Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Engineering Cybernetics

Karoline Malene Nylænder and Victoria Taklo Kenworthy

Motor Imagery-based Brain-Computer Interfaces: Exploring Optimization and Transfer Learning Techniques for Multiclass Classification

Master's thesis in Cybernetics and Robotics
Supervisor: Marta Molinas
Co-supervisor: Luis Alfredo Moctezuma
May 2023

NTNU
Norwegian University of Science and Technology

Karoline Malene Nylænder and Victoria Taklo Kenworthy

Motor Imagery-based Brain-Computer Interfaces: Exploring Optimization and Transfer Learning Techniques for Multiclass Classification

Master's thesis in Cybernetics and Robotics
Supervisor: Marta Molinas
Co-supervisor: Luis Alfredo Moctezuma
May 2023

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Engineering Cybernetics

NTNU
Norwegian University of Science and Technology

Preface

This Master's thesis completes a Master of Technology at the Norwegian University of Science and Technology (NTNU), under the Department of Engineering Cybernetics, in the spring semester of 2023. The idea for the project was proposed by Marta Molinas, who also acted as our supervisor and provided us with relevant literature. The work in this thesis is a continuation of two specialization projects completed in the fall semester of 2022 [1, 2].

Our 2nd supervisor, Luis Alfredo Moctezuma, suggested and provided code for the EEGNet. Other than that, the software implementation was done by ourselves, and methods were found from various sources, through a literature search carried out by us.

Parts of Chapter 2, 3, and 4 are extended and updated versions of what was presented in our specialization projects [1, 2]. The reason for including the same topics was to provide the reader with a comprehensive understanding of our work. The results, discussion, and conclusion presented in Chapter 5 and 6 are our own original work.

The code used in this thesis can be shared upon request to us at vickenwo@gmail.com or karoline_nylaender@hotmail.com.

We had no prior experience with EEG signals, Motor Imagery, Transfer Learning, or Deep learning when starting this project in the fall of 2022. Hopefully, the new skills and knowledge we have gained over the last year are reflected in this master's thesis.

Trondheim, 2023-05-31

Karoline Malene Nylænder and Victoria Taklo Kenworthy

Acknowledgement

We would like to express our gratitude to our supervisor, Marta Molinas, for introducing us to this interesting field, and for providing continuous supervision and support throughout this project. Her enthusiasm and valuable insight were encouraging and inspiring.

A special thanks also to our co-supervisor, Luis Alfredo Moctezuma, who gave us guidance and help when needed, as well as well-appreciated feedback on our thesis.

Finally, we want to thank our friends and family, for their love, support, and encouragement throughout our studies. Particularly thanks to our fellow student and friend, Armon, for keeping us company and entertained during the course of our master's.

Abstract

This thesis focuses on addressing some challenges of Motor Imagery-Based Brain-Computer Interfaces (MI-BCIs), including enhancing Motor Imagery (MI) detection and classification, identifying and handling Brain-Computer Interface (BCI) illiteracy, reducing costs associated with data acquisition and processing, as well as reducing the initialization and calibration time related to BCIs. It proposes viable solutions to improve the overall efficiency, reliability, and user experience of MI-BCI systems.

The thesis explores multiple techniques to tackle these challenges, including MI detection, Transfer Learning (TL), optimization of time windows and frequency bands, channel reduction, analysis of Even-Related Potential (ERP) and Power Spectral Density (PSD), and the use of Random Forest (RF) and Convolutional Neural Network (CNN) classifiers. These techniques were tested on multiple class combinations on four different datasets, namely dataset IV2b, IV2a, NTNU, and SMR.

Through experiments, several solutions to these challenges were identified. BCI illiteracy can be addressed by utilizing CNN, optimizing time windows and frequency bands, or performing channel reduction, as these techniques were particularly efficient for low-performing subjects, enabling them to achieve accuracies above the chance level. Further, a reliable BCI system could be obtained by using RF with Frequency Band Extraction (FBE) for MI detection and CNN for the classification of MI tasks, as these methods demonstrated robustness and achieved high accuracies.

Moreover, the issue of time-consuming initialization and calibration can be resolved by leveraging pre-trained models using TL, as the method was robust and achieved comparable results to the subject-dependent classification. Wearability and usability could be improved through channel reduction, as some channels were proven redundant during this process, however, this process should be done in a subject-specific manner. To reduce the overall cost of the BCI system, TL and channel reduction proved effective, as pre-trained models reduce the need for extensive data acquisition, while fewer channels lead to lower computational costs and fewer processors. In addition, it was found that an analysis of the subject's ERP and PSD signals can be used to predict performance prior to resource-intensive activities, as they were significantly different for low and high-performing subjects.

Overall, this thesis contributes to the field by addressing critical challenges in MI-BCI technology and proposing solutions to enhance its performance and usability. The findings suggest further advancements in the field, bringing the development of practical and efficient MI-BCI systems for real-world applications a step closer to realization.

Sammendrag

Dette arbeidet fokuserer på å håndtere noen utfordinger innen hjerne-datamaskin-grensesnitt (BCI) basert på forestilte bevegelser (MI), kalt MI-BCI, inkludert forbedring av deteksjon og klassifisering av MI, identifisering og håndtering av BCI-inkompetanse, reduksjon av kostnader knyttet til datainnsamling og -behandling, samt redusering av initialiserings- og kalibreringstid knyttet til BCIer. Arbeidet foreslår løsninger for å forbedre den generelle effektiviteten, pålideligheten og brukeropplevelsen til MI-BCI-systemer.

Flere teknikker utforskes for å håndtere disse utfordringene, inkludert MI-deteksjon, overføringslæring (TL), optimalisering av tidsvinduer og frekvensbånd, kanalreduksjon, analyse av hendelsesrelaterte potensialer (ERP) og effektspektraltetthet (PSD), samt bruk av Random Forest (RF) og konvolusjonelle nevrale nettverk (CNN) til klassifisering. Disse teknikkene ble testet på flere klassekombinasjoner på fire ulike datasett, kalt IV2b, IV2a, NTNU og SMR.

Gjennom eksperimenter ble flere løsninger på disse utfordringene identifisert. BCI-inkompetanse kan håndteres ved å bruke CNN, optimalisere tidsvinduer og frekevensbånd eller utføre kanalreduksjon, da disse teknikkene var spesielt effektive for lavt ytende deltakere, og gjorde det mulig for dem å oppnå nøyaktigheter over sjansenivået. Videre kan et pålitelig BCI-system oppnås ved å bruke RF med frekvensbåndekstraksjon (FBE) til MI-deteksjon og CNN til MI-klassifisering, ettersom disse metodene var robuste og oppnådde høye nøyaktigheter.

I tillegg kan problemet med tidkrevende initialisering og kalibrering løses ved å bruke forhåndstrente modeller ved hjelp av TL, da denne metoden var robust og oppnådde sammenlignbare resultater med individavhengig klassifisering. Brukskomfort og brukbarhet kan forbedres ved hjelp av kanalreduksjon, ettersom noen kanaler viste seg overflødige, men denne prosessen må utføres på en individspesifikk måte. For å redusere den overordnedde kostnaden til BCI systemet, viste TL og kanalreduksjon seg effektive, ettersom forhåndstrente modeller reduserte behovet for omfattende datainnsamling, og færre kanaler leder til lavere beregningskostnader og færre prossessorer. I tillegg ble det funnet at analyse av deltakerens ERP og PSD signaler kan brukes til å predikere ytelsen før ressurskrevende aktiviteter utføres, da de var betydelig forskjellige for høyt og lavt ytende deltakere.

Samlet sett bidrar dette arbeidet til feltet ved å addressere kritiske utfordinger innen MI-BCI-teknologi og foreslå løsninger for å forbedre ytelsen og brukbarheten. Funnene antyder videre fremskritt innen feltet, og bringer utviklingen av praktiske og effektive MI-BCI-systemer for anvendelser i virkeligheten ett skritt nærmere realisering.

Contents

List of Figures

List of Tables

List of Abbreviations

BCI Brain-Computer Interface

CL Chance Level

CNN Convolutional Neural Network

DL Deep Learning

DT Decision Tree

DWT Discrete Wavelet Transform

EEG Electroencephalography

ERD Event-Related Desynchronization

ERP Event-Related Potential

ERS Event-Related Synchronization

FBE Frequency Band Extraction

HC Hjorth Complexity

HFD Highuchi Fractal Dimension

HM Hjorth Mobility

IE Instantaneous Energy

IIR Infinite Impulse Response

LOO Leave-One-Out

LSO Leave-Some-Out

MDI Mean Decrease in Impurity

ME Motor Execution

MI Motor Imagery

MI-BCI Motor Imagery-Based Brain-Computer Interface

ML Machine learning

PFD Petrosian Fractal Dimension

PSD Power Spectral Density

RF Random Forest

RMS Root Mean Square

SMR Sensorimotor Rhythms

SNR Signal-to-Noise Ratio

TL Transfer Learning

TE Teager Energy

Chapter 1

Introduction

1.1 Background

In science fiction movies, a common futuristic technology illustrated is a method of controlling or interacting with machines using only the power of the mind. For example, in the movie "Avatar" the protagonist controls his avatar using a Brain-Computer Interface (BCI) that allows him to move, run, and jump in a virtual world by just thinking about it. Comparably, in the TV series "Black Mirror" several episodes consist of characters that enter a virtual reality through a BCI. Although these depictions of *brain-controlled technology* may seem far-fetched, they are based on real advances in neuroscience and BCI technologies.

BCI is just as the name describes, an interface that interacts and creates communication between the human brain and an external device. The human intention is captured through sensors, such as electrodes, that detect brain activity, using techniques like Electroencephalography (EEG) [3]. The measurement is then transformed into a control signal that can be used for a physical application. More formally, the human intention can be acquired through neuro-paradigms, with the purpose of generating control messages for the BCI system. One practical neuro-paradigm commonly used for BCI is Motor Imagery (MI), which involves imagining the movement of a muscle without physically moving it [4]. Capturing the intention with EEG is a non-invasive procedure where electrodes placed on the scalp detect brain activity. The number and locations of the electrode channels can be adapted to the brain activity of interest [3].

Considering that anyone can produce an MI, regardless of their motor abilities, Motor Imagery-Based Brain-Computer Interface (MI-BCI) is an enabling technology that can improve the quality of life for disabled people as well as increase their independence. MI-BCI can also be used to train and rehabilitate motor functions for individuals with neurological disorders, for example, can stroke patients, with training, regain control of their movements over time. Since the MI is translated into a control signal, the possible applications are endless, and can for example be used for gaming, healthcare, and entertainment [3].

Although MI-BCI seems great in theory, being based on measurements from EEG has its challenges. EEG signals are noisy, non-stationary, and non-linear, and the signals are often highly affected by the

person producing them. Some people are also BCI illiterate, where they lack the ability to produce a usable MI [3]. Translating an imagined movement into a response of a device is ergonomic and natural, as it is similar to how we naturally control our environment. However, it requires considerable individual training and calibration time. In addition, the effectiveness and performance of BCI are dependent on intra- and inter-subject factors, where the methods of signal processing must consider the subject-specific differences and MI abilities [5]. Due to the cost related to data acquisition, it is difficult and expensive to acquire enough data for each subject, to get sufficient data to build a robust classifier. Thus, another BCI related problem is the lack of data. For a BCI to be useful, the classification has to be reliable and efficient, as well as able to work for online classification. To compensate for the EEG signal properties, extensive signal processing, and classification make the MI-BCI more costly, in both computational complexity, processing time, and processors, which in turn makes the system less available for the intended user [3].

1.2 Objectives

The purpose of this work is to experiment with various methods to obtain a more reliable, usable, and practical MI-BCI system, and attempt to solve and reduce some of the presented challenges related to such a system. This includes experiments with classification methods, Transfer Learning (TL), channel reduction, and optimization of time window size, interval, and frequency bands for processing. Some state-of-the-art methods are included in the experiments, in addition to exploration and combinations of other methods.

This work can be summarized by the following objectives in terms of its contribution to MI-BCI research:

1. Create a model that can accurately detect the occurrence of an MI.
2. Experiment with inter-subject and inter-dataset TL to reduce calibration time and data acquisition needed for each subject.
3. Find the optimal time window sizes and frequency bands for each subject and dataset to get higher accuracy and reduced computational cost.
4. Explore different channel subsets to create a more usable and practical BCI.
5. Explore the possibility of predicting a subject's performance prior to processing and classification.
6. Explore different classification and signal processing methods.

1.3 Approach

The challenges with MI-BCI were approached by exploring a variety of methods. First, feature analysis and selection were conducted using a Mean Decrease in Impurity (MDI) feature importance method with Random Forest (RF) built with features extracted from Frequency Band Extraction (FBE) decomposed data. The results were used to make the signal processing less complex, as well as remove

redundant features. Then, a binary MI versus No-MI classifier was created using the RF classifier to detect the occurrence of an MI, here No-MI is the rest data.

To address the issue of limited training data and long calibration time related to each dataset and subject, TL was explored in both an inter-subject and inter-dataset manner. To examine the inter-subject performance of TL, two approaches were explored. The first approach was the Leave-One-Out (LOO) approach, where the source model is trained without including the target subject at all. The second approach was the Leave-Some-Out (LSO) approach, where some samples from the target subject were included during the training of the source model. Similarly, for the inter-dataset investigations, the same two TL approaches were used, along with two approaches where only the high-performing subjects were included in the source model.

To enhance the MI ability and performance of each subject and dataset, their optimal time windows and frequency bands were identified. This optimization process was also performed with the aim of reducing the computational cost and processing time related to each subject and dataset. The optimal time window and corresponding frequency band were determined by using a sliding window approach with various window sizes, combined with the individual frequency bands. The performance of each band-window combination was evaluated using the RF classifier, and the combination with the highest performance was deemed the most optimal for the subject or dataset. Channel reduction was explored with the same goal and motivation, and was investigated using the greedy backward elimination algorithm.

Additionally, the possibility of predicting BCI illiteracy and performance prior to classification was explored by examining the signals Event-Related Potential (ERP) and corresponding Power Spectral Density (PSD) for low and high-performing subjects. This was also done to understand the BCI illiteracy and lack of MI ability in some individuals. Finally, Deep Learning (DL) using CNN with EEGNet was examined as both an alternative feature extraction and classification method. The classification process was explored with both subject-dependent and subject-independent approaches, as well as for a variety of class combinations, resulting in the exploration of multiple binary and multiclass classifiers.

1.4 Limitations

The experiments were limited by the amount and quality of publicly available MI datasets. As there are few MI datasets publicly available with sufficient documentation, the TL experiments were performed with limited training data, which may have affected the results. In addition, the channel reduction experiment was also constrained because all the available datasets already contained a small subset of channels.

1.5 Outline

The thesis is structured as follows. Chapter 2 presents the relevant background theory for this thesis, from brain activity and MI to the processing and classification methods used. In Chapter 3 a review

of relevant literature is presented. Chapter 4 gives an overview of the materials utilized, such as the datasets, as well as the methods used for signal pre-processing, classification, experimentation, and performance evaluation. Then, in Chapter 5, the results obtained from the research are presented. Lastly, in Chapter 6 a discussion and conclusion on the obtained results are given, together with suggestions for future work in the field, based on our findings.

Chapter 2

Theoretical Background

This chapter is an extension of the theoretical background given in [1] and [2]. First, the relevant knowledge about the human brain and its underlying processes is presented. Then, some techniques for processing, translating, and utilizing brain signals are presented. Lastly, a transfer learning tool leveraging knowledge from previous tasks for use in a new task is described.

2.1 The Human Brain

The human brain plays an essential part in bodily and mental processes. The brain can be divided into four main parts: *the cerebral*, *the diencephalon*, *the cerebellum*, and *the brainstem*. The cerebral is the largest region of the brain, and it is responsible for handling complex functions, such as information processing, language, and memory. It can be further divided into the left and right hemispheres, each containing four lobes: the *frontal*, *parietal*, *temporal*, and *occipital* [3]. Each lobe has its own specialized functions and their location within the cerebral can be seen in Figure 2.1.

2.1.1 The Motor Cortex

The functions of motor skills and higher-level cognition are associated with the frontal lobe. At the back of the frontal lobe lays the *motor cortex*, which plays a critical part in planning, controlling, and executing voluntary movements. This cortex can be further divided into the regions *primary motor cortex*, *premotor area*, *somatosensory cortex*, and *supplementary motor area*, illustrated in Figure 2.1 [6]. The primary motor cortex is responsible for the control and execution of movements, while the premotor cortex is involved in the planning and preparation. The somatosensory cortex processes the sensory information related to pain, touch, temperature, and proprioception, then sends this information to other parts of the motor cortex, to aid in movement control [3]. The area which coordinates physical movement and tactile perception is called the *sensorimotor cortex*, which incorporates the primary motor cortex and the somatosensory cortex.

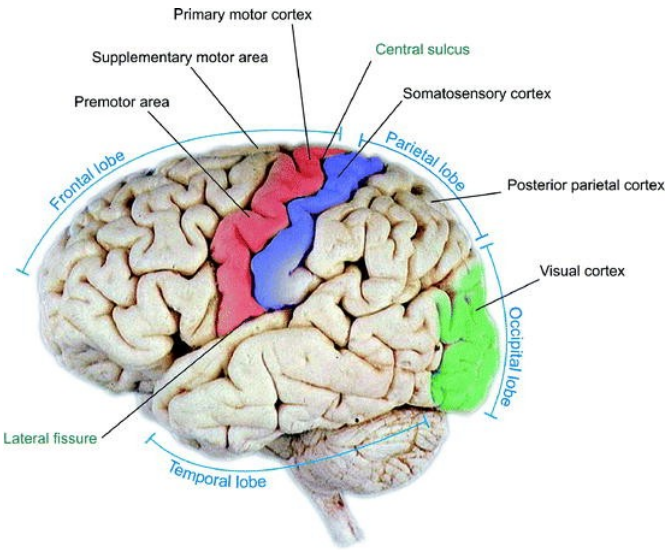


Figure 2.1: Overview of the lobes of the brain's cerebral hemisphere. Reprinted from [3].

2.1.2 Brain Activity

The central nervous system communicates through chemical and electrical signals between neurons. These neurons transmit signals when their membrane potential surpasses a certain threshold, which is triggered by a depolarization process due to the movement of ions in and out of the neuron. The electrical signal transmitted when the threshold is reached, thus activating the neuron, is called an *action potential* [7].

The human brain is an essential part of the central nervous system of the body, composed of billions of interconnected neurons. The action potential generated by these neurons enables the brain to receive, process, and transmit information. In general, the activity in the brain arises from the simultaneous activation of a neuron population, initiating the spread of an electrical signal across the brain and throughout the body. This signal propagation continues within the nervous system until it elicits a response at the intended site [7]. The various areas of the brain control different bodily processes and limbs, as illustrated in Figure 2.2. Thus, by looking at specific parts of the brain one can get an overview of the processes occurring within the brain and body at a given time.

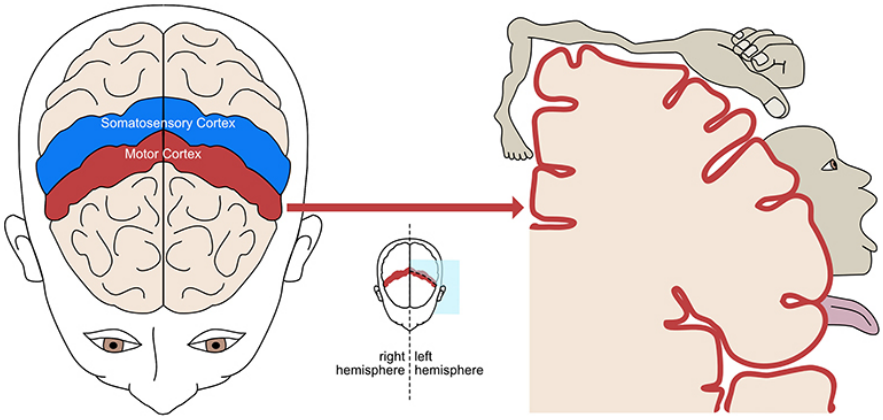


Figure 2.2: The homuncular organization of the brain, demonstrating the mapping of the body's limbs to the different areas of the motor cortex. Reprinted from [8].

2.1.3 Frequency Bands of the Brain

Neurons within the brain generate action potentials through various frequency waves, or rhythms, which can be categorized into five distinct frequency bands. These frequency bands, in order of increasing frequency, are known as Delta (δ), Theta (θ), Alpha (α), Beta (β), and Gamma (γ). Each frequency band has been linked to particular mental states, as shown in Table 2.1, and as the complexity and mental load of a task increases, there is a corresponding increase in frequency [9, 10]. The most interesting and relevant frequencies for motor planning and movement are in the Theta, Alpha, and Beta bands.

It should be noted that the frequency bands are unique to each individual, and there is currently no consensus on the precise frequencies associated with each band [11].

Frequency band	**Frequencies**	**Brain states**
Delta (δ)	0.5 – 4 Hz	Sleep
Theta (θ)	4 – 8 Hz	Inward focus
Alpha (α)	8 – 12 Hz	Passive attention
Beta (β)	12 – 30 Hz	Active, external attention
Gamma (γ)	> 30 Hz	Concentration

Table 2.1: Frequency bands of the brain. Taken from [11].

The changes in the frequency bands can be split into two categories, namely Event-Related Desynchronization (ERD) and Event-Related Synchronization (ERS). ERD is the decrease of oscillatory activity in a specific frequency band, related to attenuation, while ERS is the increase of the oscillatory activity, related to enhancement [3]. The ERD/ERS patterns follow a homuncular organization and are also contralateral. That is, brain activity on one side of the brain invokes movement on the other side of the body [12]. As illustrated in Figure 2.2, the right hand is controlled by the left hemisphere

Mu Band

In addition to the above frequency bands, there is also a specific type of frequency band known as the mu (μ) band, or mu wave. The mu band is a non-sinusoidal oscillation that is primarily used to refer to the Alpha band when it is recorded specifically from the sensorimotor regions of the brain [13]. This band occurs in the frequency range of 8 to 12 Hz and is often studied in relation to sensorimotor processes such as motion planning and execution [14]. The mu band is also related to Sensorimotor Rhythms (SMR), which are oscillations characterized by temporal, spectral, and spatial changes in the mu and Beta bands at the sensorimotor cortex.

During motor planning, an ERD event occurs in the mu band, suppressing the mu activity in the sensorimotor cortex. This suppression is related to the inhibition of motor output during motor planning, preventing actual movement execution. In relation to motor planning, the mu band can be further split into a lower and higher band ranging between 8-10Hz and 10-12Hz, respectively. The lower mu band is associated with more general motor planning, while the higher mu band is more task-specific [3]. Thus the intended movement may be identified from the upper mu band [15].

2.2 Electroenchephalography

The action potential, caused when the neurons transmit a signal, creates a measurable small electrical voltage or current. These voltage changes can then be measured by using EEG, which is a non-invasive technique for measuring the electrical activity of the neurons in the brain. The EEG signals are characterized by being nonlinear and non-stationary in the time-domain [10].

EEG utilizes metal electrodes on the scalp's surface together with a conducting medium to acquire the brain signals in real-time. These electrodes are placed on a pre-determined scheme and each of them measures a separate signal based on their placement [11]. An electrode in combination with a reference electrode creates a *channel*, which represents the change in measured electrical potential. After the channel has acquired the signal, but before being displayed on a computer, the signal is amplified and converted from an analog to a digital signal which can be used as commands or for analysis [16]. By analyzing the characteristics of the EEG signal, including its amplitude and frequency, it is possible to gain insight into the underlying activity taking place within the brain [10].

2.2.1 Electrode Placement

As mentioned, the placement of the electrodes determines which brain activity is measured. The International Federation of Societies for EEG specified a 10/20 electrode placement system, shown in Figure 2.3. This system is based on the relationship between the cerebral cortex's area and an electrode's location. The numbers 10 and 20 indicate the distance between adjacent electrodes, measured in percentage of the skull. Each electrode is given a number and a letter, where the letters F, P, T, O, and C represent brain lobes, respectively Frontal lobe, Parietal lobe, Temporal lobe, Occipital lobe, and Central lobe. Since a central lobe does not exist in the brain, this is only used for identification. The corresponding number refers to the side of the brain where the electrode is placed. Odd numbers, from 1 to 7, refer to the left hemisphere, while the even numbers, from 2 to 8, refer to the electrode position on the right hemisphere. The electrodes placed on the line between the left and right hemisphere, i.e. the center, are marked by the letter z [11, 17]. Other electrode placement schemes include systems such as the 10-10 [18], 10-5 [19], and the Laplace derivations [20], and there is no international standard for EEG measurements.

2.2.2 Topographic Map

To get an overview of the brain activity around each electrode channel, a topographic map can be used. A topographic map for EEG data works similarly to a geographical topographic map and works as a visual representation of the spatial distribution of the EEG measurements. The combination of multiple channels forms an electrode configuration. The electrical activity is represented by colors and contour lines, where the darker the color the higher the intensity. Usually, red is used to indicate positive measurements, while blue indicates negative measurements. In regard to brain signal propagation, the negative measurements can be considered as ERD while the positive measurements are ERS. By examining the topographic maps, one can observe patterns of electrical activity across different regions of the scalp. The brain activity visualized with a topographic map can either be from specific time instances or calculated from the average propagation of the signal [22].

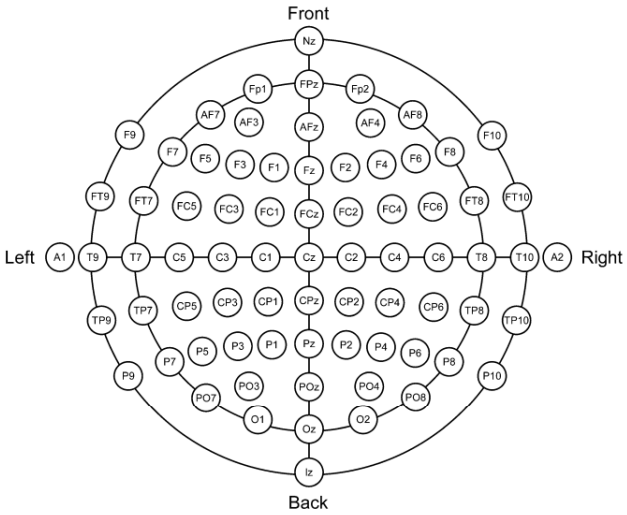


Figure 2.3: The international 10/20 electrode placement system. Reprinted from [21].

2.2.3 Disadvantages with EEG

One of the drawbacks of measuring on the scalp's surface is that there are many layers, such as skin and the skull, from the signals' origin to the electrode. This causes the already small amplitude to weaken. This will also make the signal prone to artifacts, which are unwanted changes and interferences to the signal. Such artifacts may be physiological and stem from the body's processes such as body or muscle movements and eye blinking, or they can be external, such as power-line interference [10]. Thus, the cerebral activity which is intended to be acquired is overwhelmed by other signal sources. However, many of these artifacts have characteristic properties and can be removed through visual inspection or by employing techniques that remove certain amplitudes and frequencies [23].

Another drawback with using EEG measurements is that they are highly subject-dependent. Due to individual differences, such as skull and scalp characteristics, brain complexity and anatomy, and electrode placement, the measurements will be affected by the person creating them [3]. Although this can be useful in for example subject identification [24], it makes it difficult to make a generalized classification model.

2.2.4 Event-Related Potential

ERPs are small voltages that are generated in the brain as a response to a specific stimulus or event. The stimuli can be visual, auditory, or sensory, and the event may be a cue for motor preparation and execution or cognitive events, such as MI [25]. ERPs reflect changes in EEG signals and represent the combined effect of postsynaptic potentials produced when a group of neurons fire synchronously when processing information, such as the stimuli [26]. The ERP can be obtained by extracting and averaging the temporal segments from the EEG signal that is connected to the event or stimuli of interest. It is the amplitude, latency, and topography of the ERPs that are used to assess the underlying mental operations [25].

2.2.5 Power Spectral Density

PSD is a method used to visually interpret the distribution of frequency components in a signal. It represents the relative contribution of each frequency component to the total signal power of a voltage signal [27]. In relation to EEG, PSD is often used to analyze the power of the signal for the individual frequency bands of the brain [28].

2.3 Motor Imagery

MI is the process of imagining a motor movement without any actual movement occurring and can be considered as the preparation of movement, which happens before a Motor Execution (ME). The process of motor representation, which involves intending and preparing for a movement, is generally an unconscious process. However, in MI, this process requires conscious effort and awareness in order to imagine a specific movement [29].

The concept behind MI is that imagining a movement will desynchronize the same areas and frequencies of the brain as executing an actual movement [15]. The brain areas related to MI are mainly the premotor cortex, somatosensory motor cortex, and the primary motor cortex, i.e. the sensorimotor areas. This is due to MI being accompanied by a decrease in SMR power over these areas, thus activating them and causing replicable EEG pattern to be produced [15]. These areas are also activated during a physical movement, but during MI only a mental simulation of planning, execution, and sensory feedback of the movement is occurring, while the actual movement is suppressed. Further, the frequencies related to MI are mostly present in the mu and beta bands, causing these bands to be of significance when differentiating between intended movements. This is because the ERD/ERS patterns are most prominent in these frequency ranges, causing these bands to be suppressed during MI [30].

The most common classes of MI movement are left and right hand, feet, and tongue movements. This is because they produce signal changes that are easy to discriminate from the EEG signals. Their corresponding brain areas are also topographically different and large, i.e. they correspond to different areas of the motor cortex [3]. Figure 2.2 illustrates how a different body party maps to a specific part of the motor cortex and that the mapping is contralateral. For example, since MI is a contralateral process, right-hand MI is most prominent at the electrode placement C3 while electrode C4, on the other side of the skull, has the most prominent signal for left-hand MI [12]. Foot movement is primarily located at the center and is measured best by electrodes located at Cz. A distinction between the left and right foot is considered impossible since their corresponding areas are too close. The tongue, in contrast, is mapped further from the center of the brain and closer to the ears anatomically, as illustrated in Figure 2.2. To measure patterns that can be detected, the corresponding areas have to be large enough so that the specific MI activity can be distinguished from the background EEG [3].

