative channel numbers in the dataset can be seen in Figure 4.5.

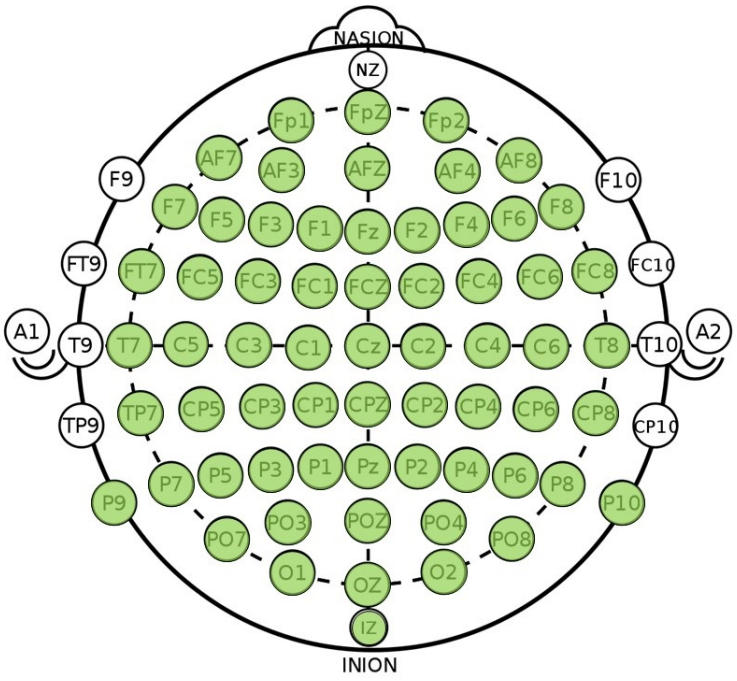

Figure 4.5: Electrode placement used in GIST-MI. Green-coloured electrodes are the electrodes which were used for recording EEG signals.

The significance of this dataset is such that it in addition to motor execution and imagination recordings, it also contains non-task continuous EEG data for six 'noise' signal types: eye blinking, eyeball movement (horizontal and vertical), head movement, jaw clenching and resting state for all 52 subjects recorded in 20 trials . The motor execution recordings only covered a two class problem between left and right hand movement; while during motor imagery recordings the subjects were asked to imagine finger movements as shown in Figure 4.6, other publications treat the data as a two-class problem between left and right hand movement, disregarding individual finger movement as separate classes. However, each class was recorded with 100 or 120 trials depending on a subject.

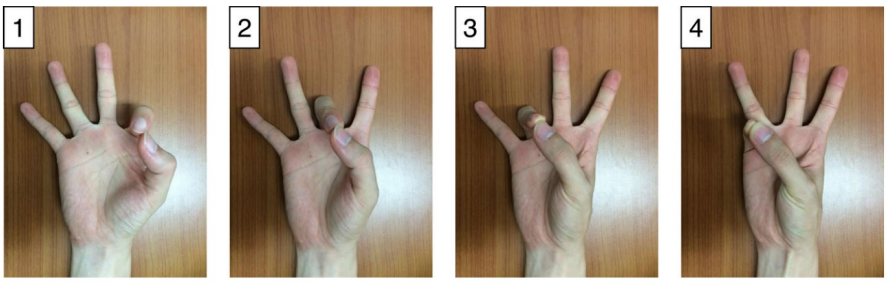

Figure 4.6: Instructions for the motor imagery recordings as seen in Cho et al. (2017).

The recording paradigm used in this study is synonymous to the one used in Brunner et al. (2008) and Ofner et al. (2017), although a different time notation is used. Each trial started with a fixation cross on the black screen in the front of the seated subject at $t = -2s$ with a cue appearing on the screen at $t = 0s$. The cue would specify which hand movement is to be moved (left or right) and would stay on the screen for the duration of 3 seconds, during which the subject would maintain the shown task. At $t = 3s$ the fixation cross reappeared and the subject could return to neutral position and have a few second break before the start of the next trial. Figure 4.7 shows the described paradigm.

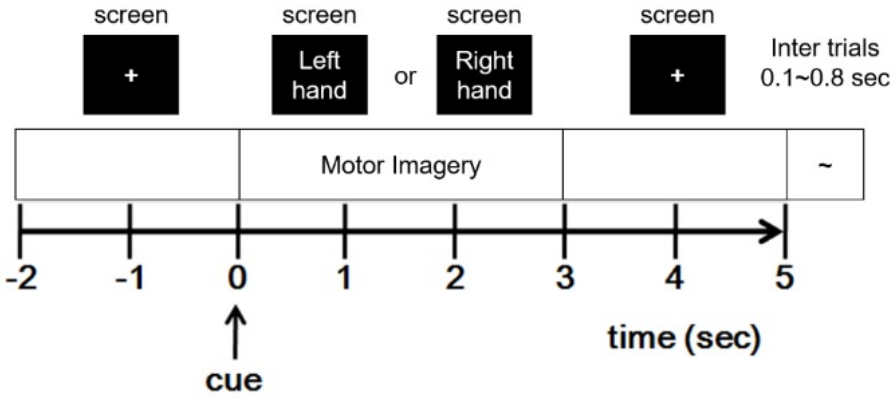

Figure 4.7: Paradigm used in GIST-MI for the EEG recordings. Sourced from Cho et al. (2017).

4.1.4 Syam (2017)

The author also had access to data previously recorded at the local laboratory. This data was part of a previous doctoral study concerning BCIs for spinal cord injury patients (Syam, 2017), in which a total of 29 subjects were involved. Eleven of those were healthy subjects, four were paraplegic and fourteen were tetraplegic patients. The recordings were split into two types: motor execution and motor imagination, during which various movements of the right fist were investigated. A specifically designed manipulandum was used to investigate extension, flexion and ulnar/radial deviation of the right wrist. EEG signals were recorded with 28 Ag/AgCl sinterted ring electrodes placed according to the international 10-10 system (shown in Figure 4.8), with reference electrode placed on the earlobe and the ground placed at AFz electrode. Signals were recorded with 2000 Hz sampling rate and a bandpass filter between 0.05 Hz and 500 Hz. In addition, 4 EMG channels were used to monitor muscular activity in the subject's forearm.

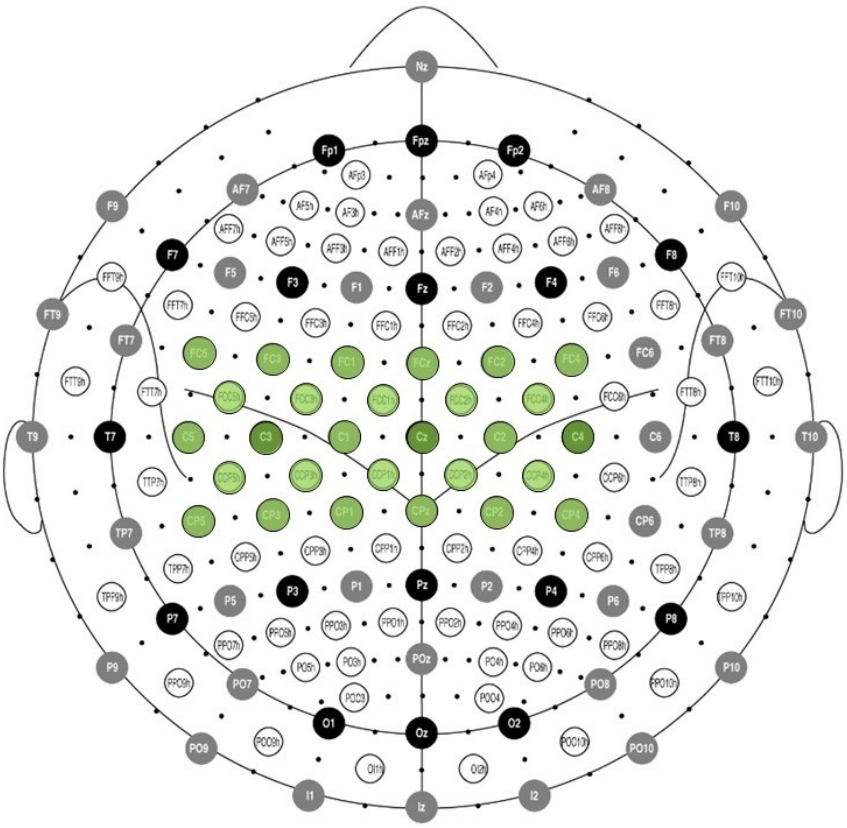

Figure 4.8: Electrode placement used in Syam (2017). Green-coloured electrodes are the electrodes which were used for recording EEG signals.

The paradigm employed in this study slightly differed from the others presented before. The trials were split into two phases A and B. During Phase A an instruction was shown indicating which direction the fist should move, while during Phase B the subject was resting. In terms of timing the paradigm proceeded as follows: 7 seconds after the display of Cue B (representing return to neutral position), Cue A would be displayed with the cue for the direction of movement. This cue would last for 3 seconds, after which Cue B (return to neutral position) would be displayed. The next movement trial would take place as described in Figure 4.9. This paradigm was used during both motor execution and imagination trials. In total every user performed 50 trials for each movement class (200 trials in total).

The original intent was to utilise this dataset in the thesis to extend the viability of the proposed DMD methods on signals recorded from patients who suffered from either tetraplegia or quadriplegia, as the brain signatures tend to be much harder to discern. However, during the investigation of this dataset few serious issues were noticed, making

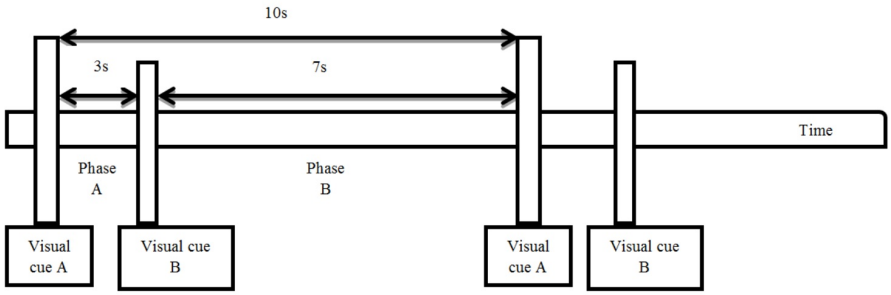

Figure 4.9: Timeline of the implemented paradigm. Sourced from Syam (2017).

the dataset unfit for further analysis. Most importantly, almost half of the data was contaminated with wire movement artefacts during the activity window, which meant that features extracted from them would not be accurate. In addition to that, other artefacts were also present in the activity window, which could be attributed to poor connection between the electrodes and the scalp area.

4.2 Preprocessing

The processing pipeline used in this thesis follows the standard structure of a BCI, as outlined in the previous chapters, and its complete structure can be seen in Figure 4.10. However, before the EEG data could have been used for extracting valuable features, the trials from the used datasets were firstly put through standard preprocessing steps: epoching and filtering, where filtering was implemented through a combination of various spectral and spatial filters. This preprocessing pipeline was fully implemented in MATLAB 2020b.

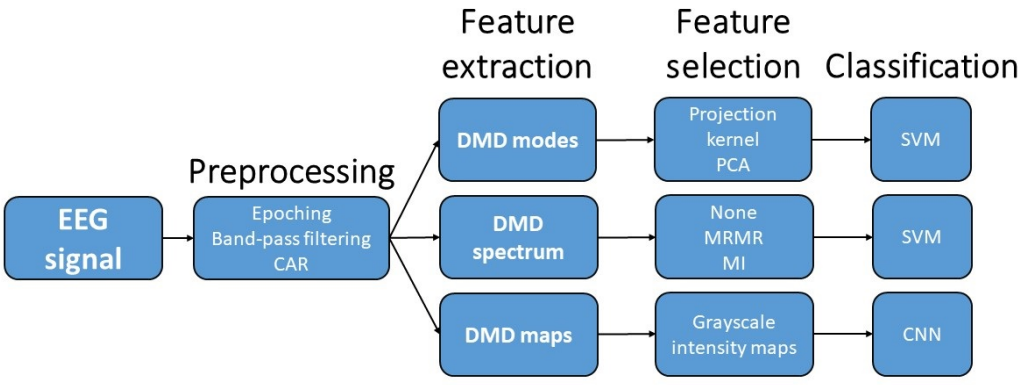

Figure 4.10: The processing pipelines used in this thesis

Before any of the preprocessing operations took place the datasets had to be firstly imported to MATLAB environment. Since the available datasets were either stored in `*.gdf` or `*.cnt` file format, `sload()` function from BioSig toolbox was used to load the files into MATLAB.

4.2.1 Epoching

All of the datasets used in this thesis were provided as continuous EEG signals therefore epoching was necessary so that exact windows of interest i.e. when motor imagination took place, could be extracted and used for the subsequent feature extraction. The initial window for a complete trial and the extracted activity window varied between datasets thus those values were collected and recorded in Table 4.1 for ease of reference. Once the trials were epoched, they were stored in a `cell array` which had size of $s \times l$, where s was the number of subjects and l was the number of classes in a given dataset. Each entry in the `cell array` contained a 3-D matrix of epoched data from a single run such that its dimensions were $c \times m \times n$, where c was the number of EEG channels, m was the number of recorded samples and n was the number of trials.

Table 4.1: Time values used for window extraction from datasets, where the $0s$ reference point is set to the appearance of cue.

	Trial window	**Extracted activity window**
BCI IV 2a	$-2s - 4s$	$1s - 4s$
BNCI 2020	$-2s - 3s$	$0s - 3s$
GIST-MI	$-2s - 3s$	$0s - 3s$

4.2.2 Filtering

This thesis analysed motor imagery as an oscillatory EEG signal recorded at specific electrode positions, where such activity could be found in distinctive frequency bands (α- and β-bands). The author would like to note here that while this interpretation of motor imagery signals might suggest use of ERD/S it is critical to understand that at no point in this thesis conventional methods of calculating ERD/S are used, such as

the ones seen in Pfurtscheller & Lopes Da Silva (1999). Instead, the effects of ERD/S phenomenon on the previously mentioned frequency bands were exploited following the example of Blankertz et al. (2007), where it has been shown that CSP algorithm was well fit to detect ERD/S effects in α and β waves if filtered accordingly. Based on that evidence the choice of applying a bandpass filter between 7Hz and 30Hz was made.

Following that, spatial filtering was applied to investigate the effect of spatial filtering on DMD modes. For that purpose `spatialfilter()` function from BioSig toolbox with the appropriate input arguments was used. After completing the preprocessing steps the windowed signals were narrowed down to the activity window size (as noted in Table 4.1) to trim any discontinuities (transients) usually left over after applying Fourier-based filtering. The trimmed data was then moved to the feature extraction module, where it was processed by the DMD algorithm to extract three different types of features: DMD modes, DMD spectrum and DMD maps. The obtained features were then paired with appropriate feature selection methods, before finally being classified with a fitting technique. Each of the feature extraction processes is described in the following sections.

4.3 DMD modes processing protocol

The first type of features which were investigated in this thesis were DMD modes. These modes also served later as the basis for the other explored features, DMD spectrum and DMD maps. The extracted modes had $hc \times r$ size, where h is the stacking factor, c is the number of channels and r is the r-rank truncation value; however $c \times \frac{r}{2}$ matrix was used for subsequent analysis, as additional rows past $c - th$ row were shift-stacked copies and every second mode was selected since DMD modes come in conjugate pairs (Section 3.2.8). These modes were then passed through two different feature selection methods: projection kernel and PCA. Lastly, the final features were used to train an SVM classifier. The full processing pipeline is depicted if Figure 4.11

Firstly, DMD modes were extracted using both natural and SVD-energy scaling methods in order to determine which technique produced better features. Speculatively, if natural scaling was to be used, then the spatial characteristics of motor imagery could

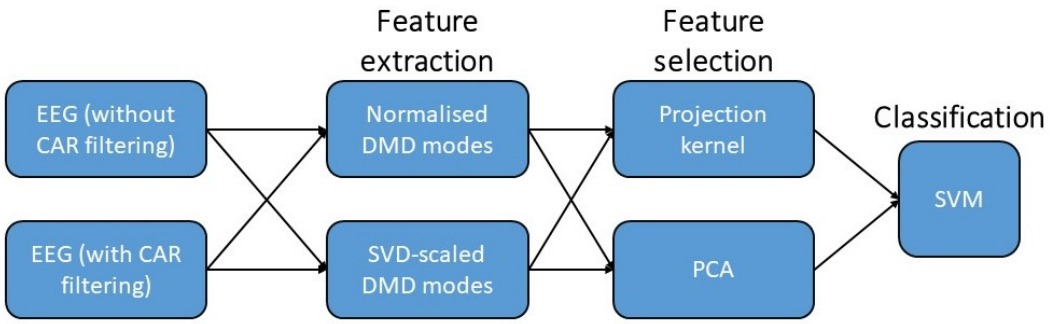

Figure 4.11: Diagram showing the processing pipeline for DMD modes. The effect of spatial filtering is assessed by extracting both normalised and SVD scaled DMD modes. Projection kernel and PCA techniques are then used to transform DMD modes features before being used to train an SVM classifier.

be exploited, where most dynamic importance should be found in central belt of electrodes, especially C1-C3 and C2-C4 electrodes. Alternatively, assessing modes scaled by their SVD energy should theoretically uncover information regarding ERD/S phenomenon, as the 'power' changes of modes located at α and β bands can be examined. Therefore, investigating both scaling techniques would allow to assess the two stated conjectures.

With the two scaling techniques applied, exact DMD modes were calculated as shown in (31) with the addition of a vector of eigenvalue frequencies (see (32)), so that relative DMD modes could be sorted in an ascending frequency order. After those operations, DMD modes for a given trial were a matrix of $hc \times r$ size. Lastly, conjugate DMD modes were removed from the matrix and c number of rows was extracted, so that the final matrix of DMD modes for a given trial was $c \times \frac{r}{2}$.

Secondly, the extracted DMD modes were subjected to two feature selection methods which have been reviewed in the previous chapter: PCA and a Grassmanian manifold projection techniques. In the case of PCA method, the process was followed just as described earlier, however when using Grassmanian manifold a small modification to the process had to be made. Originally, it has been stated that for the QR decomposition to be valid, the size of the matrix has to satisfy $c \geq r/2$ condition. Given the size of the calculated DMD matrices was $c < r/2$, the author applied a mode binning technique to reduce the number of $r/2$. Region spanning between 7Hz and 31Hz was split into equal 2Hz bins and modes which fell into each bin were averaged, such that

each bin was described by an average mode. By doing so, r was reduced to 12, thus satisfying the $c \geq r/2$ condition.

Lastly, the selected features were classified using an SVM classifier. For the multiclass problem, firstly a binary classifier template had to be created. The binary SVM classifier was specified to use RBF kernel which calculated the elements of the Gram matrix based on the supplied features, and the scale of the kernel was set to be automatically determined during the training process. The best-fit hyperplanes used to separate the data during the training were optimised with the help of $l1$ soft-margin minimisation solver.