Although the usage of MI sounds great in theory, the practical implementation of this technique is challenging. This is due to the weak, nonlinear, and non-stationary properties of EEG signals, as well as individual differences among subjects. Every individual has unique physiological and physi-

cal characteristics which will affect their produced EEG signals, as well as their ability to produce an MI. Some subjects are in need of extensive training to acquire the skill of MI, while some subjects are still unable to produce an MI, even after training. Another challenge of MI involves the necessity to acquire large amounts of high-quality data from several sessions, for each subject, in order to provide sufficient data for classification training and gain the ability to accurately distinguish between different motor tasks [31].

2.4 Brain-Computer Interfaces

A BCI is a direct communication pathway between the brain and an external device. The signals from the human body are translated into commands for the device. The signals can for example be EEG, Magnetoencephalography (MEG), or Eye-Tracking, and have different levels of invasiveness, [3]. With EEG measurements, one can use MI as commands, which requires no muscular stimulation and has shown great promise when used for rehabilitation and for augmentation of human functions, among other things [32]. The translation of the commands is independent of the human intention, which means that BCIs can be used for countless applications, from rehabilitation, control of robots, and clinical implementations.

BCI systems can be passive, active, or reactive, based on how they use the brain. Passive BCI is used for detecting unintentional cognitive states, for example for detecting drowsiness of a driver. Active BCI systems involve brain activity that is voluntary, and a control or action should be detected when the user intends it. An example of an active BCI system is BCI based on intentional MI. Reactive BCI is used when the user is responding to an external stimulus, for example visually evoked P300[32]. In addition, the system can either be invasive or non-invasive, where EEG-based BCI is an example of a non-invasive system.

A general-purpose BCI system consists of three modules. These are the source module, the signal processing module, and the user application module. The *source module* records, digitizes, and stores the EEG signal, and consists of a data acquisition and a data storage component. The *signal processing module* then converts these acquired brain signals into a signal which can control the application. This module consists of the three stages of signal pre-processing, feature extraction, and feature translation, i.e. classification. The control signals obtained from this module are then used by the *user application module* to drive an application [33].

As the BCI system is a user-dependent system, the best combination of classifier, features, and model parameters should be selected and optimized based on the subject. This is especially important to conquer the problem of *BCI illiteracy*, which occurs when a subject is unable to reach an accuracy above 70% due to the inability to produce distinctive brain waves related to the task, even after extensive training [14].

2.5 Data Pre-Processing

From the introduction of EEG signals in section 2.2 it is known that the EEG signals are non-linear, non-stationary, complex, and prone to artifacts. These characteristics make it difficult to obtain the signal pattern and extract useful information about the MI for translation into a usable command. Thus the Signal-to-Noise Ratio (SNR) is low in raw EEG signals. The goal with data pre-processing is to increase the SNR and reduce artifacts. This is often done by applying filters to the signal. Another part of pre-processing is epoching, where the signal is split into suiting epochs based on the task.

2.5.1 Filters

Filtering involves removing specific parts of the signal, which can be used to reduce what is considered noise and thus increase the SNR. To attenuate part of a signal below or above a certain frequency, one can use a high or low pass filter, respectively.

Infinite Impulse Response

An Infinite Impulse Response (IIR) bandpass filter is a digital filter that combines a high and a low pass filter. It has two stop-frequencies as parameters, one low and one high, and only the parts of the raw signal within this range are able to pass. Thus, the filter output is a signal which solely contains the desired frequencies. This type of filter is useful for EEG as low-frequency noise is often present due to movement of the subject's head and electrode wires, and high-frequency noise is present due to nearby electromagnetic fields and muscle contractions [34].

Notch Filter

A filter that works opposite to the IIR bandpass filter is the notch filter, which attenuates parts of the signal in a given frequency range. A commonly known artifact in a EEG signal is the 50Hz European power-line noise which is prominent in the EEG signal due to the utilized AC power supply. A notch filter with a narrow stopband at the power-line frequency can improve the SNR by almost exclusively removing the power-line frequency [34].

2.6 Signal Decomposition

The EEG signal is a complex composition of various patterns and components. The objective of signal decomposition is to sort and present the signal based on its distinct patterns, while also transforming the signal from one domain to another. The decomposition process works to diminish noise, reduce signal complexity, and simplify the extraction of patterns by splitting the signal into different sub-signals. In order to obtain the distinct signal components, methods such as Discrete Wavelet Transform (DWT) and Frequency Band Decomposition (FBE) can be employed to decompose the data.

Frequency Band Extraction

The five frequency bands of the brain, introduced in section 2.1.3, can be seen as the basic components of the brain signal. A decomposition into these bands will indicate which frequencies are

prominent when the task is performed. In this thesis, the decomposition method that extracts these bands will be named FBE. The method is used for analyzing and decomposing non-linear and non-stationary signals and splitting them into their basic components. These components are extracted by the use of IIR bandpass filters. The choice of which frequency bands to extract can be adapted to the application and the individual subject or dataset and is not necessarily limited to the five brain frequency bands. The number of bands affects the number of features and, consequently, the complexity of the processing method.

Discrete Wavelet Transform

The DWT is a method for signal representation and decomposition. The DWT representation exists in both the time and frequency domains, allowing the extraction of frequencies at which a signal event occurs [35]. With this method, time-frequency analysis can be performed while still in the time domain.

DWT uses a *mother wavelet* as its basis function. A wavelet has a characteristic shape and is a finite, short, and oscillating function, which windows the signal. There exists an infinite set of wavelets and a set of wavelets with the same shape but with different scales is called a wavelet family. Daubechies, Coiflet, Morlet, and biorthogonal wavelets are four examples of wavelet families [36]. The choice of wavelet depends on the characteristics of the signal, and one can adapt the window size to the frequency by scaling the wavelet. The shape of the wavelet corresponds to the quality of the resolution of the features, and the size of the window of which the wavelet is used, i.e. its scale, decides the size of the features which are seen. A wavelet with a wide window will have good frequency resolution but poor time resolution, and vice versa for a wavelet with a short window function [35]. The starting position of the wavelet also affects which features are obtained. To find the optimal mother wavelet for the specific signal, trial and error or optimization of parameters can be performed. The output of the decomposition is a shifted and scaled version of the mother wavelet.

The method works by comparing the signal to the mother wavelet with various scales and positions. The algorithm for the DWT can be seen as a step-wise high- and low-pass filtering of the signal until a desired level is achieved. The mother wavelet can be considered a high-pass filter, and the output is called the detail coefficient. The mirrored version of the wavelet can be considered a low-pass filter, and the output is again high- and low-pass filtered. The output from the low-pass filter at the desired level is called the approximation coefficients [37].

The result of the algorithm is a decomposed signal with high- and low-frequency sub-bands, where the number of bands is decided by a predetermined number of levels. The sub-bands go from high to low frequencies, so sub-band 1 contains the highest frequencies, while sub-band N contains the lowest frequencies of the signal[38].

2.7 Feature Extraction

The technique used to obtain a set of features from a signal is called *feature extraction*. This is a dimensionality and complexity reduction technique commonly used in Machine learning (ML), which

results in an improved representation of the signal characteristics. The objective is to create new features from the existing features in a dataset, thus reducing the number of features present and removing redundant information. The new sets of features should summarize the contained information in the original features [39]. The performance of the classification is dependent on the quality of the features.

The various decomposition methods transform the signal into frequency and time-frequency domains, giving the signal different characteristics. The level of information obtained from a feature is dependent on the domain the feature is extracted from. The features are computed from data extracted from each epoch of the decomposed signals.

2.7.1 Energy Distribution Features

Energy distribution features are, as the name implies, used to characterize and describe how the energy is distributed across the signal.

Teager Energy

The Teager Energy (TE) is a feature that indicates instantaneous variations in both the amplitude and frequency of a signal. Despite the amplitude being sensitive to small and subtle changes, TE is considered a robust feature as it effectively attenuates noise [40]. Equation (2.1) presents the formula for obtaining the TE of a signal. Here the wavelet coefficient in frequency band j at time r is denoted $w_j(r)$, and N is the number of samples in a segment [41, 42].

$$f_j^{TE} = log_{10}\left(\frac{1}{N_j}\sum_{r=1}^{N_j-1}|(w_j(r))^2 - w_j(r-1)\,w_j(r+1)|\right) \tag{2.1}$$

Instantaneous Energy

The Instantaneous Energy (IE) is a feature that reflects the amplitude of a signal. IE provides insights into the energy distribution across the individual frequency bands of the signal[41]. The formula for calculating IE is given in Equation (2.2), with N, j and w as denoted with the same definition as in TE [40].

$$f_j^{IE} = log_{10}\left(\frac{1}{N_j}\sum_{r=1}^{N_j}[w_j(r)]^2\right) \tag{2.2}$$

2.7.2 Fractal Dimension Features

The concept behind fractal dimension features is to exploit fractal geometry to describe a signal. Fractal dimensions represent the complexity detail of a pattern and self-similarity of a signal [43]. A smaller fractal dimension indicates simpler patterns while a higher fractal dimension indicates greater complexity.

Petrosian Fractal Dimension

Petrosian Fractal Dimension (PFD) is a feature that effectively obtains the fractal dimension of a signal. The first step is to binarize the point sequence $y_1, y_2, ..., y_N$ to obtain the binary matrix z_i. This matrix is then used to obtain the number of sign changes present in the point sequence, N_Δ. Lastly, the fractal dimension, PFD, is rapidly computed [44] using Equation (2.3) [40].

$$\text{PFD} = \frac{log_{10}^{N}}{log_{10}^{N} + log_{10}\left(\frac{N}{N+0.4N_\Delta}\right)} \tag{2.3}$$

Here N is the number of samples and N_Δ is the number of sign changes in the binary sequence. N_Δ and the binary matrix, z_i is given by

$$N_\Delta = \sum_{i=1}^{N-2} \left| \frac{z_{i+1} - z_i}{2} \right|, \qquad z_i = \begin{cases} 1, & x_i > mean(y) \\ -1, & x_i \leq mean(y) \end{cases}, i = 1, 2, ..., N$$

Highuchi Fractal Dimension

The Highuchi Fractal Dimension (HFD) feature approximates the mean length of the curve of a time-varying signal in the time domain using k number of samples and estimates its dimension [40]. This is an iterative method and it is useful for handling wave-formed objects. To apply HFD, the algorithm starts by reconstructing the original time sequence, x(1), x(2), ..., x(N), where N is the number of samples in the sequence. The reconstructed time sequence will have the format shown in Equation (2.4), where k represents the interval from a time sequence to the adjacent one and m=1,2,...k is the initial time value of the sequences [44].

$$x_m^k = \left\{ x(m), x(m+k), x(m+2k), ..., x\left(m + \left[\frac{N-m}{k}\right] * k\right) \right\} \tag{2.4}$$

Next, the total average length of the signal is obtained by taking the sum of the average lengths for each time series. This is obtained by using the formula in Equation (2.5).

$$L(k) = \sum_{m=1}^{k} L_m(k) \tag{2.5}$$

where

$$L_m(k) = \left(\sum_{i=1}^{\frac{N-m}{k}} |x(m+i*k) - x(m+(i-1)*k)| \right) \frac{N-1}{\left[\frac{N-m}{k}\right] * k} \tag{2.6}$$

The fractal dimension, HFD, is obtained by using the logarithmic transformation on the total average [44]. The formula is given in Equation (2.7) [40].

$$HFD = \frac{ln[L(k)]}{ln(\frac{1}{k})} \tag{2.7}$$

2.7.3 Statistical Based Features

From the statistical characteristics of the signal, one can compute various statistical-based features.

Root Mean Square

The Root Mean Square (RMS) of a signal is obtained by taking the square root of the arithmetic mean value squared. While the arithmetic mean is used to describe the central tendencies of a data set that contains positive values [45], the RMS is useful to obtain the central tendencies when the data consists of both positive and negative values.

$$RMS(x) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2} \tag{2.8}$$

Variance

Variance (Var) is a statistical measure that represents the overall deviation of the data points from the signals' mean [46]. It determines the spread of the data points and a greater variance value corresponds to a larger deviation. The variance is obtained by using the formula given in Equation (2.9), where N is the number of samples and $\bar{x}$ is the mean.

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N} \tag{2.9}$$

Skewness

The Skewness (Skew) refers to the third standardized moment of the signal [47]. The skewness value indicates how asymmetric the signal is compared to a Gaussian distribution and indicates the direction of outliers. Skewness can have both a negative, positive, or zero value. If a signal is normally distributed it is given zero skewness as it is symmetrical [48].

Kurtosis

Kurtosis refers to the fourth standardized moment of a signal, which is a measure of the amplitudes degree of deviation from a Gaussian distribution [49]. Essentially, it is a measure of the peakedness of a signal, where a high kurtosis value indicates a more peaked distribution. A large value indicates that the signal contains significant deviations or outliers [47].

Zero Crossing

A zero-crossing (Cross) feature is determined by obtaining the number of times a signal changes its sign, i.e. crosses the zero-axis. A zero-crossing is detected by comparing the signs of the previous and current values.

Hjorth Mobility

Hjorth Mobility (HM) is a time domain feature that gives a measurement of the central frequency of the signal. The mobility is defined as the square root of the ratio between the variances of the first derivative and the amplitude. HM indicates the number of sign changes, i.e. the frequency of zero-crossings, in the signal [50]. The formula used to obtain the HM is given in Equation (2.10) [51].

$$Mobility(x) = \sqrt{\frac{\sigma^2(\dot{x})}{\sigma^2(x)}} \tag{2.10}$$

Hjorth Complexity

Hjorth Complexity (HC) parameter is a time domain feature that gives a measurement of the frequency spread of the signal. This parameter is a measure of similarity between the signal and a pure sine wave and gives an estimate of the signal bandwidth [50]. It is defined as the ratio between the mobility of the first derivative of the signal and the mobility of the original signal. To obtain the HC parameter, the formula given in Equation (2.11) is utilized [51].

$$Complexity(x) = \frac{Mobility(\dot{x})}{Mobility(x)} \tag{2.11}$$

2.8 Classification

Classification is a categorization process where the goal is to find patterns in a dataset that can be used to make a prediction of which class some future data belongs to. The classifier performance depends on the quality of the training data combined with the choice of a classifier. Two separate classification approaches are applied. These are the ML approach and the DL approach.

When using ML, the features extracted from the signal are used as input to the classifier. Thus, this approach requires extensive signal processing and feature engineering. On the other hand, usually, DL approaches can automatically extract spatiotemporal features from the signal, thus eliminating large parts of the data processing [14].

There are three types of classification methods: supervised learning, unsupervised learning, and reinforcement learning. Supervised learning trains the classifier with known labels, thus these classifiers use label pairs to learn the mapping between input and output[52]. Unsupervised learning, however, performs classification without available labels. Reinforcement learning is trained by interacting with its environment and is given rewards [53].

2.8.1 Machine Learning

ML is a form of artificial intelligence that uses automated algorithms to learn and identify structures of the data, through statistical techniques. It uses data-driven approaches, which means that it utilizes data to learn and make predictions [53]. One commonly used ML classifier is the RF which is based on decision trees.

Decision Trees

A Decision Tree (DT) is a simple supervised classification algorithm. The algorithm is based on the construction of tree-like structures by recursively partitioning the available features through binary questions. As the tree's depth increases, i.e. as more feature splits occur, the decision rule grows in

complexity and correspondingly, the model fits better to the training data. There exist various techniques for determining the initial feature to partition and in which sequence, as well as the optimal depth of the tree [53].

A problem with DTs is that they are prone to overfitting and have poor generalization. *Overfitting* occurs when the model has achieved a perfect fit to the training data, but was not able to obtain a generalized fit when working for unseen data. It is considered a weak learner, and its performance is at times unsatisfying. Therefore, it is more usual to use classifiers that are decision tree-based and build stronger learners from multiple weak learners. One such classifier is the RF.

In addition, tree-based methods offer a straightforward approach to obtaining feature importance by averaging the importance values from each tree in the ensemble [54]. A commonly used method is the MDI method which calculates the importance scores based on the feature's contribution to the decision-making and split-criterions in the trees [55]. Essentially, the features are assigned an importance value based on how it contributes to the homogeneity of classes in the nodes of the trees [56].

Random Forest

RF is a supervised classification algorithm that consists of a combination of various decision trees. RF is a bagging ensemble classifier, which means that it combines several weak learners in parallel to make a strong learner. DTs will create *decision rules* based on the best features and feature values, during training. Here, the best feature will be at the root of the tree, i.e. the first node. Each tree in the forest will predict a class, and the final classification will be based on majority voting [57].

The idea behind this learning algorithm is that several learning models combined will produce a more accurate prediction [57]. RF selects the best features from a random subset of features, rather than choosing them while splitting a node. Thus, each DT is trained with different subsets of the dataset, resulting in different classifiers. This also introduces more randomness to the model, which again adds diversity to the classifiers, making it less prone to overfitting and creating a better model [53]. RF also performs well in dealing with samples where there is a large class imbalance, due to its concentration on local features [58], as well as on high-dimensional data. Additionally, the algorithm efficiency is not affected by missing values [59].

2.8.2 Deep Learning

DL is a subset of machine learning, which uses a specific type of ML algorithm called artificial neural networks (ANN). DL is inspired by the learning processes of the human brain and uses the neurons to learn the structure of the data. This makes the relation between each input nonlinear and more complex. Unlike ML, the classifier does not require signal processing and feature extraction to work. A disadvantage with DL is that it requires a large amount of data to learn, and is computationally complex, which results in a long training time [60]. DL algorithms are generally more difficult to interpret compared to ML and work as a black box that can be trained using different hyperparameters.

Convolutional Neural Network

Convolutional Neural Network (CNN) is a type of DL algorithm and can be considered as an adapted fully connected neural network. In a fully connected neural network, every neuron in each layer is connected to every neuron in the previous layer, whereas in CNN, every neuron in one layer is only connected to a small subset of neurons in the previous layer. The neurons in the layers within the network are organized in three dimensions. These are the spatial dimensionality of the input, which is the width and the height, as well as the depth[60]. The weight between each neuron in every layer is updated using a backpropagation approach.

A CNN consists of three types of layers: convolutional, pooling, and fully-connected. The first two layers of the network, convolution, and pooling, are used to perform feature extraction, whereas the last layer, fully-connected, performs classification, i.e. maps the features to an output [61].

The *convolutional layer* is an essential component of a CNN, as it performs feature extraction. The layer receives an input tensor from the input layer and transforms it to a feature map by the use of learnable *kernels*. This is done by sliding the filter across the spatial dimensionality of the input, and the network learns kernels when a specific feature at a given spatial position is detected [60]. The convolutional layer also reduces the model complexity by optimizing its output, through three hyperparameters; the *stride*, the *depth*, and adding *zero-padding*. The stride is the distance between two successive kernels, and a greater stride will reduce overlapping and generate an output of lower spatial dimensions, thus achieving a downsampled feature map [60, 61]. The depth should be chosen such that the computational complexity is minimized while still retaining significant pattern recognition capabilities [60]. Zero-padding is necessary to ensure that the height and width of the produced feature map are the same as the input tensor.

The purpose of the *pooling layer* is to further reduce the model dimensionality, thus reducing computational complexity and the number of parameters. *Max pooling* is the most popular pooling operation, which only outputs the maximum value of each extracted patch from the input feature maps[61].

The output feature maps of the previous layers are transformed into a one-dimensional array. Then, the *fully-connected layers* maps the features to the networks' final output, usually the probabilities for each classification class [61].

CNNs have shown promising results in pattern recognition tasks, such as computer vision and speech recognition[62]. Since CNN mimics the structure of the human brain, it is assumed that it worked well with pattern recognition of EEG signals. A limitation is that building a CNN model requires a lot of data, and an EEG measurement is limited both in time and number of sessions.

2.9 Transfer Learning

TL is a technique that exploits knowledge from a different, but related, task to enhance the learning process of a new task. Instead of starting the learning process from scratch, TL allows the model to

benefit from the previously learned patterns, features, and presentations of a source model. It can, for example, be to use a pre-trained model from a few subjects to classify data from a new, unseen subject. More formally, the new task is in the target domain, D_T, while the previous knowledge is in the source domain, D_S. Then, knowledge obtained from D_S can be used to improve the learning ability of the target predictive function and improve the performance and generalizability of the predictive model [12].

In some TL approaches, the knowledge can be used directly on the new subject, which is the case for instance knowledge TL, while in other cases, such as the feature representation and model parameters approach, the knowledge needs to be modified and adapted to the new domain or subject [12].

TL can be used to improve three important problems related to MI and EEG classification. The first is the problem of differences across subjects/sessions, which is caused by variations in feature distribution occurring in the EEG signal across subjects or sessions. This can result in reduced classification performance. The second problem is related to the small sample size, which is due to the restricted availability of subjects and experimental environments. A lack of samples may lead to overfitting or other issues, leading to reduced performance. Lastly, TL can assist in the problem of time-consuming calibration needed when a new subject should access and use the BCI system. By using a pre-trained model together with a transfer learning approach, this calibration time can be reduced and the efficiency of the BCI system can be improved [12].

TL has gained more recognition and attention in the last few years, primarily due to the emergence of pre-trained models and an increase in data availability. Due to its flexibility, transfer learning has been used in numerous fields and real-world applications, including computer vision, natural language processing, healthcare, robotics, and finance [63]. However, the extent of research and usage in each field varies based on the availability of data containing similar tasks or domains.

Chapter 3

Literature Review

This chapter presents an extended review, building upon the literature reviews presented in [1] and [2]. It gives an overview of the state-of-the-art within decomposition and classification methods in MI-BCIs, TL approaches, and optimization methods. Additionally, it provides an overview of the previous performance of the utilized datasets.

3.1 State-of-the-Art in MI Classification

Over the past decade, the use of MI in controlling BCIs has frequently gained more recognition, and several studies of translating these MI signals into meaningful commands have been conducted [64]. This requires the selection of decomposition methods, feature engineering, and classifiers.

Many methods of decomposition have been researched and explored in the study of MI classification. The authors of [65] compared the methods of Empirical Mode Decomposition (EMD) and DWT on several subjects. It was concluded that DWT was superior to EMD with an accuracy increase ranging from $4.7 - 22.2\%$. The authors state that this is due to information lost in EMD, which is present in detail 1 (D1) generated by DWT, and the features extracted from this detail are highly affecting the classification performance [65].

In order to achieve satisfactory performance, the features must be chosen and adapted based on the use case. The features can be extracted from various domains, including the time domain and statistical features, features from the frequency domain, nonlinear and complexity-based features, and geometrical-based features. The authors of [66] explored a total of 54 features from the domains and used feature selection to determine the most significant ones. The selection of features was based on their contribution to the classification performance. The article concludes that the best performance is achieved when a combination of the feature domains is used [66].

Classification plays a critical role in the translation of signals into commands for use in BCI systems. Lotte et al. [64] made a comprehensive review of the state-of-the-art classification approaches for EEG-based BCIs over the span of ten years, focusing on the methods employed between 2007 and 2017. The authors evaluated the performance of various algorithms and discuss some advantages and disadvantages for the different approaches, in order to identify the most efficient approaches.

Lotte et al. [64] found that the most used classifiers in recent years have been ML algorithms such as Support-Vector Machines (SVM) and Linear Discriminant Analysis (LDA), and DL approaches such as Convolutional Neural Networks (CNNs). These classification approaches have increased in performance in the last few years, with recent studies achieving a performance above 90% for binary classification and around 80% for multiclass classification. Here, the most used MI classes have been right-hand and left-hand, followed by feet, tongue, and movement of imaginary objects [64].

Among these state-of-the-art classifiers, the DL approaches are frequently showing promising potential, partially due to their effectiveness in learning features and classifies simultaneously, thus eliminating the need for a separate feature extraction process [64]. CNNs have also shown promising results when used in an adaptive manner [67], in addition to tackling the problem of BCI illiteracy [14, 68]. According to [14], using DL approaches will increase the accuracy as compared to standard ML approaches. The study concluded that using CNN would increase the accuracy for each subject within the range 2.37% – 28.28%. This increase was more significant in subjects with initially poor performance, thus simultaneously tackling the BCI illiteracy problem and the performance challenge [14]. However, these networks may need to be quite shallow as deeper networks require a significant amount of training data, which is rarely present in BCIs based on MI [64].

In addition to the state-of-the-art classifiers, Lotte et al. [64] also suggest that the Random Forest (RF) classifier has been shown to perform well in BCI applications, and has been used in several studies for feature selection and classification. This classifier has shown great results in terms of interpretability, computational efficiency, and accuracy, as well as performing well when used on both small and large samples [64]. A study by Sadiq et al [66], has also found that RF, combined with feature selection, can achieve a performance increase of 26.7% as compared to other literature, and it can be used to create a more flexible development of BCI systems, both for subject-independent and subject-dependent models. However, it is pointed out that further research is necessary in order to explore the potential of RF in EEG-based BCIs and be able to compare its performance with other classification algorithms [64].

Lastly, Lotte et al. [64] also discuss the challenges associated with EEG-based BCI systems, including the high degree of inter-subject variability, the limited amount of publicly available EEG datasets, and the calibration time of the system associated with each individual user. They suggest that TL approaches should be developed further in future research to improve classification performance and reduce the need for individual calibration [64].

Overall, there are currently many state-of-the-art approaches, which makes it difficult to conclude on a superior one. The ideal method to choose is highly dependent on the use case, thus this metastudy does not hold the ground truth [64]. There is still a need to seek new methods and further develop and adapt the existing ones. Therefore, this thesis also explores methods that were not considered in the previous review. As decomposition methods, DWT is further explored in this thesis as it has proven the superior decomposition method over EMD [65, 69], and the method of FBE presented in [1] and [2] is also further explored due to its promising results and simplicity. For classification, RF and CNN

are further explored due to their promising results, as well as the necessity to further research their potential. Notably, upon analyzing the findings in [1] and [2], it was observed that Gradient Boosting (GB) exhibited a level of performance comparable to that of RF. Consequently, the decision was made to exclude GB from the current thesis.

3.2 Transfer Learning Approaches

In the field of BCI, the process of MI-EEG decoding contains many challenges. Limited amount of training data, inter-subject and inter-session variability, as well as long calibration times, are among these. In recent years, several TL approaches have been researched in order to undertake these challenges. According to [12], TL addresses these challenges by leveraging pre-trained models and data from other domains and tasks and applying these to a new subject.

TL can be utilized in various approaches, where the three main approaches are based on instance knowledge (ITL), feature representation (FTL), and model parameters (MTL). Which approach to use must be chosen and adapted based on the use case and data availability [63]. The authors of [12] compared multiple TL solutions, using various methods of ITL, FTL, and MTL on various datasets, classes, and patterns. Each method yielded an accuracy ranging between $62.50-89.21\%$, $61.00-92.00\%$, and $69.71-99.60\%$, respectively[12]. For MI classification the highest obtained result was 89.3%, which was achieved when using SVM with Fuzzy TL based on generalized hidden-mapping Ridge Regression, which is an FTL approach [12, 70]. Thus, all the explored approaches yield satisfying results which are mainly above the chance level of 60%, deeming TL an effective and robust technique for classification. However, many of these TL solutions are specific to their given applications, and cannot be generalized for use in different applications [64].

Research has suggested that TL can be applied in order to increase the performance of an unseen subject, thus also reducing or removing the calibration time of a BCI [68]. In [67] an adaptive TL approach utilizing MTL was tested with a deep CNN to create a subject-independent model and compared it to subject-specific approaches, where the model is created using solely data from one subject. Their results implied that TL could increase the performance by 32.50% in accuracy. However, the deeper layers of the CNN are prone to inter-subject variability, but this issue could be improved by fine-tuning the parameters with data from the target subject [67].

In conclusion, TL has shown great promise in increasing subject performances and improving the challenge of long calibration time and lack of data, by utilizing pre-trained models [12, 67]. Despite this, the field of TL still requires extensive research in order to find new and improved solutions. Due to the limited amount of data acquired from each subject, this thesis looks into TL to investigate if a subject-independent model may be used on an unseen subject to improve the performance and investigate if the calibration time can be reduced by comparing the TL performance and the subject-dependent performance.

3.3 Sliding Window and EEG Frequency Band Optimization

In many previous studies, a fixed time window has been used for all subjects. This time window was usually between 0.5 seconds to 2.5 seconds after the MI cue. However, in an article by Song and Epps, in [71], it was brought to light that some subjects had a time latency causing the MI-patterns to be more present in the post-imagination period between 3.5 to 4 seconds. This indicated that the use of a more subject-specific window would increase the performance of the classifier, which would also counteract the challenges of subject-specific MI factors [71].

In relation to this, the authors of [30], Malan and Sharma, hypothesized that the reason for subject-dependent MI capabilities might be due to the fact that the timing of ERD/ERS is inconsistent between subjects. Thus, a fixed time window would lead to suboptimal MI classification performance. The authors of [30], thus explored different time windows and how they affect the performance of the classifier, in a subject-dependent manner. The authors segmented the EEG signal into multiple time windows using a sliding time window approach with a width of 2 seconds and a 0.5 time shift to find the appropriate time window for each subject. The results from this approach showed that in the time windows, -0.5 s to 1,5 s, 0.5 s to 2.5 s, and 2 s to 4 s the features indicated the presence of ERD/ERS. In the other time windows, the presence was not as significant.

In addition to the time window optimization, Malan and Sharma [30] simultaneously optimized the EEG frequency bands. This approach was motivated by previous studies which suggested that optimizing the frequency bands could improve the accuracy of MI classification. The frequency bands they explored were 4-8Hz, 8-16Hz, and 16-32Hz, and it was concluded that the frequencies of 8-16Hz and 16-32Hz were the most significant ones.

Malan and Sharma [30] proposed and explored an optimization approach of combining sliding window and frequency band optimization on datasets IV2a and IV2b from BCI competition IV and dataset IIIa from BCI competition III. The exploration was conducted by comparing three different CSP feature extraction algorithms with and without shifted time windows and optimized frequency bands, using SVM as the classifier. This method achieved an accuracy of 82.1%, 84.5%, and 91.7%, respectively for the datasets. This indicated that their proposed method achieved better classification performance than other existing methods, highlighting the importance of optimizing both the time window and frequency band for improving the classification of EEG signals [30]. Additionally, their experiments resulted in a pre-processing method that was adapting and identifying the individual subject's ability to produce an MI [30]. However, this method was only tested for various binary classifications, including Left-Hand vs. Right-Hand, Left-Hand vs Foot, and Right-Hand vs. Foot classification. In addition to binary classification, this thesis explores a similar method for a multiclass dataset.