Using the 'one versus one' approach for multiclassification mean that the classification model stored $n(n-1)/2$ binary SVM classifiers; a classifier for each possible binary combination of the supplied classes. The optimal hyperparameters used for training of the multiclass model were found with bayesian optimisation technique. The performance of the trained models was monitored using an acquisition function which evaluated the expected amount of the improvement in the optimiser, where the additional '*plus*' parameter allowed escaping a local minima during computation. The list of the parameters can be found in Table 4.2.

Table 4.2: Parameters used for the binary SVM classifier and the multiclass classification model.

binary SVM classifier		**Multiclass model**	
Kernel function	'rbf'	**Coding**	'onevsone'
Kernel scale	'auto'	**Optimizer**	'bayesopt'
Solver	'L1QP'	**Acquisition Function Name**	'expected-improvement-plus'

4.4 DMD spectrum processing protocol

The second type of features investigated in this thesis are the power values of the DMD spectrum. As per Chapter 8 and 12 of Kutz et al. (2016) and earlier sections of this thesis, power spectrum of DMD modes which are scaled by their SVD energy resembles the average FFT of the same windowed signal. Therefore, similarly to the motivation presented when looking into DMD modes, the author proposes assessment of DMD spectrum with the aim of discovering valuable features which could be attributed to ERD/S phenomena. The experimental procedure involving DMD power spectrum features is shown in Figure 4.12 below.

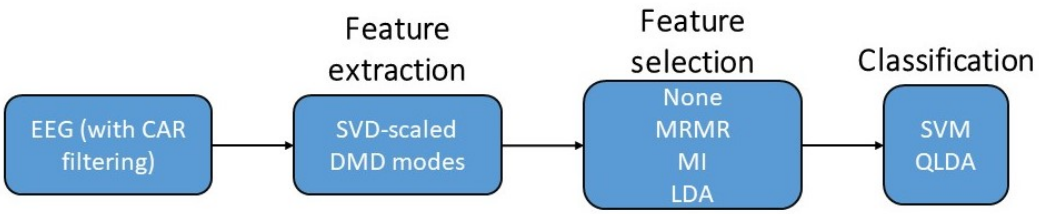

Figure 4.12: Diagram showing the processing pipeline for DMD spectrum.

SVD-scaled and frequency sorted DMD modes obtained from Section 4.3 were inspected further to extract the DMD power spectrum for Φ_t which can be calculated by using (37) such that $P_t = ||\Phi_t||$. The resultant power vector P_t is a $1 \times r/2$ row-vector, which can be plotted against the related frequency values to obtain a spectrum similar to average FFT power spectrum. Since the characteristic frequency of each mode can vary between trial, the calculated power values were averaged in 1Hz bins between 7-30Hz bandwidth, to keep frequencies consistent over different trials. The above binning process reduces P_t vector to 1×24 size allowing to create a feature space $\mathcal{F}_P$ by horizontally stacking power vectors P_t such that

$$\mathcal{F}_P = \begin{bmatrix} P_1 \\ P_2 \\ \vdots \\ P_t \end{bmatrix} \in \mathbb{R}^{t \times 24} \tag{66}$$

Three different approaches have been employed for feature selection in addition to supplying unmodified $\mathcal{F}_P$ as it is to the classifier. Firstly, MRMR technique was implemented to find the most important features from $\mathcal{F}_P$; the in-built Matlab function `fscmrmr()` was used to extract 5 most important features. In a similar fashion an MI algorithm was deployed to extract 5 most important features. For this an external function `MI()` provided by Pohjalainen et al. (2015) was used. Both of the functions operate as stated in Section 3.3.1.

The author initially planned to also include LDA as an additional feature selection, which throughout the literature is found to be a common choice for dealing with bandpower or FFT power spectrum feature vectors. An LDA model was created by supplying $\mathcal{F}_P$ features and the associated class labels to a `fitcdiscr()` function. To maximise the accuracy of the classifier, a 10-fold cross validation was additionally implemented. After the training was completed, the structure of the best performing model was explored to obtain `DeltaPredictor` values, which measure importance of the supplied predictors.

The values from `DeltaPredictor` were further checked for statistical importance in order to extract the most meaningful features. Firstly, the mean was removed from all the extracted values. This was followed by finding predictors which value was equal or more than that of double of standard deviation. The choice of double of standard deviation follows the idea of normal distribution, where 95% of data is contained within two standard deviations, meaning that data with higher standard deviation values would be of importance, and thus could be selected as meaningful features. Originally, the predictors which met the above criteria indices of the three best performing features were supposed to be chosen as the final ones; however during testing stages the author discovered that this method was highly unreliable and very often `DeltaPredictor` would be empty, meaning that the LDA mode failed to create a correct classifier from the provided features. Because of that the author decided to not include LDA as a feature selection method.

4.5 DMD maps processing protocol

Throughout the thesis, the author has highlighted the importance of spatial relations in EEG signals multiple times. The highly localised nature of ERD/S phenomena and its almost exclusive appearance at specific areas of the scalp should be enough to encourage development of some form of image-based analysis for motor imagery applications. As it was noted in Section 3.4.4, the rise of neural networks, and especially CNNs, allowed such analytical approaches to be researched and developed, since CNNs excel at looking for spatial features and relations in images. Recalling further, it has also been shown that maps extracted from FFT, STFT or scalogram features were indeed quite successful in classifying motor imagery problems. Therefore, the author decided to investigate whether maps produced by DMD modes could be equally well understood by CNNs and yield a satisfactory performance in terms of accuracy of classification. Thus, the last explored feature type assessed in this thesis are maps produced by DMD modes. The processing pipeline is shown in Figure 4.13. Both normalised and energy-scaled DMD modes were used to assess which scaling method performs better. Intensity maps were then obtained for both scales, before being utilised as input to CNN.

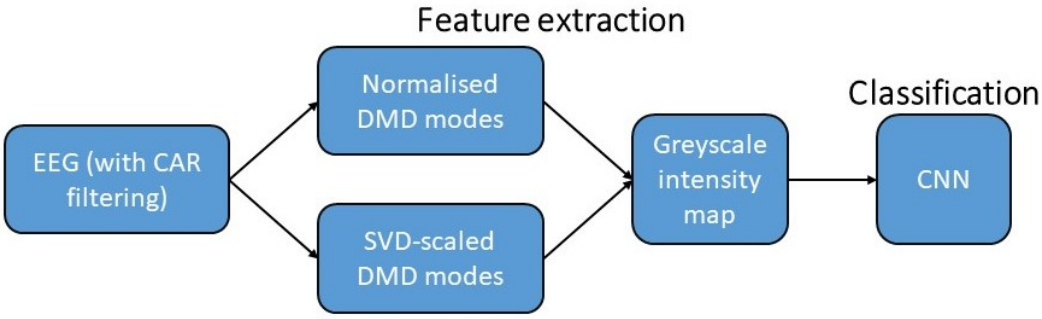

Figure 4.13: Diagram showing the processing pipeline for DMD maps.

Matrices containing both normalised and energy-scaled DMD modes were obtained as in Section 4.3, ensuring that complex conjugates were removed, as they do not provide valuable information in the case of maps; their magnitudes are the same, with the only difference being opposite phases of the modes. Furthermore, the modes have also been ensured to be sorted in ascending order of their characteristics frequencies;

however, they were not binned as in Sections 4.3 and 4.4. Each mode Φ_t was separated into two matrices: matrix containing absolute values $\Phi_{t,abs}$ and the second matrix containing phase information $\Phi_{t,\angle}$. The phases were extracted using `angle()` function in Matlab, which returned a matrix of phase values.

The matrices containing magnitude and phase information of Φ_t were then converted into greyscale images with the help of `mat2gray()` Matlab function. This function also rescaled the incoming matrices to $0-1$ value range so that all $\Phi_{t,abs}$ and $\Phi_{t,\angle}$ had consistent scale across different trials. The output of `mat2gray()` is the same as the size of the supplied matrix which in this case was $c \times r/2$. Lastly, each greyscale image was saved as '`*.png`' file with the correct corresponding name using `imwrite()` function. The saved image had the same $c \times r/2$ size so that each pixel corresponded to a single entry in the DMD mode matrix.

In order to investigate the usefulness of phase maps, the author proposes two different networks: one allowing investigating the effectiveness of using absolute maps on their own and another one combining information extracted from absolute and phase maps. The author constructs a 'processing' layer which is a recurrent block used in both of the implemented networks. This layer, seen in Figure 4.14, is comprised of a convolutional layer which contains five filters of 3×3 size, batch normalization layer, ReLU layer and max pooling layer of 2×2 size. A batch normalization layer is a recommended addition between convolutional layers and non-linearities, which helps to stabilise the network and speed up the training process. ReLU layer (Rectified Linear Unit) is chosen over *sigmoid* or *tanh* activation functions as it greatly helps omitting the vanishing gradient problem, often experienced when using the other two activation functions.

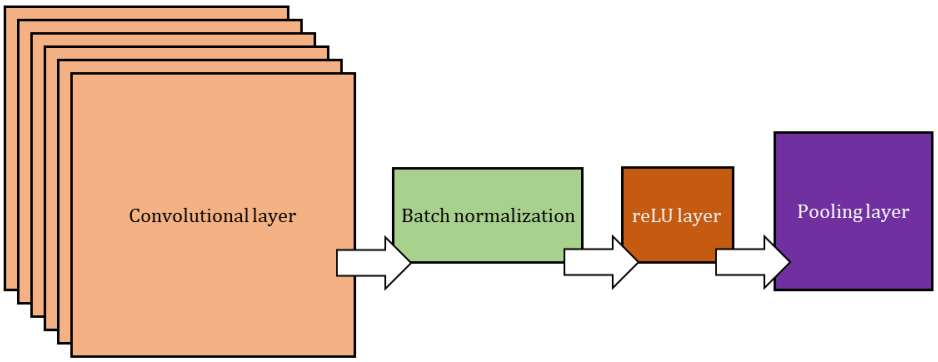

Figure 4.14: The structure of the proposed processing layer used as a building block in the implemented neural networks.

Using the described processing block, the network used for processing absolute maps is constructed as shown in Figure 4.15. The input layer accepted images with single channel (greyscale) data in $c \times r/2$ size, which then were processed by the three aforementioned processing layers. The output of the last max pooling layer is connected to a fully-connected layer with size equal to the number of classes present in the supplied data. Before reaching the output neurons, the data is processed by the softmax layer, which calculates the probability for every possible class present in data and assigns the obtained values accordingly.

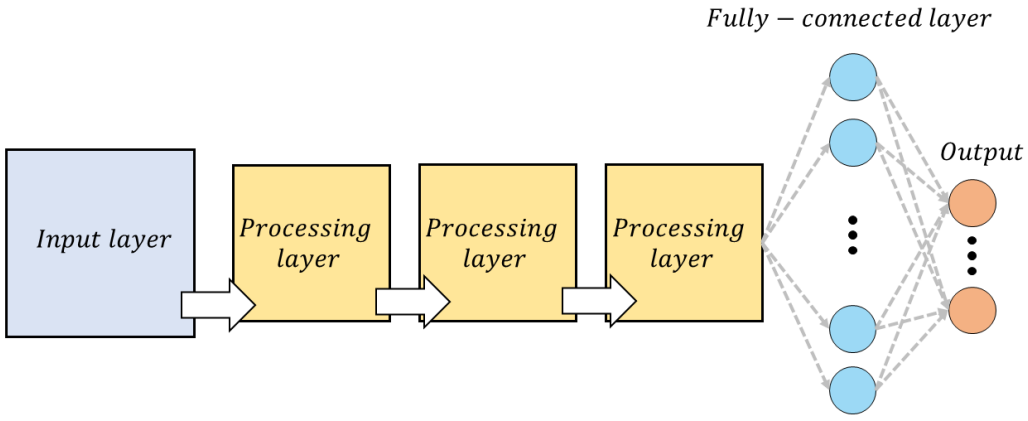

Figure 4.15: The structure of the neural network used for processing absolute DMD maps.

When combining absolute and phase maps, the structure of the implemented neural network is similar to the first one; all layers up to the fully-connected layer are implemented twice as seen in Figure 4.16, to form two neural network branches and thus

allow multi-image input. The outputs of the last max pooling layers are multiplied together and then connected to the fully-connected layer. During the implementation stages the author found out that multiplying the two outputs provided the best classification results compared to adding the outputs or simply concatenating the outputs. Following the common structure of neural networks, the outputs of the fully-connected layer are fed to the softmax layer before terminating at the classification layer.

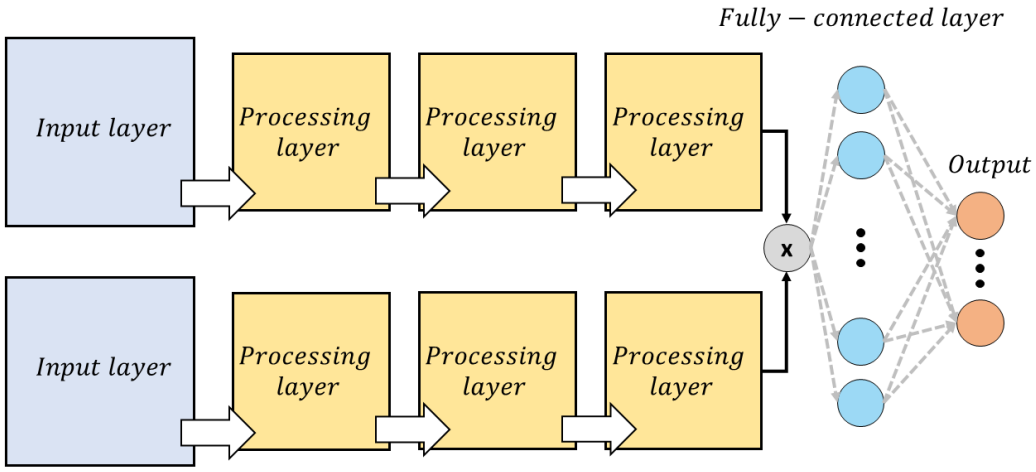

Figure 4.16: The structure of the neural network used for processing absolute and phase DMD maps.

Lastly, the author provides the parameters used for training the proposed neural networks. In order to select the best parameters for the network a series of different processes have been used. Firstly, data was split into 70-30% ratio, such that 70% of data was used for training and 30% of data was used for validation of the network. Following this, the initial learning rate was optimised by using the "Experiment Manager" module in MATLAB. Exhaustive sweeps were run between 0.001 and 0.01 values for two solver types: stochastic gradient descent with momentum (SGDM) and adaptive moment estimation (ADAM). Lastly, the number of max epochs was found through observing 20 training processes and finding the most common epoch at which the network would stop improving further. The optimised parameters can be found in Table 4.3.

Table 4.3: Optimised values for the CNN used in the experiment concerning DMD maps.

Solver name	Max epochs	Initial learning rate
ADAM	150	0.0042

4.6 Conclusion

The author outlined different datasets used in the study presented in this thesis and has shown the preprocessing steps taken to prepare them for the subsequent analysis by the three proposed pipelines. After epoching EEG data, temporal filtering was applied between 7-30Hz to emphasize ERD/S phenomenon. As mentioned, the literature was not fully clear upon the effect of spatial filtering on DMD features, therefore the author decided to investigate this problem in the initial study concerning absolute values of DMD modes as the extracted features. In this case PCA and projection kernel methods were employed as feature translation methods and their implementation was fully described. The second experiment aimed to investigate the viability of DMD spectrum as features for detecting motor imagery. The feature vector was formed as an average mode power between specific frequency bins to keep consistency between trials. This vector was also reduced in dimension through the use of MInf and MRMR approaches with the intent on finding the most suitable technique for DMD spectrum features. Lastly, a novel approach has been proposed based on creating intensity maps of DMD modes and using those maps as input to a CNN. A range of performance metrics has also been introduced which will be used in the next section, when presenting the findings obtained from the three above experiments.

Chapter 5

Results

Following the implementation of the experimental designs outlined in the previous methodology chapter, the results for the protocols exploring the proposed different DMD features were obtained, and are presented here accompanied by a thorough discussion concerning the findings. This chapter is comprised of four subsections: initial investigation into the best classifier for DMD modes features, and then individual results for each proposed feature extraction approach: DMD modes, DMD spectrum and DMD maps. These subsections all follow the same layout: accuracy results are shown first, followed by sensitivity, specificity and lastly kappa values along with brief description of the findings.

5.1 Selection of the appropriate classifier

Before testing of the proposed processing routes was carried out, the author carried out an initial investigation to assess what classifier would be the most suitable for the later experiments. Three candidate classifiers were tested in MATLAB: quadratic linear discriminant analysis (QLDA), naive Bayes (NB) and RBF-SVM using processing routes 2a and 2b (refer to Table 5.3 for the description of the processing route).

Average classification accuracy was measured and then used to identify the best performing classifier for both projection kernel and PCA features. Tables 5.1 and 5.2 show the obtained results; all of the classifier tests were validated with 10-fold cross validation technique to ensure that the results were more accurate. RBF-SVM was

found to be the best performing classifier overall, achieving the best results in 4 out of 6 tests. NB achieved best results in 2 tests and QLDA performed at chance level; a clear sign that this classifier is not reliable in distinguishing classes in high-dimensional feature spaces. It was also observed that in some cases, during the testing, QLDA classifier would show warning about returning empty confusion matrices, which meant that it was not able to discern between multiple classes.

Following those findings, the author implemented RBF-SVM classifier in the final processing pipeline. This concluded this small investigation and allowed for the main investigation of DMD modes and spectrum based features to be carried out.

Table 5.1: Average classification accuracy obtained for the three tested classifiers on features extracted using projection kernel method.

	Average accuracy for projection kernel, (%)		
	RBF-SVM	QLDA	NB
BCI IV 2a	**48.85**	25.79	33.82
GIST-MI	**62.9**	50.85	55.35
BNCI 2020	17.2	16.7	**17.8**

Table 5.2: Average classification accuracy obtained for the three tested classifiers on features extracted using PCA method.

	Average accuracy for PCA, (%)		
	RBF-SVM	QLDA	NB
BCI IV 2a	35.92	25.79	**36.5**
GIST-MI	**57.89**	50.85	57.44
BNCI 2020	**17.8**	16.7	17.3

5.2 Using DMD modes as features

Implementing the steps outlined in section 4.3 as depicted in Figure 4.11 allowed to firstly explore the viability of DMD modes as features describing motor imagery in EEG, and secondly, investigate what is the effect, if any, of spatial filtering using CAR method on the performance of the system. Surprisingly, the initial experiments revealed an unexpected behaviour after mode binning and the subsequent feature selection using projection kernel method. In total, 8 different processing combinations were explored for DMD modes. To refer to those combinations more easily, Table 5.3 contains shorter reference names used in this chapter.

Table 5.3: Average classification accuracy for the proposed processing routes with normalised DMD modes.

	Processing combination name			
	Projection kernel, CAR filtering, mode binning	PCA, CAR filtering, mode binning	Projection kernel, CAR filtering, no mode binning	PCA, CAR filtering, no mode binning
Reference name	1a	1b	2a	2b
	Processing combination name			
	Projection kernel, no CAR filtering, mode binning	PCA, no CAR filtering, mode binning	Projection kernel, no CAR filtering, no mode binning	PCA, no CAR filtering, no mode binning
Reference name	3a	3b	4a	4b

Accuracy. Figures 5.1 and 5.2 show the distribution of the subjects' classification accuracy results in different datasets in the form of boxplots. Examining the plots reveals that neither normalised nor scaled DMD modes perform particularly well with mean accuracy, suggesting performance of just above chance level (25%, 50% and 16.7% for BCI IV 2a, GIST-MI and BNCI 2020 respectively). However, it is fairly clear that the processing combinations which use the projection kernel as the feature selection tend to perform slightly better than PCA selection method.