A similar experiment was conducted by Riascos Salas [72]. The author used an FBCSP algorithm to explore time window sizes of 100, 200, 400, 600, 800, 1000, and 2000ms, and EEG frequency bands of Alpha (8-12Hz), low Beta (12-16Hz), middle Beta (16-24Hz), high Beta (24-30Hz), and whole Beta (12-30Hz), to identify the optimal combination for each subject. The results indicated that the most significant frequencies were usually around 7-8Hz, 12-14Hz, and 18-24Hz, and the optimal window

sizes were around 800, 1000, and 2000ms, thus the largest sizes, yielding an accuracy of 76%, 78%, and 91%, respectively.

As seen, an optimization process of the time window and frequency bands has shown great promise in increasing the classification performance [30, 72]. Therefore, this thesis explores a similar approach, but with some changes to the sliding window approach and which frequency bands are investigated.

3.4 Overview of the Published Results With the Used Datasets

In this thesis, various techniques are tested on four publicly available datasets. These are dataset IV2b and dataset IV2a from BCI competition IV [73], dataset NTNU created in relation to a previous master at NTNU [69], and dataset SMR which was acquired by Graz University [74]. As IV2b and IV2a were created in relation to a BCI competition, they are well known and have been experimented with by many researchers in the field, while dataset NTNU and dataset SMR are less known. Some previously tested techniques and the datasets' resulting performances are shown in Table 3.1.

3.4.1 Overview of Published Results With the Dataset IV2b

As seen from Table 3.1 several methods have been tested on dataset IV2b, including RF+FBE, RF+DWT, LDA with and without sliding window optimization, SVM+FBCSP with and without band-window optimization, and CNN+SAE. Among these methods, SVM+FBSCP with band-window optimization performed the best, with an accuracy of 84.50%, followed by SVM+FBCSP with shifted time window achieving an accuracy of 81.30% [30]. This result shows the efficiency of using frequency domains and spatial filters as well as optimization for the classification of dataset IV2b. The SVM+FBSCP without optimization, RF+FBE, and CNN+SAE methods also showed promising results for dataset IV2b, with accuracies of 78.20%, 78.07% and 78.00%, respectively [30, 1, 76]. Thus, optimization techniques, RF+FBE and CNN are further explored for this dataset.

3.4.2 Overview of Published Results with the Dataset IV2a

Many methods for classification have been tested on dataset IV2a, both for binary classification of LH vs. RH and RH vs. F, as well as multiclass classification of LH, RH, F, and T, as seen from Table 3.1. For binary classification, SVM+CSP has been tested with and without optimization. For both binary classification types, i.e. LH vs. RH and RH vs. F, the method which both optimizes the frequency bands and time window achieves the best accuracy, yielding a performance of 82.13% and 85.70%, respectively [30]. This proves the importance of optimization for binary classification and is thus further explored in this thesis. For multiclass classification the methods of FBCSP and LDA+CSP+ITL are explored, yielding performances of kappa=0.569 and an accuracy of 77.00%, respectively [75, 77]. Due to the good performance of the TL approach, i.e. ITL, this is further explored in this thesis.

3.4.3 Overview of Published Results With the NTNU Dataset

Multiple methods of classification have also been tested for dataset NTNU, for both binary classification of LH vs. RH, and three-class classification of LH, RH, and F, as seen in table 3.1. For binary classification this includes RF+FBE, RF+DWT, CSP+SVM, and CNN, where CSP+SVM achieved the best

Dataset	Classes	Method	Performance	Reference
IV2b	LH, RH	RF+FBE	78.07%	[1, 2]
	LH, RH	RF+DWT	77.25%	[1, 2]
	LH, RH	GB+FBE	76.17%	[1, 2]
	LH, RH	GB+DWT	75.56%	[1, 2]
	LH, RH	Naive Bayes Parzen Window + FBCSP	kappa=0.60	[75]
	LH, RH	LDA+Window Optimization+CSSD	kappa=0.58	[73]
	LH, RH	LDA or SVM + 1-s Sliding Window + NTSPP	kappa=0.43	[73]
	LH, RH,	SVM+FBCSP	78.20%	[30]
	LH, RH	SVM+Shifted Time Window+FBCSP	81.30%	[30]
	LH, RH	SVM+Band-Window Optimization+FBCSP	84.50%	[30]
	LH, RH	CNN+SAE	78.00%	[76]
IV2a	LH, RH	SVM+CSP	72.04%	[30]
	LH, RH	SVM+Shifted Time Window+CSP	74.93%	[30]
	LH, RH	SVM+Band-Window Optimization+CSP	82.13%	[30]
	RH, F	SVM+CSP	71.20%	[30]
	RH, F	SVM+Shifted Time Window+CSP	77.70%	[30]
	RH, F	SVM+Band-Window Optimization+CSP	85.70%	[30]
	LH, RH, F, T	FBCSP	kappa=0.569	[75]
	LH, RH, F, T	LDA+CSP+ITL	77.00%	[77]
NTNU	LH, RH	RF+FBE	54.12%	[1, 2]
	LH, RH	RF+DWT	52.22%	[1, 2]
	LH, RH	GB+FBE	53.69%	[1, 2]
	LH, RH	GB+DWT	51.09%	[1, 2]
	LH, RH	SVM+CSP	$63.02 \pm 16.51\%$	[69]
	LH, RH	CNN	$60.63 \pm 19.58\%$	[69]
	LH, RH, F	SVM+CSP	$54.90 \pm 19.82\%$	[69]
	LH, RH, F	SVM+Cov+Tangent Space	63.75%	[69]
	LH, RH, F	LDA+Cov+FGDA+R-CSP	75.00%	[69]
SMR	RH, F	RF+DFT	$66.82 \pm 14.63\%$	[74]
	RH, F	RF+CSP	$79.30 \pm 12.56\%$	[74]
	RH, F	sLDA+CSP	$77.15 \pm 12.42\%$	[74]

Table 3.1: Overview of previous classification and processing methods and their results for the used datasets; IV2b, IV2a, NTNU, and SMR. With classes left-hand (LH), right-hand (RH), Feet (F), and Tongue (T). The *Performance* is either given in accuracy, with or without standard deviation, or kappa value.

performance, with a mean accuracy of 63.02%, followed closely by CNN which yielded a performance of 60.63% [69, 1]. The multiclass classification was performed by using CSP+SVM, SVM+Cov+Tangent Space, and LDA+Cov+FGDA+R-CSP, where the latter clearly outperformed the two other methods by achieving an accuracy of 75.00% [69].

3.4.4 Overview of Published Results With the Dataset SMR

Three methods for classifying dataset SMR were tested by the researchers at Graz University [74]. These are RF+DFT, RF+CSP, and sLDA+CSP, which yielded a mean performance accuracy of 66.82%, 79.30% and 77.15% [74]. As classification with RF yields the best performance, further exploration of this classifier is done in this thesis.

Chapter 4

Material and Methods

The purpose of this chapter is to give an overview of the materials and methods utilized in this thesis. First, the structure and content of the datasets are described together with their corresponding processing parameters. Then, the algorithm pipeline for MI classification is illustrated, followed by a description of the methods used to implement its various steps. Next, the experimental parameters and techniques are presented. Lastly, the methods used for the evaluation of the results are described.

4.1 The Datasets

The experiments are conducted on four different MI datasets. The datasets consist of datasets 2a and 2b from BCI competition IV [73], dataset SMR from Graz University [74], and dataset NTNU created by two master students from NTNU [69]. These will be referred to as respectively datasets IV2a, IV2b, SMR, and NTNU.

Datasets IV2b and SMR consist of two classes, while datasets NTNU and IV2a consist of three and four classes, respectively. For datasets IV2a, IV2b, and SMR, the MI recordings were conducted both with and without feedback, thus each subject is regarded as two separate subjects, even though they are technically the same. The recordings without feedback are indicated with a 'T', and the recordings with feedback are indicated with an 'E'. The decision to split each subject into two separate subjects was made to account for potential differences in performance due to the presence or absence of feedback.

Dataset IV2b

Dataset IV2b is a binary MI dataset consisting of left-hand and right-hand MI recorded from 9 subjects over the course of 5 sessions. For the first 3 sessions, the subjects went through 120 trials without feedback, and for the last 2 sessions, there were 160 trials, performed with feedback. Thus, each subject is split into two separate subjects, 'T' and 'E', and experiments are conducted on 18 subjects. Equal amounts of trials containing right-hand and left-hand MI were acquired in each session, but some trials contain artifacts and are removed [78].

The dataset was recorded by the use of 3 bipolar EEG electrodes and a sampling frequency of 250Hz. The electrodes utilized were C3, Cz, and C4, placed using the 10/20 system, as shown in Figure 4.1.

Each channel is filtered with a bandpass filter of 0.5-100Hz and a 50Hz notch filter.

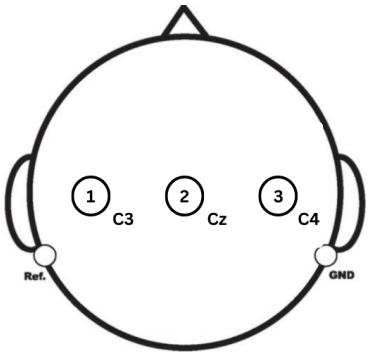


Figure 4.1: Electrode placement for dataset IV2b. Based on 10/20 electrode scheme in Figure 2.3.

The trial epochs have the structure shown in Figure 4.2. Each trial starts with a fixation cross which lasts for 3 seconds. Then a visual cue lasting for 1.25 seconds appears on the screen to alert the subject that the imagery task should be completed. As the cue appears, the imagery period of 4 seconds starts where the subject should evoke the corresponding hand MI and perform the task. After the imagery period, there is a pause before the next trial starts and a new cue appears. Only the *imagery period* will be used as task data for classification, while the pause will be used to classify the rest data. For pre-processing and extraction of trial data, the parameter values given in Table 4.1 are generally used unless specified otherwise.

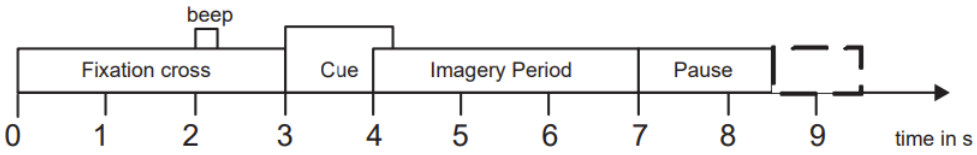


Figure 4.2: Timing scheme of one trial from dataset IV2b. Reprinted from [78].

Parameter	**Value**
Number of subjects	18: 9T and 9E
MI classes	LH, RH
Sampling frequency	250 Hz
Task window interval	3-7 sec
Task number of datapoints	1000 samples
Rest window interval	7-8.5 sec
Rest number of datapoints	375 samples
Frequency bands	Theta, mu_{low}, mu_{high}, Beta
Electrode channels	C3,Cz, C4

Table 4.1: Predefined data extraction parameters for one trial in dataset IV2b.

In Figure 4.3 the raw EEG data with each MI task evoked is illustrated together with the raw EEG data of the resting state. This is obtained from the first trial of each class in channel C3 for subject B1T.

As observed, by the use of only visual inspection it is difficult to distinguish an MI task being evoked from when it is not, and also identify which MI task is being performed. Thus there is a need for other inspection methods.

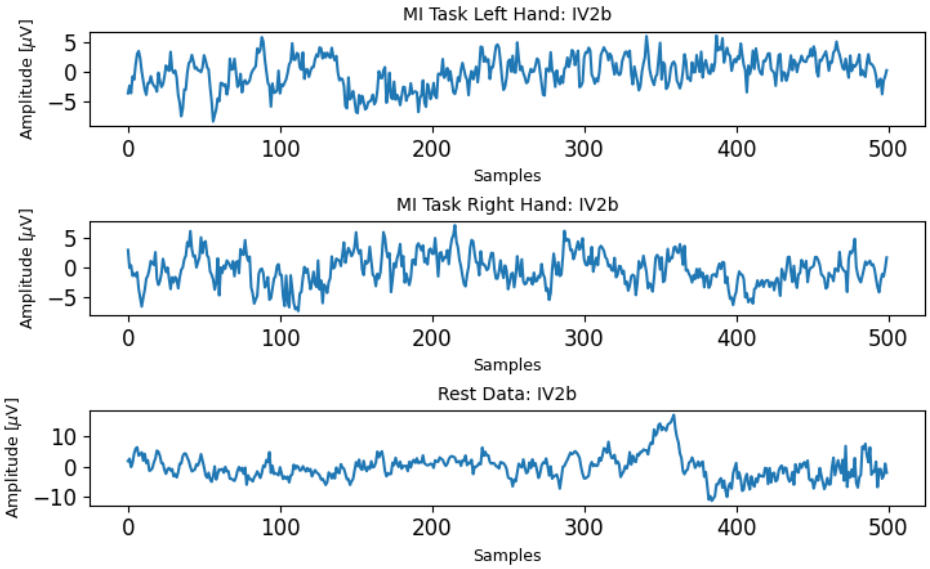
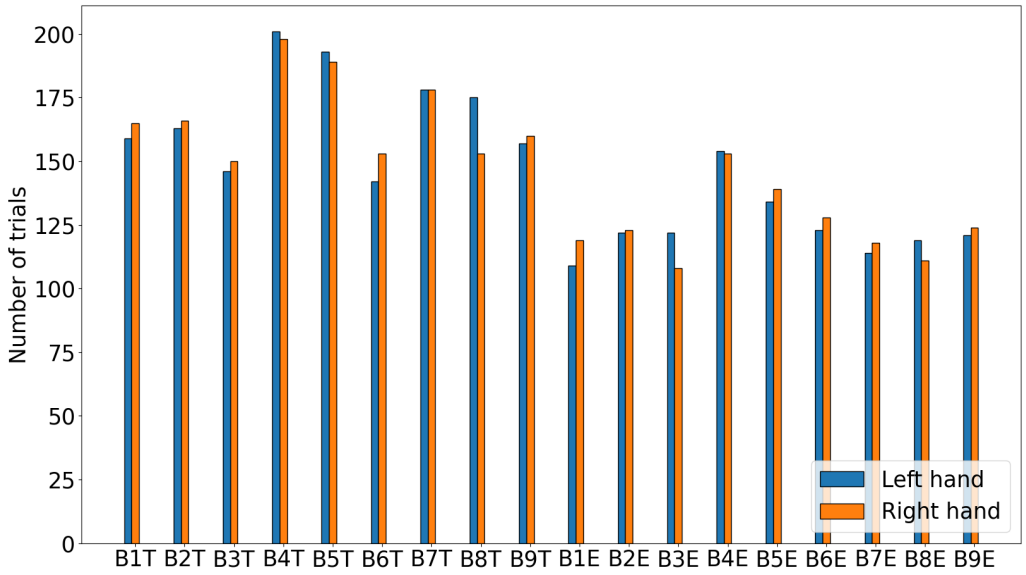


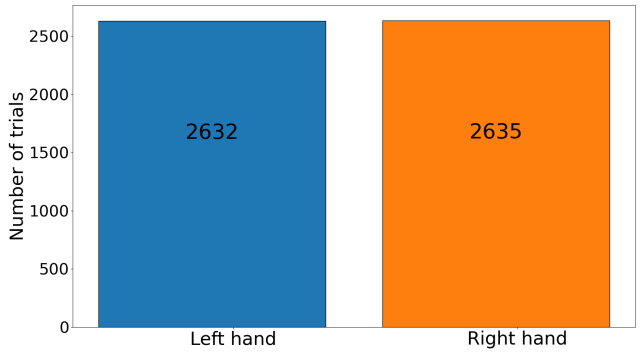
Figure 4.3: The raw EEG signal of dataset IV2b with and without MI evoked, respectively. Taken from the first trial of each class from channel C3 in subject B1T.

Class Distribution

The dataset includes information about the occurrence of artifacts during a trial. If an artifact occurred during a trial it is marked by a 1, and this specific trial can be removed during signal preprocessing. This affects the number of trials for each subject and class, which is illustrated in Figure 4.4. All of the subjects have over 100 trials per class, whereas some subjects have close to 200 trials per class. By combining all the data from every subject, both of the classes have over 2630 trials each, illustrated in Figure 4.4b.



(a) Class distribution for individual



(b) Class distribution for combined subjects

Figure 4.4: Class distribution for individual and combined subjects after removing trials with artifacts, for dataset IV2b.

Dataset IV2a

Dataset IV2a is a multiclass MI dataset recorded from 9 subjects. Each subject was asked to perform four different MI tasks which included left-hand, right-hand, feet, and tongue MI. The data was recorded over the course of 2 sessions, each containing 6 runs consisting of 48 trials each. This yields a total of 288 trials per session, distributed equally between the 4 classes. The first session was conducted without feedback, while the second was conducted with feedback, thus each subject is split into 'T' and 'E' [73]. Due to errors during the processing of subject A4T, this subject was removed, and the final dataset consists of 17 subjects. In practice, some of these trials are removed due to artifacts.

For recording, 22 electrodes based on the 10/20 scheme were utilized at a sampling frequency of 250Hz, shown in Figure 4.5. After the acquisition, the dataset was filtered with a 0.5Hz-100Hz bandpass filter and an additional notch filter of 50Hz.

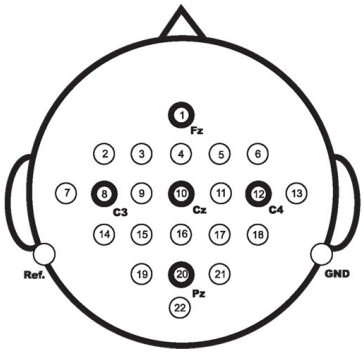


Figure 4.5: Electrode placement for dataset IV2a. Reprinted from [73].

The timing scheme of the trials in dataset IV2a is shown in Figure 4.6. The start of a new trial is indicated by a beep together with a fixation cross lasting for 2 seconds. Then a visual cue is shown to indicate which MI task should be performed, and the subject should evoke the MI and perform the task for 3 seconds. After the MI period is done, a break follows before the new trial is indicated by a beep. The parameters used to extract data from a single trial are given in Table 4.2. These are used unless specified otherwise.

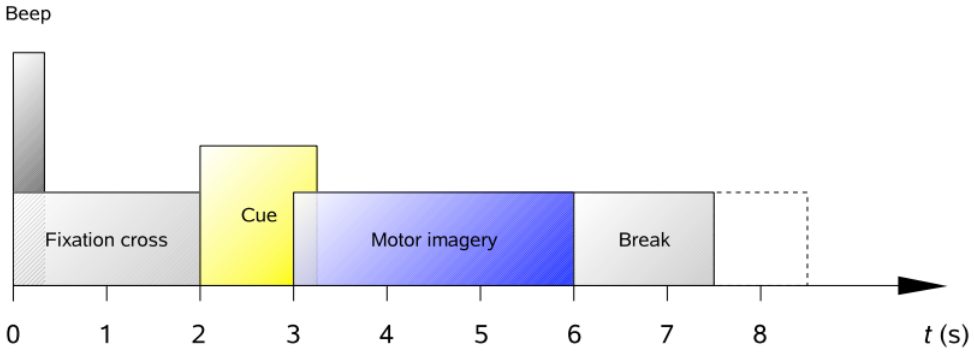


Figure 4.6: Timing scheme of one trial from dataset IV2a. Reprinted from [73].

Parameter	Value
Number of subjects	17: 8T and 9E
MI classes	LH, RH, F, T
Sampling frequency	250 Hz
Task window interval	3-6 sec
Task number of datapoints	750 samples
Rest window interval	6-7.5 sec
Rest number of datapoints	375 samples
Frequency bands	Theta, mu_{low}, mu_{high}, Beta
Number of electrode channels	Fz, C3, Cz, C4, Pz and 17 surrounding 10/20 electrodes

Table 4.2: Predefined data extraction parameters for one trial in dataset IV2a.

In Figure 4.7 a snapshot of left-hand, right-hand, feet, and tongue MI together with resting data is shown. This is generated from the first trial of each class in channel C3 for subject A1T. Again, the

structure of each MI is visually similar.

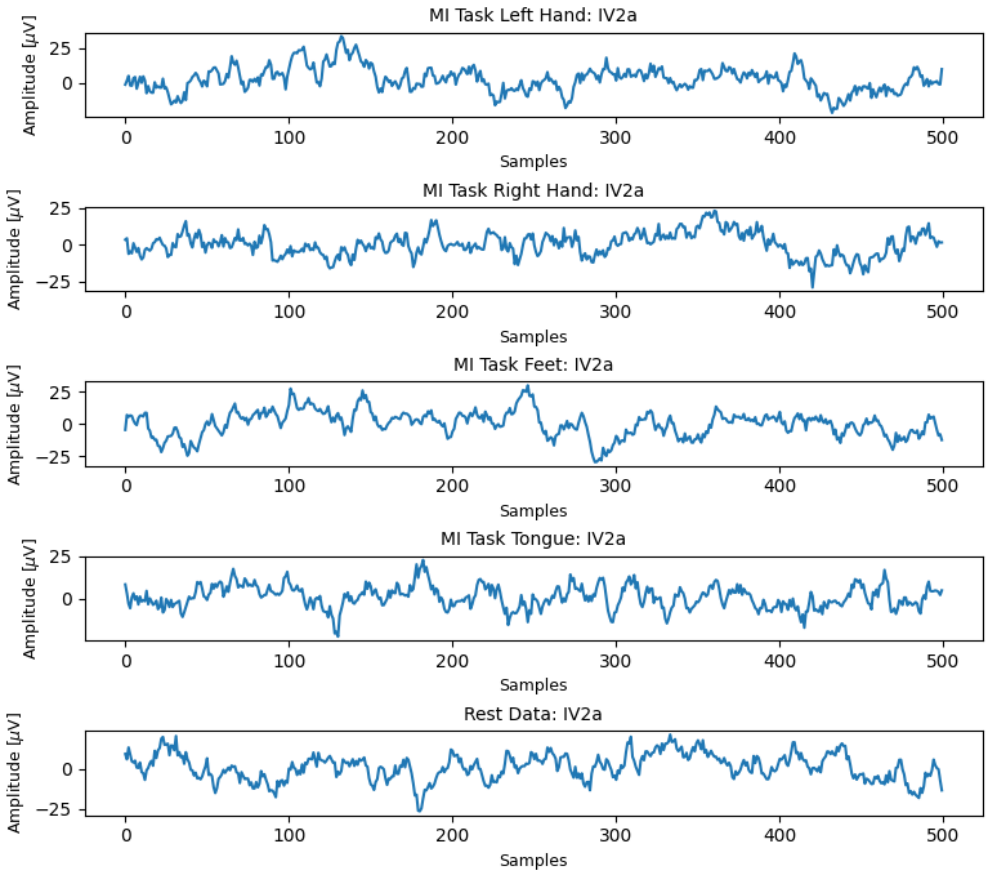
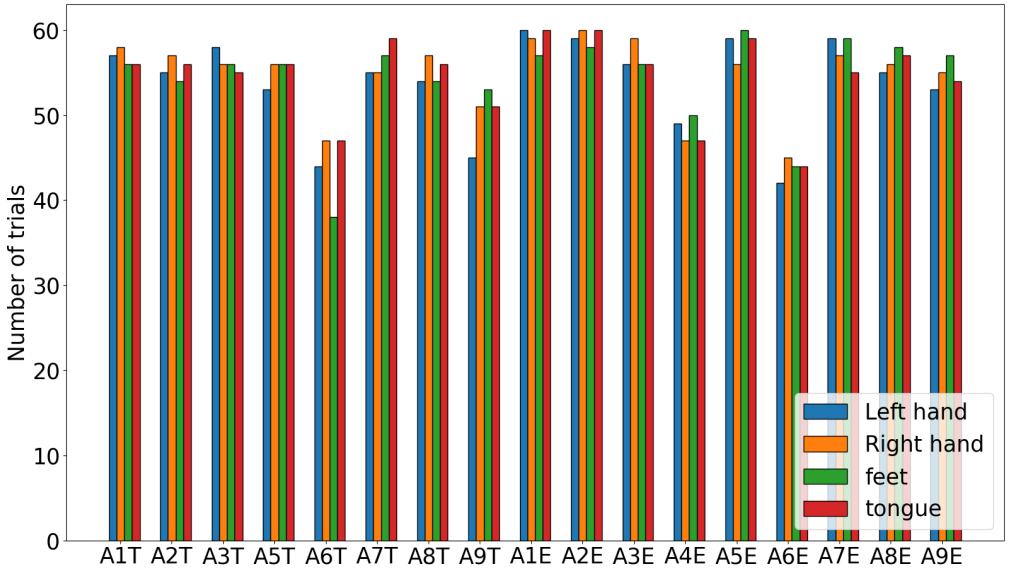


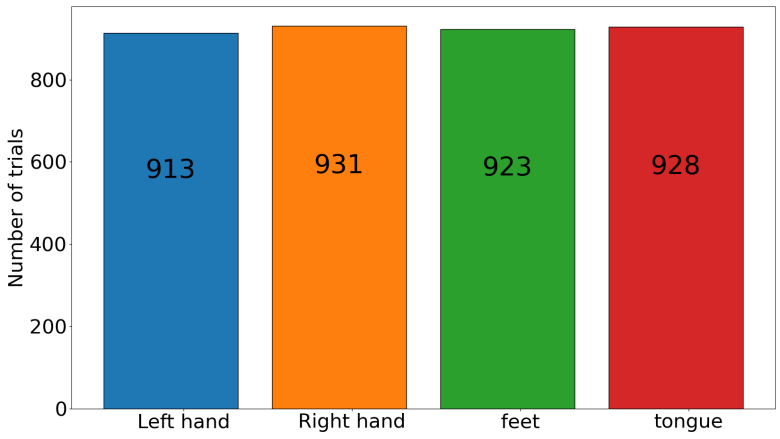
Figure 4.7: The raw EEG signal from dataset IV2a with and without MI evoked, respectively. Taken from the first trial of each class in channel C3 for subject A1T.

Class Distribution

After the recording, all of the data was visually inspected to detect the occurrence of artifacts. The trials which contained artifacts were marked, and are later removed during processing. This reduces the number of trials per subject and also affected the distribution of the classes. Figure 4.8 illustrates the new class distribution for each subject and the entire dataset, respectively. Every subject has under 60 trials per class, and each class has over 900 trials when combining all subjects.



(a) Class distribution for each subject



(b) Combined subjects

Figure 4.8: Class distribution for individual and combined subjects after removing trials with artifacts, for dataset IV2a.

Dataset SMR

Another dataset that is explored is obtained by Graz University [74], and is further referred to as dataset SMR, as the data is based on Sensorimotor Rhythms (SMR). This dataset contains binary MI of right hand and feet, acquired from 14 participants. The data was acquired through a single session consisting of 8 runs of 20 trials with equal amounts of each MI task, yielding a total of 160 trials [74]. The first 5 runs were used for training, without feedback, while the last 3 runs were used for validation, with feedback. For the 14 subjects and two recording methods, this resulted in 28 subjects that can be considered different from each other. The subjects marked with 'T' contains 100 trials, while the ones marked with 'E' contains 60 trials.

The dataset was acquired by the use of 15 electrodes, using Laplacian derivations. Channels C3, Cz, and C4 were used as centers, with the remaining electrodes positioned around them at a distance of 2.5cm, as shown in Figure 4.9. The data were sampled at a frequency of 512Hz [74]. After the acquisition, the dataset was filtered with a 50Hz notch filter and a 0.5-100Hz bandpass filter.

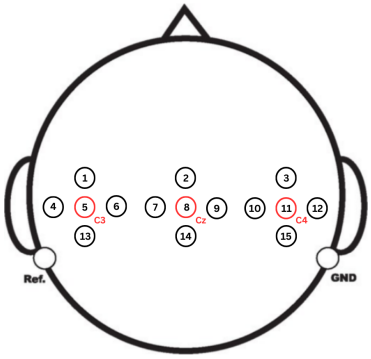


Figure 4.9: Electrode placement for dataset SMR. Based on Laplacian derivations.

In Figure 4.10 the timing scheme of one epoch of dataset SMR is illustrated. The initialization of an MI task is instructed with a cue, also indicating which MI task should be performed. The subject should then perform the specific MI task for 5 seconds. After the imagery period is done, a resting period is entered until a new cue is given. It should be noted that the first second after the cue has started is defined as reaction time, and is excluded when extracting the task data. The parameters used to extract one single trial from the dataset are shown in Table 4.3. These are used unless specified otherwise.

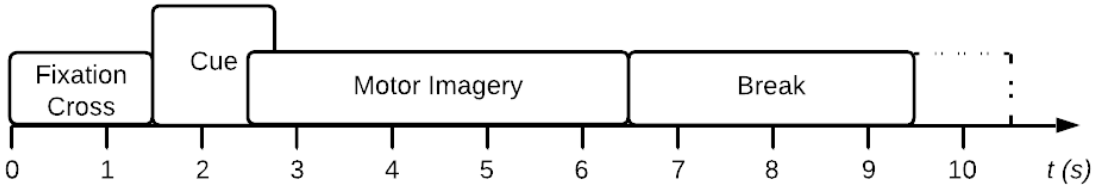


Figure 4.10: Timing scheme of one trial from dataset SMR.

Parameter	Value
Number of subjects	28: 14T and 14E
MI classes	RH, F
Sampling frequency	512 Hz
Task window interval	2.5-6.5 sec
Task number of datapoints	2048 samples
Rest window interval	7-9 sec
Rest number of datapoints	1024 samples
Frequency bands	Theta, mu_{low}, mu_{high}, Beta
Electrode channels	C3, Cz, C4, and 12 unlabeled, surrounding electrodes

Table 4.3: Predefined data extraction parameters for one trial in dataset SMR.

Figure 4.11 illustrates the raw EEG signals of right-hand MI and feet MI, as well as rest data, taken from a single trial for each class in the fifth channel, C3, of subject S1T. Again, it is not feasible to separate one task from another using visual deduction.