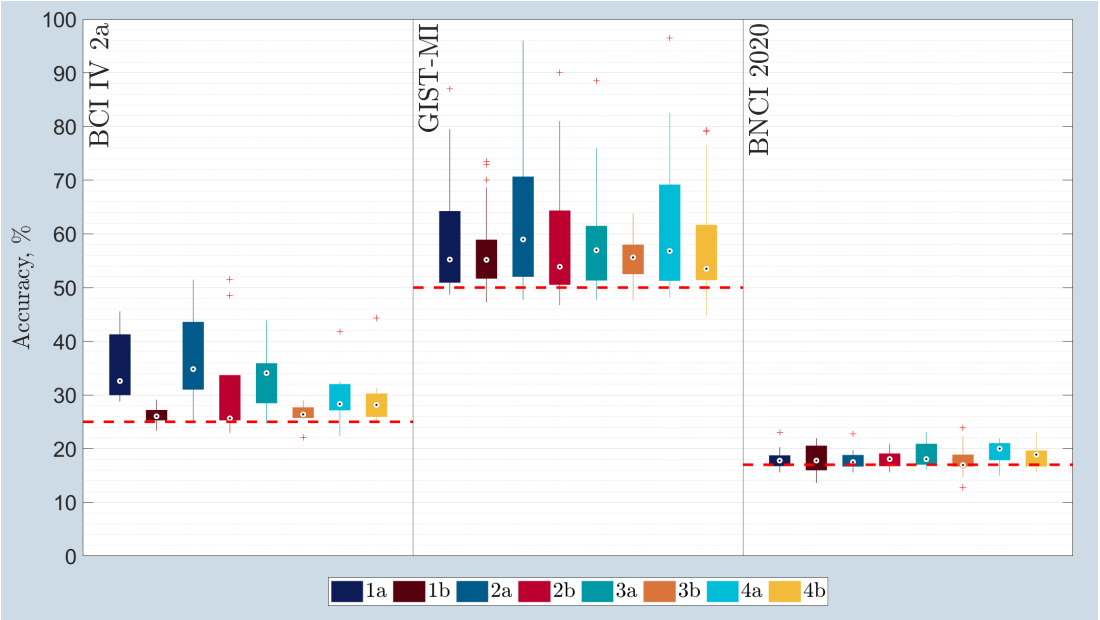

Figure 5.1: Classification accuracy for normalised DMD modes. Red dashed line indicates the chance level for each dataset (25%, 50%, 16.7% respectively).

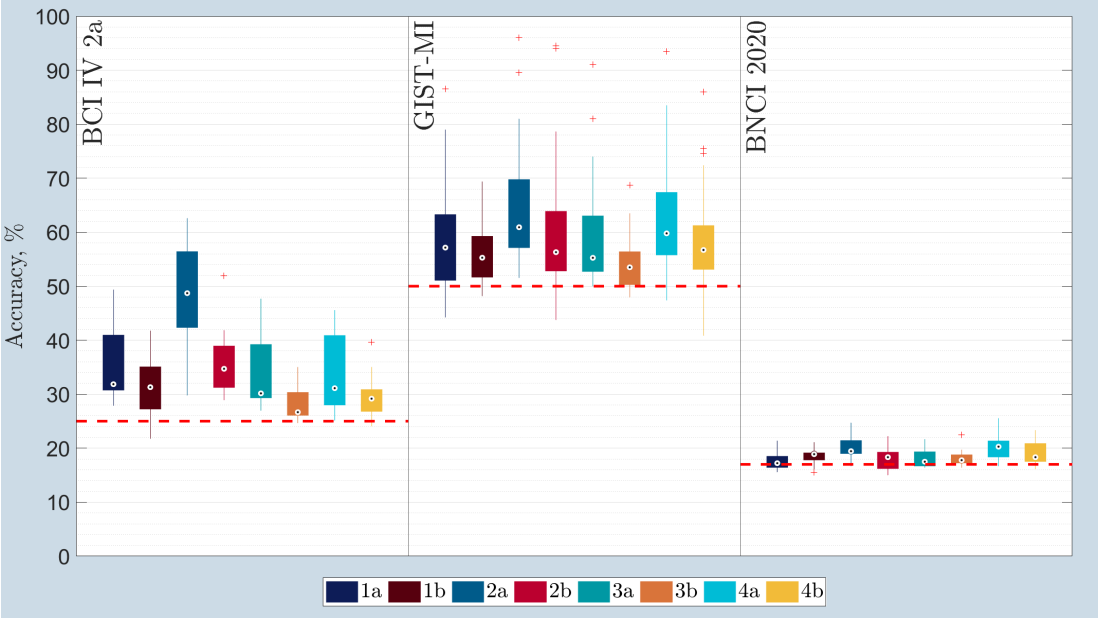

Figure 5.2: Classification accuracy for scaled DMD modes. Red dashed line indicates the chance level for each dataset (25%, 50%, 16.7% respectively).

To have a clearer understanding of the effect of CAR filtering on the performance of the proposed processing combinations, and better investigate the differences between the different methods, the mean classification accuracies were extracted from the box-plots and tabulated in Table 5.4 and 5.5. The best performing processing routes are highlighted in bold face.

Overall, the best performing processing combination based on the average classification accuracy is method 2a. However, an exception to that was found when testing BNCI 2020 dataset, where in case of normalised DMD modes, combination 4a performs the best. Nonetheless, it can be seen that scaled DMD modes perform better than normalised DMD modes offering +11.5%, +1.7% and +0.8% (respectively to the order of datasets seen in Tables 5.4 and 5.5) increase in the average classification accuracy, which was the expected outcome. On that note, the accuracy of scaled DMD modes is 23.5%, 14.3% and 3.5% higher than the chance level respectively for each dataset used.

Analysing average classification accuracy also shows that implementing CAR filtering has an overall positive impact on the performance, offering the highest boost of 14.4%, 2.4% and 0.1% (respectively for each dataset) in the case of scaled DMD modes (2a vs 4a). This positive trend is followed by the other processing combinations, however it is not as impactful as in the case of the aforementioned processing method.

The most surprising finding is the effect of the mode binning process on the performance of DMD modes features. Recalling the requirements for orthogonal subspaces on Grassmannian manifolds from Section 3.3.3.3 and further description of how DMD modes can be orthogonalised outlined in Section 4.3, the results obtained in the experiments concerning DMD modes are conflicting with the requirements presented in the literature. In terms of normalised DMD modes, not applying mode binning and thus disregarding orthogonal requirement for calculating projection kernel induced a +1.6%, +4.3% and -0.1% change in the classification accuracy (CAR filtered data), while for scaled DMD modes the change is even more impactful: +13.1%, +5.4% and +3.7%(CAR filtered data), respectively for each dataset.

Table 5.4: Average classification accuracy for the proposed processing routes with normalised DMD modes.

	Average accuracy for normalised DMD modes, (%)							
	1a	1b	2a	2b	3a	3b	4a	4b
BCI IV 2a	35.4	26.1	**37**	31.3	33.1	26.3	29.8	29.7
GIST-MI	58.3	56.1	**62.6**	58.6	57.8	55.3	61.5	57.7
BNCI 2020	18.1	18	18	17.9	18.7	17.8	**19.4**	18.5

Table 5.5: Average classification accuracy for the proposed processing routes with scaled DMD modes.

	Average accuracy for scaled DMD modes, (%)							
	1a	1b	2a	2b	3a	3b	4a	4b
BCI IV 2a	35.4	31.7	**48.5**	36.5	34.4	28.5	34.1	29.7
GIST-MI	58.9	56.1	**64.3**	59.9	58.5	53.7	61.9	57.9
BNCI 2020	17.5	18.4	**20.2**	18	18	18.2	20.1	19.1

Sensitivity. Observing results for normalised and scaled modes in Figures 5.3 and 5.4 reveals that the sensitivity closely follows the trend seen in the accuracy plots (5.1, 5.2). While it can be seen that the sensitivity increases if the dataset contains smaller number of classes, it generally still remained low and, in the case of the BNCI 2020 dataset, the difference in sensitivity between different processing routes is almost non-distinguishable.

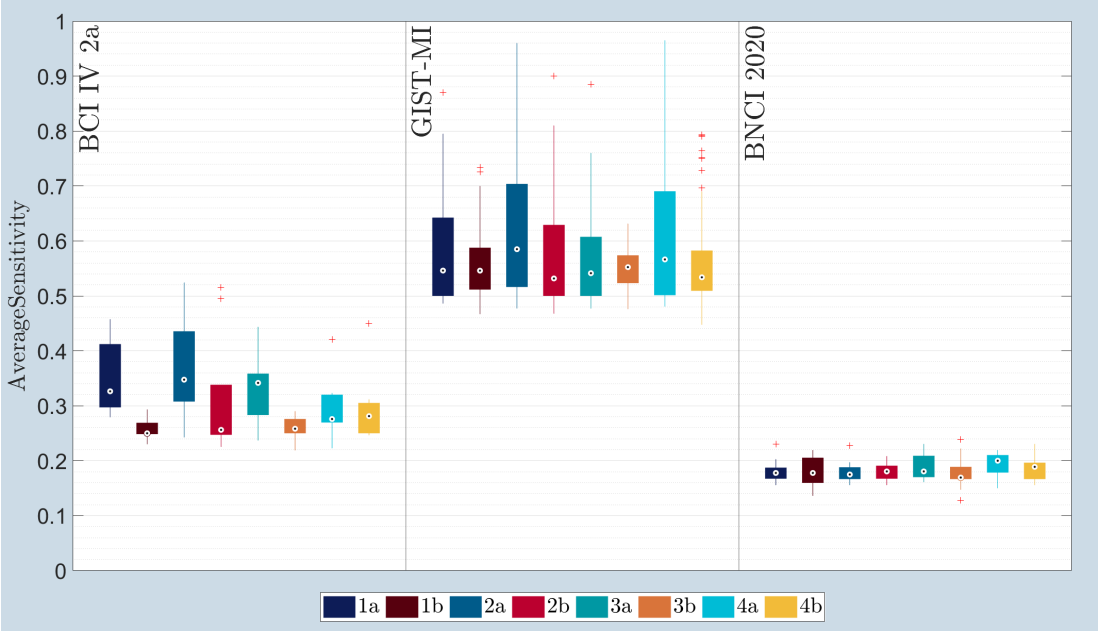

Figure 5.3: Sensitivity for normalised DMD modes.

Investigating the average sensitivity values which are presented in Tables 5.6 and 5.7 supports the initial findings from the accuracy data, which revealed that using method 2a yields the best results from all the proposed processing combinations. Features extracted from scaled DMD modes provide slightly better sensitivity than features

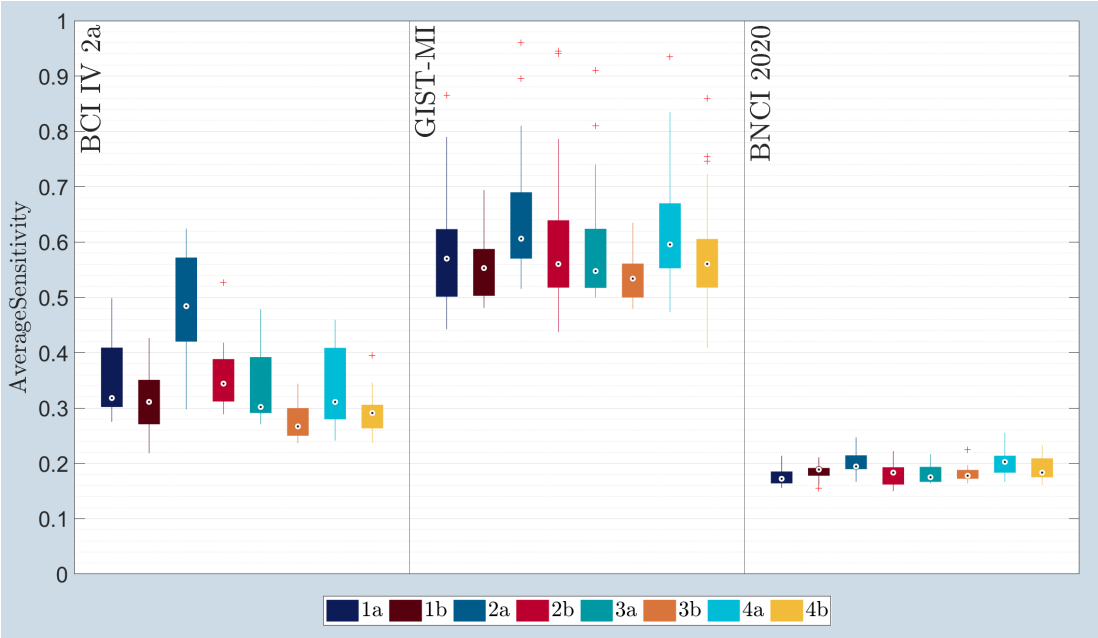

Figure 5.4: Sensitivity for scaled DMD modes.

obtained from normalised DMD modes, offering an +0.12, +0.02 and +0.02 increase respectively for each dataset.

While analysing the effect of CAR filtering, taking the best performing processing route into consideration shows that, when compared to a route which did not use CAR filtering, the sensitivity increased by +0.14 and +0.02 for BCI IV 2a and GIST-MI datasets and in the case of BNCI 2020 dataset no change in the sensitivity was noted. In the case of mode binning, it can be seen that for normalised DMD modes the sensitivity increased by 0.02 and 0.04 for the first two datasets and no change was noted for the last dataset, whereas for scaled DMD modes the sensitivity increased by 0.14, 0.06 and 0.02 for each dataset.

Table 5.6: Average sensitivity for the proposed processing routes with normalised DMD modes.

	Average sensitivity for normalised DMD modes							
	1a	1b	2a	2b	3a	3b	4a	4b
BCI IV 2a	0.35	0.26	**0.37**	0.31	0.33	0.26	0.3	0.3
GIST-MI	0.58	0.56	**0.62**	0.58	0.57	0.55	0.61	0.57
BNCI 2020	0.18	0.18	0.18	0.18	**0.19**	0.18	**0.19**	**0.19**

Table 5.7: Average sensitivity for the proposed processing routes with scaled DMD modes.

	Average sensitivity for scaled DMD modes							
	1a	1b	2a	2b	3a	3b	4a	4b
BCI IV 2a	0.35	0.32	**0.49**	0.37	0.34	0.28	0.35	0.3
GIST-MI	0.58	0.56	**0.64**	0.6	0.58	0.53	0.62	0.58
BNCI 2020	0.18	0.18	**0.2**	0.18	0.18	0.18	**0.2**	0.19

Specificity. Figures 5.5 and 5.6 contain boxplots showing specificity for every processing route when using normalised or scaled DMD modes respectively. Analysing the figures yields interesting findings: BCI IV 2a and BNCI 2020 datasets have high specificity, while GIST-MI dataset has a mediocre specificity, almost at the exact same level as sensitivity. A reason for such behaviour is because of the particular way that specificity and sensitivity are calculated in multiclass problems as described in section 4.3, leading to higher specificity number being calculated.

The effect on the system's specificity of either of the three investigated components cannot be clearly seen on Figures 5.5 and 5.6, thus the average specificities were extracted for both normalised and scaled DMD modes and tabulated in Tables 5.8 and 5.9.

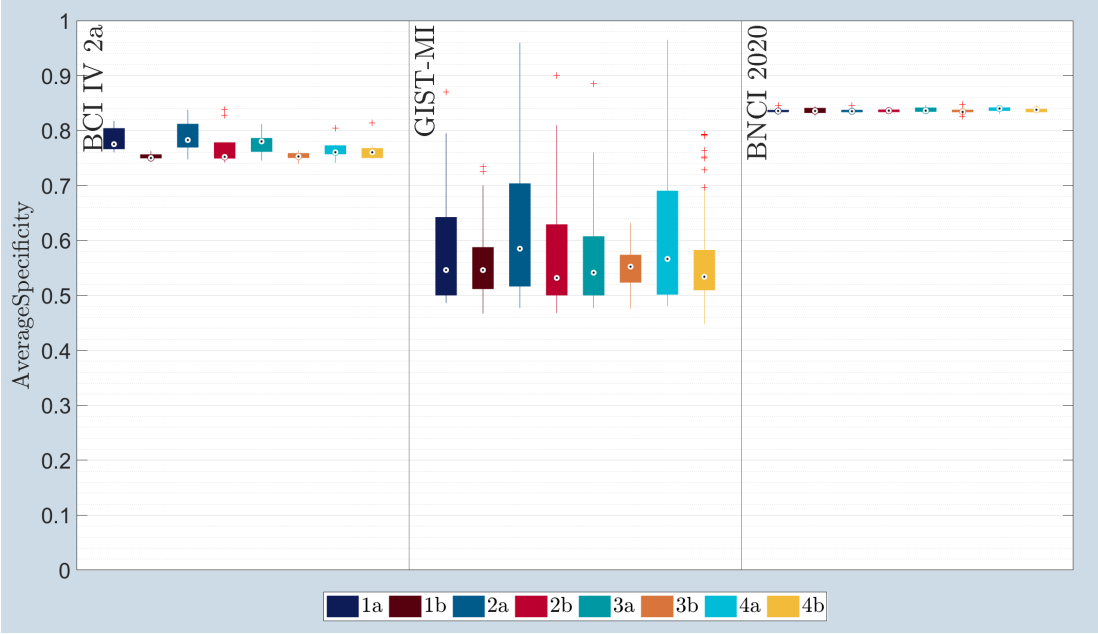

Figure 5.5: Specificity for normalised DMD modes.

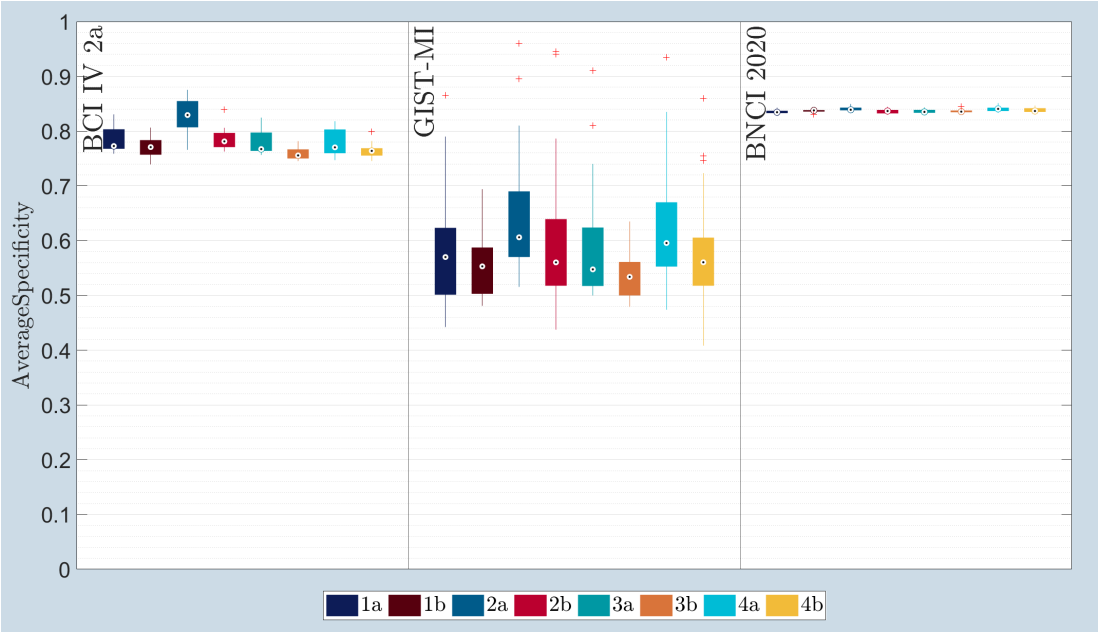

Figure 5.6: Specificity for scaled DMD modes.

The tabulated results reveal straight away that in the case of the BNCI 2020 dataset the specificity is constant and is completely unaffected by any of the processing routes, nor by using normalised or scaled DMD modes. In the case of BCI IV 2a and GIST-MI datasets the difference between different processing routes and normalised or scaled modes is minimal. Nonetheless, extracting scaled DMD modes from CAR filtered EEG signals and subsequently using projection kernel without mode binning (method 2a), displayed the highest performance out of all the explored approaches again, offering +0.04 and +0.02 increase in specificity when compared to the exact processing route, when extracting normalised DMD modes.

Investigating the average specificity values further shows that implementing CAR filtering on EEG signals provides only a +0.02 and +0.01 increase in specificity for normalised DMD modes for the first two datasets, while scaled DMD modes benefited slightly more, gaining +0.05 and +0.02 average specificity. Lastly, analysing the effect of mode binning further indicates that not implementing mode binning provides a specificity increase of +0.01 and +0.04 for normalised DMD modes and +0.05 and +0.06 for scaled DMD modes.

Table 5.8: Average specificity for the proposed processing routes with normalised DMD modes.

	Average specificity for normalised DMD modes							
	1a	1b	2a	2b	3a	3b	4a	4b
BCI IV 2a	0.78	0.75	**0.79**	0.77	0.78	0.75	0.77	0.77
GIST-MI	0.58	0.56	**0.62**	0.58	0.57	0.55	0.61	0.57
BNCI 2020	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84

Table 5.9: Average specificity for the proposed processing routes with scaled DMD modes.

	Average specificity for scaled DMD modes							
	1a	1b	2a	2b	3a	3b	4a	4b
BCI IV 2a	0.78	0.77	**0.83**	0.79	0.78	0.76	0.78	0.76
GIST-MI	0.58	0.56	**0.64**	0.59	0.58	0.53	0.62	0.58
BNCI 2020	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84

Kappa value. The last metric investigated for this experiment is Cohen's kappa value; the results obtained from each dataset are shown in Figures 5.7 and 5.8. Low classification accuracy seen in Figures 5.1 and 5.2 is reflected in the calculated kappa values, which prove that the supplied features in BCI IV 2a and BNCI 2020 dataset are almost not reliable for multiclass classification at all; while in the case of binary classification present in GIST-MI dataset some subjects' data was almost perfect (attaining kappa value above 0.9), while other subjects performed just as poorly as in the case of the other two datasets (kappa value below 0.1). Despite this wide range of performance, the boxplots indicate that the average kappa value remained fairly low for GIST-MI dataset (≈0.2). Additionally, one can see a clear advantage of using scaled DMD modes instead of the normalised DMD modes, as scaled DMD modes offer a substantial increase in the data reliability, particularly for BCI IV 2a dataset.

Looking closer at the average kappa values for normalised and scaled DMD modes in Tables 5.10 and 5.11, the differences between using scaled DMD modes becomes more visible especially for the first dataset; data obtained through those modes is shown to be +0.15, +0.03 and +0.02 more reliable on average for the best performing processing route. This route has been shown again to be the one which used projection kernel method with no mode binning on CAR filtered EEG signals (method 2a).

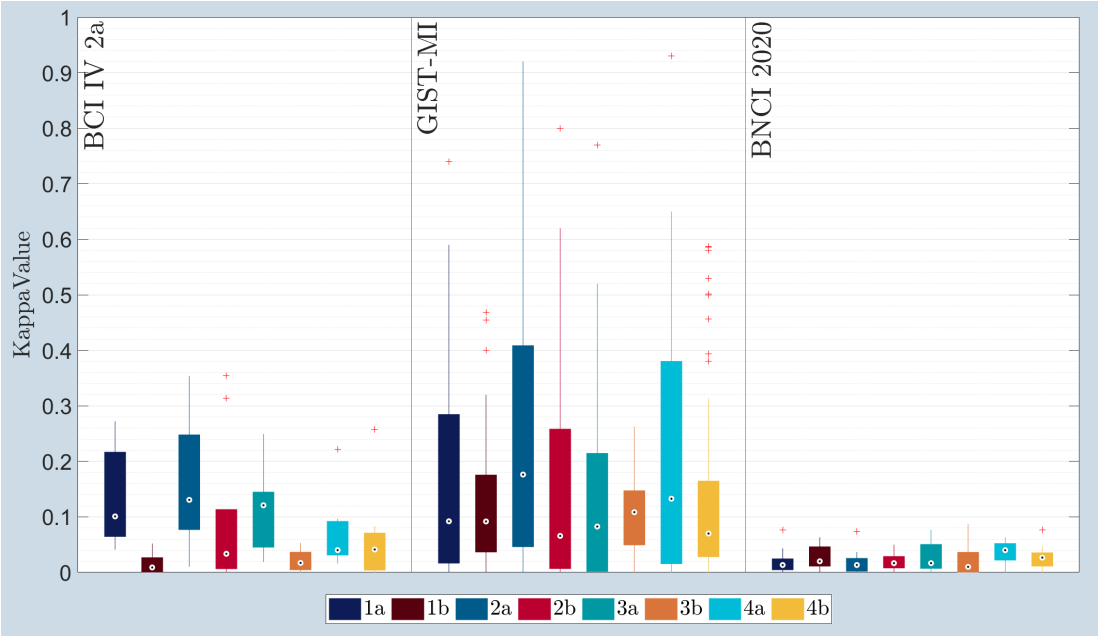

Figure 5.7: Kappa values for normalised DMD modes.

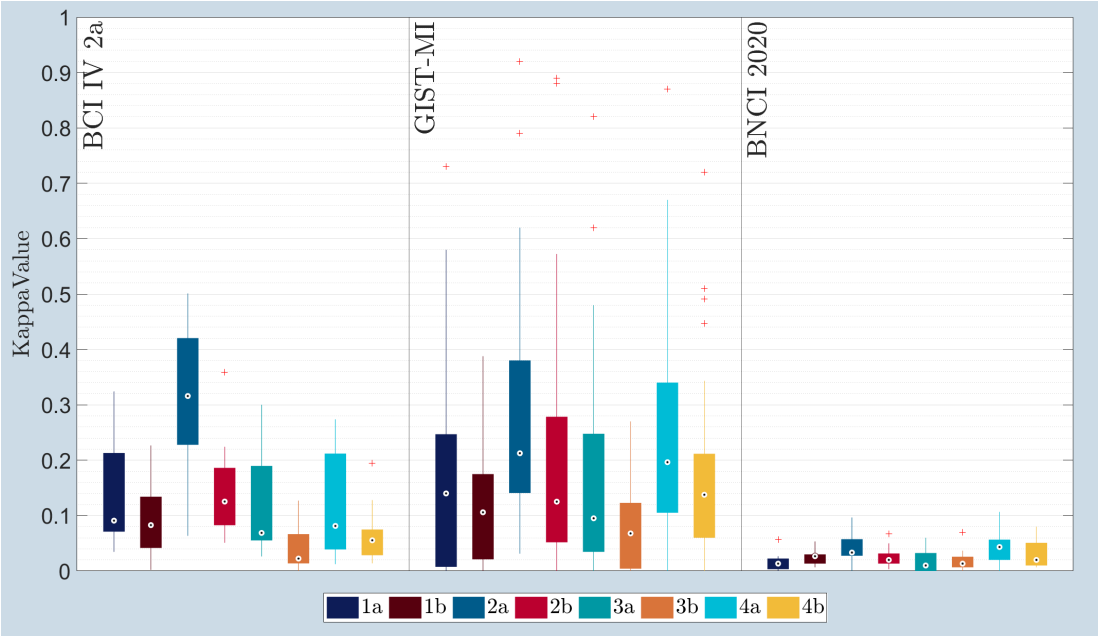

Figure 5.8: Kappa values for scaled DMD modes.

Investigating the difference in performance between CAR filtered and unfiltered processing routes solidifies the advantage of spatial filtering for DMD modes as it yields an increase in data reliability. For normalised DMD modes, when using projection kernel with no mode binning (method 2a), it has been found that kappa value increases by +0.09 and +0.03 points for the first two datasets, while for BNCI 2020 dataset the

absence of CAR filtering (method 4a) yielded better performance by +0.02 points. In the case of scaled DMD modes, filtering signals with CAR improved kappa value by +0.19 and +0.05 points for BCI IV 2a and GIST-MI datasets, while BNCI 2020 dataset was unaffected.

Lastly, it can be seen from Tables 5.10 and 5.11 the positive impact of omitting mode binning step. While for normalised DMD modes this only resulted in slight increase in kappa value, i.e. +0.02, +0.09 and 0 for each dataset respectively, examining scaled DMD modes shows that kappa value doubled for almost all datasets: +0.17, +0.11 and +0.03 kappa value increase.

Table 5.10: Average kappa values for the proposed processing with normalised DMD modes.

	Average kappa values for normalised DMD modes							
	1a	1b	2a	2b	3a	3b	4a	4b
BCI IV 2a	0.14	0.02	**0.16**	0.09	0.11	0.02	0.07	0.06
GIST-MI	0.16	0.12	**0.25**	0.17	0.15	0.1	0.22	0.15
BNCI 2020	0.02	0.03	0.02	0.02	0.03	0.02	**0.04**	0.03

Table 5.11: Average kappa values for the proposed processing routes with scaled DMD modes.

	Average kappa values for scaled DMD modes							
	1a	1b	2a	2b	3a	3b	4a	4b
BCI IV 2a	0.14	0.1	**0.31**	0.15	0.12	0.05	0.12	0.07
GIST-MI	0.17	0.12	**0.28**	0.19	0.16	0.07	0.23	0.16
BNCI 2020	0.01	0.02	**0.04**	0.02	0.02	0.02	**0.04**	0.03

5.3 Using DMD spectrum as features

Implementation of CAR filtering on EEG signals is based on the findings from the previous experiment, which have shown the positive performance impact of CAR filtering on DMD modes. Each dataset had three different processing routes which were based on three different feature selection methods, as shown before in Figure 4.12: raw spectrum, MRMR and MInf methods.

Accuracy. The performance of DMD spectrum features in terms of classification accuracy is shown in Figure 5.9. Investigating the boxplots reveals a poor performance of DMD spectrum features with accuracy reaching just above chance levels again, similarly to DMD modes features. Furthermore, the choice of feature selection has almost negligible effect on the performance; however MInf seems to provide the best accuracy alas by a very small margin.

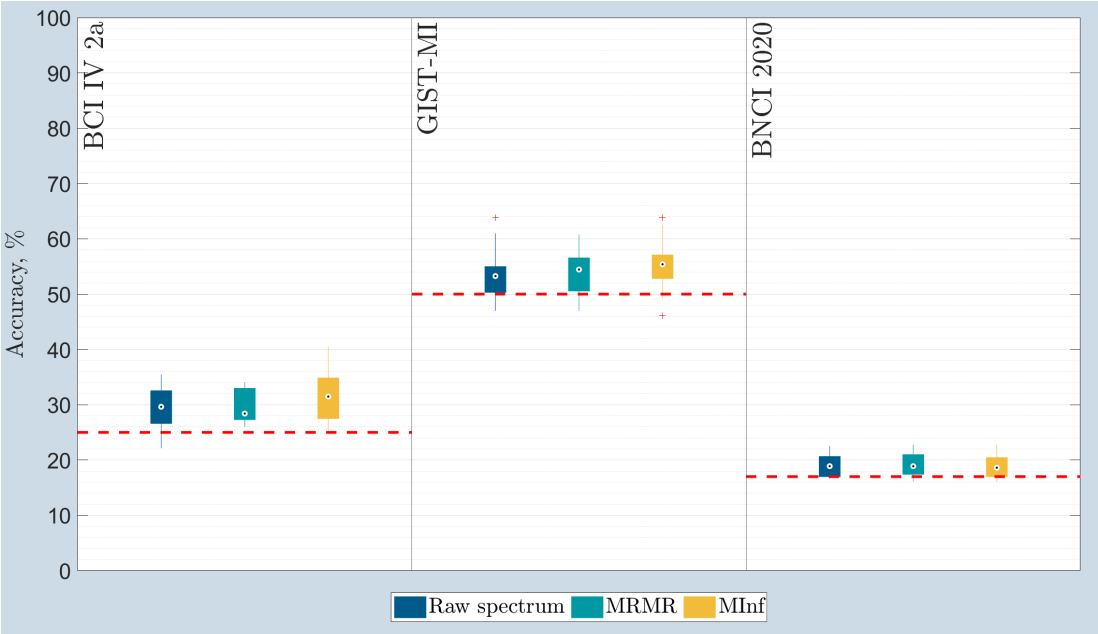

Figure 5.9: Classification accuracy for DMD spectrum. Red dashed line indicates the chance level for each dataset (25%, 50%, 16.7% respectively)

Inspecting the average accuracy values found in Table 5.12 confirms that MInf method does indeed provide higher accuracy, albeit the increase is small: 2.3% and 1.8% increase over raw features and MRMR approach respectively for BCI IV 2a dataset, 2% and 1.1% increase for GIST-MI dataset. For BNCI 2020 dataset, MRMR performed better although only by 0.4% over raw features and MInf.

Table 5.12: Average accuracy for the proposed processing routes with DMD spectrum features.

	Average accuracy for DMD spectrum, %		
	Raw features	MRMR	MInf
BCI IV 2a	29.5	29.9	**31.8**
GIST-MI	53.1	54	**55.1**
BNCI 2020	19	**19.4**	19

Sensitivity. Sensitivity results for DMD spectrum experiment presented in Figure 5.10 show a similar distribution to the one seen for DMD modes. Similarly to the accuracy plots, it is hard to notice any significant differences in sensitivity between the three feature selection methods, although MInf seems to perform slightly better.

Figure 5.10: Sensitivity for DMD spectrum.

After extracting average sensitivity from the boxplots data and tabulating it in Table 5.13, it becomes more evident that DMD spectrum features are almost unaffected by the feature selection method. While MInf still turned out to produce the best results, although it only provided a very modest +0.03 and +0.02 sensitivity gain for the BCI IV 2a dataset when compared to raw features and MRMR method respectively. The performance gain is almost the same for the GIST-MI dataset, offering an increase of +0.03 and +0.01 over raw features and MRMR respectively. Notably, no change in sensitivity was noted in the case of the BNCI 2020 dataset.

Table 5.13: Average sensitivity for the proposed processing routes with DMD spectrum.

	Average sensitivity for DMD spectrum		
	Raw features	MRMR	MInf
BCI IV 2a	0.29	0.30	**0.32**
GIST-MI	0.52	0.54	**0.55**
BNCI 2020	0.19	0.19	0.19

Specificity. Figure 5.11 shows the calculated specificity values for experiments using DMD spectrum features. Higher specificity observed for BCI IV 2a and BNCI 2020 dataset is again attributed to the way in which specificity is calculated for mutliclass classification problems. After inspecting the boxplots, it can be safely said that the choice of feature selection method has close to no effect on the specificity, as it was the case with sensitivity.

Investigating the average values found in Table 5.14 confirms the observations made from Figure 5.11. For the BCI IV 2a dataset, choosing either MRMR or MInf increases the specificity by 0.01 points over using raw DMD spectrum. GIST-MI dataset favours use of MInf as it provided a gain of 0.03 and 0.02 specificity compared to raw features and MRMR selected features respectively. In the BNCI 2020 dataset specificity remained completely unaffected by the choice of feature selection method, following the behaviour noticed in sensitivity results.

Figure 5.11: Specificity for DMD spectrum.

Table 5.14: Average specificity for the proposed processing routes with DMD spectrum.

	Average specificity for DMD spectrum		
	Raw features	MRMR	MInf
BCI IV 2a	0.76	**0.77**	**0.77**
GIST-MI	0.52	0.53	**0.55**
BNCI 2020	0.84	0.84	0.84

Kappa value. Lastly, the obtained kappa values from each dataset are shown in Figure 5.12. The observed classification accuracies which are just slightly above the chance level when using DMD spectrum as features are appropriately reflected in equivalently low kappa values, with none of the processing routes seemingly breaking an average kappa value of 0.10, clearly indicating that features extracted from DMD spectrum are not reliable at all ($\leq 1\%$ reliability).

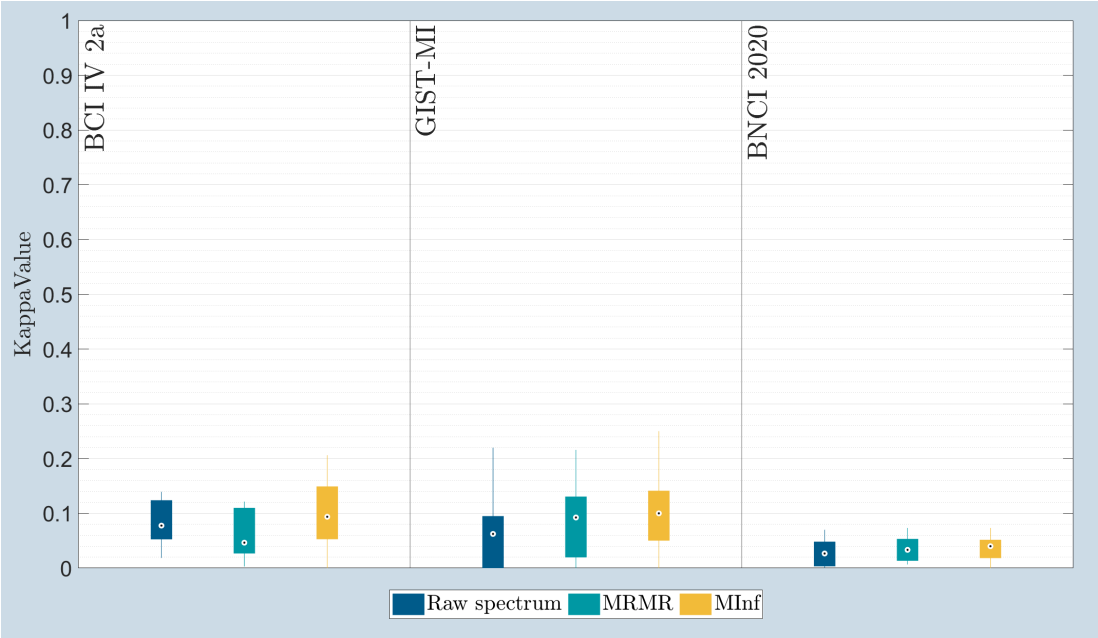

Figure 5.12: Kappa values for DMD spectrum.