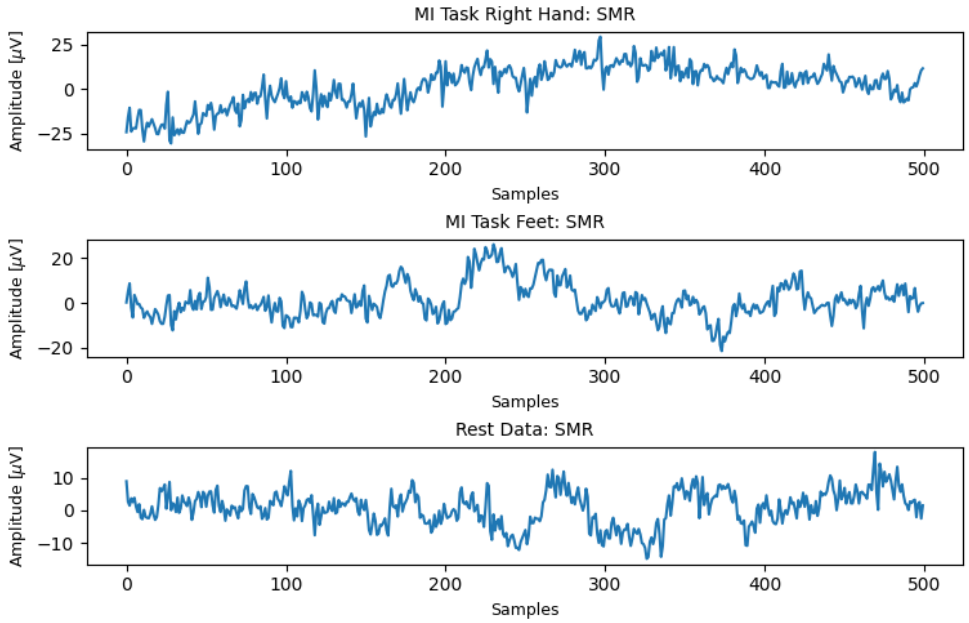
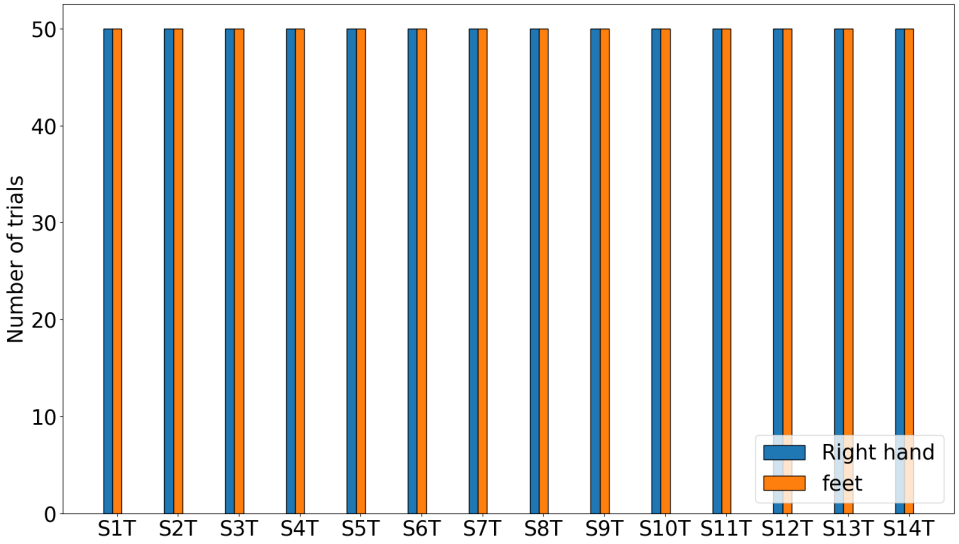


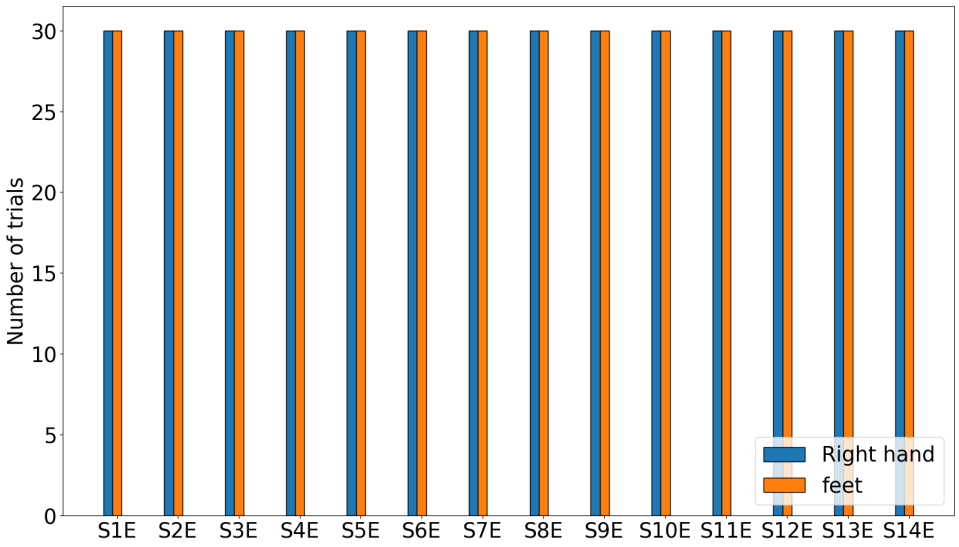
Figure 4.11: The raw EEG signal from dataset SMR with and without MI evoked, respectively. Taken from the first trial of each class in channel C3 for subject S1T.

Class Distribution

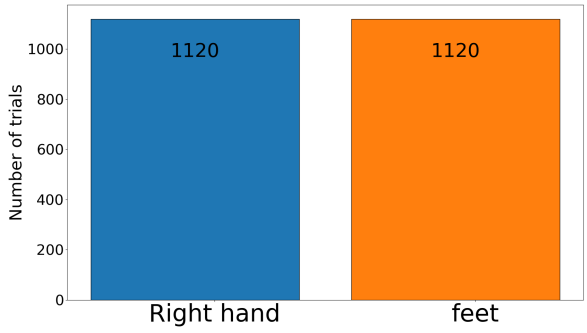
In contrast to the BCI competition datasets, the SMR dataset does not contain any artifact detection or labeling. Therefore, the number of trials per class and subject is balanced. For the without feedback recordings, 'T', the subjects have 50 trials per class, and for the with feedback recordings, 'E', the subjects have 30 trials per class, which is illustrated in Figures 4.12a and 4.12b respectively. In total for the entire dataset, each class consists of 1120 trials, as illustrated in figure 4.12c



(a) Subjects S1T to S14T



(b) Subjects S1E to S14E



(c) Combined subjects

Figure 4.12: Class distribution for individual and combined subjects, for dataset SMR.

Dataset NTNU

Dataset NTNU was acquired by two previous master students at NTNU, as part of their master's thesis on MI-BCI. Dataset NTNU contains binary MI data acquired from 16 subjects, where only 3 had previous MI experience [69]. All 16 subjects went through a total of 240 trials distributed between 4 sessions, thus 60 trials per session. There are equal amounts of trials from right-hand imagery movements and left-hand imagery movements, i.e. a total of 120 trials per MI. The best 6 of the 16 subjects were also recorded with three MI classes consisting of left-hand MI, right-hand MI, and feet MI task[69].

To collect the data, the OpenBCI Ultracortex Mark IV EEG Headset was used. This headset is based on the 10/20 system for electrode placement. The data was collected with 8 electrodes placed at the positions of C3, C4, Cz, FC1, FC2, CP1, CP2, and Fpz, with sampling frequency 250Hz [69]. The electrode placement is shown in Figure 4.13. However, electrode Fpz is not used for classification and was only included for possible artifact removal.

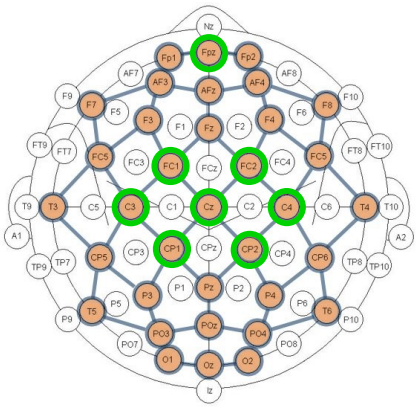


Figure 4.13: Electrode placement for dataset NTNU. Electrodes used are marked in green. Reprinted from [69].

The epochs of each trial last for 7.5 seconds, structured as shown in Figure 4.14. The trial is initialized by a cue indicating which MI to perform, followed by a task period with a duration of 2.5 seconds. After the cue is gone, the subject enters a rest period which lasts for 5 seconds until the next trial is initialized. Note that only the 2 last seconds of the task period are used as task data, as the first 0.5 seconds are labeled as reaction time. The parameters used to extract data from a single task, are shown in Table 4.4. These are used unless specified otherwise.

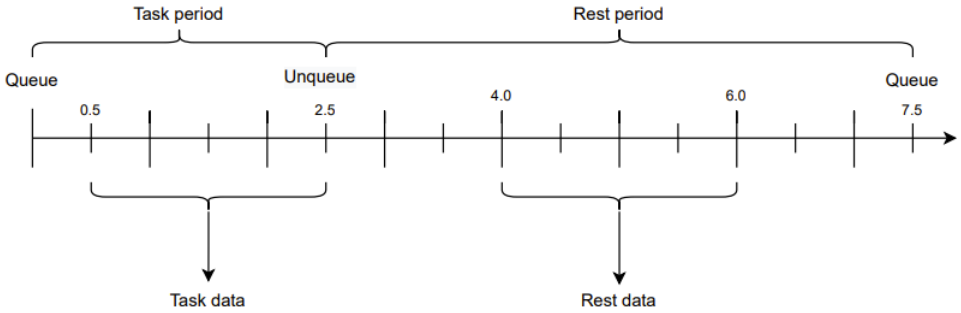


Figure 4.14: Timing scheme of one trial from dataset NTNU. Reprinted from [69].

Parameter	Value
Number of subjects	16
MI classes	LH, RH, F
Sampling frequency	250Hz
Task window interval	0.5-2.5 sec
Task number of datapoints	1000 samples
Rest window interval	4-6 sec
Rest number of datapoints	1000 samples
Frequency bands	Theta, mu_{low}, mu_{high}, Beta
Electrode channels	CP1, C3, FC1, Cz, FC2, C4 and CP2

Table 4.4: Predefined data extraction parameters for one trial in dataset NTNU.

In Figure 4.15 a snapshot of raw EEG data with each MI task, i.e. left-hand, right-hand, and feet, evoked is illustrated together with raw EEG data of rest data, i.e. no MI is evoked. The data is taken from the first trial of each class in channel C3 for subject 15. This Figure demonstrates that it is difficult to determine with visual inspection if an MI task is evoked or not, and also distinguish which MI task is evoked.

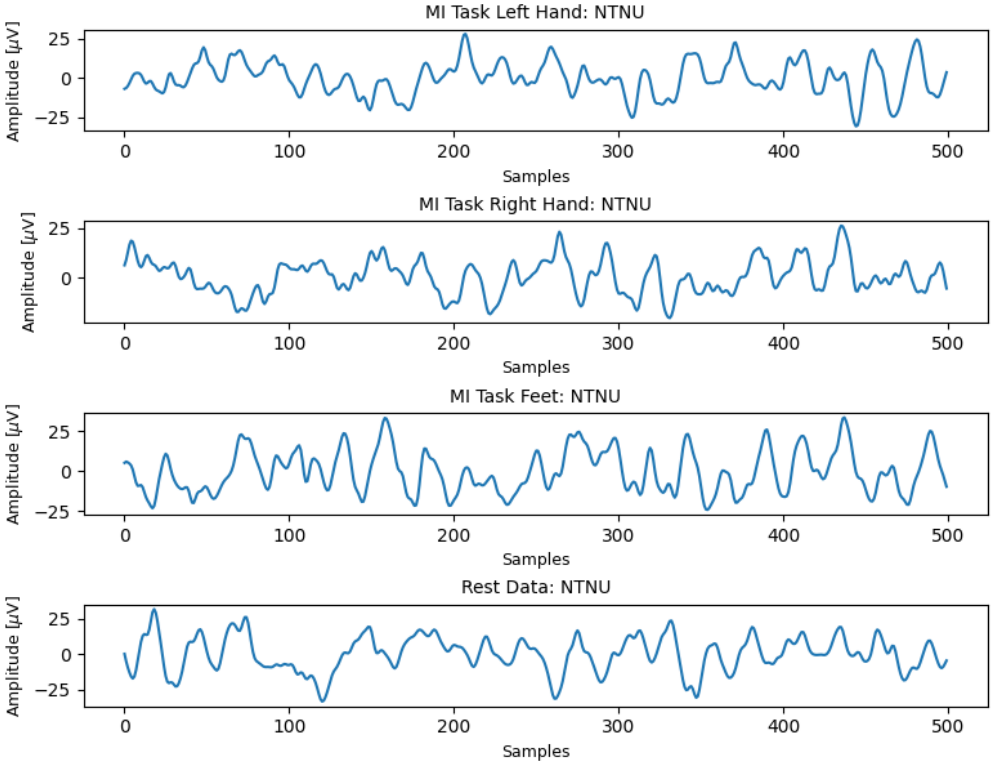
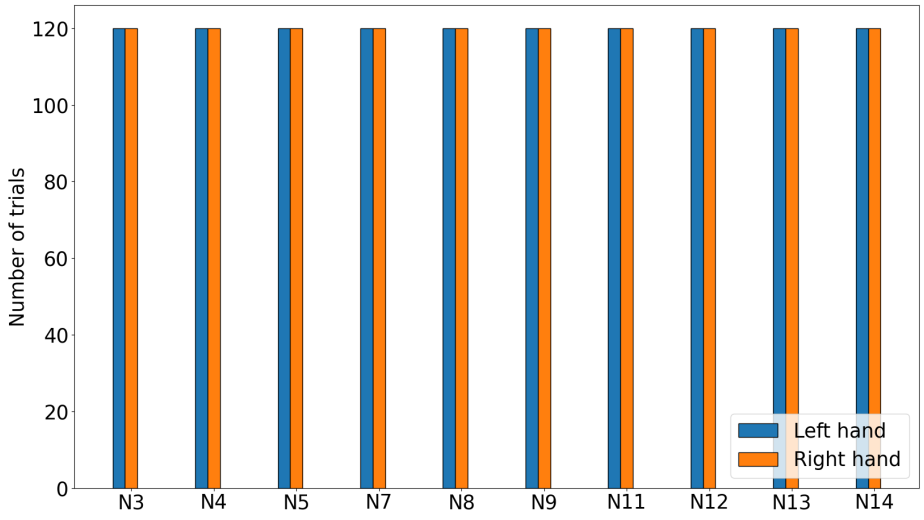


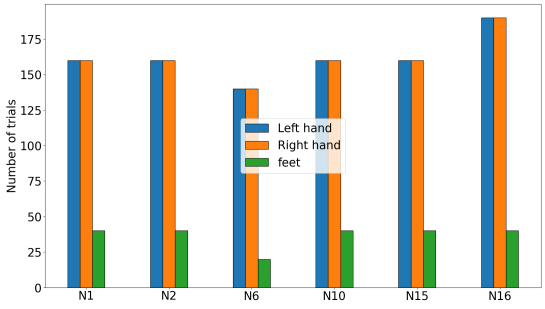
Figure 4.15: Raw EEG signal from dataset NTNU with left and right hand MI evoked and no MI evoked, respectively. Taken from the first trial of each class in channel C3, for subject N15.

Class Distribution

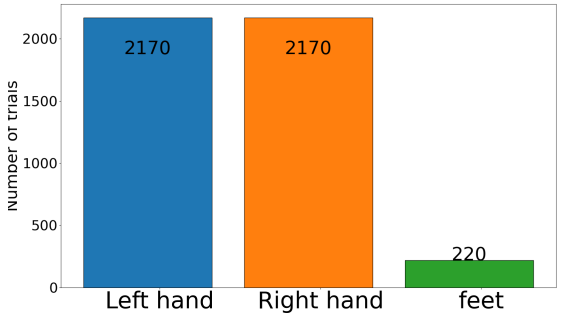
The number of trials per subject with two and three classes is illustrated in Figure 4.16a and 4.16b, respectively. The combined trials for each subject are illustrated in Figure 4.16c. For left and right-hand MI this resulted in 2170 trials each, while feet MI only consisted of 220 trials.



(a) Class distribution for subjects with 2 classes



(b) Class distribution for subjects with 3 classes



(c) Combined subjects

Figure 4.16: Class distribution for individual and combined subjects, in dataset NTNU.

4.2 Algorithm Pipeline

The proposed classification algorithms used in this thesis follow the pipelines illustrated in Figure 4.17. This flow chart visualizes the steps utilized in transforming the raw, unprocessed EEG signal into a classified MI task that can be used as a command. The pipeline is separated into an ML and two DL approaches to highlight their processing differences. The ML approach and the first DL approach consist of five steps, including pre-processing, feature engineering, and classification, while the second DL approach creates epochs from the data, then passes the raw EEG data directly into the classifier.

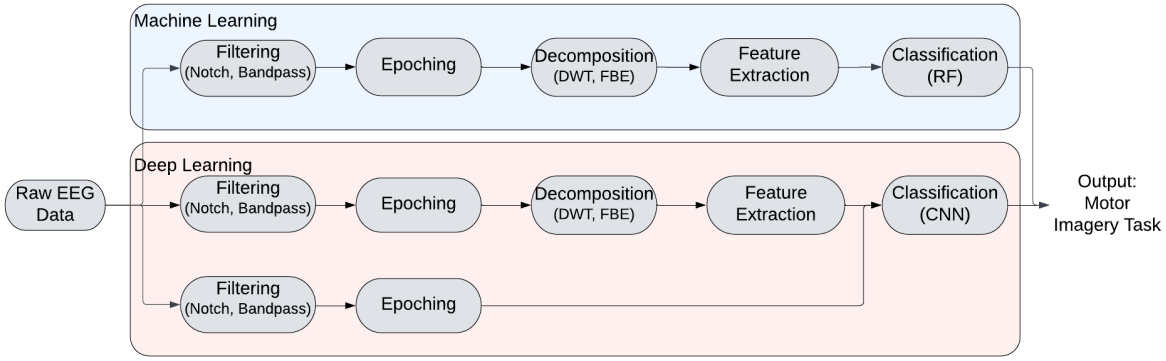


Figure 4.17: The data processing and classification pipeline for the ML and DL algorithms.

4.2.1 Filtering

For dataset NTNU, filtering is required. First, a 50Hz notch filter is applied to remove European power-line noise, then a IIR Butterworth bandpass filter from 1 to 50Hz is applied to remove additional noise and unwanted frequencies. Since datasets IV2b, IV2a, and SMR had been pre-filtered by their creators, this step is skipped in the pre-processing of these datasets.

4.2.2 Epoching

From the signal structures described in Section 4.1, one can split the raw data into parts that contain task data and rest data. The task data contains the MI, and by only extracting the task data the amount of data is reduced, and irrelevant information is excluded. The shape of the task data is the same for each subject and MI type in the individual datasets. The new shape of the raw data is dependent on the dataset's number of electrode channels, sampling frequency, and task duration. The rest data, also referred to as No-MI data, can be used to construct a binary classifier to detect if a signal contains an MI or not.

For most experiments the epoch shapes are consistent. This shape is customized to each dataset, based on their structure. The start of the epoch coincides with the cue, excluding some reaction time, and it will last for the entirety of the task period. However, there is also conducted experiments with varying epoch shape, in order to optimize it for each subject or dataset.

4.2.3 Decomposition and Feature Extraction

After each trial is reshaped into same-shaped epochs, the raw data is further decomposed to extract and enhance different properties of the raw signal. All epochs in the datasets are decomposed, and features are extracted from the decomposed signal. The experiments are conducted with two different decomposition methods, FBE and DWT.

Decomposition

With *FBE* some of the brain frequency bands presented in Section 2.1.3, as well as variations of these are extracted. This is done using a IIR Butterworth bandpass filter. The amount of decomposition

levels created is dependent on the number of frequency bands extracted from the signal. As a general basis the FBE decomposition is used to extract the brain waves of Theta, mu, and Beta. These frequency bands are chosen as mu and Beta have been extensively used in MI classification, and Theta is added as increased power in this band can be seen over the motor cortex during MI [79, 30]. Further in the report, these bands are split into subbands which are extracted for optimization analysis.

For *DWT*, four levels of decompositions are used. The resulting decomposition levels for datasets with a sampling frequency of 250Hz and 512Hz are given in Table 4.5. Due to the average decomposition, this results in 5 decompositions. The utilized mother wavelet is *Biorthogonal 2.8* (bior2.8), due to its satisfying results in previous research on EEG, [80], in addition to its biorthogonal and symmetric properties, which are also useful in EEG, as concluded in previous studies [81].

Decomposion Level	**Frequency Range [Hz]**	
	Sampling Frequency 250Hz	**Sampling Frequency 512Hz**
Detail coefficient 1 (D1)	62.5-124	128-256
Detail coefficient 2 (D2)	31.25-62.5	64-128
Detail coefficient 3 (D3)	15.625-31.25	32-64
Detail coefficient 4 (D4)	7.81-15.625	16-32
Approximation coefficients (A4)	0-7.81	0-16

Table 4.5: Frequency range of DWT with four levels of decomposition for datasets using a sampling frequency of 250Hz and 512Hz, respectively. Reprinted from [82].

Feature Extraction

The amount of features extracted from each epoch depends on the number of decomposition levels utilized or the number of frequency bands extracted. The number of extracted features for the DWT and FBE decomposition methods are given in Equations (4.1a) and (4.1b), respectively.

$$\#\text{Extracted Features} = \#\text{Channels} * \#\text{Feature Types} * \#\text{Decomposition Levels} \tag{4.1a}$$

$$\#\text{Extracted Features} = \#\text{Channels} * \#\text{Feature Types} * \#\text{Extracted Frequency Bands} \tag{4.1b}$$

From previous experiments with the IV2b and NTNU datasets, [1, 2], the features chosen to be extracted consist of TE, IE, Var, Skew, kurtosis, RMS, and HC. TE and IE were calculated using the formula presented in Equations (2.1) and (2.2), respectively, while the four other features were calculated using the functions defined in the tool package $mne\_features.univariate$ [83]. However, a new feature selection experiment, using all twelve features from Section 2.7 is conducted to verify this selection for the additional datasets. Thus HFD, PFD, mean, Cross, and HM are also calculated, using Equations (2.7) and (2.3) to calculate HFD and PFD, and equations from $mne\_features.univariate$ to calculate the remaining features.

4.2.4 Classification

In this thesis, classification methods using RF and CNN classifiers are used for experimentation. These are chosen to compare the performance of ML and DL in relation to MI.

Random Forest

To perform classification with RF a model is trained on a training set, and the performance is tested on a test set. In this thesis, the data is randomly split into 67% training data and 33% test data.

Convolutional Neural Networks

CNN can be used as a combination of a feature extraction and a classifier, thus it can use the raw unprocessed data as input. However, to get a better comparison of ML and DL, the performance of CNN is tested on two types of inputs: raw unprocessed data and feature data obtained from DWT or FBE decomposition. The data is split into 50% training data, used to train the classifier, 25% to validate during training, and the remaining 25% is used to test the classification performance.

A CNN model created specifically for EEG data is EEGNet. EEGNet is formally composed of two blocks and a classifier. The architecture of the EEGNet is given by Table 4.8 and is also illustrated in Figure 4.18. For binary classification, binary cross entropy is used as the loss function, while for multiclass classification categorical cross entropy is used as the loss function. Adam is used as the optimizer, with accuracy used as metric [62]. The input parameters are defined in 4.6. Some changes from the original implementation are introduced to compensate for the different sampling frequencies and different number of classes in each dataset, these parameters are given in Table 4.7.

Parameter	**Definition**
N	Number of classes
C	Number of electrode channels
T	Number of time samples for one trial
K	Kernel length, the length of temporal convolution in the first layer
F_1	Number of temporal filters to learn. Default = 8
F_2	Number of pointwise filters to learn. Default $F_2 = F_1$*D
D	Number of spatial filters to learn within each temporal convolution. Default D = 2
DropoutType	SpatialDropout or Dropout
DropoutRate	Dropout percentage

Table 4.6: Definition of the parameters that are used to build EEGNet [62].

Parameter	**Subject-dependent**	**Subject-independent**
Dropout rate	0.5	0.25
Number of epochs	200	200
Kernel length	$\frac{\text{Sampling Frequency}}{2}$	$\frac{\text{Sampling Frequency}}{2}$

Table 4.7: Default parameter values for EEGNet for the subject-dependent and subject-independent approach.

Block	Layer	#Filters	Size	#Parameters	Output	Activation	Options
1	Input				(C, T)		
	Reshape				(1,C,T)		
	Conv2D	F_1	(1,K)	K*F_1	(F_1, C,T)	Linear	padding = same
	BatchNorm			2*F_1	(F_1,C,T)		
	DepthwiseConv	D*F_1	(C,1)	C*D*F_1	(D*F_1, 1,T)	Linear	depth=D, max norm = 1, depth multiplier = 0
	Batchnorm			2*D*F_1	(D*F_1,1,T)		
	Activation				(D*F_1, 1,T)	ELU	
	AveragePool2D		(1,4)		(D*F_1, 1,T//4)		
	Dropout				(D*F_1, 1,T//4)		p=Dropout rate
2	SeparableConv2D	F_2	(1,16)	16*D*F_1 + F_2*(D*F_1)	(F_2,1,T//4)	Linear	padding = same
	Batchnorm			2*F_1	(F_2,1,T//4)		
	Activation				(F_2,1,T//4)	ELU	
	AveragePool2D		(1,8)		(F_2,1,T//32)		
	Dropout				(F_2,1,T//32)		p = Dropout rate
	Flatten				(F_2,(T//32)		
Classifier	Dense			N*(F_2*T//32)+N	N	Softmax	max norm= 0.25

Table 4.8: The EEGNet architecture with the parameters defined in Table 4.6.

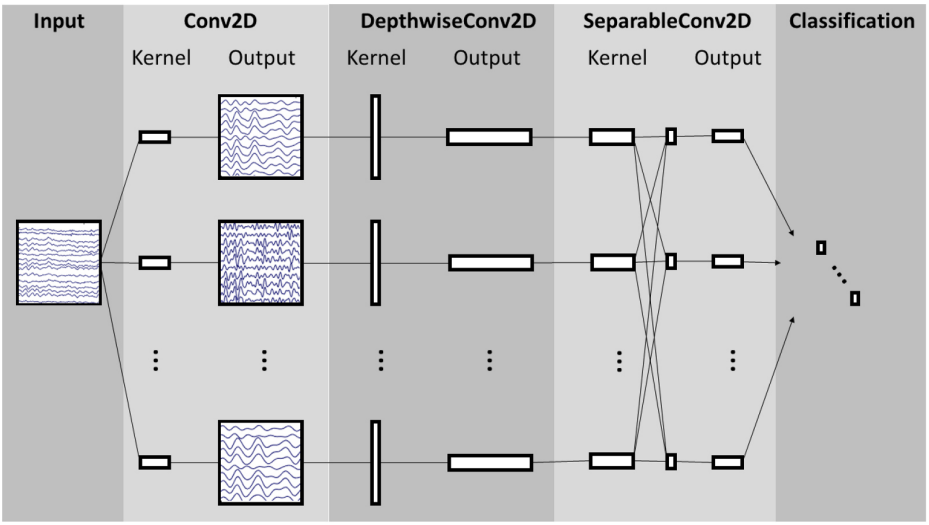


Figure 4.18: Visualization of the architecture of the EEGNet. Reprinted from [62].

4.3 Transfer Learning

In this thesis, an instance-based TL approach is used. Samples from subjects and datasets are used as sources to train the model. Then the model is tested on samples from the target, which is an unseen

subject or dataset.

TL is performed with various approaches. The first is a LOO approach where one subject or dataset is left out of the training stage and only used for testing. Here the model is only trained on source data and then tested on target data. Another method is a LSO approach, where some randomly chosen samples from the target subject or dataset are included in the training stage, and the remaining is used for testing the generated model. In this thesis, these approaches are used in *inter-subject* and *inter-dataset* TL.

Inter-subject TL is TL between subjects in the same dataset, acquired with the same protocol and equipment and by the same organization. With this approach, samples from all but one subject, the target subject, in the dataset are used as source to build a model, which then is tested on the target subject.

Inter-dataset TL is TL between datasets acquired by different organizations in different experimental environments, and in some cases with different equipment. Here, multiple datasets are used to build a source model, and after the model is built, it is tested on the target subject, which is a subject in a different dataset than the ones used to create the model. The approach to this type of TL is more complex than for inter-subject TL as it requires some changes in the pre-processing stage, such as adjusting the channel selection for data extraction to only include common channels. However, the actual TL approach using instance knowledge remains the same.

4.4 Sliding Window and Frequency Band Optimization

In this thesis, a frequency band and time window optimization are performed for each subject and each dataset. This combined optimization of the frequency band and the time window can be referred to as *band-window optimization*.

This proposed optimization process is conducted by testing various window sizes, window offsets, and frequency bands, then determining the optimal combination for each subject, or dataset, based on classification performance. Here, the *window offset* is the starting point of the task relative to the original given in Section 4.1, and the process of shifting the starting point of the task extraction is referred to as the *sliding window* technique. The optimization is performed using FBE decomposition and the optimal parameters are chosen by evaluating the performance of the RF classifier.

The parameter values which are tested during the optimization process are given in Table 4.9, where the new frequency bands with their respective values are given in Table 4.10.

Parameter	Values
Window size	2, 3, 4, 5 seconds
Window offset	−0.5 - 1.0 seconds (step 0.1)
Frequency Bands	Theta, mu_{low}, mu_{high}, Beta_1, Beta_2, Beta_3, Beta_4, Beta_5

Table 4.9: Parameter values tested in the frequency band and time window optimization. The values for the frequency bands are given in Table 4.10.

Name	Frequencies
Theta	4-8 Hz
mu_{low}	8-10 Hz
mu_{high}	10-12Hz
Beta_1	12-14 Hz
Beta_2	14-16 Hz
Beta_3	16-20 Hz
Beta_4	20-24 Hz
Beta_5	24-30 Hz

Table 4.10: Frequency band values used for optimization.

4.4.1 Electrode Channel Reduction

To find an optimal electrode channel subset, the greedy backward elimination algorithm is performed. The idea behind the algorithm is to iteratively eliminate the least informative channels until either the desired number of channels or level of accuracy is achieved. The algorithm starts with all the available channels and then removes each channel, iteratively. The channels that remain after the algorithm stops are considered the most informative for the classification task. For a dataset with N channels, the processing, and classification process is performed $N*(N-1)$ times [84]. The algorithm compares a subset of possible channel combinations, the results are not guaranteed to be the global optima, but a local optima.

The performance in each iteration is calculated by classifying features extracted from decomposed FBE data using standard extraction parameters and using accuracy as the performance measure. The only change in the exploration is the subset of channels that are being extracted from the data.

Topographic Map Analysis

Topographic maps are used in this thesis to analyze and get insight into which regions in the brain are activated, thus also which electrodes are most relevant, for the different MI tasks, in relation to low and high-performing subjects. The data used to create the topographic map is based on the average magnitude for every channel.

4.5 Methods for Evaluation

In order to evaluate the performances of the algorithms it is necessary to establish some methods and metrics for evaluation. This will help determine the algorithms' ability to classify the tasks of the epochs correctly.

Classification Performance

All experiments are evaluated by the same methods and metrics. This is to contain a consistent evaluation of the results and obtain the ability to compare between results. In this thesis, the metrics utilized are accuracy, F1-score, precision, and recall. All metrics evaluate the classification performance based on True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). These results are obtained by comparison between actual and predicted labels. The metrics will assign each classifier a value between 0 and 1 to indicate their performance. Here, 0 is a completely incorrect classification and 1 is the desired value, indicating perfect classification results. These values are then turned into percentages.

Additionally, a Chance Level (CL) is defined to indicate that a performance underneath this level may have occurred by chance, which is defined a little above a statistical guess, to illustrate that even a few percentages above a guess, can still have occurred by chance. For binary classification this level is set to 60%, taken from [85], 3-class classification has a CL of 40%, while 4-class classification has a CL of 30%. The subjects can be split into *high* and *low performers* based on their ability to get a classification performance above this CL.

To visualize the results, tables, plots, and confusion matrices are used. The confusion matrices for binary and multiclass classification are shown in Table 4.11, placed in order.

	Actual Positive	**Actual Negative**
Predicted Positive	True Positive	False Positive
Predicted Negative	False Negative	True Negative

		Predicted Class			
		C_1	C_2	…	C_N
Actual Class	C_1	$C_{1,1}$	**FP**	…	$C_{1,N}$
	C_2	**FN**	**TP**	…	**FN**
	…	…	…	…	…
	C_N	$C_{N,1}$	**FP**	…	$C_{N,N}$

Table 4.11: Confusion matrix for binary and multiclass classification, respectively.

To measure the correctness of the classifier, *accuracy* is used. From the accuracy, the portion of correctly predicted classes can be obtained.

$$\text{Accuracy} = \frac{\text{Number of correct samples}}{\text{Number of samples}}$$

Precision is used to find the portion of true positives in the samples classified as positives. This is a measure of observational error.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Recall is used to determine the classifiers' ability to identify all the positive samples in the set.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives + False Negatives}}$$

F1-score is a good metric for comparing different methods predicting the same quantity. It is defined as the balanced mean between precision and recall.

$$\text{F1-score} = 2 * \frac{\text{Precision*Recall}}{\text{Precision + Recall}}$$

Feature Importance

To assess the significance of each feature type, their scores obtained after classification using RF combined with FBE decomposition, are considered. The importance value for each feature is determined using the mean and standard deviation of the mean impurity decrease accumulation in each tree, i.e. the MDI method is used [54]. The feature types are then ranked based on their importance value, and the most and least significant features are identified. All these importance values sum up to 1.

Chapter 5

Results

In this chapter, the obtained experimental results for the four datasets are presented. This includes experiments involving feature selection, MI detection, transfer learning, band-window optimization, channel reduction, and classification approaches.

5.1 Experiment 1: Feature Selection

Before further experiments can be conducted, the features chosen in our previous work, [1, 2], must be validated for the new datasets, SMR and IV2a, and frequency bands. This is done by conducting a new feature importance experiment for all datasets by using the 12 features described in Section 2.7, and only extracting the Theta, mu, and Beta band from the signal through FBE decomposition. The importance values are assigned using the RF classifier and the results of the experiment are given in Figure 5.1. Only the most significant features are used in further experiments.

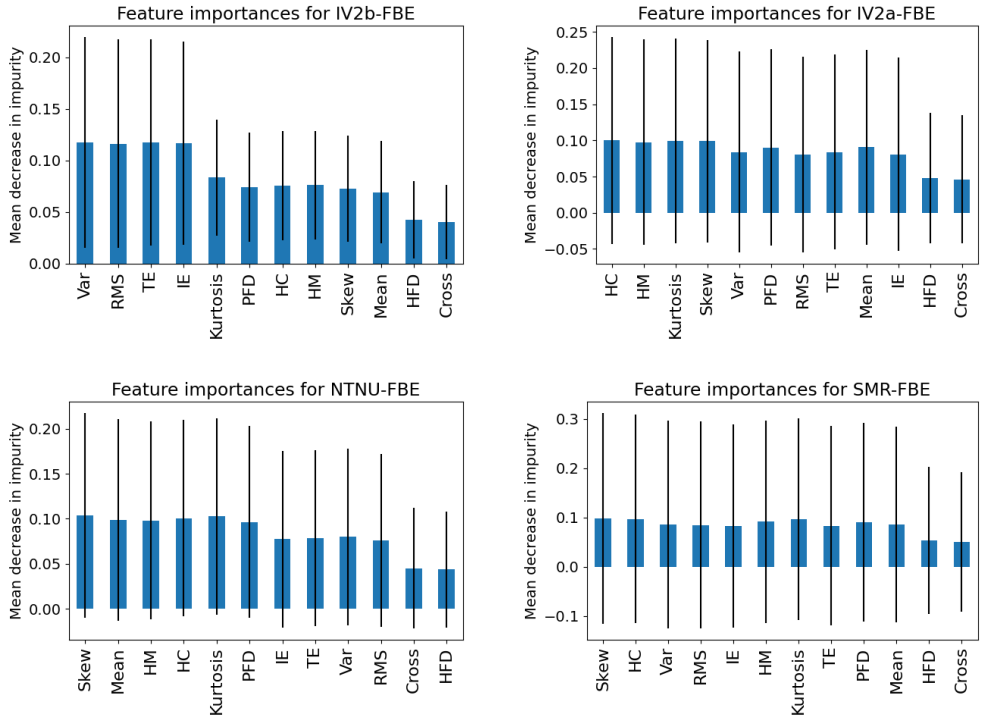


Figure 5.1: Subject-independent Feature Importance from RF, for all four datasets.

The selected features from [1] were TE, IE, PFD, Var, RMS, HM, and HC. From the new results in 5.1, it can be observed that some new choices should be made. HFD and Cross are consistently the worst features and are definite exclusions. As IV2b was proven the superior dataset over NTNU in [1] and [2], its top four features, Var, RMS, TE, and IE, are included further. Additionally, Skew, HC, and Kurtosis are persistently among the most important feature for the last three datasets. Thus, further experiments are conducted using these seven features.

5.2 Experiment 2: MI vs No-MI

In order to evaluate the proposed algorithm's ability to detect the occurrence of an MI, an experiment involving classification between MI task data and rest data, labeled No-MI, is conducted. The task data is the combined data of all the MI classes in the individual datasets, while the rest data is the break between each MI task, extracted from their respective time intervals given in Table 4.1, 4.2, 4.3, and 4.4. In this experiment, the goal is to obtain the classification performance when differentiating between an MI occurring against when no MI is present, thus what *kind* of MI is performed is ignored. The application of such a classifier in a BCI system would enable continuous detection of an MI occurrence, and upon detecting an MI, a subsequent classification could be performed to identify which MI task was performed.

The classification was performed using the RF classifier with FBE decomposition, and the experiment was conducted for all datasets to check robustness. The performance of the classification is given in Table 5.1, and the resulting confusion matrices for each dataset are given in Figure 5.2.

Dataset	Accuracy	F-Score	Precision	Recall
IV2b	89.89%	89.89%	89.96%	89.94%
IV2a	94.87%	94.87%	94.93%	94.85%
NTNU	95.12%	95.12%	95.15%	95.15%
SMR	90.75%	90.74%	90.97%	90.72%

Table 5.1: The average subject-independent performances for binary classification of MI and No-MI, for all datasets. Using RF with FBE.

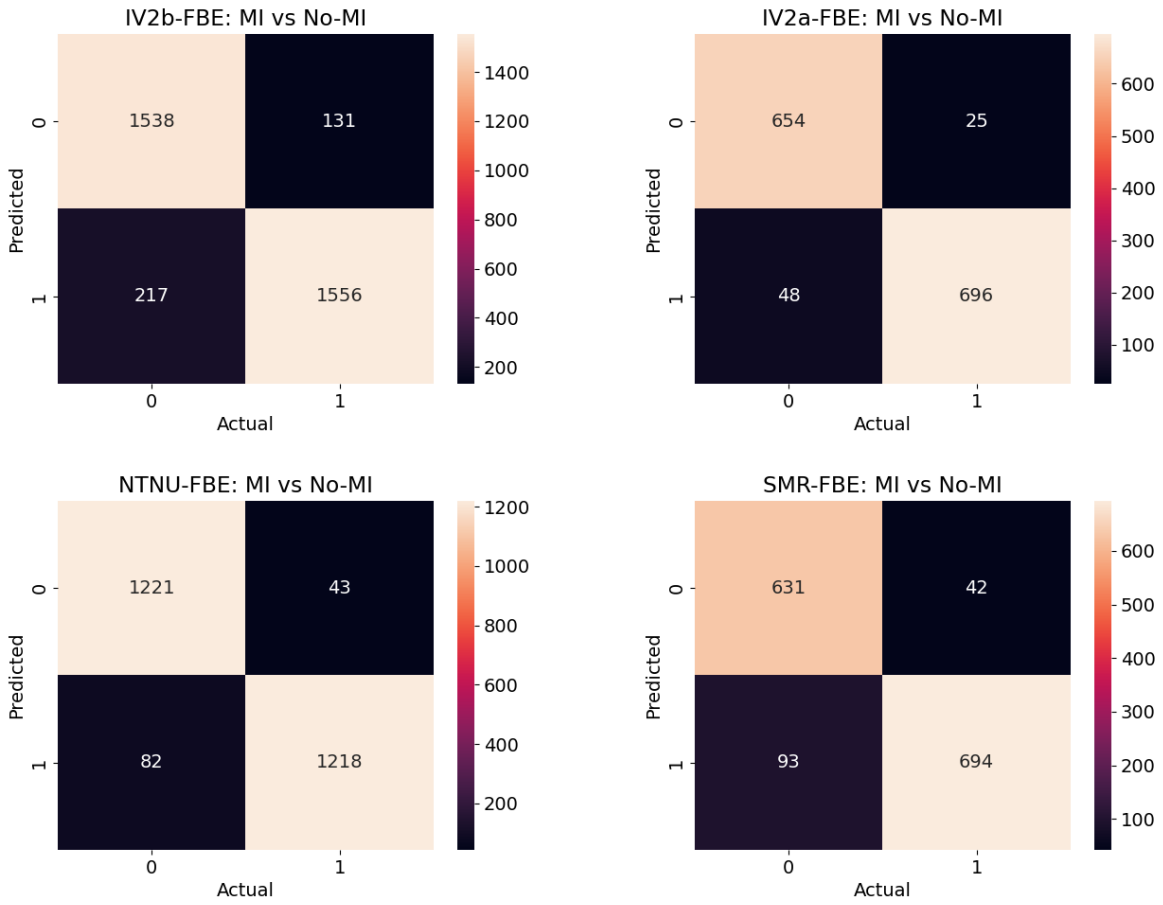


Figure 5.2: Confusion matrices resulting from RF classification between MI and No-MI for all datasets, using a subject-independent approach. Label 0 refers to MI data, while 1 refers to No-MI data.

The results in Table 5.1 indicate that the proposed MI versus No-MI classification method achieves a high accuracy, above 89% and similar score for F-score, precision, and recall across all datasets, where dataset NTNU performs the best, and dataset IV2b the worst. This implies effectiveness of the method in detecting MI across different datasets.

5.3 Experiment 3: Transfer Learning

Due to the limited amount of data for each subject and dataset, the training process may be insufficient and incomplete, thus leading to poorer performance. In addition, there is a long calibration time linked to the training of a new model for a subject. TL experiments are conducted to see if data and pre-trained models from other subjects and datasets can be used to save time on calibration, increase the amount of training data, and possibly improve performance. This is done by using inter-subject and inter-dataset TL.

5.3.1 Inter-Subject

To investigate the effectiveness of inter-subject TL in improving the predictive abilities of a model and reducing calibration time, an experiment is conducted. This experiment creates a model based on all the subjects in a dataset, excluding a target subject. Thus, the approach is building a source

model based on data that was acquired with the same equipment and in the same experimental environment as the target data. This model is then tested on the target subject, and its RF performance is compared to the RF performance obtained when a model is trained and tested on data from the target subject only, i.e. the subject-dependent model, labeled "Own data". Two experiments are conducted, where the first is excluding the whole subject in a LOO manner, labeled "TL LOO", while the other includes some target samples in the training phase in a LSO manner, labeled "TL LSO".

This TL approach is tested for all datasets, including IV2b, IV2a, binary IV2a, NTNU, and SMR. For dataset IV2a both multiclass classification, using all four MI classes, and binary classification, using only left-hand and right-hand MI, is performed. For dataset NTNU only binary classification using left-hand and right-hand MI is performed. The results from these experiments are given in Figures 5.3, 5.4, 5.5, 5.6, and 5.7.

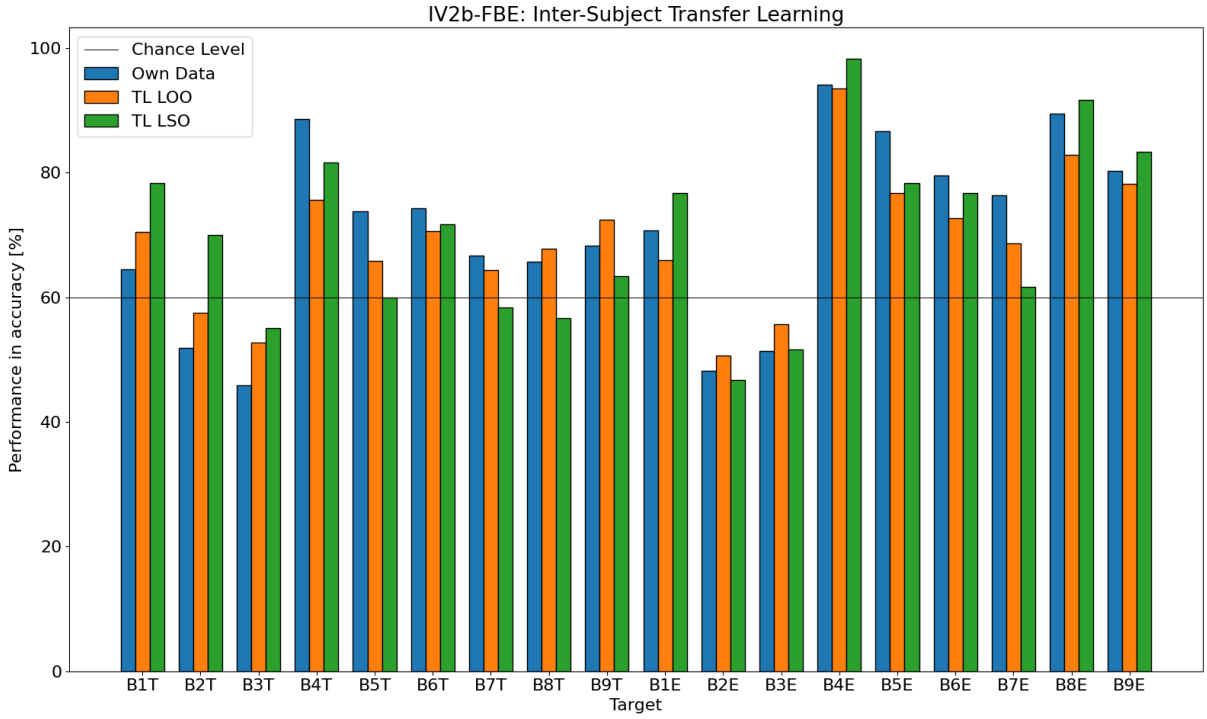


Figure 5.3: Results inter-subject TL for subjects in dataset IV2b using RF with FBE decomposition.

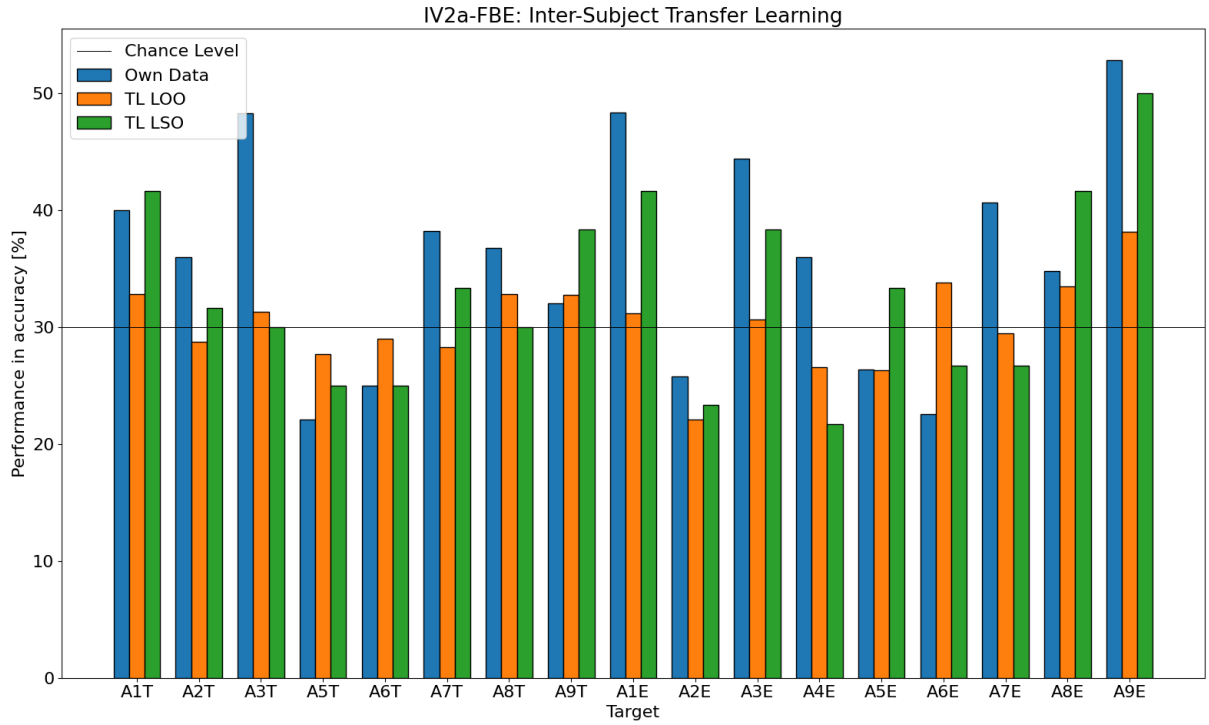


Figure 5.4: Results inter-subject TL for subjects in dataset IV2a using RF with FBE decomposition. Using all four MI classes.

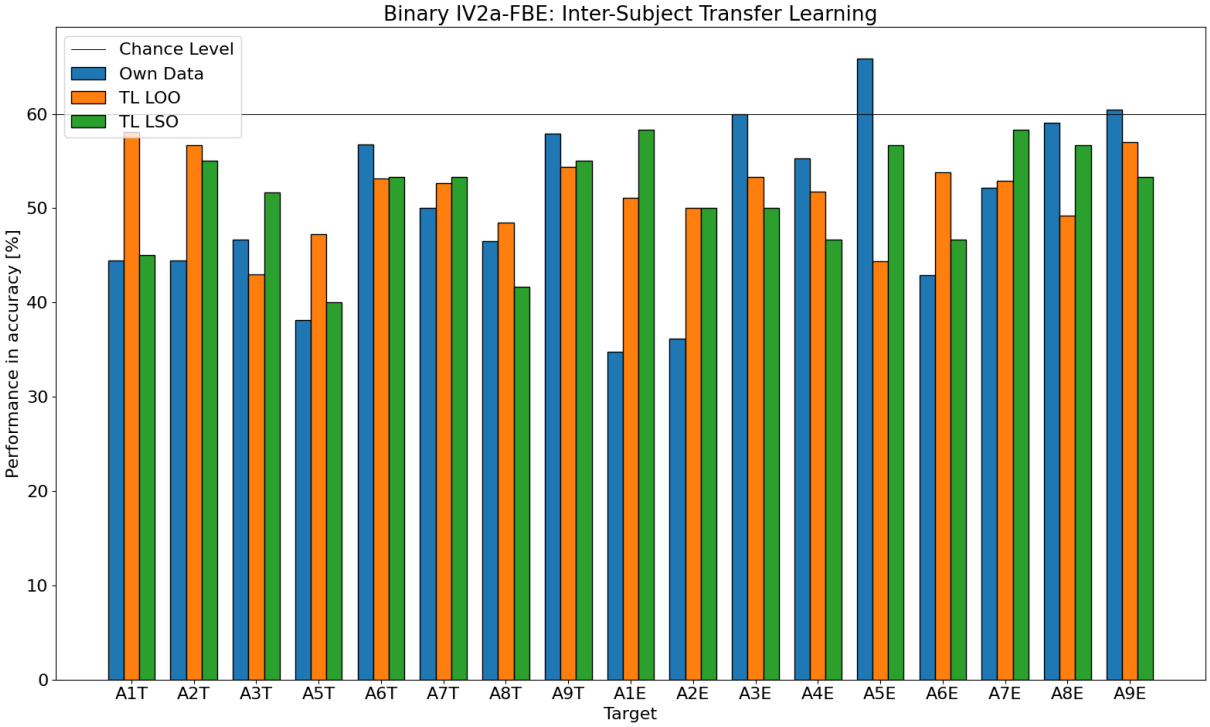


Figure 5.5: Results inter-subject TL for subjects in binary dataset IV2a RF with using FBE decomposition. Classes used are left and right-hand MI.

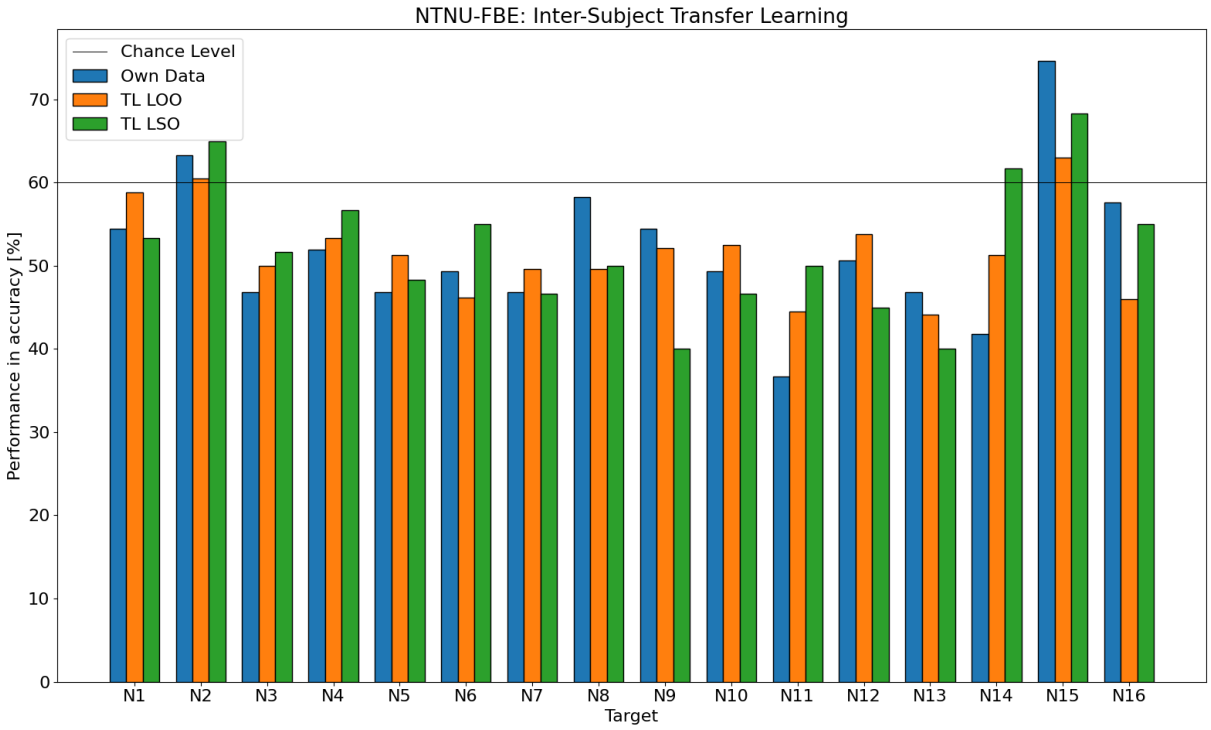


Figure 5.6: Results inter-subject TL for subjects in dataset NTNU using RF with FBE decomposition. Classes used are left and right-hand MI.

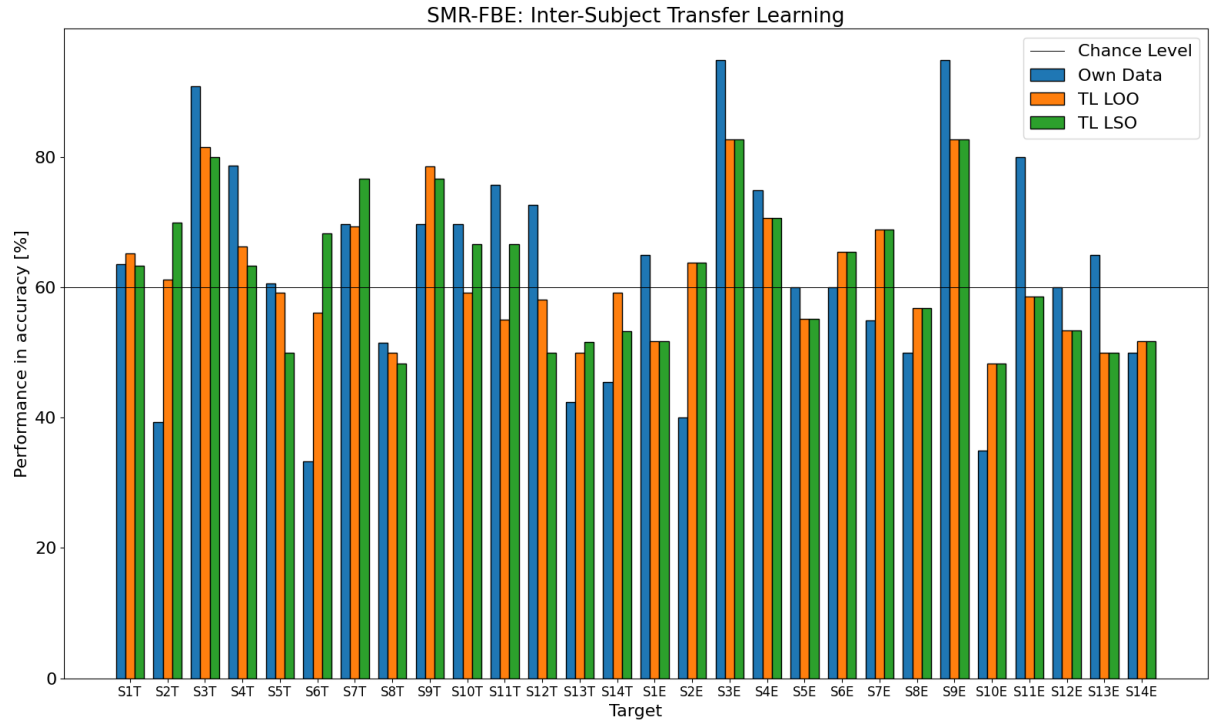


Figure 5.7: Results inter-subject TL for subjects in dataset SMR using RF with FBE decomposition.

From these results, it can be observed that TL is a well-performing approach. Even though there are cases where the TL approaches do not reach the performance of the subject-dependent model, in most cases their performance is in close proximity to this performance. As the purpose of this exper-

iment was to see if a TL approach could be used as a substitute for a subject-dependent model, to reduce calibration time, and aid in solving the problem of limited data, it can be deemed successful, as its performance was satisfactory.

Moreover, it can be observed that the best method varies between the subjects. To get a better overview of the effect of TL, as well as the occurrences of which a TL approach yields the best result, the frequencies and occurrence rates of each method as the best one are given in Table 5.2. Additionally, the table is divided into low and high performers to investigate if the impact of TL is influenced by the subject prior performance levels.

	Total		**High Performers**		**Low Performers**	
	Frequency	**Rate**	**Frequency**	**Rate**	**Frequency**	**Rate**
Own data	44	45.83%	30	68.18%	14	26.92%
TL LOO	27	54.17%	4	31.82%	23	73.08%
TL LSO	32		10		22	

Table 5.2: The frequencies in number and percentage of which each classification or TL method yields the highest performance. *Frequency* represents the number of times the corresponding method achieves the best performance, while *Rate* represents the percentage of subjects for whom the corresponding method is deemed the best.

Observe from the table that TL outperforms the method utilizing the subject's own data in 54.17% of the cases. It can also be seen that TL is especially efficient for low performers, as it is the best approach for 73.08% of the subjects. Additionally, it can be seen that the *LSO* approach outperforms the *LOO* approach for the high performers, while the two approaches perform similarly for the low performers.