Investigating the average kappa values closer in Table 5.15 confirms the previous results where MInf feature selection method was seen performing the best out of the three approaches. For BCI IV 2a dataset, MInf offered an increase by 0.02 and 0.04 kappa compared to raw features and features proposed by MRMR respectively. In GIST-MI dataset, kappa increased by 0.04 and 0.01 while in BNCI 2020 dataset MInf gained 0.01 kappa compared to raw features and no change was noted between average kappa values when using MInf or MRMR.

Table 5.15: Average kappa values for the proposed processing routes with DMD spectrum.

	Average kappa values for DMD spectrum		
	Raw features	MRMR	MInf
BCI IV 2a	0.08	0.06	**0.10**
GIST-MI	0.06	0.09	**0.10**
BNCI 2020	0.03	**0.04**	**0.04**

5.4 Using DMD maps as features

The last experiment performed in this thesis focused on the novel exploitation of DMD modes. As described in Section 4.5 the author proposed to represent DMD modes as intensity maps and use neural networks to extract features and classify them accordingly. While use of heatmaps to present DMD modes is common in the literature, the author has not seen any academic publication using such maps as features in any way or form; such heatmaps were previously only used as visual aid to concisely present the matrices of DMD modes.

In total four different processing routes were proposed which utilised DMD maps with the aim of finding the best performing approach. Two pipelines used normalised or scaled absolute mode values as maps while the other two pipelines used a combination of phase maps and absolute maps for both normalised and scaled DMD modes.

Accuracy. Figure 5.13 shows the classification accuracy reached by each of the proposed processing routes in the three datasets used in this thesis. Initial observation of the resultant boxplots firstly indicates that maps acquired from energy-scaled DMD modes provide better accuracy than the ones obtained from normalised DMD modes, demonstrating once again that such scaled modes are better at discovering MI-EEG related features. Secondly, contrary to the author's expectations, the addition of phase maps has a notably negative impact on the classification accuracy. This trend seems to hold true for BCI IV 2a and GIST-MI datasets, however in the case of BNCI 2020 dataset it is not entirely clear.

Further examination of the average classification accuracies collected in Table 5.16 reveals the following: using only absolute maps from energy-scaled DMD modes provides highest average accuracy of 39.5% and 61.7% for the first and second datasets while including phase maps in the processing route provides highest average accuracy of 17.3% for the last dataset. That said, in the case of the last dataset the difference between presence and absence of phase maps in the processing route is only +0.4% and −0.8% for normalised and energy-scaled DMD maps respectively, where positive percentage indicates difference favouring absence of phase maps and negative percentage differences favours presence of phase maps. This difference is much bigger for the first

two datasets as noted when examining Figure 5.13: +5.1% and +10.2% for the first dataset and +4.6% and +5.1% for the second dataset.

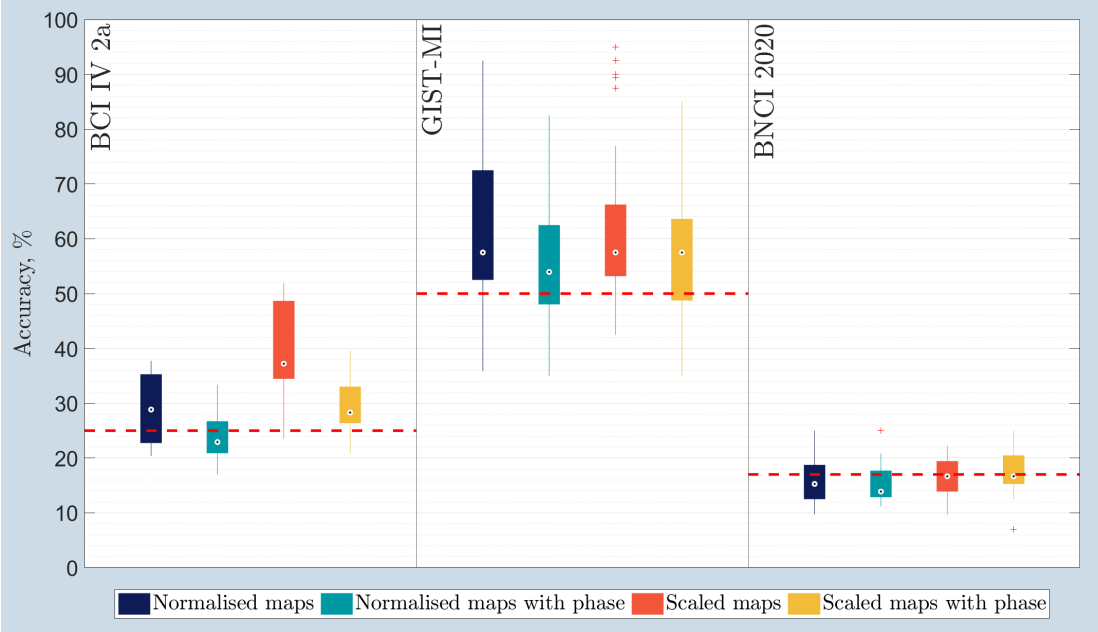

Figure 5.13: Classification accuracy for DMD maps. Red dashed line indicates the chance level for each dataset (25%, 50%, 16.7% respectively)

Table 5.16: Average classification accuracy for the proposed processing routes with DMD maps.

	Average accuracy, %			
	Normalised absolute maps	Normalised absolute & phase maps	Scaled absolute maps	Scaled absolute & phase maps
BCI IV 2a	29	23.9	**39.5**	29.3
GIST-MI	60.9	56.3	**61.7**	56.6
BNCI 2020	16	15.6	16.5	**17.3**

Sensitivity. The sensitivity of the proposed processing routes is investigated. From Figure 5.14 it can be seen that the sensitivity distribution follows the trend seen in accuracy results. Maps obtained from energy-scaled DMD modes have the highest sensitivity out of all the proposed approaches. As it was the case with accuracy of the explored processing routes, including phase maps lowered the sensitivity of the system.

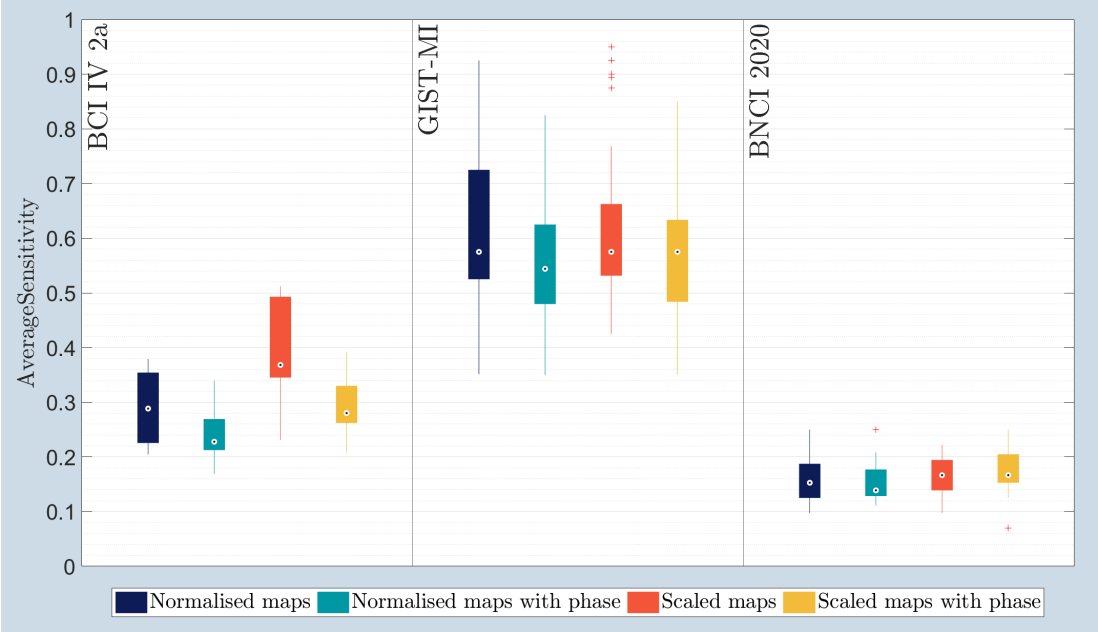

Figure 5.14: Average sensitivity for DMD maps.

Average sensitivity results can be seen in Table 5.17. The advantage of using energy-scaled maps over normalised ones is very clear in BCI IV 2a dataset, where a sensitivity increase of 0.10 is noted. This is not the case for GIST-MI and BNCI 2020 dataset where no change is observed at all. Furthermore in BNCI 2020 dataset, both absolute and phase maps were the most sensitive, offering a minimal increase of 0.01.

Table 5.17: Average sensitivity for the proposed processing routes with DMD maps.

	Average sensitivity			
	Normalised absolute maps	Normalised absolute & phase maps	Scaled absolute maps	Scaled absolute & phase maps
BCI IV 2a	0.29	0.24	**0.39**	0.29
GIST-MI	**0.61**	0.56	**0.61**	0.56
BNCI 2020	0.16	0.16	0.16	**0.17**

Specificity. Figure 5.15 shows the obtained specificity results from DMD maps experiments. Investigating the plots reveals more supportive evidence in favour of not combining phase and absolute maps. While this observation is noticeable for BCI IV 2a and GIST-MI datasets, it is completely unclear if that is the case for BNCI 2020 dataset. In addition, the author notes that in the case of the second dataset some of the subjects achieved very high specificity values (seen as red crosses on the plots).

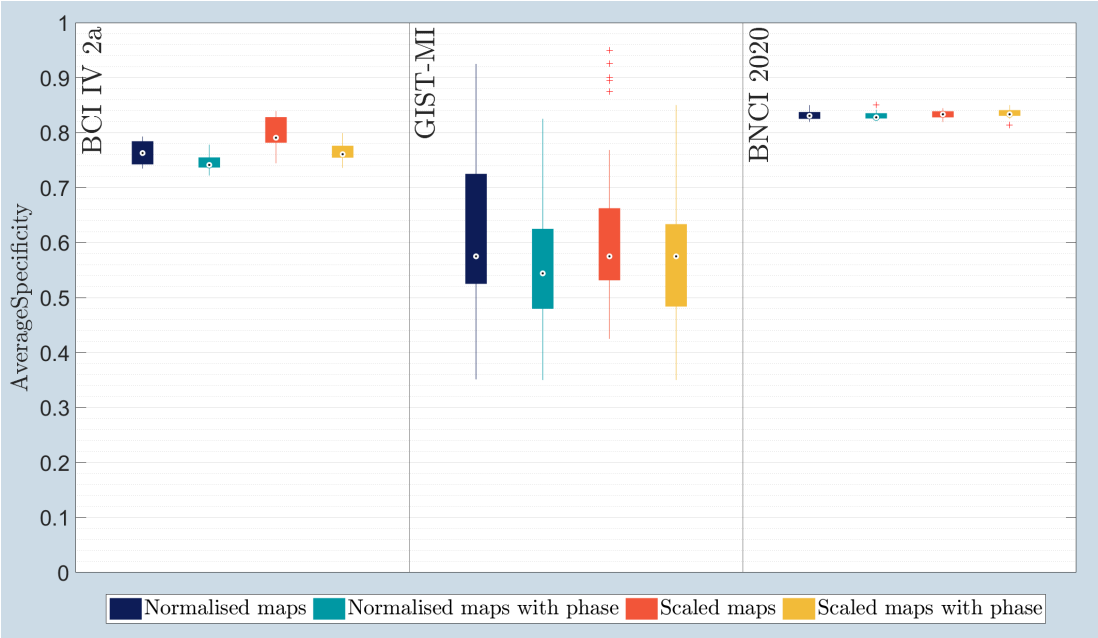

Figure 5.15: Average specificity for DMD maps.

Average specificity results presented in Table 5.18 clarify that in the case of BNCI 2020 dataset, none of the proposed routes offer any advantage in terms of specificity. The results also revealed some other interesting findings. While overall, using absolute maps from energy-scaled DMD modes yielded the highest specificity, in the first dataset the increase from normalised maps was only noted to be 0.04 and in the second dataset no difference between normalised and energy-scaled maps was noted.

Table 5.18: Average specificity for the proposed processing routes with DMD maps.

	Average specificity			
	Normalised absolute maps	Normalised absolute & phase maps	Scaled absolute maps	Scaled absolute & phase maps
BCI IV 2a	0.76	0.75	**0.80**	0.76
GIST-MI	**0.61**	0.56	**0.61**	0.56
BNCI 2020	0.83	0.83	0.83	0.83

Kappa value. Kappa value was the last metric explored in the experiment concerning DMD maps and the obtained results are presented in Figure 5.16. Observing the results for BCI IV 2a dataset clearly confirms that scaled absolute maps provide much more reliable features than the other proposed approaches. The situation however changes when looking at GIST-MI dataset, where it can be seen that scaled absolute maps obtained lower median kappa value than normalised maps or scaled maps with phase information. In the case of BNCI 2020 dataset the plots indicate no difference between either of the proposed approaches.

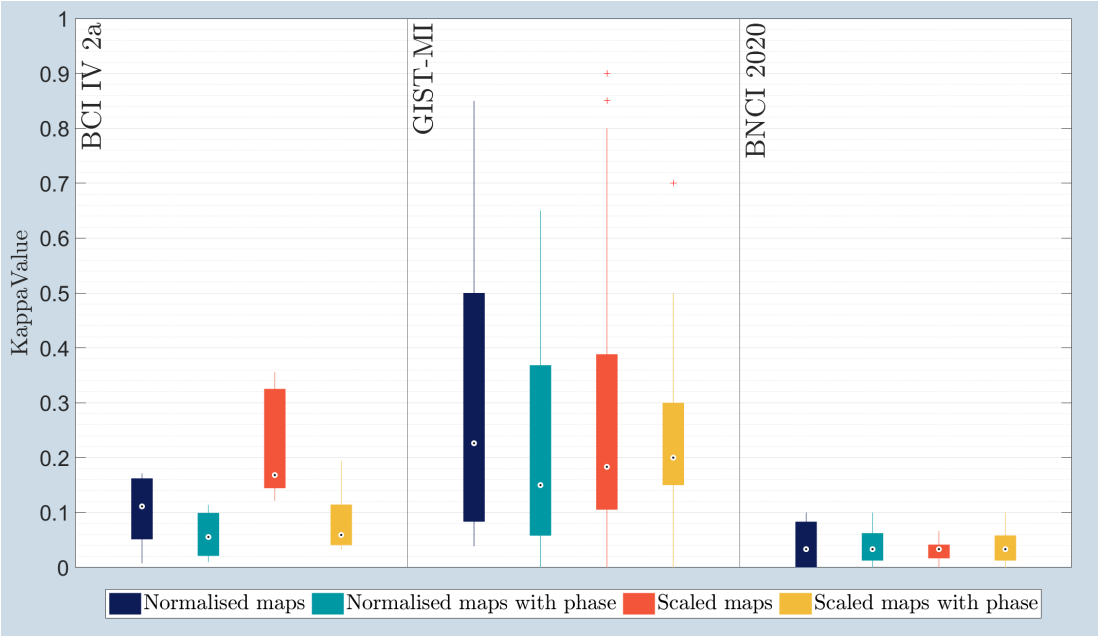

Figure 5.16: Kappa values for DMD maps.

Closer examination of average kappa values found in Table 5.19 reveals a significant kappa gain for scaled absolute maps over any other method (+0.12 over normalised absolute maps and +0.14 over combination of absolute and phase maps) for the first dataset. Table 5.19 also confirms that for the second dataset average kappa value for the normalised absolute maps is higher than scaled absolute maps by 0.03 points. Furthermore, in the case of the last dataset scaled absolute maps have the lowest kappa value (0.03), albeit it is only lower by 0.01 compared to other three methods.

Table 5.19: Average kappa values for the proposed processing routes with DMD maps.

	Average kappa value			
	Normalised absolute maps	Normalised absolute & phase maps	Scaled absolute maps	Scaled absolute & phase maps
BCI IV 2a	0.10	0.06	**0.22**	0.08
GIST-MI	**0.31**	0.22	0.28	0.23
BNCI 2020	0.04	0.04	0.03	0.04

5.5 Conclusion

The author has showed the justification for the selected RBF-SVM classifier through a comparison study between other two classifiers: QLDA and NB. RBF-SVM outperformed the other classifiers and was able to classify supplied features in cases where other approaches were not able to do so. Following that, the results for the three proposed DMD pipelines were presented and brief overview was provided which serves as the basis for the discussion in the next chapter.

Chapter 6

Discussion

Observations made in the results chapter regarding the performance of the three proposed features based on DMD method are discussed in the following subsections, where each individual behaviour is scrutinised by the author offering the best possible explanation for the functioning of the proposed processing pipelines. At the end of this discussion chapter a comparison between the best performing approaches from DMD modes, spectrum and maps are compared to the state-of-the-art techniques used for each assessed dataset.

The best processing routes from each experiment are firstly listed below and the metrics for average classification accuracy and kappa values have been gathered and tabulated in Table 6.1. Most interestingly 'plain' DMD modes performed the best out of all three proposed types of features. They were closely followed by the novel use of DMD maps and DMD spectrum coming last.

- DMD modes: CAR filtering, energy-scaled DMD modes, no mode binning and selecting features with projection kernel (combination 2a)
- DMD spectrum: use of MInf approach to select best features from the provided DMD spectrum vector
- DMD maps: using only absolute maps obtained from energy-scaled DMD modes

Table 6.1: Best results obtained from the three proposed experiments.

	Average accuracy %, (kappa value)		
	DMD modes	DMD spectrum	DMD maps
BCI IV 2a	**48.5** (0.31)	31.8 (0.10)	39.5 (0.22)
GIST-MI	**64.3** (0.28)	55.1 (0.10)	61.7 (0.28)
BNCI 2020	**20.2** (0.04)	19 (0.04)	16.5 (0.03)

6.1 DMD modes

The presented results have provided valuable information regarding the validity of DMD modes as features, their general performance as well as currently the best possible processing route maximising the performance. Furthermore, performed experiments provided numerical evidence showing the importance of CAR spatial filtering and the surprising negative influence of mode binning on the system performance.

All the evidence shown unanimously agrees that filtering EEG signals with CAR method, followed by extracting energy-scaled DMD modes and skipping mode binning process, with final feature projection using projection kernel achieves the best performance out of all the proposed processing routes. Despite this, the achieved classification accuracy and kappa values are lower than state-of-the-art techniques, leading the author to reject the hypothesis that DMD modes provide meaningful features. However, the author notes that the proposed method utilising DMD modes is still at very early stages of development, especially in EEG and MI-BCI problems, and suggests that DMD has a potential to perform better if researched further in this particular application.