5.3.2 Inter-Dataset

To perform inter-dataset TL between datasets using different amounts of channels, the data must be modified to only extract the features corresponding to the common channels of the datasets. As dataset SMR is acquired using a different electrode scheme than the other datasets, as well as the unlabeled nature of its channels, inter-dataset TL is performed without including this dataset.

The only shared channels between datasets IV2b, IV2a, and NTNU are C3, Cz, and C4, thus the inter-dataset TL is performed by solely extracting data from these channels. In addition, only data from the common classes, left-hand and right-hand, are extracted from the datasets, thus the inter-dataset TL is tested for binary classification using the RF classifier with FBE.

This experiment is performed by using a model created from two datasets as source, and then using the individual subjects from the remaining dataset as target. Four different versions of this TL approach are tested. The first method creates a model based on data from all subjects in the source datasets, this is labeled "LOO", while the second method also includes some samples from the target subject in the model training, labeled "LSO". The third method only includes the subjects of the source datasets who can achieve a performance above the CL, when building the source model, la-

beled "LOO>CL". The last method is similar to the third method, but it also includes some samples from the target subject, labeled "LSO>CL". These methods are also compared to the performance obtained when the model is trained and tested on data from the target subject only, labeled "Own", i.e. a subject-dependent model.

Dataset IV2b

To perform inter-dataset TL on dataset IV2b, the source models are built by using data from dataset NTNU and dataset IV2a. Then TL is performed by iteratively using every subject in IV2b as target. The resulting performance accuracy is shown in Figure 5.8.

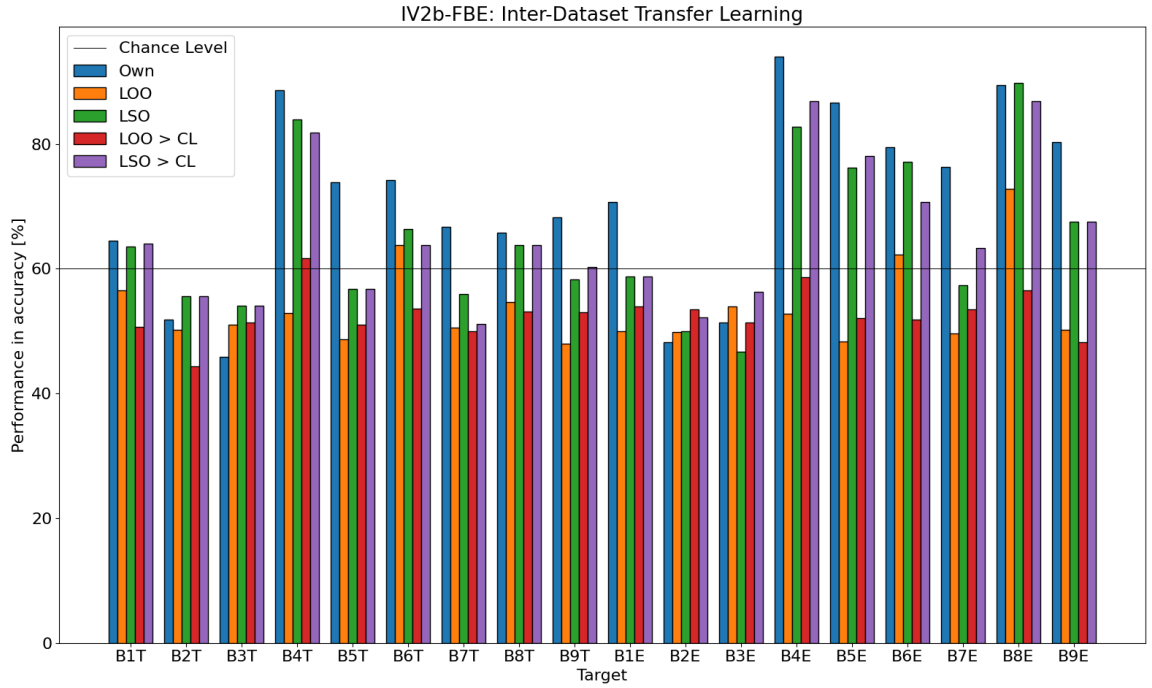


Figure 5.8: Accuracies for inter-dataset TL on subjects from dataset IV2b using source models pre-trained on data from IV2a and NTNU. Using FBE decomposition and RF classifier.

Dataset IV2a

When performing inter-dataset TL on dataset IV2a, source models are created using data from dataset IV2b and dataset NTNU. Then TL is performed using each subject in dataset IV2a as target, which resulted in the performances shown in Figure 5.9.

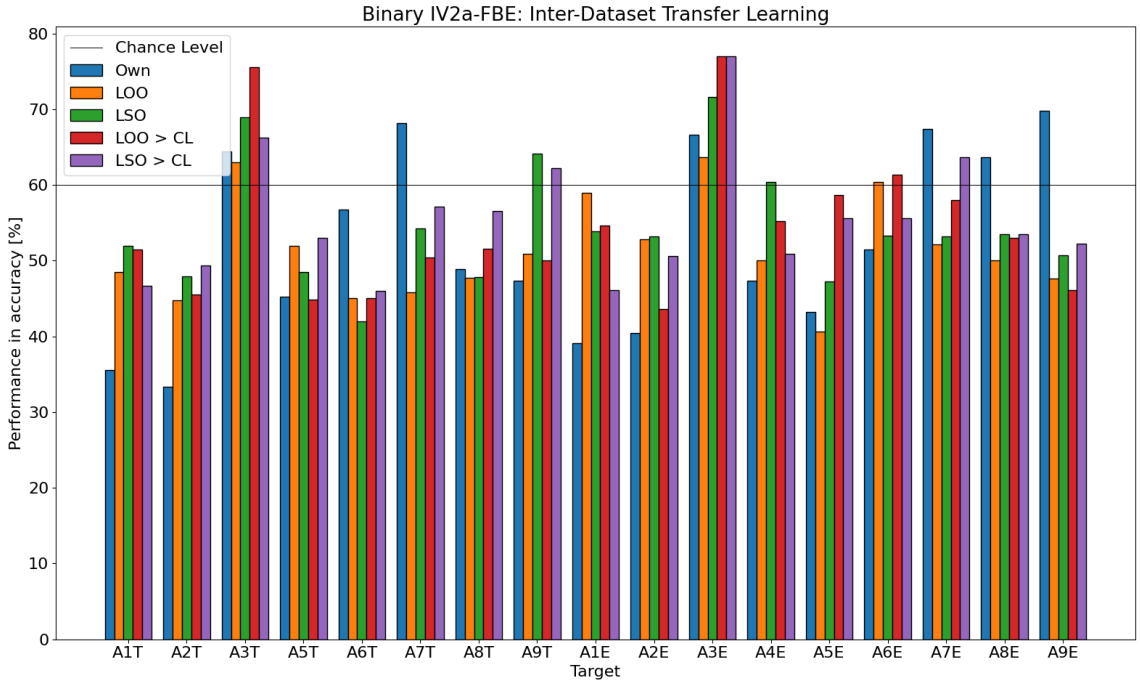


Figure 5.9: Accuracies for binary inter-dataset TL on subjects from dataset IV2a using source models pre-trained on data from IV2b and NTNU, Using FBE decomposition and RF classifier.

Dataset NTNU

Inter-dataset TL for dataset NTNU is performed by using source models created from dataset IV2b and dataset IV2a. The results when performing TL using the individual subject in dataset NTNU as target, are given in Figure 5.10.

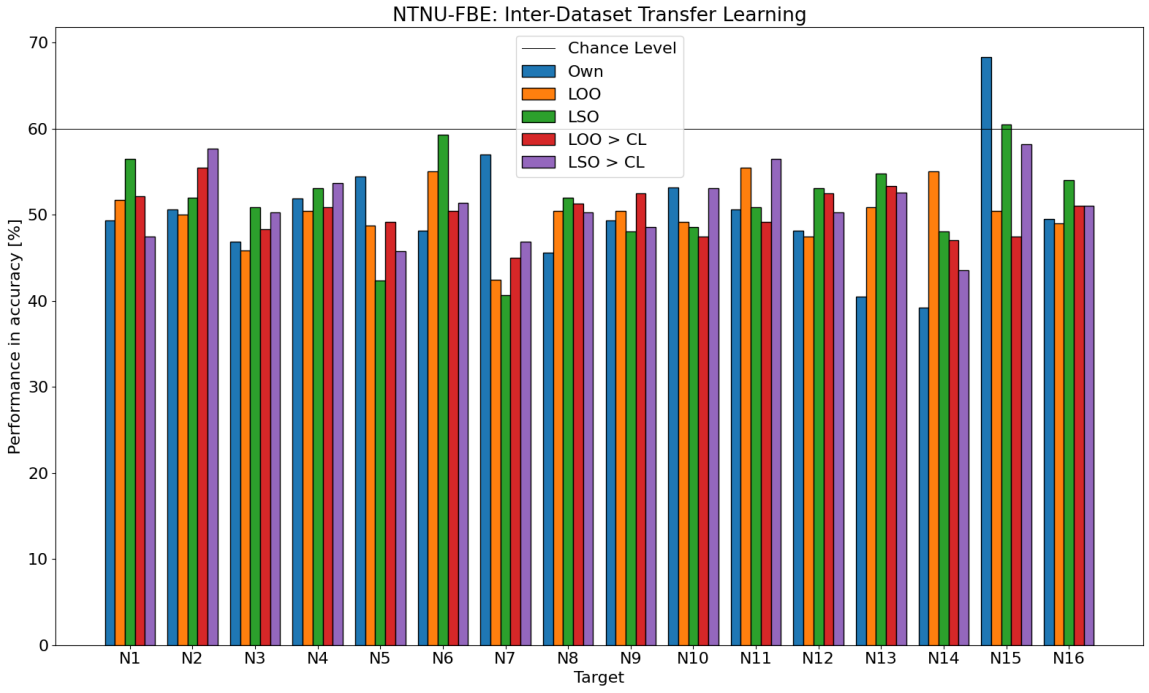


Figure 5.10: Accuracies for binary inter-dataset TL on dataset NTNU using source models pre-trained on data from IV2a and IV2b. Using FBE decomposition and RF classifier.

From the results in Figures 5.8, 5.9, and 5.10, it can be deduced that the TL approaches which only contains data from the high-performing subjects, i.e. "LOO>CL" and "LSO>CL" in the figures, obtains the best performance for 65.38% of the subjects, as compared to the TL approaches which use data from all subjects, i.e. "LOO" and "LSO". Also, notice that the LSO approaches obtains the best performance for 84.62% of the subjects. This indicates that these approaches are the preferred ones when building a TL model which should be used on subjects from unseen datasets. It can also be seen from results that TL is an especially effective classification approach for the low-performing, as compared to the high-performing, subjects.

5.4 Experiment 4: Sliding Window and Frequency Band Optimization

The following experiments are performed to investigate the effectiveness of an optimization process in improving the classification of MI tasks. The processing parameters time window and frequency band are optimized, and an optimal band-window combination is obtained. These parameters are interesting to optimize as it is crucial to select a time window that covers the significant ERD/ERS patterns and choose a frequency band that contains the frequencies at which these patterns are occurring. The frequency bands are extracted using the FBE decomposition method.

This optimization process is performed both *subject-dependently*, where the optimal combination is chosen for each subject, and *subject-independently* where the combination is determined by evaluation on a whole dataset.

5.4.1 Subject-Dependent Optimization

Given the significant variability amongst subjects, the individual characteristics of each subject should be considered when optimizing the processing parameters. This is also important to investigate if the problems of BCI illiteracy and lack of MI ability can be conquered. Consequently, a subject-dependent experiment is conducted where the optimal frequency band and time window are determined for each subject individually, taking into account their individual performance. For each subject the performance of each band-window combination is evaluated using the RF classifier, and the optimal combination is determined accordingly. The resulting optimal performance is evaluated against the performance obtained using the standard combinations given in Section 4.1 to see the effect and efficiency of the optimization process.

This experiment is performed for all datasets, but for dataset IV2a the optimization process is performed both on multiclass, using all four classes, and on binary classification, using left-hand and right-hand MI only. For dataset NTNU, only binary classification using left-hand and right-hand MI is performed. The Tables 5.3, 5.4, 5.5, 5.6, and 5.7 contains the resulting optimal frequency bands and time windows for each subject. The Figures 5.11, 5.12, 5.13, 5.14, and 5.15 presents the performance of the optimal band-window combination compared to the performance using the standard parameter values.

Dataset IV2b

Subject	Optimal window [sec]	Optimal band
B1T	3.5-7.5	mu_{high}
B2T	3.0-6.0	mu_{high}
B3T	3.7-6.7	Beta_5
B4T	2.8-5.8	mu_{high}
B5T	2.9-5.9	Beta_5
B6T	3.4-7.4	Beta_1
B7T	3.5-7.5	Beta_1
B8T	2.9-6.9	mu_{high}
B9T	3.6-7.6	mu_{high}
B1E	3.2-6.2	mu_{high}
B2E	2.6-6.6	mu_{high}
B3E	2.8-4.8	Theta
B4E	2.4-4.4	mu_{high}
B5E	2.6-4.6	Beta_5
B6E	3.0-7.0	Beta_1
B7E	3.4-6.4	Beta_1
B8E	2.8-6.8	mu_{high}
B9E	2.5-6.5	Beta_1

Table 5.3: Optimal time window and frequency band values for the individual subjects in dataset IV2b.

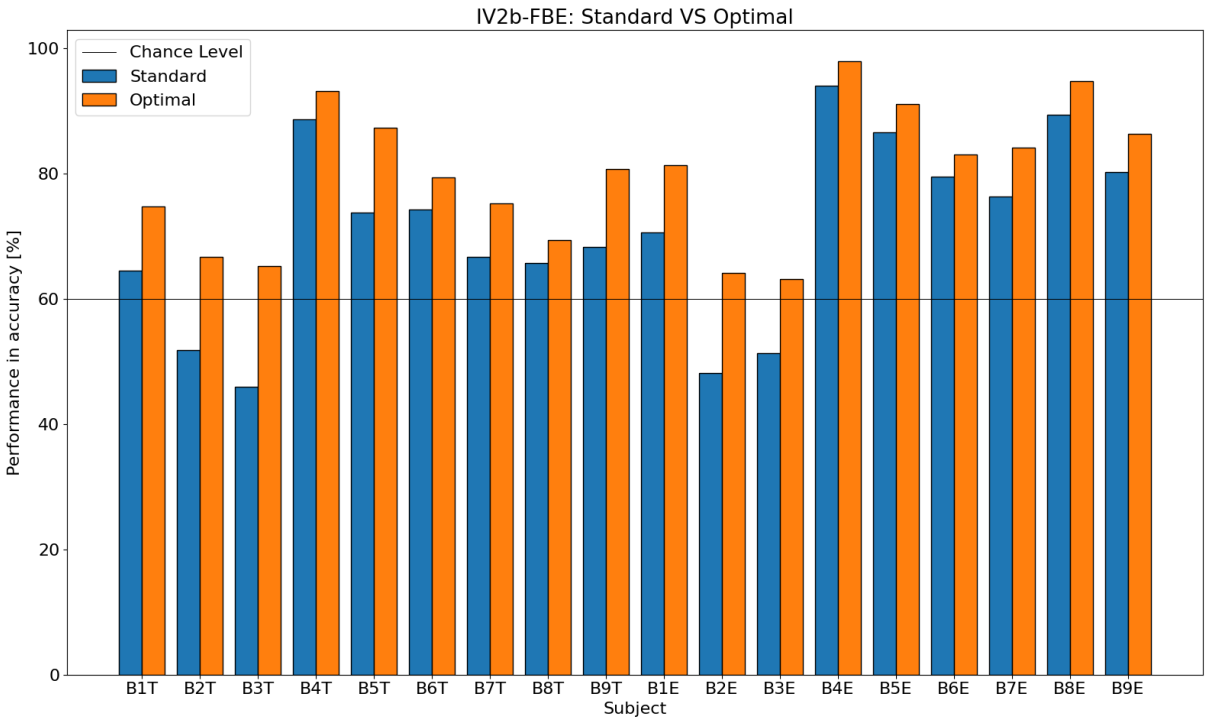


Figure 5.11: Subject-dependent classification performance using standard vs optimized parameter values, for dataset IV2. Using RF with FBE.

Dataset IV2a

Subject	Optimal window [sec]	Optimal band
A1T	2.5-4.5	mu_{high}
A2T	3.1-6.1	Beta_5
A3T	2.9-5.9	mu_{high}
A5T	3.4-7.4	Beta_4
A6T	4.0-8.0	Beta_1
A7T	2.5-4.5	Beta_4
A8T	3.6-5.6	mu_{high}
A9T	2.7-4.7	Beta_4
A1E	2.9-5.9	mu_{high}
A2E	2.8-5.8	Beta_1
A3E	3.0-5.0	mu_{high}
A4E	2.9-4.9	mu_{high}
A5E	2.7-5.7	mu_{high}
A6E	3.6-6.6	Theta
A7E	3.4-5.4	Beta_4
A8E	2.5-4.5	mu_{low}
A9E	2.9-4.9	mu_{low}

Table 5.4: Optimal time window and frequency band values for the individual subjects in dataset IV2a, for multiclass classification.

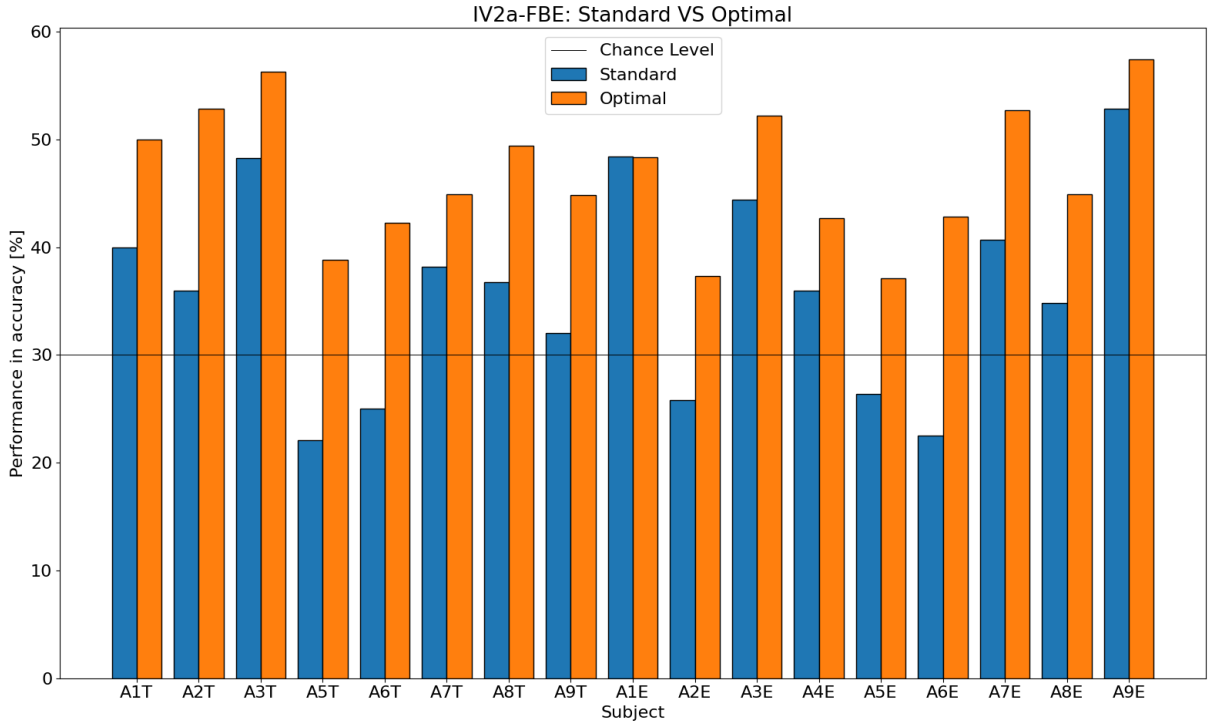


Figure 5.12: Subject-dependent classification performance using standard vs optimized parameter values, for dataset IV2a. Using RF with FBE.

Subject	Optimal window [sec]	Optimal band
A1T	2.5-4.5	mu_{high}
A2T	3.8-5.8	Beta_1
A3T	2.5-4.5	Beta_1
A5T	3.2-7.2	Beta_3
A6T	3.8-7.8	Theta
A7T	2.5-4.5	Beta_3
A8T	2.8-4.8	mu_{low}
A9T	3.4-6.4	Beta_5
A1E	3.2-7.2	mu_{low}
A2E	3.6-7.6	Beta_3
A3E	2.6-5.6	mu_{high}
A4E	3.1-6.1	mu_{high}
A5E	2.5-5.5	mu_{low}
A6E	2.8-5.8	Beta_1
A7E	2.7-6.7	mu_{high}
A8E	3.1-7.1	mu_{low}
A9E	2.8-4.8	Beta_4

Table 5.5: Optimal time window and frequency band values for the individual subjects in dataset IV2a, for binary classification using left-hand and right-hand MI.

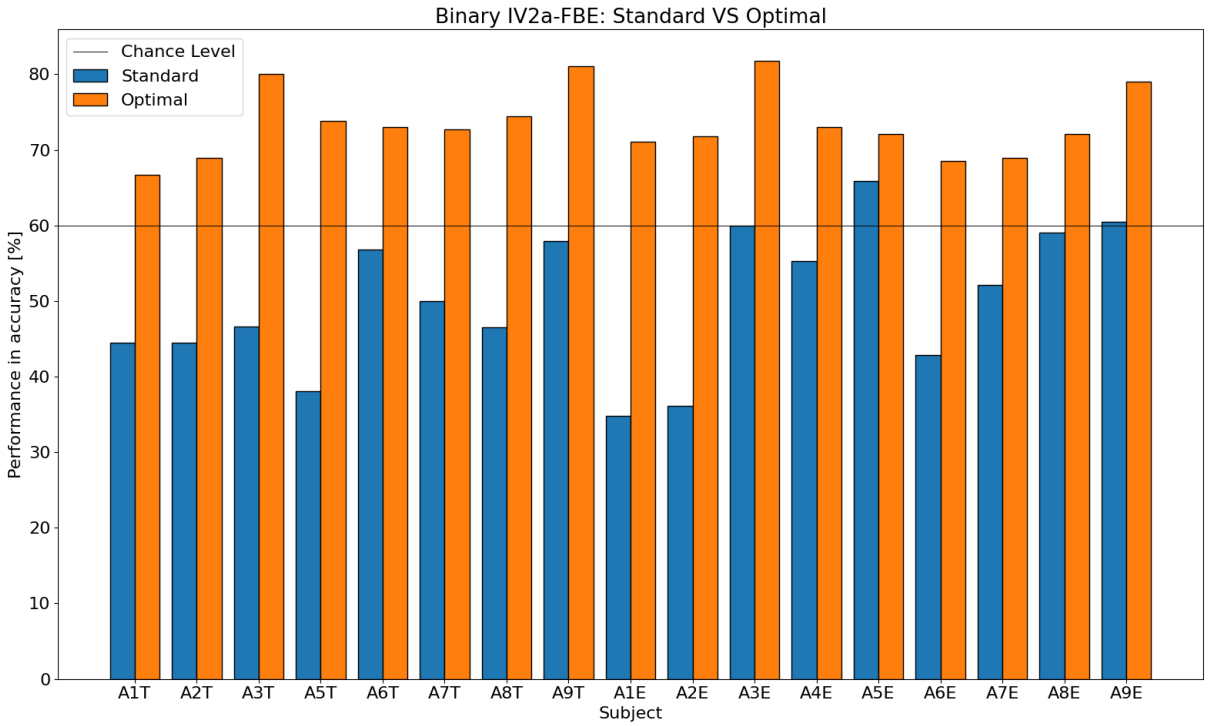


Figure 5.13: Subject-dependent binary classification performance using standard vs optimized parameter values, for dataset IV2a, for left-hand and right-hand MI. Using RF with FBE.

Dataset NTNU

Subject	Optimal window [sec]	Optimal band
N1	1.0-3.0	mu_{high}
N2	0.2-3.2	mu_{high}
N3	0.8-2.8	mu_{high}
N4	0.6-2.6	mu_{high}
N5	1.3-3.3	Beta_5
N6	0.1-2.1	mu_{high}
N7	0.9-4.9	Beta_1
N8	0.0-3.0	Beta_5
N9	1.1-5.1	Beta_4
N10	1.1-4.1	Beta_5
N11	0.3-4.3	Beta_1
N12	0.3-3.3	Beta_4
N13	1.5-5.5	Beta_2
N14	0.1-2.1	Theta
N15	0.3-3.3	mu_{high}
N16	1.1-4.1	Beta_5

Table 5.6: Optimal time window and frequency band values for the individual subjects in dataset NTNU, for binary classification using left-hand and right-hand MI.

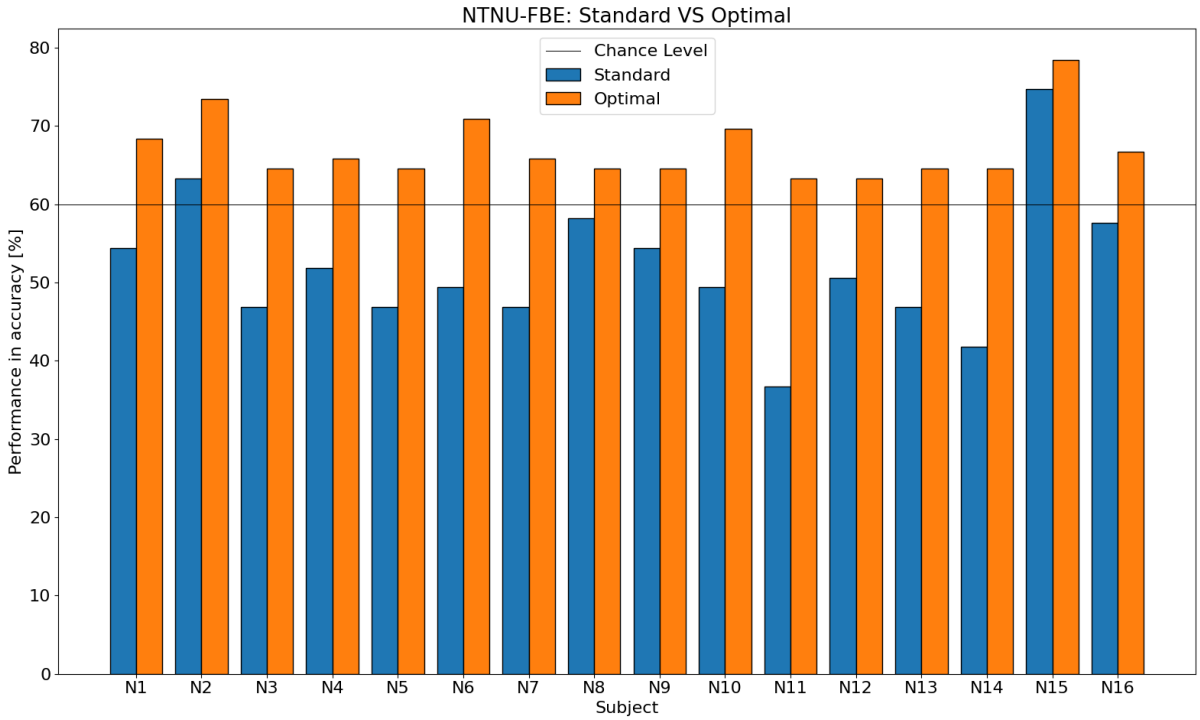


Figure 5.14: Subject-dependent binary classification performance using standard vs optimized parameter values, for dataset NTNU, using left-hand and right-hand MI. Using RF with FBE.

Dataset SMR

Subject	Optimal window [sec]]	Optimal band
S1T	2.0-6.0	Theta
S2T	2.3-5.3	Theta
S3T	3.2-7.2	Beta_4
S4T	2.5-6.5	Beta_1
S5T	3.0-6.0	Beta_1
S6T	3.1-5.1	mu_{low}
S7T	2.9-5.9	Beta_4
S8T	2.6-6.6	mu_{high}
S9T	3.5-7.5	Beta_3
S10T	2.5-5.5	Beta_4
S11T	3.3-7.3	Beta_1
S12T	2.9-4.9	mu_{low}
S13T	2.3-5.3	Beta_2
S14T	2.0-6.0	Beta_2
S1E	2.2-6.2	Beta_5
S2E	3.4-7.4	Beta_4
S3E	2.5-6.5	Beta_5
S4E	3.2-5.2	Beta_5
S5E	2.2-6.2	Theta
S6E	2.3-4.3	mu_{high}
S7E	2.3-6.3	mu_{high}
S8E	2.3-4.3	Theta
S9E	3.2-7.2	mu_{high}
S10E	2.1-5.1	Beta_3
S11E	3.3-7.3	mu_{high}
S12E	2.0-6.0	Beta_4
S13E	2.4-5.4	Beta_4
S14E	2.1-5.1	Beta_5

Table 5.7: Optimal time window and frequency band values for the individual subjects in dataset SMR.