In terms of the differences between individual variables in the processing pipeline, the clearest change was seen in the type of DMD modes used as features; scaled DMD modes provided better performance than normalised DMD modes. Initially, the author speculated that normalised DMD modes should provide better performance, since individual normalised modes show relative influence of each electrode within a particular mode. This, in theory, should have produced a matrix of modes in which electrodes within sensorimotor cortex would have carried the most influence on modes, due to

strong spatial characteristics of the ERD/S phenomenon. However in practice, as seen from the experimental data, it was the energy-scaled DMD modes which provided better features in terms of classification accuracy and kappa value. This leads the author to infer that DMD approach is better suited in finding the energy-related changes in MI-EEG signals rather than the spatial ones.

The evidence seen in the presented results shows that applying CAR filtering to EEG signals before extracting DMD modes has a positive impact and increases overall performance of the system. This finding coincides with the similar findings made in Hirsh et al. (2020) and Seenivasaharagavan et al. (2021). While the mentioned studies did not look specifically on the effect of spatial filtering on DMD modes, they investigated the effect of mean subtraction and data centring processes on the extracted DMD modes. Since, CAR filtering subtracts global mean at an electrode of interest from a given trial, the resultant filtered signal has zero mean meaning that data is centred (Section 3.1.2). The existing knowledge of the benefits related to spatial filtering combined with the information found in the two mentioned studies, and the numerical results obtained in this thesis allows the author to fully recommend usage of CAR filtering when extracting DMD modes from MI-EEG signals.

The most perplexing finding however was discovered in the feature selection part of the system. As mentioned in the results section, at the early stage of experiments the author has noticed that skipping mode binning step increased the classification accuracy, meaning that features used for classification contained more meaningful data. This is further supported by the calculated kappa values for different processing routes. While this type of behaviour would be certainly expected when using PCA as the feature selection method, as it performs dimensionality reduction on its own through selection of only first few principal components, in the case of projection kernel the observed behaviour was against the constraints present in the subspace projection process.
In order to find the distance (through principal angle) between two subspaces on a Grassmannian manifold, the two subspaces have to be orthogonal. Furthermore, one can recall that extracted DMD modes are stated to be non-orthogonal. As a workaround, previous literature suggested that matrices of DMD modes can be orthogonalised through employing QR decomposition process such that $\Phi = QR$, where Q is

the orthonormal representation of the matrix Φ. In the case of Bito et al. (2019), the number of modes preserved for analysis seems to have been chosen arbitrarily ($r = 9$), which allowed to satisfy the requirement for a valid QR decomposition: $c \geq r$. Conveniently the code from the said publication along with the used dataset are publicly available. The author investigated the available code and compared it with the version used in the current thesis, and found no differences between the two, thus leading to conclusion that problem does not lie in the code interpretation of the DMD algorithm, but potentially in the nature of signals. Keeping that remark in mind, investigating another recent publication by Shiraishi et al. (2020), which used DMD to extract MI features from ECoG signals, reveals another technique for orthogonalisation of DMD modes. The researchers in the aforementioned study seem to completely disregard orthogonalisation through QR decomposition and they firstly normalize each DMD mode by its $l2$-norm; then use such normalised matrices of DMD modes, as the new subspaces meaning that (54) effectively becomes:

$$kp(\Phi_i, \Phi_j) = ||\Phi_i' \Phi_j||_F^2$$

The research presented in this thesis is not able to fully explain the discrepancies found in the literature as the obtained results are conflicting with the reviewed literature. Most notably, using the above equation with non-orthogonalised DMD modes provided better performance than following the orthogonalisation procedure described in Bito et al. (2019). The author offers some potential explanations for such peculiar behaviour.

Firstly, the author notes that during mode binning process all modes within a certain frequency bin are averaged, which can lead to some data loss, especially if there is a wide range in data i.e. modes with both low and high energy. Therefore, the author speculates that application of another metrics such as higher-order statistics could potentially better describe the data within the frequency bin. A suggestion would be to potentially look at the skewness and kurtosis of the modes within a certain frequency bin and find which modes hold more significance.

Additionally, the author suggests that the issue might lie within the core idea of mode projection itself. Orthogonalising DMD modes through mode binning and sub-

sequent QR decomposition could potentially remove the previously present separation between the modes in the process. Following this reasoning, the author offers another explanation as to why, generally, energy-scaled DMD modes perform better than normalised DMD modes: the differences between relative influences between electrodes in modes are not as clearly separable as the differences between individual modal energies.

Lastly, the author acknowledges that Bito et al. (2019) used phone sensor signals, which are much clearer, and differences between different activities can be seen with small amount of further signal processing. That cannot be said about EEG recordings, particularly the ones measuring MI activity, as the signals are almost blended with background brain activity or noise. This argument gains more strength when investigating Shiraishi et al. (2020) and even the original publication in Brunton et al. (2016), as both of the papers implemented DMD in ECoG recordings, which are known to have a much greater SNR and overall readability compared to EEG.

6.2 DMD spectrum

Building upon the findings from the experiment concerning use of 'pure' DMD modes, the author implemented a novel approach by extracting DMD spectrum with the purpose of utilising it as features from MI-EEG signals. The author notes that according to his current knowledge, it is the first ever such utilisation of DMD method in this field.

From the three proposed routes used for processing DMD spectrum features, using MInf to select best features yields the best results overall. Alas, as noted in Section 5.3 the difference between either of the approaches is minimal. While accuracy results are comparable to the ones obtained for DMD modes as seen in Table 6.1, comparing kappa values instead reveals that features from DMD spectrum are the least useful ($\kappa \leq 0.10$), indicating that less than 1% of data is reliable, and consequently meaning there is no agreement in the supplied features. With such low kappa values the author confidently rejects the hypothesis that DMD spectrum is able to extract any meaningful features from MI-EEG signals based on the ERD/S phenomenon.

The author offers some explanation as to why, despite extracting DMD spectrum

from energy-scaled DMD modes obtained from CAR filtered EEG signals (the best processing route as shown in the first experiment), the measured performance metrics for DMD spectrum features are unsatisfactorily low. The potential problem lies within the DMD spectrum itself. Recall that the DMD spectrum qualitatively resembles average FFT spectrum (over all electrodes), as described in Section 3.2.8 and Kutz et al. (2016). As such, calculated power in the DMD spectrum is equivalent to the contribution of all electrodes in a particular mode, which as a result leads to loss of vital information from individual electrodes. Indeed previous literature forms feature vectors by calculating FFT or bandpower (average or logarithmic) for each of the electrodes independently (López-Larraz et al., 2014; Sburlea et al., 2015; Majkowski et al., 2017).

The author also notes that another factor negatively impacting the performance of DMD spectrum features is the use of mode binning. As it was the case in the experiment concerning DMD modes, mode binning leads to loss of information. Unfortunately, the author is not able to recommend an effective alternative to mode binning. Simply extracting DMD modes from a certain frequency range could be a viable alternative allowing to preserve all the modes, however during the experiments the author noted that each trial produced modes at different characteristic frequencies, leading to situations that the number of modes within 7-30Hz for one trial was different to number of modes within the same frequency range for another trial. Thus mode binning through simple averaging was necessary to ensure that each feature vector has the same length.

6.3 DMD maps

The results obtained in the culminating experiment of this thesis provided solid evidence allowing the author to reject the last hypothesis formed in this study, which theorised if the spatio-temporal maps extracted from DMD modes would produce valuable features and provide satisfactory performance. Nonetheless, few remarkable discoveries were made which, in the author's opinion, are a valuable addition to the knowledge and could help shaping the future research. Two main comparisons under scrutiny are the differences found between the use of normalised or energy-scaled maps and the effect of inclusion of phase maps in the system.

Similarly to the first experiment, the author initially postulated that maps obtained from normalised DMD modes could be more beneficial and perform better than maps from energy-scaled modes, especially when using a convolutional neural network to look for patterns in the most influential electrodes and modes. However, as seen in the results, information contained within the energy-scaled DMD modes still provided the best performance, even when transformed into intensity maps.

From average classification accuracy found in Table 5.16, energy-scaled maps noticed a 10.5%, 0.8% and 0.5% increase across the tested datasets compared to normalised maps. However, looking at the obtained kappa values in Table 5.19 reveals some anomalies which do not fully align with the classification accuracy results. While for the first dataset energy-scaled maps have more than a double kappa value than this of normalised maps (+0.12 gain), which is in agreement with the previous accuracy results, in the second and third dataset normalised maps attained higher kappa values, alas by minimal margin (+0.03 and +0.01), despite having lower classification accuracy. The author cannot explain this peculiarity, as assessing average sensitivity and specificity shows that both normalised and energy-scaled maps have the exact same values in second and third dataset.

Nonetheless, the general trend is in favour of energy-scaled maps, which in the author's view this solidifies the premise that the most suitable DMD modes for MI-EEG signals are energy-scaled ones. Furthermore, this would indicate that DMD approach describes ERD/S phenomenon better by its energy distribution across the modes (following the spectral characteristics), rather than the localised influence of individual electrodes (strictly spatial characteristic).

Assessing the provided results also revealed that the proposition made by the author concerning the inclusion of phase maps in the system turned out to have an overall negative impact on the performance. From Table 5.16 it can be seen that across the tested datasets accuracy dropped by 5.1%, 3.6% and 0.4% respectively for normalised maps. Whereas, for the energy-scaled maps including phase maps resulted in a 10.2%, 5.1% drop for the first two datasets. Unexpectedly for the third dataset, the inclusion of phase maps increased classification accuracy by 0.8%. Apart from this single measured abnormality, the negative trend associated with the inclusion of phase maps is clearly

seen across the used data. Kappa values confirm that as recalling Table 5.19 shows that normalised maps saw a kappa decrease of 0.04 and 0.09 for the first two datasets, while no change was noted for the last dataset.

Considering the discussed points regarding type of scaling used and incorporation of phase information, the author suggests that a possible rearrangement of the DMD modes, specifically the order of channels, and slight adjustments to the filter parameters in the proposed CNN, could potentially be of benefit. Ensuring that the data channels are arranged in a more 'anatomically correct' order e.g., frontal, central and parietal areas, could possibly bring more cohesion to data and thus allow the CNN to find better spatial relations in the provided maps. Consequently, the vertical stride of the convolutional filters should be altered such that it would appropriately accommodate the aforementioned areas specifically.

6.4 Comparison to the state-of-the-art

The performance of the proposed approaches is compared to the current state-of-the-art methods observed in the literature. This process is completed per dataset basis, which allows to discuss specifics of the analysed dataset.

Firstly, the BCI Competition IV Dataset 2a from Brunner et al. (2008) is assessed. As observed in Table 6.2, all three proposed DMD based methods have almost half the performance of the current state-of-the-art. DMD modes which produced the highest results in this thesis fall behind CSP approach by -25.2% and by -28.9% compared to EEG-TCNet (Ingolfsson et al., 2020), a CNN approach using raw EEG signals. Interestingly enough the approach used by Hersche et al. (2018) is conceptually quite close to the idea of utilising DMD modes with the projection kernel. There, a multiscale CSP approach is applied by splitting EEG signal into multiple temporal and spectral components and using CSP to calculate a set of features, clearly tackling all three domains to maximise gathered information just as proposed in this thesis. Furthermore, Hersche et al. (2018) also applies Riemannian manifold projection to spectral information extracted from EEG signals to obtain Riemannian covariance matrices. On the contrary, the approaches presented in Shallow ConvNet by Schirrmeister et al. (2017),

EEGNet by Lawhern et al. (2018) and the previously mentioned EEG-TCNet only use raw EEG data and allow the convolution filters present in neural networks to find their own features in EEG signals, rather than supply 'man-made' CSP or spectral features. The results show that CNNs are able to perform just as well as classic CSP approach and much better than the proposed DMD based methods. The supplied kappa values for the three CNNs also reveal that features extracted from raw EEG signals have more than double the value, and are much more reliable.

The results for dataset from Cho et al. (2017) reveal an interesting case for DMD modes as shown in Table 6.3. Compared to the original approach utilising CSP features in Cho et al. (2017), DMD modes suffered only a -3.2% decrease in accuracy with DMD maps falling behind by -5.8% which is much smaller than the difference observed in the first dataset. Furthermore, DMD modes managed to outperform both Shallow ConvNet and EEGNet by $+1.3\%$ and $+0.3\%$ respectively, while DMD maps fell behind only by -1.3% and -2.3%. The author notes that Shallow ConvNet and EEGNet with the addition of Parallel CRNN (originally by Zhang et al. (2018)) were implemented in Ko et al. (2021) to specifically explore their performance on the examined dataset as the original papers did not utilise it and only took into account dataset by Brunner et al. (2008). Parallel CRNN ends the good streak of DMD modes by providing an impressive 79% accuracy, only to be superseded by MSNN proposed by Ko et al. (2021) with 81% accuracy. Unfortunately, none of the explored studies provided any kappa values therefore it is not possible to assess the reliability of the features used in those publications. Those findings are of significance because the dataset under investigation only offers 2 classes of signals, which would mean that DMD is almost as capable as CSP approach in the case of simple binary classification problems.

Comparing the performance of the proposed DMD based features in the the last dataset could not be accomplished fully due to some certain issues. Recalling from Chapter 4, dataset by Ofner et al. (2017) recorded a staggering number of 6 different classes with addition of separate rest (inactivity) class. In this thesis the author chose to follow the original paper by Ofner et al. (2017) and train a classifier which would be able to deal with 6 different classes (omitting the rest class). During the literature review, the author was able to find only one other publication which followed the same

methodology, i.e. including all movements in the classifier (Mammone et al., 2020), while the other two publications included in Table 6.4 use specific pairings of actions and classify them against rest class. Examining the results in Table 6.4, DMD modes lose 4.8% and 6.8% accuracy compared to the two original methods used by Ofner et al. (2017): single time point and time window, respectively. It is important to note that the two mentioned methods used time signals which were examined and processed specifically to extract MRCPs instead of exploiting ERD/S phenomena as it has been done in this thesis. The deep CNN implemented by Mammone et al. (2020) outperforms DMD modes by more than triplefold (42.3% gain over DMD modes). The Hierarchical Flow CNN (HF-CNN) achieved an accuracy of 51%; however it only used the forearm supination and pronation actions and compared them against rest class, therefore it cannot be directly compared to the score achieved by DMD modes. Similarly, DCNN approach (Ieracitano et al., 2021) only used opening and closing hand actions against rest class; this method reached an accuracy of 90% but, again, cannot be directly compared to DMD modes. As it was seen in the second dataset, the publications reviewed here did not provide any kappa values, therefore the author was not able to compare the reliability of the extracted features from the presented state-of-the-art methods.

Table 6.2: Comparison of the proposed techniques to the current state-of-the-art approaches in literature for the BCI IV 2a dataset

	DMD modes	DMD spectrum	DMD maps	CSP (Hersche et al., 2018)	Riemannian (Hersche et al., 2018)	Shallow ConvNet (Schirrmeister et al., 2017)	EEGNet (Lawhern et al., 2018)	EEG-TCNet (Ingolfsson et al., 2020)
Accuracy %	48.5	31.8	39.5	73.7	74.8	74.3	72.4	77.4
κ	0.31	0.10	0.22	-	-	0.66	0.63	0.70

Table 6.3: Comparison of the proposed techniques to the current state-of-the-art approaches in literature for the GIST-MI dataset

	DMD modes	DMD spectrum	DMD maps	CSP (Cho et al., 2017)	Shallow ConvNet	EEGNet	Parallel CRNN	MSNN (Ko et al., 2021)
Accuracy %	64.3	55.1	61.7	67.5	63	64	79	81
κ	0.28	0.10	0.28	-	-	-	-	-

Table 6.4: Comparison of the proposed techniques to the current state-of-the-art approaches in literature for the BNCI 2020 dataset

	DMD modes	DMD spectrum	DMD maps	Single time point (Ofner et al., 2017)	Time window (Ofner et al., 2017)	Deep CNN (Mammone et al., 2020)	HF-CNN* (Jeong et al., 2020)	DCNN** (Ieracitano et al., 2021)
Accuracy %	20.2	19	16.5	25	27	62.5	51	90
κ	0.04	0.04	0.03	-	-	-	-	-

* - only comparing forearm supination and pronation actions against rest, ** - only comparing opening and closing hand actions against rest

6.5 Conclusion

Despite the initially observed poor performance by all three proposed DMD based features, a number of valuable discoveries have been made for which the author has presented discussion in this chapter. Even though none of the approaches i.e. DMD modes, DMD spectrum and DMD map have attained a satisfactory level of performance, the author argues that both DMD modes and DMD maps cannot be completely disregarded. Given the very early stage of DMD-based research in general, the author sees a potential in this technique, particularly for DMD modes. The author suggests that further investigation into the projection kernel method would be beneficial to the DMD method, and would allow it to reach higher performance. However, in the case of DMD spectrum the author stands by discouraging the use of this method, as the results clearly show that it is not suitable for extracting MI-EEG based features. Through the examination of current state-of-the-art and comparing their performance in the three used datasets, the author found out that the performance of DMD modes was quite close to state-of-the-art in two out of three datasets. In the case of the GIST-MI dataset, DMD modes managed to even outperform two state-of-the-art methods.

Chapter 7

Conclusion and future work

This research utilised a novel signal decomposition technique DMD to the field of MI-EEG in order to fill out the identified lack of spatial domain based approaches when investigating ERD/S phenomenon. DMD produces the so-called DMD modes which are in fact spatio-temporal patterns describing the low-rank dynamics present in the examined window of signal. After researching the use of this method in the literature, the author proposed an in-depth investigation into three different types of features: DMD modes, DMD spectrum and DMD maps and presented the findings regarding classification performance.

7.1 Conclusion

A number of experiments were performed while using DMD modes, in order to firstly identify the importance of spatial filtering of the original EEG signal and its effect on the classification. This investigation revealed that applying CAR spatial filter was able to greatly enhance the signal and the resultant DMD modes, leading to a notable increase in classification accuracy and kappa value. Secondly, the author compared the two scaling methods used for DMD modes: normalised method and scaling by SVD energy. While initially the author made the assumption that normalised scaling would lead to better performance, as ERD/S phenomenon is heavily localised across the electrodes and such scaling enhances the relative influence of channels (electrodes) on the calculated modes, the obtained results disproved that claim and showed a sig-

nificant increase in performance when SVD-energy scaling was used instead. Lastly, an appropriate feature selection method for DMD modes was studied. The analysed literature revealed two methods suggested for such process, namely PCA and projection kernel used on Grassmannian manifold. The author confirmed through thorough investigation that the use of projection kernel produced more accurate and reliable features compared to the ones selected by PCA.

Additionally, the author discovered an anomaly in the process involved with the preparation of DMD modes, so they can be used with projection kernel. Specifically, DMD modes which are used as subspaces for the later projection are required to be orthogonal, which is problematic as by nature DMD modes are non-orthogonal. The literature suggests to use Q matrix obtained from QR decomposition as an orthogonal representation of the DMD modes; however it was found out that for such decomposition to be valid the number of channels in DMD modes has to be higher than the rank number. This was challenging as in this study the number of channels was smaller than the rank number. Implementing mode binning process which allowed to satisfy the QR decomposition requirement was shown to have drastically negative impact on the final classification accuracy. Consequentially, breaking that requirement and using DMD modes without orthogonalisation to calculate projection kernels provided a superior performance.

In the case of DMD spectrum, the author wished to explore the spectral information contained in the DMD modes, to look for patterns in the modal energy distribution, with the intent of finding patterns reflecting ERD/S phenomenon. In addition to using raw DMD spectrum, MRMR and MInf feature selection methods were employed with hopes of further narrowing down the feature vector to only contain the most useful data. However, the experimental results revealed little to no difference between using one of the aforementioned selection techniques and supplying raw spectrum to the classifier. Nonetheless, MInf method performed marginally better, but the author stated that in general DMD spectrum method is not useful due to the obtained very low kappa values, clearly suggesting DMD spectrum did not contain any reliable information regarding ERD/S phenomenon.

The last representation of DMD based features explored in this thesis were DMD

maps where, initially the author suggested that spatial patterns would be discovered by a developed CNN on the su pplied maps reflecting the absolute and phase information of DMD modes. As it was the case with DMD modes investigation, the difference between normalised and energy-scaled maps was examined, where at the start the author was in favour of the normalised approach. The results showed again that scaling DMD modes by SVD energy provided more reliable maps than normalised ones. Furthermore, the results also revealed that using absolute maps on their own instead of combing map information from absolute and phase data was more accurate and reliable.

Gathering the findings from the experiments on the three proposed features revealed that DMD modes were able to achieve the highest performance out of all investigated approaches. While this performance did not seem very satisfactory at the beginning, comparing it to the current state-of-the-art led to some surprising discoveries. In the case of the BCI IV 2a dataset, all DMD modes were nowhere near the performance achieved by the classic CSP approach, or much more modern CNN methods. This however changed when GIST-MI dataset was investigated, as DMD modes only minimally fell behind CSP method while also outperforming other state-of-the-art techniques, albeit by small fraction. Although DMD maps are not being able to beat any of the state-of-the-art, they still managed to perform relatively well. Similarly in the case of the BNCI 2020 dataset, DMD modes were able to perform very closely to the method used in the original paper, but then fell off by a big margin compared to a modern CNN approach.

7.2 Future work

Lastly, the author wishes to share a number of suggestions for future work, which are the result of the observations made throughout the research discussed in this thesis. The most obvious suggestion is further development of the DMD technique. As the research community has been gaining a better understanding of DMD and its associated processes, more robust techniques for extracting modes can be developed (Scherl et al., 2020; Abolmasoumi et al., 2021). For certain, the current approach to calculate DMD is not suitable for real-time applications, which is limited by the lack of efficient and

reliable methods to update eigenvalues in real-time, and the computational speed of calculating SVD for bigger windows of data. Thus, to allow DMD to be used in real-time BCIs, more work would have to be carried out to address the described limitations.

From the outcomes presented in the results section, the author can make more specific suggestions for future work. As the results showed, DMD spectrum does not produce any reliable features for ERD/S classification, and therefore the author would stand by the claims made earlier in the thesis, and discourage any work in that direction. However, the experiments involving DMD modes and maps revealed a number of potential future improvements.

From the studies concerning DMD modes, the author would like to propose a number of ideas. It is evident that feature selection through projection kernel works very well with DMD modes. However, as pointed out earlier in the thesis, the results obtained in the discussed research do not align with the literature on this subject. Therefore it would be of the highest importance to further investigate the issues involved with orthogonalisation of DMD mode matrices. This naturally pairs up with evaluating more reliable methods for extracting the most meaningful modes to reduce the size of DMD matrix, adhering to the requirements of QR decomposition.

Regarding DMD maps, the author suggests that the absolute and phase information of the extracted DMD modes should be investigated further as the acquired results indicate a potential of the proposed technique. For certain, implementing a suitable mode selection technique, as it was suggested previously with DMD modes, could be beneficial for the DMD maps as only the most important modes would be retained. For the last suggestion, the author proposes a deeper research into neural networks utilised for classifying DMD maps and further research into hyperparameter tuning of the associated networks.

Bibliography

Abolmasoumi, A. H., Netto, M., & Mili, L. (2021). Robust Dynamic Mode Decomposition.
URL http://arxiv.org/abs/2105.09869

Akhtari, M., Bryant, H. C., Mamelak, A. N., Flynn, E. R., Heller, L., Shih, J. J., Mandelkem, M., Matlachov, A., Ranken, D. M., Best, E. D., Dimauro, M. A., Lee, R. R., & Sutherling, W. W. (2002). Conductivities of three-layer line human skull. *Brain Topography*, *14*(3), 151–167.

Amin, S. U., Alsulaiman, M., Muhammad, G., Bencherif, M. A., & Hossain, M. S. (2019a). Multilevel Weighted Feature Fusion Using Convolutional Neural Networks for EEG Motor Imagery Classification. *IEEE Access*, *7*, 18940–18950.

Amin, S. U., Alsulaiman, M., Muhammad, G., Mekhtiche, M. A., & Shamim Hossain, M. (2019b). Deep Learning for EEG motor imagery classification based on multi-layer CNNs feature fusion. *Future Generation Computer Systems*, *101*, 542–554.

Anderson, C., & Sijercic, Z. (1996). Classification of EEG signals from four subjects during five mental tasks. *Advances*, (pp. 407–414).

Ang, K. K., Chin, Z. Y., Wang, C., Guan, C., & Zhang, H. (2012). Filter bank common spatial pattern algorithm on BCI competition IV datasets 2a and 2b. *Frontiers in Neuroscience*, *6*(Mar).

Ang, K. K., Chin, Z. Y., Zhang, H., & Guan, C. (2008). Filter Bank Common Spatial Pattern (FBCSP) in brain-computer interface. In *Proceedings of the International Joint Conference on Neural Networks*, (pp. 2390–2397).

Artoni, F., Delorme, A., & Makeig, S. (2018). Applying dimension reduction to EEG data by Principal Component Analysis reduces the quality of its subsequent Independent Component decomposition. *NeuroImage*, *175*, 176–187.

Barachant, A., Bonnet, S., Congedo, M., & Jutten, C. (2010). Common spatial pattern revisited by Riemannian geometry. *2010 IEEE International Workshop on Multimedia Signal Processing, MMSP2010*, (pp. 472–476).

Barachant, A., Bonnet, S., Congedo, M., & Jutten, C. (2012). Multiclass brain-computer interface classification by Riemannian geometry. *IEEE Transactions on Biomedical Engineering*, *59*(4), 920–928.

Barachant, A., Bonnet, S., Congedo, M., & Jutten, C. (2013). Classification of covariance matrices using a Riemannian-based kernel for BCI applications. *Neurocomputing*, *112*, 172–178.

Bassi, P. R. A. S., & Attux, R. (2021). FBCNN: A Deep Neural Network Architecture for Portable and Fast Brain-Computer Interfaces. *ArXiv*.

Batula, A. M., Mark, J. A., Kim, Y. E., & Ayaz, H. (2017). Comparison of Brain Activation during Motor Imagery and Motor Movement Using fNIRS. *Computational Intelligence and Neuroscience*, *2017*.

Berger, H. (1929). Über das Elektrenkephalogramm des Menschen. *Archiv für Psychiatrie und Nervenkrankheiten*, *87*(1), 527–570.

Bhattacharyya, S., Khasnobish, A., Chatterjee, S., Konar, A., & Tibarewala, D. N. (2010). Performance analysis of LDA, QDA and KNN algorithms in left-right limb movement classification from EEG data. In *International Conference on Systems in Medicine and Biology, ICSMB 2010 - Proceedings*, (pp. 126–131).

Bigdely-Shamlo, N., Mullen, T., Kothe, C., Su, K. M., & Robbins, K. A. (2015). The

PREP pipeline: Standardized preprocessing for large-scale EEG analysis. *Frontiers in Neuroinformatics*, *9*(JUNE), 1–19.

Bito, T., Hiraoka, M., & Kawahara, Y. (2019). Learning with Coherence Patterns in Multivariate Time-series Data via Dynamic Mode Decomposition. In *Proceedings of the International Joint Conference on Neural Networks*, vol. 2019-July, (pp. 1–8). IEEE.

Blankertz, B., Curio, G., & Müller, K. R. (2002). Classifying single trial EEG: Towards brain computer interfacing. *Advances in Neural Information Processing Systems*.

Blankertz, B., Dornhege, G., Krauledat, M., Müller, K. R., & Curio, G. (2007). The non-invasive Berlin Brain-Computer Interface: Fast acquisition of effective performance in untrained subjects. *NeuroImage*, *37*(2), 539–550.

Blankertz, B., Tomioka, R., Lemm, S., Kawanabe, M., & Müller, K. R. (2008). Optimizing spatial filters for robust EEG single-trial analysis. *IEEE Signal Processing Magazine*, *25*(1), 41–56.

Blokland, Y., Spyrou, L., Thijssen, D., Eijsvogels, T., Colier, W., Floor-Westerdijk, M., Vlek, R., Bruhn, J., & Farquhar, J. (2014). Combined EEG-fNIRS decoding of motor attempt and imagery for brain switch control: An offline study in patients with tetraplegia. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *22*(2), 222–229.

Brunner, C., Leeb, R., Müller-Putz, G. R., Schlögl, A., & Pfurtscheller, G. (2008). BCI Competition 2008 – Graz data set A.

Brunner, C., Naeem, M., Leeb, R., Graimann, B., & Pfurtscheller, G. (2007). Spatial filtering and selection of optimized components in four class motor imagery EEG data using independent components analysis. *Pattern Recognition Letters*, *28*(8), 957–964.

Brunton, B. W., Johnson, L. A., Ojemann, J. G., & Kutz, J. N. (2016). Extracting spatial-temporal coherent patterns in large-scale neural recordings using dynamic mode decomposition. *Journal of Neuroscience Methods*, *258*, 1–15.

Bugli, C., & Lambert, P. (2007). Comparison between principal component analysis and independent component analysis in electroencephalograms modelling. *Biometrical Journal*, *49*(2), 312–327.

Buzsáki, G., Anastassiou, C. A., & Koch, C. (2012). The origin of extracellular fields and currents-EEG, ECoG, LFP and spikes. *Nature Reviews Neuroscience*, *13*(6), 407–420.

Cahn, B. R., & Polich, J. (2006). Meditation states and traits: EEG, ERP, and neuroimaging studies. *Psychological Bulletin*, *132*(2), 180–211.

Carter, C. S., Braver, T. S., Barch, D. M., Botvinick, M. M., Noll, D., & Cohen, J. D. (1998). Anterior cingulate cortex, error detection, and the online monitoring of performance. *Science*, *280*(5364), 747–749.

Carvalhaes, C., & De Barros, J. A. (2015). The surface Laplacian technique in EEG: Theory and methods. *International Journal of Psychophysiology*, *97*(3), 174–188.

Cecotti, H., & Gräser, A. (2011). Convolutional neural networks for P300 detection with application to brain-computer interfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(3), 433–445.

Chevallier, S., Barthélemy, Q., & Atif, J. (2014). Subspace metrics for multivariate dictionaries and application to EEG. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, (pp. 7178–7182).

Chiarelli, A. M., Croce, P., Merla, A., & Zappasodi, F. (2018). Deep learning for hybrid EEG-fNIRS brain-computer interface: Application to motor imagery classification. *Journal of Neural Engineering*, *15*(3).

Cho, H., Ahn, M., Ahn, S., Kwon, M., & Jun, S. C. (2017). EEG datasets for motor imagery brain–computer interface. *GigaScience*, *6*(7), 1–8.

Cohen, M. X. (2017). Where Does EEG Come From and What Does It Mean? *Trends in Neurosciences*, *40*(4), 208–218.

Congedo, M., Barachant, A., & Bhatia, R. (2017). Riemannian geometry for EEG-based brain-computer interfaces; a primer and a review. *Brain-Computer Interfaces*, *4*(3), 155–174.

Craik, A., He, Y., & Contreras-Vidal, J. L. (2019). Deep learning for electroencephalogram (EEG) classification tasks: A review. *Journal of Neural Engineering*, *16*(3), 28.

de Cheveigné, A., & Nelken, I. (2019). Filters: When, Why, and How (Not) to Use Them. *Neuron*, *102*(2), 280–293.

Dutta, K. K. (2019). Multi-class time series classification of EEG signals with recurrent neural networks. In *Proceedings of the 9th International Conference On Cloud Computing, Data Science and Engineering, Confluence 2019*, (pp. 337–341). Institute of Electrical and Electronics Engineers Inc.

Farwell, L. A., & Donchin, E. (1988). Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and Clinical Neurophysiology*, *70*(6), 510–523.

Ferrez, P. W., & Del R. Millán, J. (2008). Error-related EEG potentials generated during simulated brain-computer interaction. *IEEE Transactions on Biomedical Engineering*, *55*(3), 923–929.

Fisher, R. A. (1936). the Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, *7*(2), 179–188.

Freer, D., & Yang, G. Z. (2020). Data augmentation for self-paced motor imagery classification with C-LSTM. *Journal of Neural Engineering*, *17*(1).

Galán, F., Nuttin, M., Lew, E., Ferrez, P. W., Vanacker, G., Philips, J., & Millán, J. d. R. (2008). A brain-actuated wheelchair: Asynchronous and non-invasive Brain-computer interfaces for continuous control of robots. *Clinical Neurophysiology*, *119*(9), 2159–2169.

Gavish, M., & Donoho, D. L. (2014). The optimal hard threshold for singular values is 4/sqrt(3). *IEEE Transactions on Information Theory*, *60*(8), 5040–5053.

Goldman, R. I., Stern, J. M., Engel Jr, J., & Cohen, M. S. (2002). Simultaneous EEG and fMRI of the alpha rhythm. *NeuroReport*, *13*(18), 2487–2492.

Graimann, B., Huggins, J. E., Levine, S. P., & Pfurtscheller, G. (2002). Visualization of significant ERD/ERS patterns in multichannel EEG and ECoG data. *Clinical Neurophysiology*, *113*(1), 43–47.

Grosse-Wentrup, M., & Buss, M. (2008). Multiclass common spatial patterns and information theoretic feature extraction. *IEEE Transactions on Biomedical Engineering*, *55*(8), 1991–2000.

Guger, C., Allison, B. Z., Großwindhager, B., Prückl, R., Hintermüller, C., Kapeller, C., Bruckner, M., Krausz, G., & Edlinger, G. (2012). How many people could use an SSVEP BCI? *Frontiers in Neuroscience*, *0*(NOV), 169.

Hamm, J. (2008). *Subspace-Based Learning With Grassmann Kernels*. Doctoral thesis, University of Pennsylvania.

Hamm, J., & Lee, D. D. (2008). Grassmann discriminant analysis: A unifying view on subspace-based learning. In *Proceedings of the 25th International Conference on Machine Learning*, (pp. 376–383).

Hamm, J., & Lee, D. D. (2009). Extended Grassmann kernels for subspace-based learning. *Advances in Neural Information Processing Systems 21 - Proceedings of the 2008 Conference*, (pp. 601–608).

Heldman, D. A., & Moran, D. W. (2020). Local field potentials for BCI control. *Handbook of Clinical Neurology*, *168*, 279–288.

Hermes, D., & Miller, K. J. (2020). iEEG: Dura-lining electrodes. *Handbook of Clinical Neurology*, *168*, 263–277.

Hersche, M., Rellstab, T., Schiavone, P. D., Cavigelli, L., Benini, L., & Rahimi, A. (2018). Fast and accurate multiclass inference for MI-BCIS using large multiscale temporal and spectral features. In *European Signal Processing Conference*, vol. 2018-Septe, (pp. 1690–1694). European Signal Processing Conference, EUSIPCO.

Hiraiwa, A., Shimohara, K., & Tokunaga, Y. (1990). EEG Topography Recognition by Neural Networks. *IEEE Engineering in Medicine and Biology Magazine*, *9*(3), 39–42.

Hirsh, S. M., Harris, K. D., Nathan Kutz, J., & Brunton, B. W. (2020). Centering data improves the dynamic mode decomposition. *SIAM Journal on Applied Dynamical Systems*, *19*(3), 1920–1955.

Hjorth, B. (1970). EEG analysis based on time domain properties. *Electroencephalography and Clinical Neurophysiology*, *29*(3), 306–310.

Hjorth, B. (1975). An on-line transformation of EEG scalp potentials into orthogonal source derivations. *Electroencephalography and Clinical Neurophysiology*, *39*(5), 526–530.

Hobson, J. A., & Pace-Schott, E. F. (2002). The cognitive neuroscience of sleep: Neuronal systems, consciousness and learning. *Nature Reviews Neuroscience*, *3*(9), 679–693.

Hochberg, L. R., Bacher, D., Jarosiewicz, B., Masse, N. Y., Simeral, J. D., Vogel, J., Haddadin, S., Liu, J., Cash, S. S., Van Der Smagt, P., & Donoghue, J. P. (2012). Reach and grasp by people with tetraplegia using a neurally controlled robotic arm. *Nature*, *485*(7398), 372–375.

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780.

Hortal, E., Planelles, D., Costa, A., Iáñez, E., Úbeda, A., Azorín, J. M., & Fernández, E. (2015). SVM-based Brain-Machine Interface for controlling a robot arm through four mental tasks. *Neurocomputing*, *151*(P1), 116–121.

Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Snin, H. H., Zheng, Q., Yen, N. C., Tung, C. C., & Liu, H. H. (1998). The empirical mode decomposition and the Hubert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *454*(1971), 903–995.

Hudson, D. L., & Cohen, M. E. (1999). *Neural networks and artificial intelligence for biomedical engineering*. Wiley-IEEE Press.
URL https://onlinelibrary.wiley.com/doi/book/10.1109/9780470545355

Hughes, J. R. (2008). Gamma, fast, and ultrafast waves of the brain: Their relationships with epilepsy and behavior. *Epilepsy and Behavior*, *13*(1), 25–31.

Hyvärinen, A., & Oja, E. (1997). A Fast Fixed-Point Algorithm for Independent Component Analysis. *Neural Computation*, *9*(7), 1483–1492.

Hyvärinen, A., & Oja, E. (2000). Independent component analysis: Algorithms and applications. *Neural Networks*, *13*(4-5), 411–430.

Ibáñez, J., Serrano, J. I., Del Castillo, M. D., Monge-Pereira, E., Molina-Rueda, F., Alguacil-Diego, I., & Pons, J. L. (2014). Detection of the onset of upper-limb movements based on the combined analysis of changes in the sensorimotor rhythms and slow cortical potentials. *Journal of Neural Engineering*, *11*(5).

Ieracitano, C., Mammone, N., Hussain, A., & Morabito, F. C. (2021). A novel explainable machine learning approach for EEG-based brain-computer interface systems. *Neural Computing and Applications*, (pp. 1–14).

Ingolfsson, T. M., Hersche, M., Wang, X., Kobayashi, N., Cavigelli, L., & Benini, L. (2020). EEG-TCNet: An Accurate Temporal Convolutional Network for Embedded Motor-Imagery Brain-Machine Interfaces. In *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 2020-Octob, (pp. 2958–2965).

Jasper, H. H. (1958). The ten twenty electrode system of the international federation. *Electroencephalography and Clinical Neurophysiology*, *10*, 371–375.
URL https://ci.nii.ac.jp/naid/10017996828

Jeong, J. H., Lee, B. H., Lee, D. H., Yun, Y. D., & Lee, S. W. (2020). EEG Classification of Forearm Movement Imagery Using a Hierarchical Flow Convolutional Neural Network. *IEEE Access*, *8*, 66941–66950.