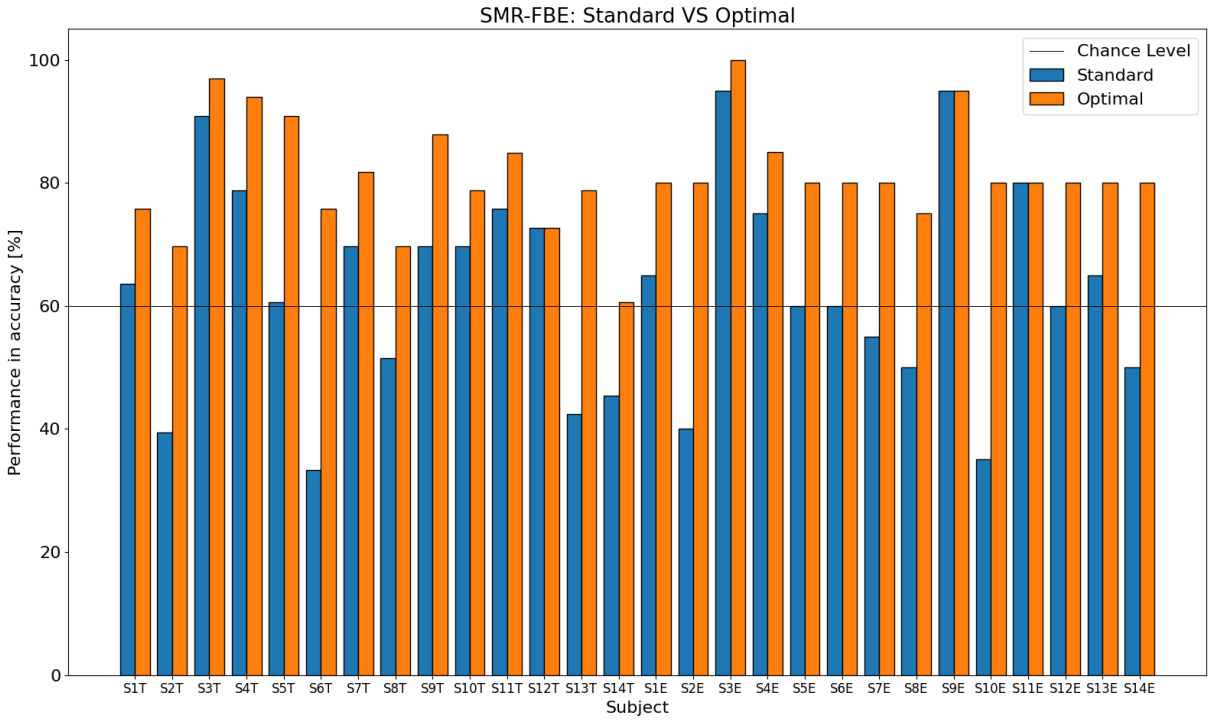


Figure 5.15: Subject-dependent classification performance using standard vs optimized parameter values, for dataset SMR. Using RF with FBE.

From the results, one can observe that all subjects obtain a performance above the corresponding the CL when the optimal band-window combination is utilized. As this is a fact for all four datasets, this optimization method shows great generalization ability, as well as efficiency. It can also be seen that there is a particularly greater increase in performance for the previously low-performing subjects, as compared to the high performers. This is especially noticeable when looking at the performance increase in the subjects in dataset IV2a for binary classification, in Figure 5.13, compared to the subjects in dataset IV2b, in Figure 5.11. On average the performance increase in low performers is 19.23%, ranging from 6.33% to 45.00%, while high performers have an average increase of 10.21%, ranging from 0.43% to 21.82%. Thus, this optimization method is robust and especially efficient for low-performing subjects.

Regarding the parameters, there is a significant variation in the optimal band-window combinations amongst the subjects, as evident from Tables 5.3, 5.4, 5.5, 5.6, and 5.7. This may be a consequence of the general subject variability in the performance prior to optimization, as well as the individual's MI ability.

5.4.2 Subject-Independent Optimization

In addition to the above subject-dependent optimizations, a subject-independent optimization is performed on all the datasets. This experiment is conducted by combining all subjects in each dataset, then performing the same procedure as for the subject-dependent optimization, i.e. evaluating the RF performance of all band-window combinations. For each dataset, the identified optimal frequency band and time window are given in Table 5.8. Dataset IV2a is optimized both for multiclass MI, using all four classes, and binary MI, using left and right-hand only, while dataset NTNU is only

optimized for binary left-hand and right-hand MI.

Dataset	Optimal Window [sec]	Optimal Band
IV2b	3.4-7.4	mu_{high}
IV2a Multi	2.5-4.5	mu_{high}
IV2a Binary	2.7-4.7	mu_{high}
NTNU	0.7-2.7	mu_{high}
SMR	3.5-6.5	mu_{high}

Table 5.8: The optimal windows and frequency bands when performing subject-independent optimization on the entire datasets. *IV2a Multi* refers to using all four classes, while *IV2a Binary* refers to only using left and right-hand MI.

The classification performances when using these optimal parameter values compared to the standard ones, are shown in Figure 5.16.

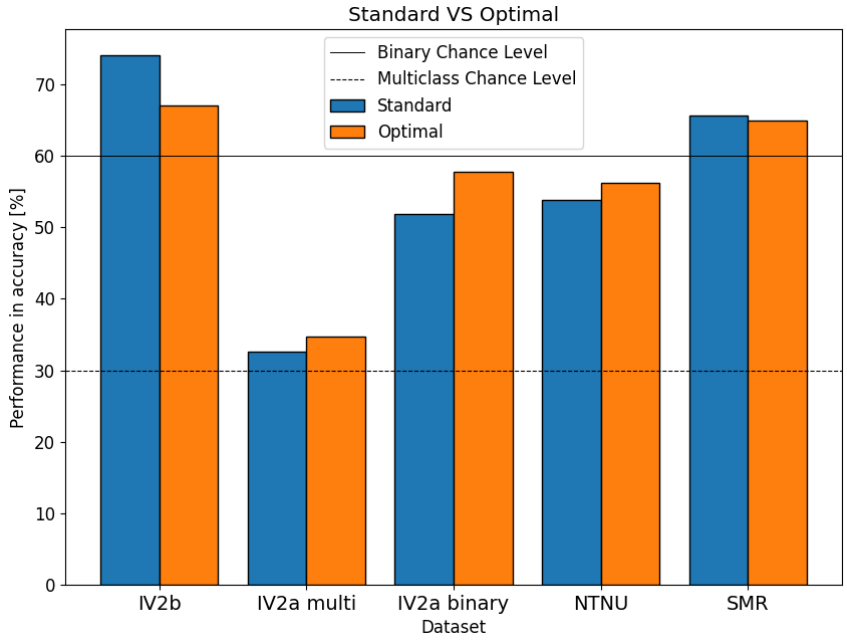


Figure 5.16: Subject-independent classification performance using standard vs optimized parameter values, for each dataset. *Binary Chance Level* corresponds to dataset IV2b, IV2a binary, NTNU, and SMR, while *Multiclass Chance Level* corresponds to dastaset IV2a multi.

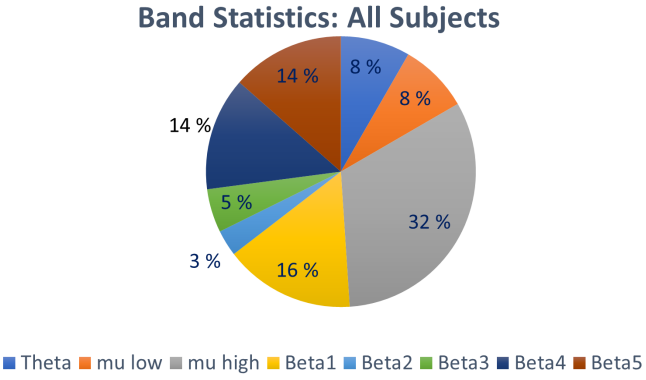
From the results, it can be observed that most datasets will get an increased performance when utilizing their optimal band-window combination. Despite dataset NTNU and dataset IV2a binary still keeping a performance beneath their respective the CL, their performances have increased, thus indicating the significance of an optimization process. On the other hand, the already high-performing datasets, IV2b and SMR, did decrease in performance, indicating that the optimization process is more effective for lower-performing datasets, as seen from the results for datasets IV2a binary, IV2a multi, and NTNU. It is also worth noting that all datasets deemed mu_{high} as their optimal band for classification.

Overall, it can be observed that this optimization process was more effective for subject-dependent optimization rather than subject-independent optimization.

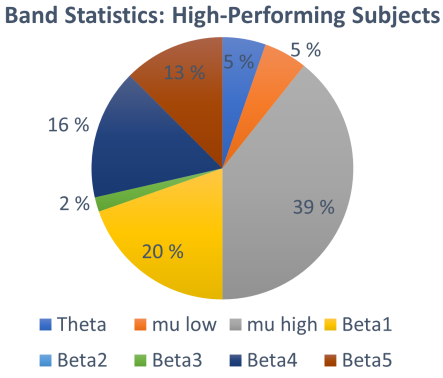
5.4.3 Optimization Statistics

To determine which frequencies are most important in the classification of MI, a visualization is created. Figure 5.17 presents a visualization of the frequency with which each band has been deemed the optimal one for a subject, expressed as percentages. Figure 5.17a presents the combined occurrences for all subjects, while Figures 5.17b and 5.17c present the occurrence of the bands in high and low performers, respectively.

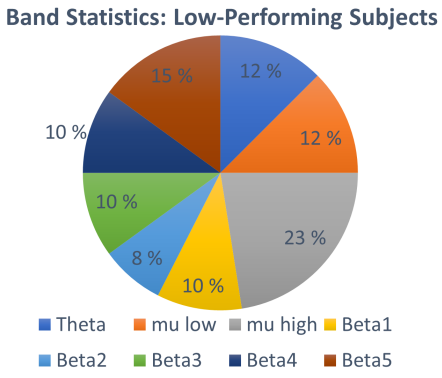
In Figure 5.17a, the numbers indicate what percentage of all the subjects have chosen the corresponding band as their optimal band, for example, 32% for mu_{high} means that 32% of all subjects, in all datasets combined, have deemed mu_{high} as their optimal band. The numbers in figures 5.17b and 5.17c, indicate the same thing, but the subjects are split into high and low performers, respectively. For example, in Figure 5.17b the number 20% for Beta_1 means that 20% of the high performers have Beta_1 as their optimal band, while 2% for Beta_3 means that only 2% of the high performers have deemed Beta_3 as their optimal band, and so on. The same goes for Figure 5.17c, only the percentages are given for the amount of low-performing subjects.



(a) Frequency band distribution for all subjects



(b) Frequency band distribution for high-performers



(c) Frequency band distribution for low-performers

Figure 5.17: Statistics of occurrence of each frequency band as the optimal band for a subject, in all datasets combined. Illustrated for all the subjects, and high- and low-performing subjects respectively. The percentages indicate the proportion of the subject group who deemed each corresponding band as their optimal one.

As seen from the statistics, most of the subjects, almost 48%, got an optimal band close to the upper mu band frequency, i.e. mu_{high} or Beta_1. There was also an accumulation around the upper Beta band, i.e. Beta_4 and Beta_5, which can be explained by the fact that higher cognitive abilities, such as concentration, are activated close to this frequency area, as seen in Table 2.1 in Section 2.1.3.

When looking at the statistics for the high performers there is a larger accumulation around the upper mu band and upper Beta band, where a total of 88% of the high-performing subjects had an optimal band in these frequency areas. In contrast, the low performance had a wider spread around the frequencies, and no clear accumulation can be observed. It is also worth mentioning that for dataset IV2b, which is the highest-performing dataset, a total of 86% of the high-performing subject had an optimal band of either mu_{high} or Beta_1. Again, this indicates a correspondence between the upper mu band, and MI task recognition, as well as a connection between high performers and the optimal frequency band.

5.5 Experiment 5: Channel Reduction using Greedy Backward Elimination Algorithm

The following experiments are performed to investigate how the performance of the classifier is affected by reducing the number of channels. The fact that dataset IV2a has a significantly poorer performance than dataset IV2b, while it used five times the number of electrodes, might be an indication that some electrodes are redundant or increase the subject variability. Reducing the number of electrodes has many benefits, which include reducing the amount of data to process, the cost of processing, and making the BCI system more comfortable for the user. Finding a reduced, and hopefully, a higher performing subset of electrodes is thus wanted. To reduce the number of channels, a backward greedy elimination algorithm is utilized, and RF built with FBE using the standard extraction parameters as presented in Section 4.1, is used to calculate and compare the performances.

The experiments are conducted for the datasets IV2a, NTNU, and SMR. For the IV2a and SMR datasets, all the classes in the datasets are used. The experiment is not conducted with dataset IV2b because of the already small number of channels.

Considering that the results from the previous experiments show significant subject variability, the individual characteristics of each subject are considered when reducing the channels. On the other hand, it is also interesting to see if channel reduction has an effect when combining the subjects as well. Therefore, channel reduction is both done for *subject-dependent* and *subject-independent* classification. The order in which the channels are removed, in the algorithm, is used to assign channel weights to each channel for each subject. Finally, for the NTNU and IV2a datasets, with defined electrode mapping, a topographic map analysis is also conducted.

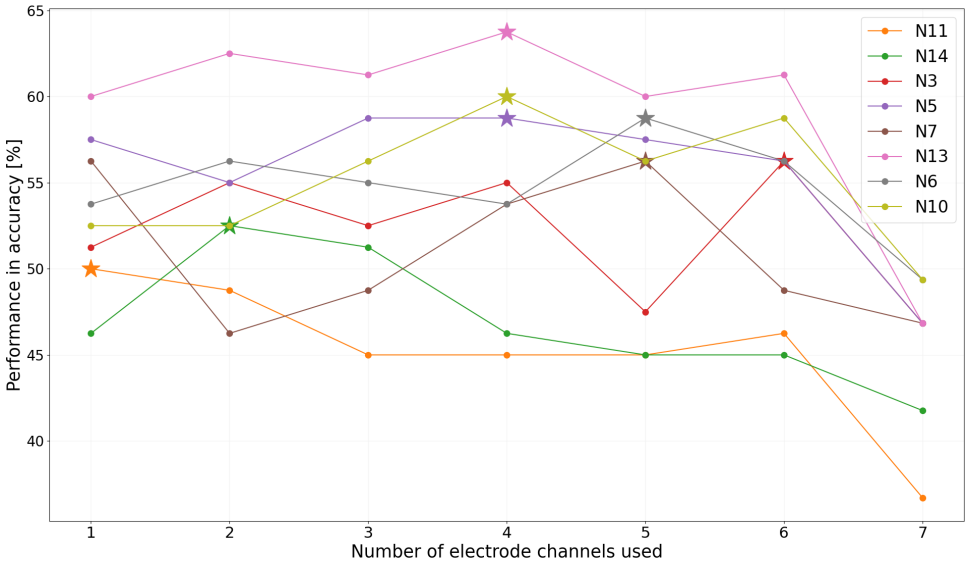
5.5.1 Dataset NTNU

The subset and placement of the channels used for channel reduction in dataset NTNU are given in Figure 4.13. Due to only six of the subjects in the NTNU dataset having recordings of feet MI and be-

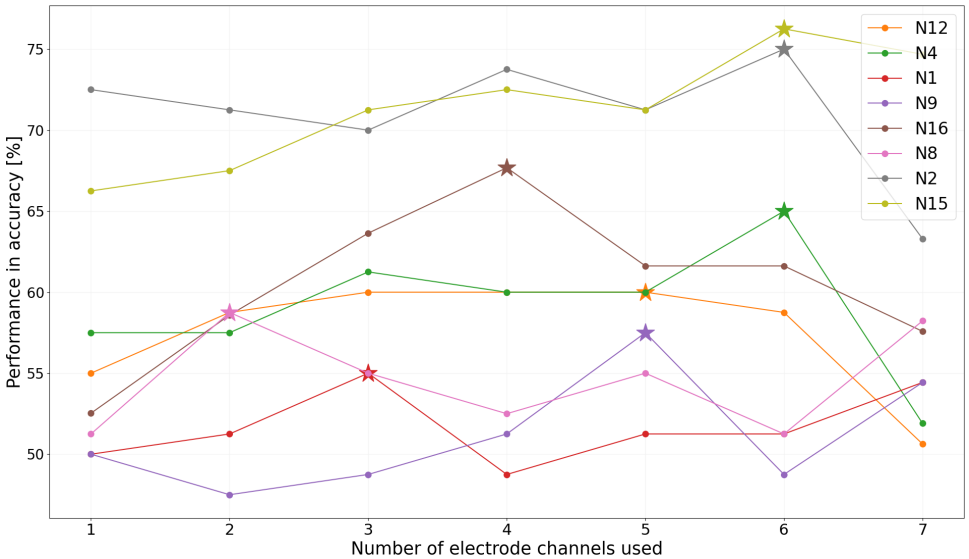
cause of class unbalance between feet MI and the other classes, the channel exploration is only done with the left-hand and right-hand classes. First channel reduction is performed for every subject, and then the channel weights are calculated from the results. Then, a topographic map analysis of a low and high-performing subject is given. Lastly, a subject-independent channel reduction is explored. The subjects are sorted, and presented, in two groups. The groups consist of a lower half performing group and a higher half performing group, based on their performance using the full set of channels.

Subject-Dependent Reduction

The accuracies for the different subsets of channels are given in 5.18, with the lower half performing subjects presented in Figure 5.18a and the remaining higher half performing subjects in Figure 5.18b. The highest achieved performance for every subject is indicated with a star.



(a) Lowest performing subjects



(b) Highest performing subjects

Figure 5.18: Obtained accuracies with different numbers of channels for the subjects in the NTNU dataset. Performance calculated using RF with FBE. The star indicates the highest achieved performance.

The performance of the subjects with the lowest performance, presented in Figure 5.18a, increases when the number of channels is reduced by one. The majority of the subjects have their highest performance when the subset of channels consists of 4 or 5 electrodes. The highest increase of performance was for subject N13, pink, and it increased with over 15%. Although the performance increased for all the subjects, only subjects N10 and N13 reached above the CL.

Every subject in the higher performing group, presented in Figure 5.18b, except N8 and N6, has an increased performance when removing one channel. For subjects N8 and N6, respectively the purple and pink lines, the performance increase only after removing more channels. The highest performance for every subject, marked with a star, is higher than the initial performance. For the majority of subjects, the peak performance is achieved when the channel subset consists of 5 or 6 channels, and the performance is reduced as the number of channels further decreases. For subjects N4 and N2, respectively the green and gray lines, the performance increase with 10% when using their optimal number of channels. Although the performances of subjects N9, N8, N12, and N1 increase, they never reach above the CL.

Channel Importance Weights

The visualization of the channel weights for all the subjects is given in Figure 5.19. The weights are based on when the channel was removed in the backward algorithm, where yellow represents the highest weight (7, i.e. the number of channels), while blue represents the lowest weight.

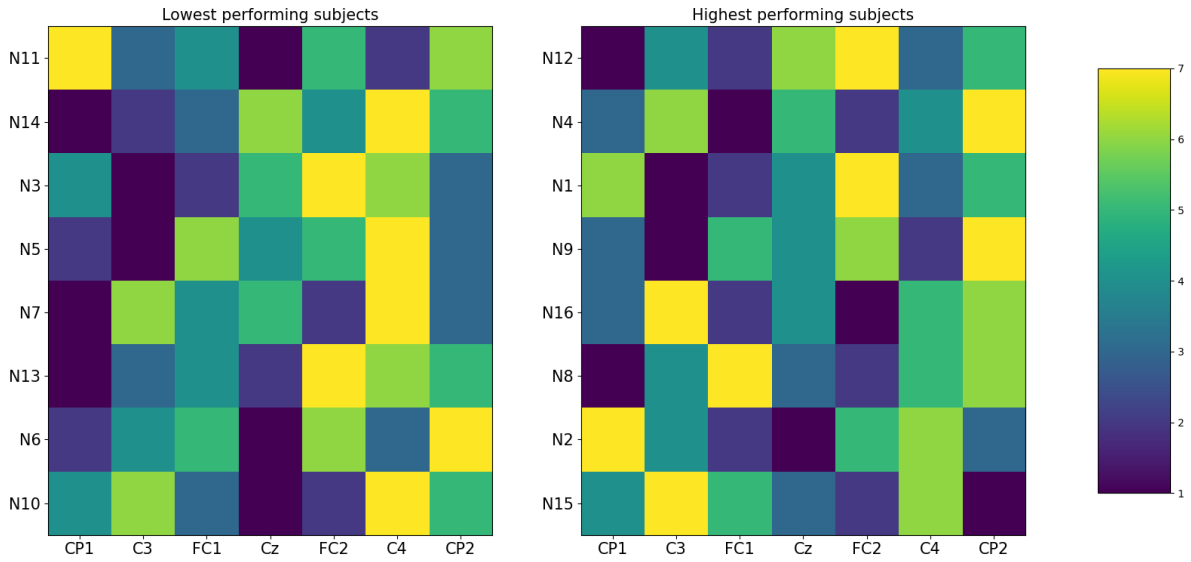
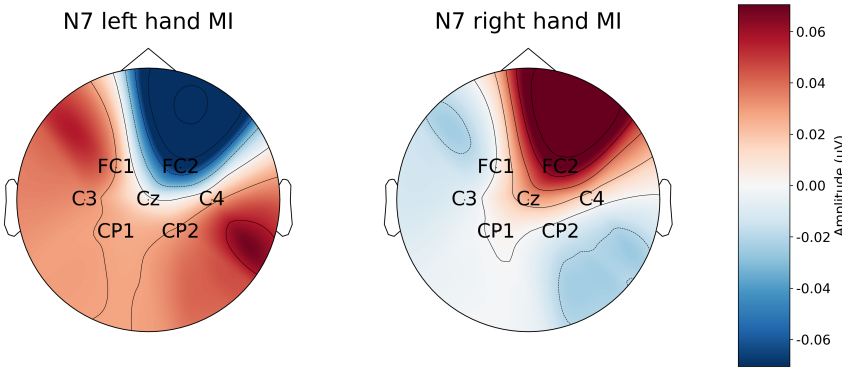


Figure 5.19: Importance weight for each electrode channel for each individual subject in dataset NTNU. Separated into lower and higher performing subjects, respectively.

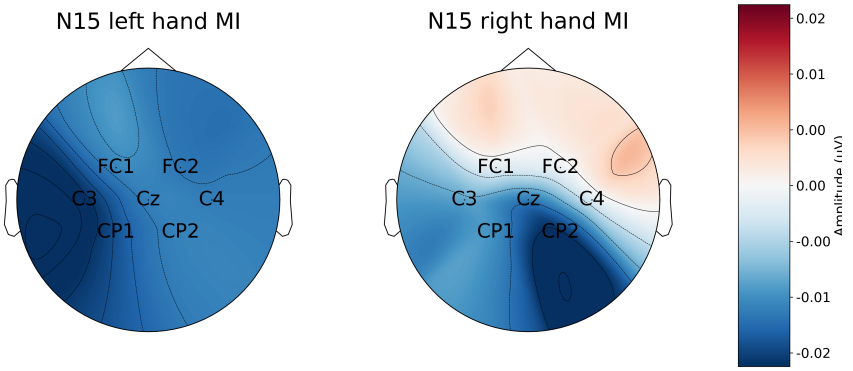
For the lowest performing subjects there is a higher weight on channels C4, GC2, and CP2, and a lower weight on channels CP1 and Cz. For the higher-performing subjects, the channels with the highest weight are CP2, FC2, and C3, while the channels Cz and FC1 are given the lowest weight.

Channel Topographic Analysis

The topographic map of a low and high-performing subject, respectively N15 and N7, is presented in Figure 5.20.



(a) Low-performing subject: N7



(b) High-performing subject: N15

Figure 5.20: Topographic map of high and low-performing subject in the NTNU dataset, respectively N15 and N7.

The largest difference between the two subjects is the interval in the measurement, where the lower-performing subject has an interval four times as large as the high-performing subjects. For the low-performing subject in figure 5.20a there is higher activity around channels FC2 and CP2, which switches polarity for the two MI classes. The areas with the least amount of activity in magnitude are located at channels Cz and CP1. Comparing the regions of the brain, the right frontal region has the most activity, while the left parietal side of the brain has lower activity.

For the higher-performing subject, with the topographic map presented in Figure 5.20b, the highest intensity switch between the left and right side of the brain for the two MI tasks. The largest difference between the two classes is seen at channels C3 and CP2 which lies in the high magnitude area, while for channels FC1, FC2, and C4 they switch polarity and become more positive.

Subject-Independent Reduction

The accuracies for the different subsets of channels for the subject-independent channel reduction are presented in Figure 5.21, where the highest performance is indicated by a star.

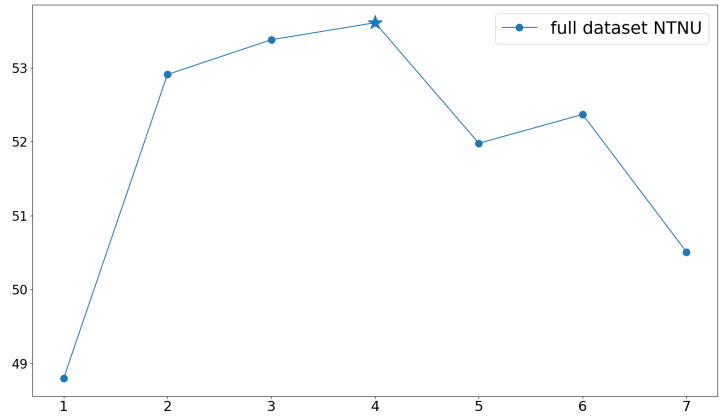


Figure 5.21: Obtained accuracies with different numbers of channels for a combination of subjects in the NTNU dataset. Performance calculated using RF with FBE. The star indicates the highest achieved performance.

The result for the subject-independent channel reduction shows that the highest performance is reached when using 4 channels. Although the performance increased, the relative increase is only 3% and still below the CL.

Comparing the effects of subject-dependent and subject-independent performance when using channel reduction, the reduction has a higher effect on the individual subjects. From the channel weights for the individual subjects, it can be observed that each subject assigns different weights to different channels, which will affect the results for the combined subjects. Additionally, the topographic map analysis shows that low and high-performing subjects use different regions of the brain, on average, when creating an MI. This could be the reason for the highly subject-dependent results.

5.5.2 Dataset IV2a

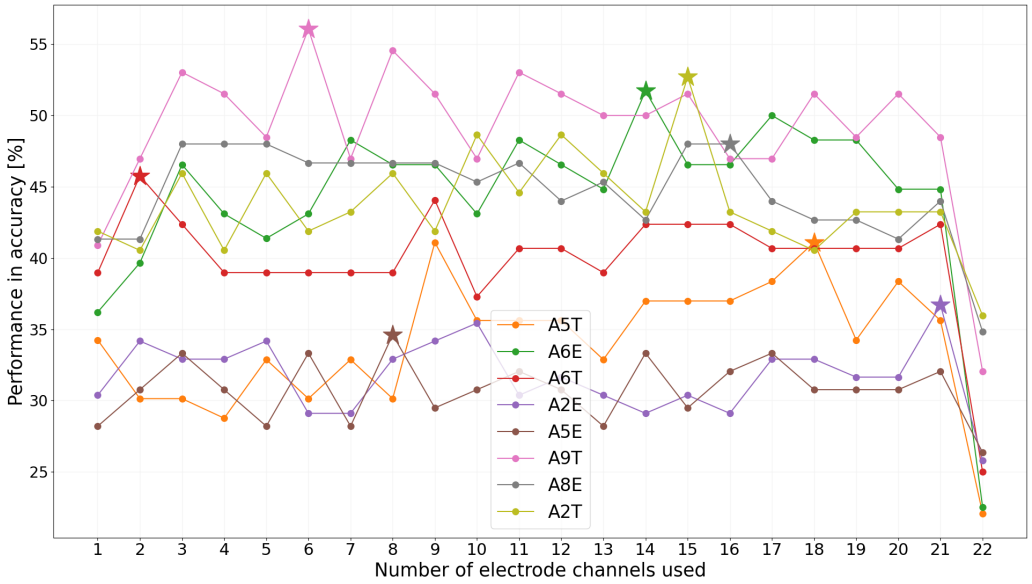
The channel exploration for the IV2a dataset is done including all four classes in the dataset, and the channel locations are given in Figure 4.5. Since more classes are explored, compared to the other datasets, the set of optimal channels and electrode placements will be different.

First, a subject-dependent channel reduction is performed, where the results are used to calculate the channel weights. Then, using a low and high-performing subject, a topographic map analysis is performed, using all four classes in the dataset. Lastly, channel reduction is performed for subject-independent classification. The order of the subjects is sorted and presented based on their performance using the full set of channels.

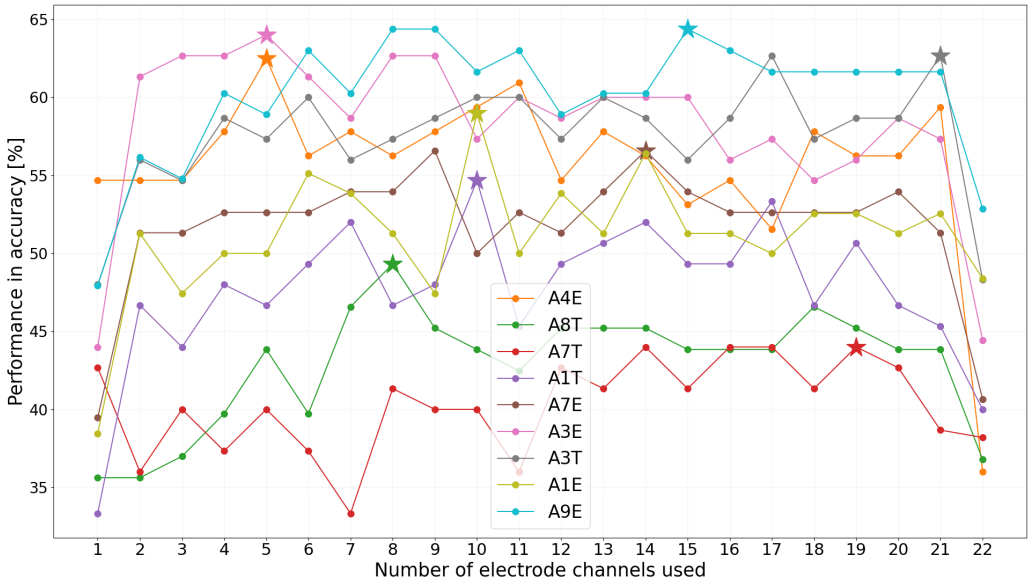
Subject-Dependent Reduction

The accuracy for the different subsets of channels is given in Figure 5.22, with the lower performing subject presented in Figure 5.22a and the higher performing subjects in Figure 5.22b. The highest

achieved performance for each subject is indicated with a star.



(a) Lowest performing subjects



(b) Highest performing subjects

Figure 5.22: Obtained accuracies with different numbers of channels for the individual subjects in the IV2a dataset. Performance calculated using RF with FBE. The star indicates the highest achieved performance.