Kevric, J., & Subasi, A. (2017). Comparison of signal decomposition methods in classification of EEG signals for motor-imagery BCI system. *Biomedical Signal Processing and Control*, *31*, 398–406.

Ko, W., Jeon, E., Jeong, S., & Suk, H. I. (2021). Multi-Scale Neural Network for EEG Representation Learning in BCI. *IEEE Computational Intelligence Magazine*, *16*(2), 31–45.

Koles, Z. J., Lazar, M. S., & Zhou, S. Z. (1990). Spatial patterns underlying population differences in the background EEG. *Brain Topography*, *2*(4), 275–284.

Kornhuber, H. H., & Deecke, L. (1965). Hirnpotentialänderungen bei Willkürbewegungen und passiven Bewegungen des Menschen: Bereitschaftspotential und reafferente Potentiale. *Pflügers Archiv für die Gesamte Physiologie des Menschen und der Tiere*, *284*(1), 1–17.

Kutz, J. N., Brunton, S. L., Brunton, B. W., & Proctor, J. L. (2016). *Dynamic Mode Decomposition*. Society for Industrial and Applied Mathematics.

Lashgari, E., Ott, J., Connelly, A., Baldi, P., & Maoz, U. (2021). An end-to-end CNN with attentional mechanism applied to raw EEG in a BCI classification task. *Journal of Neural Engineering*, *18*(4), 0460e3.

Lawhern, V. J., Solon, A. J., Waytowich, N. R., Gordon, S. M., Hung, C. P., & Lance, B. J. (2018). EEGNet: A compact convolutional neural network for EEG-based brain-computer interfaces. *Journal of Neural Engineering*, *15*(5).

Leamy, D. J., Collins, R., & Ward, T. E. (2011). Combining fNIRS and EEG to improve motor cortex activity classification during an imagined movement-based task. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *6780 LNAI*, 177–185.

Lee, H., & Choi, S. (2002). PCA-based linear dynamical systems for multichannel EEG classification. In *ICONIP 2002 - Proceedings of the 9th International Conference on Neural Information Processing: Computational Intelligence for the E-Age*, vol. 2, (pp. 745–749). Institute of Electrical and Electronics Engineers Inc.

Lee, J. H., Ryu, J., Jolesz, F. A., Cho, Z. H., & Yoo, S. S. (2009). Brain-machine interface via real-time fMRI: Preliminary study on thought-controlled robotic arm. *Neuroscience Letters*, *450*(1), 1–6.

Lew, E., Chavarriaga, R., Silvoni, S., & Millán, J. d. R. (2012). Detection of self-paced reaching movement intention from EEG signals. *Frontiers in Neuroengineering*, *5*(July).

Li, B., & Chen, X. (2014). Wavelet-based numerical analysis: A review and classification. *Finite Elements in Analysis and Design*, *81*, 14–31.

Li, X., Guan, C., Ang, K. K., Zhang, H., & Ong, S. H. (2014). Spatial filter adaptation based on geodesic-distance for motor EEG classification. *Proceedings of the International Joint Conference on Neural Networks*, (pp. 3859–3864).

Lopes Dias, C., Sburlea, A. I., & Müller-Putz, G. R. (2018). Masked and unmasked error-related potentials during continuous control and feedback. *Journal of Neural Engineering*, *15*(3), 036031.

López-Larraz, E., Montesano, L., Gil-Agudo, Á., & Minguez, J. (2014). Continuous decoding of movement intention of upper limb self-initiated analytic movements from pre-movement EEG correlates. *Journal of NeuroEngineering and Rehabilitation*, *11*(1), 153.

Lotte, F., Bougrain, L., Cichocki, A., Clerc, M., Congedo, M., Rakotomamonjy, A., & Yger, F. (2018). A review of classification algorithms for EEG-based brain-computer interfaces: A 10 year update. *Journal of Neural Engineering*, *15*(3).

Lotte, F., Congedo, M., Lécuyer, A., Lamarche, F., & Arnaldi, B. (2007). A review of classification algorithms for EEG-based brain-computer interfaces. *Journal of Neural Engineering*, *4*(2).

Lou, B., Hong, B., Gao, X., & Gao, S. (2008). Bipolar electrode selection for a motor imagery based brain-computer interface. *Journal of Neural Engineering*, *5*(3), 342–349.

Majkowski, A., Kolodziej, M., Zapala, D., Tarnowski, P., Francuz, P., Rak, R. J., & Oskwarek, L. (2017). Selection of EEG signal features for ERD/ERS classification using genetic algorithms. In *Proceedings of 2017 18th International Conference on Computational Problems of Electrical Engineering, CPEE 2017*. Institute of Electrical and Electronics Engineers Inc.

Mammone, N., Ieracitano, C., & Morabito, F. C. (2020). A deep CNN approach to decode motor preparation of upper limbs from time–frequency maps of EEG signals at source level. *Neural Networks*, *124*, 357–372.

McFarland, D. J., McCane, L. M., David, S. V., & Wolpaw, J. R. (1997). Spatial filter selection for EEG-based communication. *Electroencephalography and Clinical Neurophysiology*, *103*(3), 386–394.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, *22*(3), 276–282.

McMullen, D. P., Hotson, G., Katyal, K. D., Wester, B. A., Fifer, M. S., McGee, T. G., Harris, A., Johannes, M. S., Vogelstein, R. J., Ravitz, A. D., Anderson, W. S., Thakor, N. V., & Crone, N. E. (2014). Demonstration of a semi-autonomous hybrid brain-machine interface using human intracranial EEG, eye tracking, and computer vision to control a robotic upper limb prosthetic. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *22*(4), 784–796.

Meng, J., Zhang, S., Bekyo, A., Olsoe, J., Baxter, B., & He, B. (2016). Noninvasive Electroencephalogram Based Control of a Robotic Arm for Reach and Grasp Tasks. *Scientific Reports*, *6*(1), 1–15.

Millán, J. D. R., Renkens, F., Mouriño, J., & Gerstner, W. (2004). Noninvasive brain-actuated control of a mobile robot by human EEG. *IEEE Transactions on Biomedical Engineering*, *51*(6), 1026–1033.

Müller, K.-R., Krauledat, M., Dornhege, G., Curio, G., & Blankertz, B. (2004). Machine Learning Techniques for Brain-Computer Interfaces. *Machine Learning*, *49*, 11–22.

Müller-Gerking, J., Pfurtscheller, G., & Flyvbjerg, H. (1999). Designing optimal spatial filters for single-trial EEG classification in a movement task. *Clinical Neurophysiology*, *110*(5), 787–798.

Müller-Putz, G. R. (2020). Electroencephalography. *Handbook of Clinical Neurology*, *168*, 249–262.

Ofner, P., Schwarz, A., Pereira, J., & Müller-Putz, G. R. (2017). Upper limb movements can be decoded from the time-domain of low-frequency EEG. *PLoS ONE*, *12*(8), 1–24.

Ogawa, S., Lee, T. M., Kay, A. R., & Tank, D. W. (1990). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciences of the United States of America*, *87*(24), 9868–9872.

Olivas-Padilla, B. E., & Chacon-Murguia, M. I. (2019). Classification of multiple motor imagery using deep convolutional neural networks and spatial filters. *Applied Soft Computing Journal*, *75*, 461–472.

Oostenveld, R., & Praamstra, P. (2001). The five percent electrode system for high-resolution EEG and ERP measurements. *Clinical Neurophysiology*, *112*(4), 713–719.

Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(8), 1226–1238.

Perrin, F., Pernier, J., Bertrand, O., & Echallier, J. F. (1989). Spherical splines for scalp potential and current density mapping. *Electroencephalography and Clinical Neurophysiology*, *72*(2), 184–187.

Pfurtscheller, G. (2001). Functional brain imaging based on ERD/ERS. *Vision Research*, *41*(10-11), 1257–1260.

Pfurtscheller, G., & Aranibar, A. (1977). Event-related cortical desynchronization detected by power measurements of scalp EEG. *Electroencephalography and Clinical Neurophysiology*, *42*(6), 817–826.

Pfurtscheller, G., & Lopes Da Silva, F. H. (1999). Event-related EEG/MEG synchronization and desynchronization: Basic principles. *Clinical Neurophysiology*, *110*(11), 1842–1857.

Pfurtscheller, G., Neuper, C., Flotzinger, D., & Pregenzer, M. (1997). EEG-based discrimination between imagination of right and left hand movement. *Electroencephalography and Clinical Neurophysiology*, *103*(6), 642–651.

Pinegger, A., Hiebel, H., Wriessnegger, S. C., & Müller-Putz, G. R. (2017). Composing only by thought: Novel application of the P300 brain-computer interface. *PLoS ONE*, *12*(9).

Pohjalainen, J., Räsänen, O., & Kadioglu, S. (2015). Feature selection methods and their combinations in high-dimensional classification of speaker likability, intelligibility and personality traits. *Computer Speech and Language*, *29*(1), 145–171.

Ramoser, H., Müller-Gerking, J., & Pfurtscheller, G. (2000). Optimal spatial filtering of single trial EEG during imagined hand movement. *IEEE Transactions on Rehabilitation Engineering*, *8*(4), 441–446.

Ramsey, N. F. (2020). Human brain function and brain-computer interfaces. *Handbook of Clinical Neurology*, *168*, 1–13.

Rao, R. P., & Scherer, R. (2010). Statistical Pattern Recognition and Machine Learning in Brain-Computer Interfaces. *Statistical Signal Processing for Neuroscience and Neurotechnology*, (pp. 335–367).

Rehman, N., & Mandic, D. P. (2010). Multivariate empirical mode decomposition. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *466*(2117), 1291–1302.

Riccio, A., Simione, L., Schettini, F., Pizzimenti, A., Inghilleri, M., Belardinelli, M. O., Mattia, D., & Cincotti, F. (2013). Attention and P300-based BCI performance in people with amyotrophic lateral sclerosis. *Frontiers in Human Neuroscience*, *7*(732).

Richardson, A. (1967). Mental Practice: A Review and Discussion Part I. *Research*

Quarterly of the American Association for Health, Physical Education and Recreation, *38*(1), 95–107.

Richardson, A. (1969). *Mental Imagery*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1st ed.
URL http://link.springer.com/10.1007/978-3-662-37817-5

Sahonero-Alvarez, G., & Calderon, H. (2017). A comparison of SOBI, FastICA, JADE and infomax algorithms. *IMCIC 2017 - 8th International Multi-Conference on Complexity, Informatics and Cybernetics, Proceedings*, *2017-March*, 17–22.

Sain, S. R., & Vapnik, V. N. (1996). The Nature of Statistical Learning Theory. *Technometrics*, *38*(4), 409.

Sakhavi, S., Guan, C., & Yan, S. (2015). Parallel convolutional-linear neural network for motor imagery classification. In *2015 23rd European Signal Processing Conference, EUSIPCO 2015*, (pp. 2736–2740). Institute of Electrical and Electronics Engineers Inc.

Sakhavi, S., Guan, C., & Yan, S. (2018). Learning Temporal Information for Brain-Computer Interface Using Convolutional Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, *29*(11), 5619–5629.

Sburlea, A. I., Montesano, L., & Minguez, J. (2015). Continuous detection of the self-initiated walking pre-movement state from EEG correlates without session-to-session recalibration. *Journal of Neural Engineering*, *12*(3), 036007.

Scherl, I., Strom, B., Shang, J. K., Williams, O., Polagye, B. L., & Brunton, S. L. (2020). Robust principal component analysis for modal decomposition of corrupt fluid flows. *Physical Review Fluids*, *5*(5).

Schirrmeister, R. T., Springenberg, J. T., Fiederer, L. D. J., Glasstetter, M., Eggensperger, K., Tangermann, M., Hutter, F., Burgard, W., & Ball, T. (2017). Deep learning with convolutional neural networks for EEG decoding and visualization. *Human Brain Mapping*, *38*(11), 5391–5420.

Schmid, P. J. (2010). Dynamic mode decomposition of numerical and experimental data. *Journal of Fluid Mechanics*, *656*, 5–28.

Schomer, D. L., & Lopes da Silva, F. (2012). *Niedermeyer's electroencephalography: Basic Principles, Clinical Applications, and Related Fields*. Lippincott Williams \& Wilkins, 2012, 6th ed.

Seenivasaharagavan, G. S., Korda, M., Arbabi, H., & Mezić, I. (2021). Mean Subtraction and Mode Selection in Dynamic Mode Decomposition.

Seliya, N., Khoshgoftaar, T. M., & Van Hulse, J. (2009). A study on the relationships of classifier performance metrics. In *Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI*, (pp. 59–66).

Seo, J. H., Tsuda, I., Lee, Y. J., Ikeda, A., Matsuhashi, M., Matsumoto, R., Kikuchi, T., & Kang, H. (2020). Pattern recognition in epileptic EEG signals via dynamic mode decomposition. *Mathematics*, *8*(4), 481.

Shibasaki, H., Barrett, G., Halliday, E., & Halliday, A. M. (1980). Components of the movement-related cortical potential and their scalp topography. *Electroencephalography and Clinical Neurophysiology*, *49*(3-4), 213–226.

Shibasaki, H., & Hallett, M. (2006). What is the Bereitschaftspotential? *Clinical Neurophysiology*, *117*(11), 2341–2356.

Shih, J. J., Krusienski, D. J., & Wolpaw, J. R. (2012). Brain-computer interfaces in medicine. *Mayo Clinic Proceedings*, *87*(3), 268–279.

Shiraishi, Y., Kawahara, Y., Yamashita, O., Fukuma, R., Yamamoto, S., Saitoh, Y., Kishima, H., & Yanagisawa, T. (2020). Neural decoding of electrocorticographic signals using dynamic mode decomposition. *Journal of Neural Engineering*, *17*(3), 036009.

Solaija, M. S. J., Saleem, S., Khurshid, K., Hassan, S. A., & Kamboh, A. M. (2018). Dynamic mode decomposition based epileptic seizure detection from scalp EEG. *IEEE Access*, *6*, 38683–38692.

Song, F., Mei, D., & Li, H. (2010). Feature selection based on linear discriminant analysis. *Proceedings - 2010 International Conference on Intelligent System Design and Engineering Application, ISDEA 2010*, *1*, 746–749.

Sorger, B., & Goebel, R. (2020). Real-time fMRI for brain-computer interfacing. *Handbook of Clinical Neurology*, *168*, 289–302.

Subasi, A., & Gursoy, M. I. (2010). EEG signal classification using PCA, ICA, LDA and support vector machines. *Expert Systems with Applications*, *37*(12), 8659–8666.

Sutton, S., Braren, M., Zubin, J., & John, E. R. (1965). Evoked-potential correlates of stimulus uncertainty. *Science*, *150*(3700), 1187–1188.

Sweeney-Reed, C., Andrade, A., & Nasuto, S. (2004). Empirical mode decomposition of EEG signals for synchronisation. *Proc. of IEEE EMBSS UKRI Postgraduate Conference on Biomedical Engineering and Medical Physics*, (pp. 15–6).

Syam, S. H.-F. (2017). *Developing multi degree of freedom control brain computer interface system for spinal cord injury patients*. Ph.D. thesis, University of Strathclyde. URL https://stax.strath.ac.uk/concern/theses/pz50gw19n

Takeishi, N., Fujii, K., Takeuchi, K., & Kawahara, Y. (2021). Discriminant Dynamic Mode Decomposition for Labeled Spatio-Temporal Data Collections. *ArXiv*.

Tayeb, Z., Fedjaev, J., Ghaboosi, N., Richter, C., Everding, L., Qu, X., Wu, Y., Cheng, G., & Conradt, J. (2019). Validating deep neural networks for online decoding of motor imagery movements from eeg signals. *Sensors (Switzerland)*, *19*(1), 210.

Tu, J. H., Rowley, C. W., Luchtenburg, D. M., Brunton, S. L., & Kutz, J. N. (2014). On dynamic mode decomposition: Theory and applications. *Journal of Computational Dynamics*, *1*(2), 391–421.

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, *10*(5), 988–999.

Vidal, J. J. (1973). Toward direct brain-computer communication. *Annual review of biophysics and bioengineering*, *2*, 157–180.

Villringer, A., & Chance, B. (1997). Non-invasive optical spectroscopy and imaging of human brain function. *Trends in Neurosciences*, *20*(10), 435–442.

Wang, H., Tang, C., Xu, T., Li, T., Xu, L., Yue, H., Chen, P., Li, J., & Bezerianos, A. (2020). An Approach of One-vs-Rest Filter Bank Common Spatial Pattern and Spiking Neural Networks for Multiple Motor Imagery Decoding. *IEEE Access*, *8*, 86850–86861.

Wang, P., Jiang, A., Liu, X., Shang, J., & Zhang, L. (2018a). LSTM-based EEG classification in motor imagery tasks. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *26*(11), 2086–2095.

Wang, Z., Cao, L., Zhang, Z., Gong, X., Sun, Y., & Wang, H. (2018b). Short time Fourier transformation and deep neural networks for motor imagery brain computer interface recognition. *Concurrency and Computation: Practice and Experience*, *30*(23), e4413.

Washizawa, Y., & Hotta, S. (2012). Mahalanobis distance on Grassmann manifold and its application to brain signal processing. *IEEE International Workshop on Machine Learning for Signal Processing, MLSP*, (pp. 23–26).

Whitney, A. W. (1971). A Direct Method of Nonparametric Measurement Selection. *IEEE Transactions on Computers*, *C-20*(9), 1100–1103.

Wu, H., Niu, Y., Li, F., Li, Y., Fu, B., Shi, G., & Dong, M. (2019). A Parallel Multiscale Filter Bank Convolutional Neural Networks for Motor Imagery EEG Classification. *Frontiers in Neuroscience*, *13*, 1275.

Wu, Z., & Huang, N. E. (2009). Ensemble empirical mode decomposition: A noise-assisted data analysis method. *Advances in Adaptive Data Analysis*, *1*(1), 1–41.

Xu, B., Zhang, L., Song, A., Wu, C., Li, W., Zhang, D., Xu, G., Li, H., & Zeng, H. (2019). Wavelet Transform Time-Frequency Image and Convolutional Network-Based Motor Imagery EEG Classification. *IEEE Access*, *7*, 6084–6093.

Yu, X., Chum, P., & Sim, K. B. (2014). Analysis the effect of PCA for feature reduction

in non-stationary EEG based motor imagery of BCI system. *Optik*, *125*(3), 1498–1502.

Zhang, D., Yao, L., Zhang, X., Wang, S., Chen, W., & Boots, R. (2018). Cascade and parallel convolutional recurrent neural networks on EEG-based intention recognition for brain computer interface. In *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, (pp. 1703–1710).

Zhao, X., Zhang, H., Zhu, G., You, F., Kuang, S., & Sun, L. (2019). A Multi-Branch 3D Convolutional Neural Network for EEG-Based Motor Imagery Classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *27*(10), 2164–2177.

Ziehe, A., Laskov, P., Nolte, G., & Müller, K. R. (2004). A fast algorithm for joint diagonalization with non-orthogonal transformations and its application to blind source separation. *Journal of Machine Learning Research*, *5*, 777–800.