For all the subjects with the lowest initial performance, presented in Figure 5.22a, the performance increase when removing one channel, and every subject achieves a performance above the CL. For subject A6E, green line, the performance increases from below 25% to above 50%, which is an increase of over 25%. For the majority of the subjects the peak performance, indicated with a star, is achieved when the number of channels included in the subset is around 14.

For all the subjects with the highest initial performance, presented in Figure 5.22b, the performance increases when removing one channel and the highest increase can be seen for subject A4E, orange

line, where the performance increases from below 40% to above the CL. After the highest peak is reached, indicated with a star, the performance reduces more drastically than it increased.

Channel Importance Weights

The visualization of the channel weights for all the subjects is given in Figure 5.23. The weights are based on when the channel was removed in the backward algorithm, where yellow (weight of 22, i.e. the number of channels) has the highest and best weight, and blue has the lowest weight.

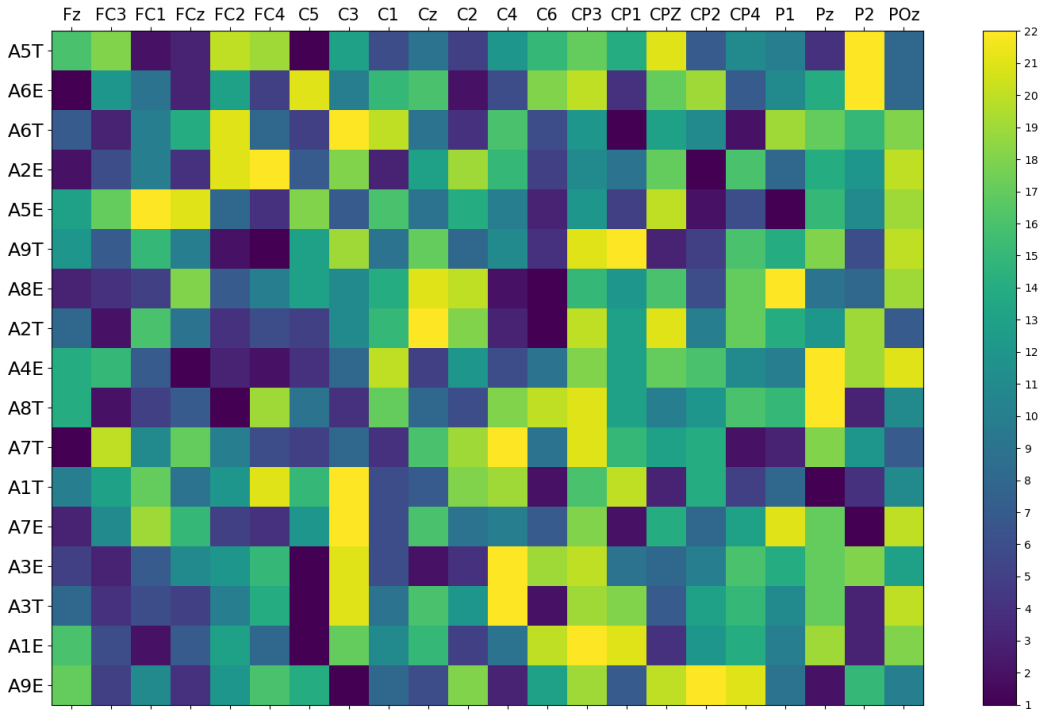
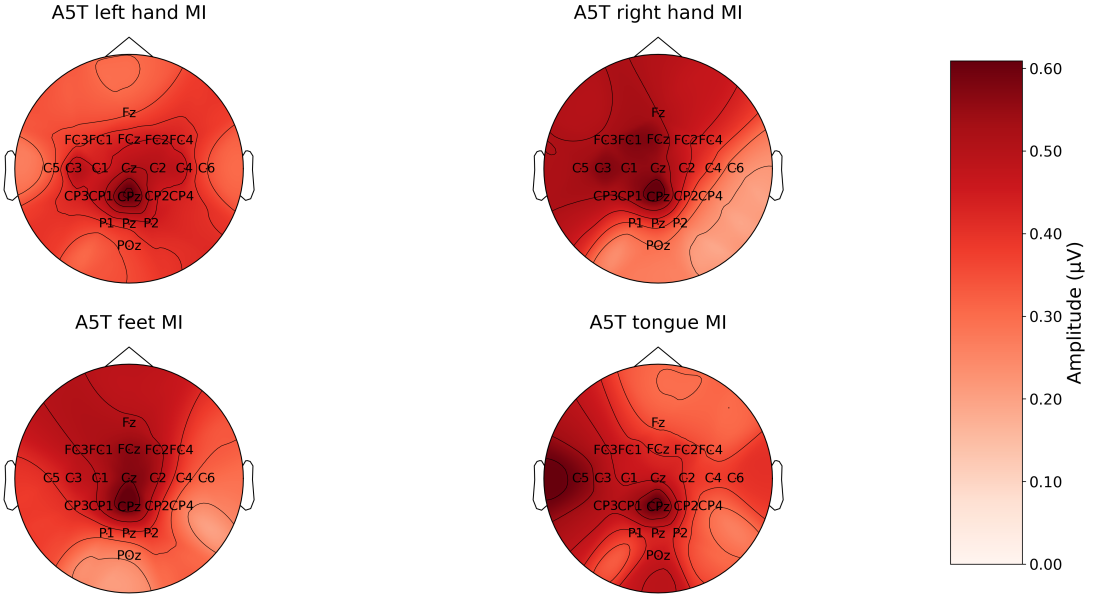


Figure 5.23: Importance weight for each electrode channel for each subject in IV2a dataset. Ranging from the lowest performing to the highest performing subject.

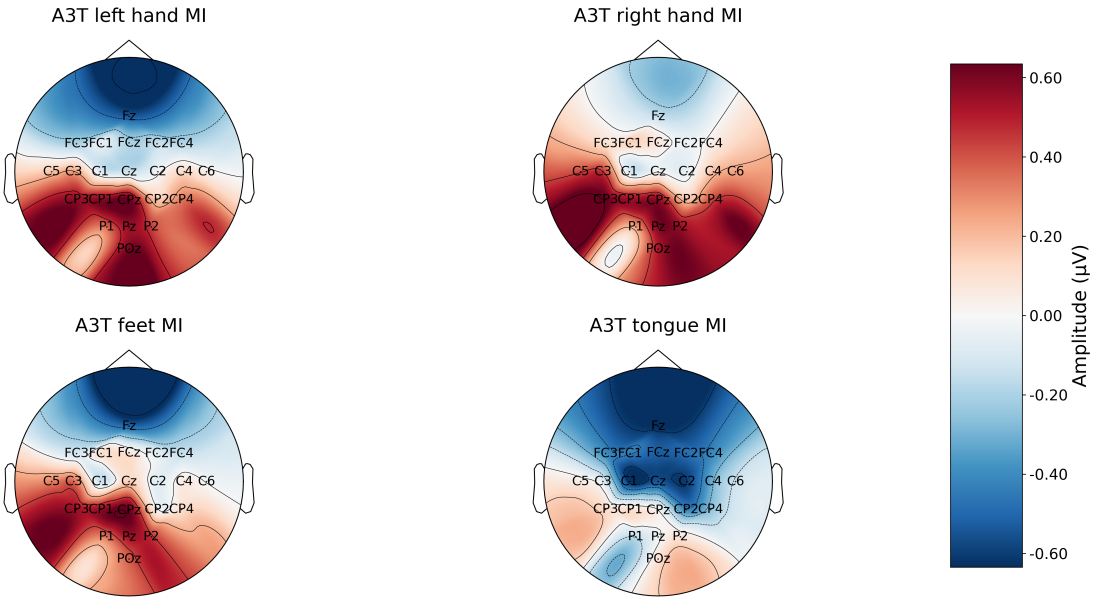
The low-performing subjects, shown in the upper half of the figure, has a higher weight on channels P2, FC2, and CP3, and a lower weight on channels C6, Fz, and CP2. The higher performing subjects, shown in the lower half of the figure, have a higher weight on channels C3, CP3, and C4, and a lower weight on channels C5, C1, and FC1.

Channel Topographic Analysis

The topographic maps for a low and a high-performing subject, respectively A3T and A5T, are given in Figure 5.24.



(a) Low performing subject: A5T



(b) High performing subject: A3T

Figure 5.24: Topographic map of high and low-performing subject in the IV2a dataset, respectively A3T and A5T.

The magnitude interval for the low-performing subject is half of the interval of the high-performing subject and spans only the positive parts. For every class for the low-performing subject, presented in Figure 5.22a, there is a high intensity around channel CPz, making the topographic map look like a geographical mountain with CPz as its peak. Visually, it is challenging to differentiate between the topographical maps for each class. For right-hand MI there is a lower magnitude in the right parietal lobe, especially for electrodes C6 and CP4. Feet MI looks similar to right-hand MI but with a more positive measurement for channel P2. Tongue and left-hand MI also look similar in intensity, where

the intensity is centered around CPz. For tongue MI the measurement is higher around POz and C5.

The topographic map for the high-performing subject in Figure 5.24b shows more variation across the brain. Each of the classes has a high positive magnitude in the parietal lobe, with a negative dip around channel P1. On the other hand, there is a high negative magnitude around the frontal lobe, especially around channel Fz. There is also a negative magnitude around the channel C1 and C2, for all the classes, while the lower surrounding channels are positive in magnitude.

Subject-Independent Reduction

The obtained accuracies from the channel reduction using combined subjects are presented in Figure 5.25, where the start indicates the highest achieved performance.

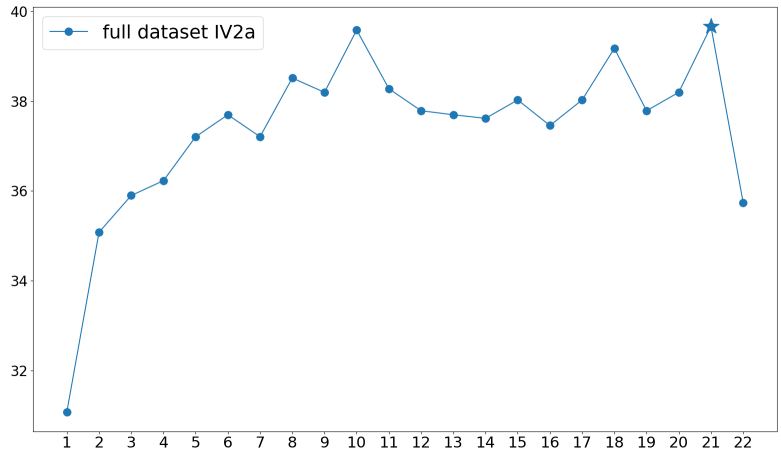


Figure 5.25: Obtained accuracies with different numbers of channels for a combination of subjects in the IV2a dataset. Performance calculated using RF with FBE. The star indicates the highest achieved performance.

The results for the subject-independent channel reduction show that the highest performance is obtained when using 21 channels. The relative increase is only 3% from the initial accuracy when using the full set of channels. The performance is below the CL for every subset of channels.

5.5.3 Dataset SMR

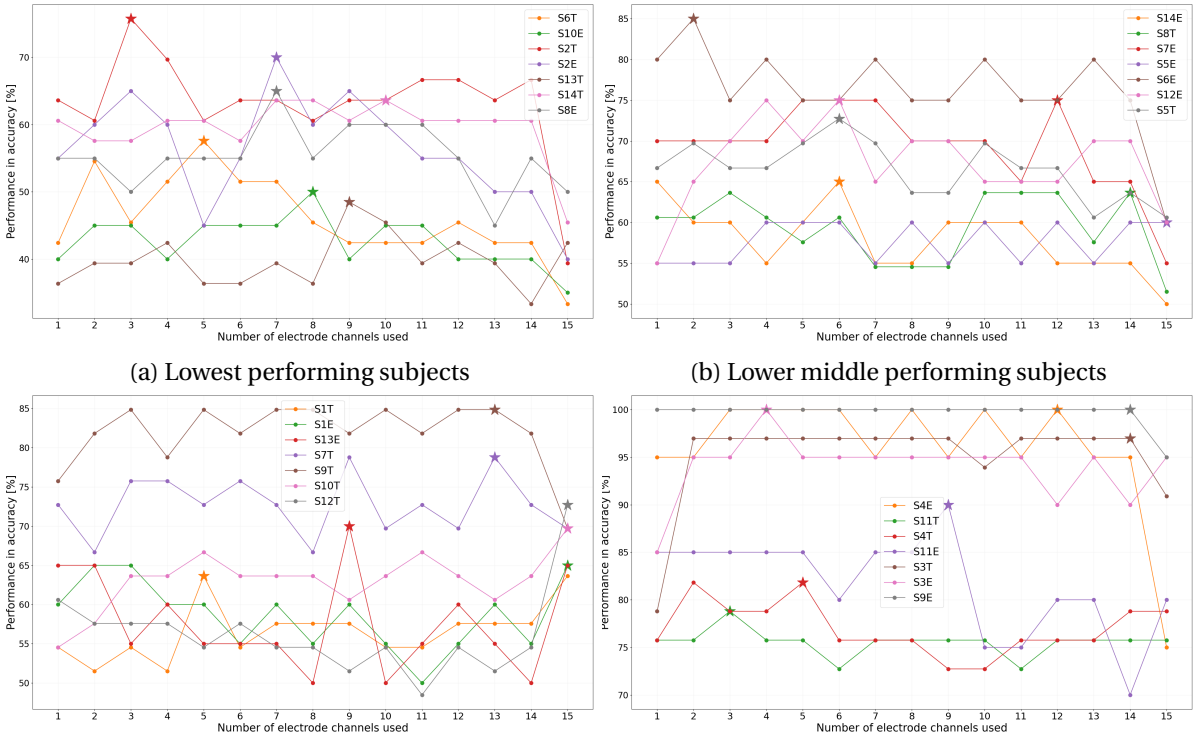
The SMR dataset contains 15 electrode channels placed as shown in Figure 4.9. The SMR dataset contains a subset of the classes that are contained in the IV2a dataset, which makes it interesting to see if the optimal number of channels and electrode placement will be different.

First, a subject-dependent channel reduction is performed, and the results are used to calculate the channel weights for each subject. Due to the SMR using the Laplace derivations electrode placement scheme as well as unlabeled channels in the data, it is unrealizable to do a topological map analysis. Lastly, channel reduction is performed for subject-independent classification. The order of the subjects is sorted and presented based on their performance using the full set of channels. Due to the amount of subjects in the dataset, the subjects are split into lowest performing, lower middle performing, higher middle performing, and highest performing subject groups, to make the plots more

readable.

Subject-Dependent Reduction

The accuracy for the different subsets of channels is given in Figure 5.22. More specifically, with the lowest performing subjects in Figure 5.26a, lower middle performing subjects in Figure 5.26b, higher middle performing subjects in Figure 5.26c and highest performing subjects in Figure 5.26d.



(a) Lowest performing subjects

(b) Lower middle performing subjects

(c) Higher middle performing subjects

(d) Highest performing subjects

Figure 5.26: Obtained accuracies with different numbers of channels for the individual subjects in the SMR dataset. Performance calculated using RF with FBE. The star indicates the highest achieved performance.

The performance for the seven lowest-performing subjects, is presented in Figure 5.26a. The performance increase for every subject, except for S13T in brown, when removing one channel. For subject S2T, in red, the performance increases from below 40% to above the CL, and then it peaks at above 70% which is a 30% increase. For all of the subjects, except S13T, S10E, and S6T, the performance increased from below the CL to above.

For every subject in the lower middle performance group, presented in Figure 5.26b, except S5T in purple, the performance increases when removing the first channel. For all the subjects starting below the CL, namely S14E, S8T, and S7T, the performance increase to above the CL when using their optimal channel subset. Subject S6E, in brown, increases from just above the CL to 85% with its optimal channel subset.

For the subject in the higher middle performing group, presented in Figure 5.26c, does the performance when removing one channel only increase for the subjects S7T and S9T, while for the other 5

subjects, the highest performance is achieved when all the channels are included. For some channel subsets, the performance of subject S12T is decreased from above 70% to below 50%. On the other hand, the performance of subject S9T increased by 15%, reaching around 85%, with its optimal channel subset.

When removing one channel from the highest performing group, presented in Figure 5.26d, the performance only increases for subjects S9E, S3T, and S4E, and the performance reaches above 95%. For the other subjects, the performance stays constant or decreases. The highest performance for each subject is nevertheless higher than the initial performance when all the channels were included.

Channel Importance Weights

The visualization of the channel weights for all the subjects is given in 5.27. The weights are based on when the channel was removed in the backward algorithm, where yellow (weight of 15, i.e. the number of channels) has the highest and best weight, and blue has the lowest weight.

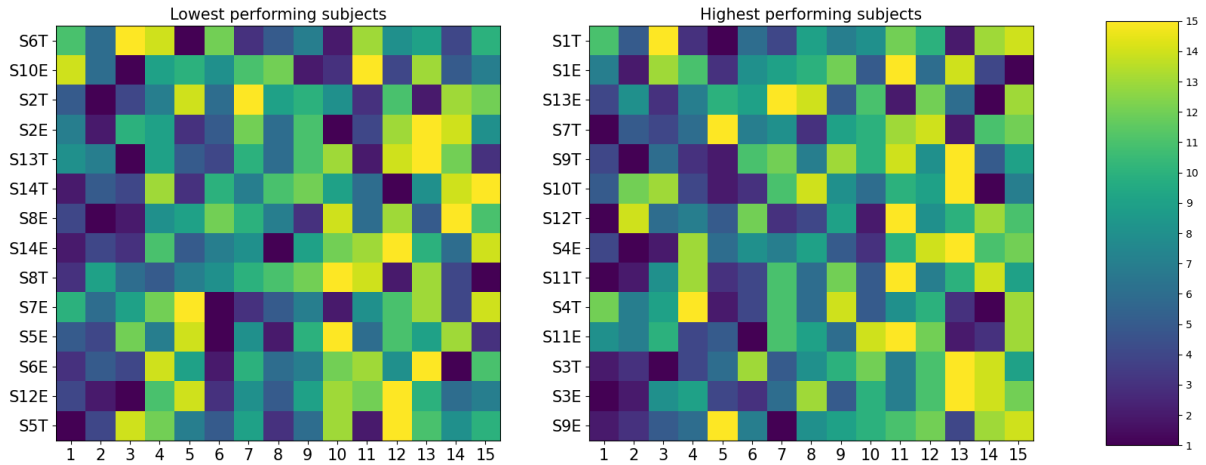


Figure 5.27: Importance weight for each electrode channel for each subject in SMR dataset. Separated into lower and higher performing subjects, respectively.

The subjects with the lowest performance weighted channels 10, 12, and 13 the highest, and channels 1, 2, and 3 the lowest. For the subjects with the highest performances, channels 11, 13, and 15 are given the highest weight, while channels 1, 2, and 6 are weighted lowest.

Subject-Independent Reduction

The obtained accuracies from the channel reduction using combined subjects are presented in Figure 5.28, where the star indicates the highest achieved performance.

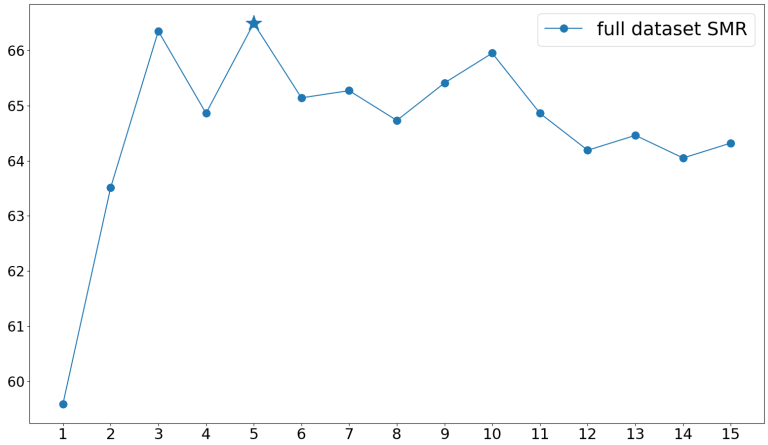


Figure 5.28: Obtained accuracies with different numbers of channels for a combination of subjects in the SMR dataset. Performance calculated using RF with FBE. The star indicates the highest achieved performance.

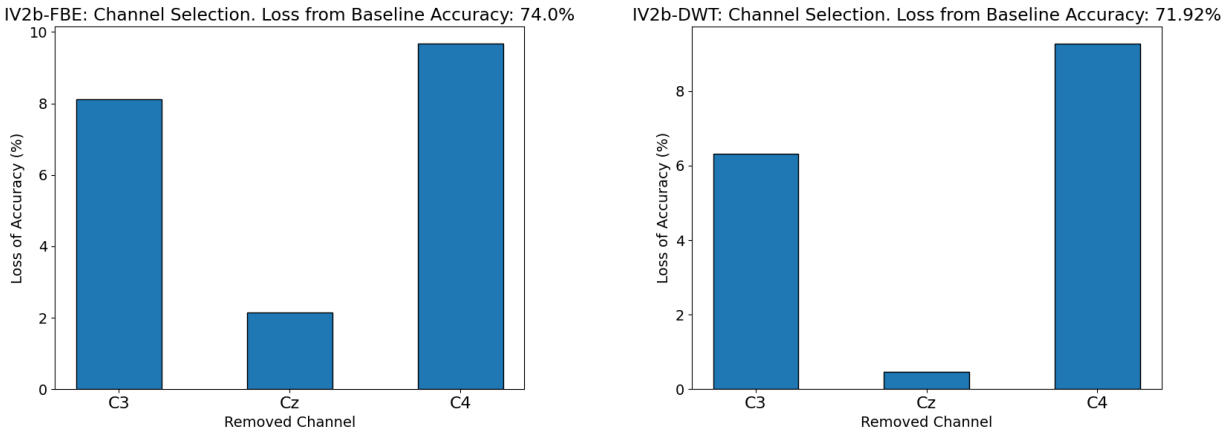
The results from the subject-independent channel reductions show that the highest performance is obtained when the number of channels is reduced to 5 channels. The performance is varying between 64-66 %, but when the number of channels is less than 3 the performance decreases just below the CL.

5.6 Experiment 6: ERP and PSD Analysis

It is advantageous to be able to predict the performance of a subject prior to extensive processing and classification, in order to save time and take relevant measures, such as adaptation of MI training. Therefore, it is interesting to see if there is a difference in the ERPs and PSD of a high-performing subject versus a low-performing subject. Thus, the ERP and PSD are calculated for the best and worst performing subjects, namely B4T and B3T, for the best-performing dataset, dataset IV2b.

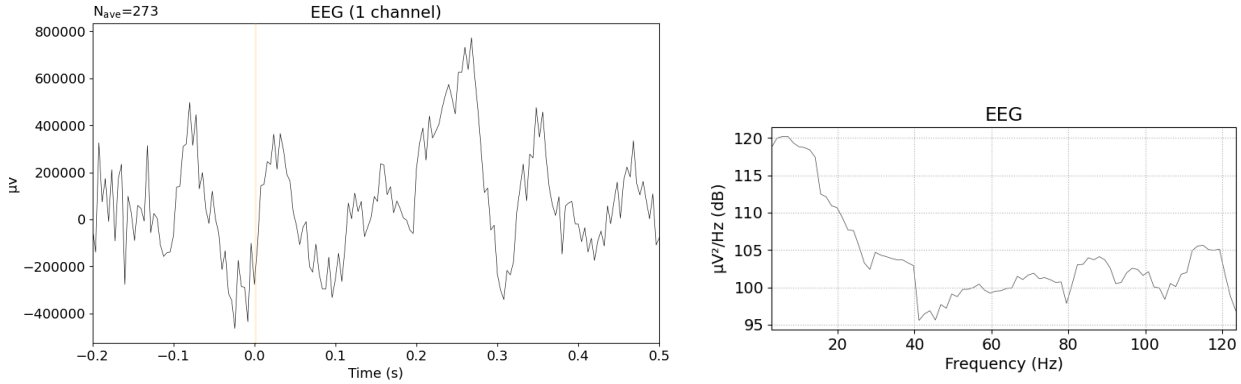
As the ERP should be calculated for the most important channel, a loss-of-accuracy analysis is performed. This is done by removing a channel, calculating the new accuracy, and comparing it to the original accuracy. This is done for each channel, and the loss-of-accuracy is given in Figure 5.29 using FBE and DWT decomposition with RF, respectively.

Based on the channel reduction results, it is evident that C4 is the most important channel, as its removal results in the greatest decrease in accuracy compared to channels C3 and Cz. Thus, an ERP and PSD analysis is performed using data from this channel for subjects B3T and B4T, i.e. the worst and best performing subjects. The event is the cue for the MI task appearing on the screen, which happens at second 3 in each trial. The resulting ERPs and PSDs are given in Figure 5.30, where the cue starts at second 0.0.

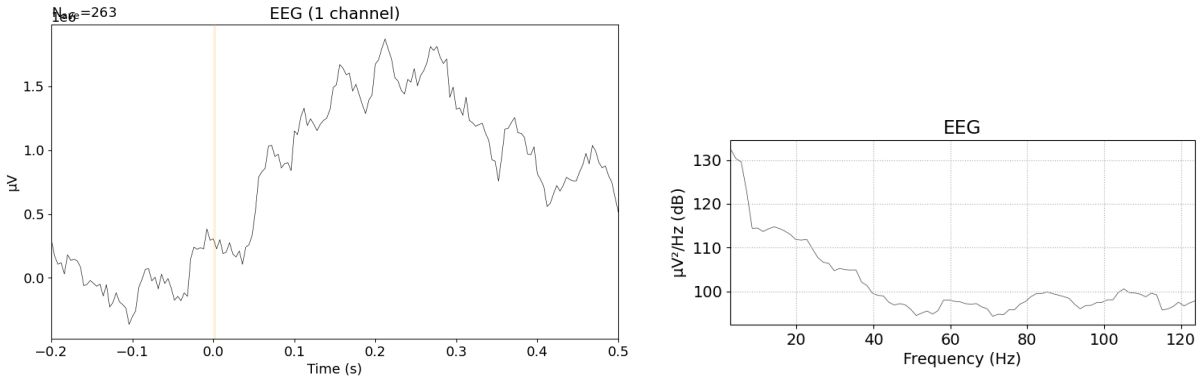


FBE DWT

Figure 5.29: Loss of accuracy after removing corresponding channel from dataset IV2b, using FBE and DWT decompositions, respectively. Channel Cz loses the least accuracy, while channel C4 loses the most accuracy, in both cases.



(a) ERP at cue and corresponding PSD for low-performing subject B3T



(b) ERP at cue and corresponding PSD for high-performing subject B4T. Note that the scale is 10^6.

Figure 5.30: ERP and PSD at the MI cue for the worst and best performing subject of dataset IV2b at channel C4. The cue is indicated by the orange line at time zero.

The results in Figure 5.30 indicate that there is a significant difference in the ERP signals for a low versus a high-performing subject. As shown in Figure 5.30b, the high-performing subject B4T exhibits an ERP signal which has a consistent increase after the cue at 0 seconds, followed by a peak and a con-

sistent decrease. On the other hand, as seen from Figure 5.30a, the ERP signal for the low-performing subject B3T exhibits no consistency and no obvious pattern, and the signal is more variable and random. Another clear difference between the two is the scaling of the signals. The high-performers have an almost doubled potential compared to the low-performers, suggesting that their MI is stronger than the low performer's MI signal, indicating a better and stronger response to the cue stimulus.

There is also a clear difference between the subject's PSDs. The high-performing subject, B4T, exhibits a PSD which is slowly decreasing until it reaches around 8Hz, stabilizes until around 25Hz, and decreases until 40Hz then it keeps a somewhat steady signal for the higher frequencies. On the other hand, the low-performing subject, B3T, has a PSD which is rapidly decreasing until 20Hz, then 40Hz, followed by a wavy and unsteady signal for the higher frequencies. These results indicate that for high-performing subjects the frequencies vary more around the mu and Beta bands, which are related to MI, and less for the higher frequencies, while low-performing subjects have a more spread distribution of their frequencies, which again may indicate that their MI abilities are poor.

5.7 Experiment 7: Classification with CNN

Considering the properties of CNN as both a feature extraction and classification method, it is interesting to explore the algorithm's performance compared to the intuitive ML classification with RF. The specific CNN model explored is EEGNet. The input to the EGGNet can both be raw unprocessed data or features extracted from decomposed data. Considering that the ML algorithm RF needs preprocessed data it is interesting to look at the performance of the DL classifier in both cases. Three different DL classifiers are built and inspected:

1. EEGNet with features extracted from decomposed data using FBE
2. EEGNet with features extracted from decomposed data using DWT
3. EEGNet with raw, unprocessed data

It is interesting to compare the various classification methods, to see if one can identify the most efficient approach. Thus the performance of the above DL approaches is compared to the performance of RF with features extracted from decomposed data using both FBE and DWT. This also includes performing the classification process in both *subject-dependent* and *subject-independent* approaches. Additionally, it is of interest to test the robustness of the proposed algorithms in Section 4.2 by inspecting how it performs on various combinations of MI classes. The different classification methods are thus tested on different combinations of left-hand, right-hand, feet, and tongue MI. This results in both a binary and multiclass exploration.

5.7.1 Subject-Dependent Binary Classification

Results from binary classification can give insight into the classification capabilities of the EEGNet, as well as which combinations of MI classes it might have difficulties with. By exploring the subject-dependent classification, one can also investigate EEGNet's ability to classify low and high-performing subjects, defined by their performance with RF. Two binary classification combinations are explored:

left- and right-hand and *right-hand and feet.*

The EEGNet classifier is constructed for all four datasets. The *left- and right-hand* classification is performed for the IV2b, IV2a, and NTNU datasets, and their performance is presented in Figures 5.31, 5.32 and 5.33, respectively. The second classification combination, *right-hand and feet*, is performed for the SMR, IV2a, and NTNU dataset, and their performance is presented in Figures 5.34,5.35 and 5.36. As the NTNU dataset only has *feet MI* recordings for six of its subjects, the classification is only performed for these six subjects. Finally, due to the class imbalance in the NTNU dataset, the confusion matrices for the classification combination are presented, in Figure 5.36, to further investigate the dataset's performance.

Dataset IV2b: Left-Hand and Right-Hand

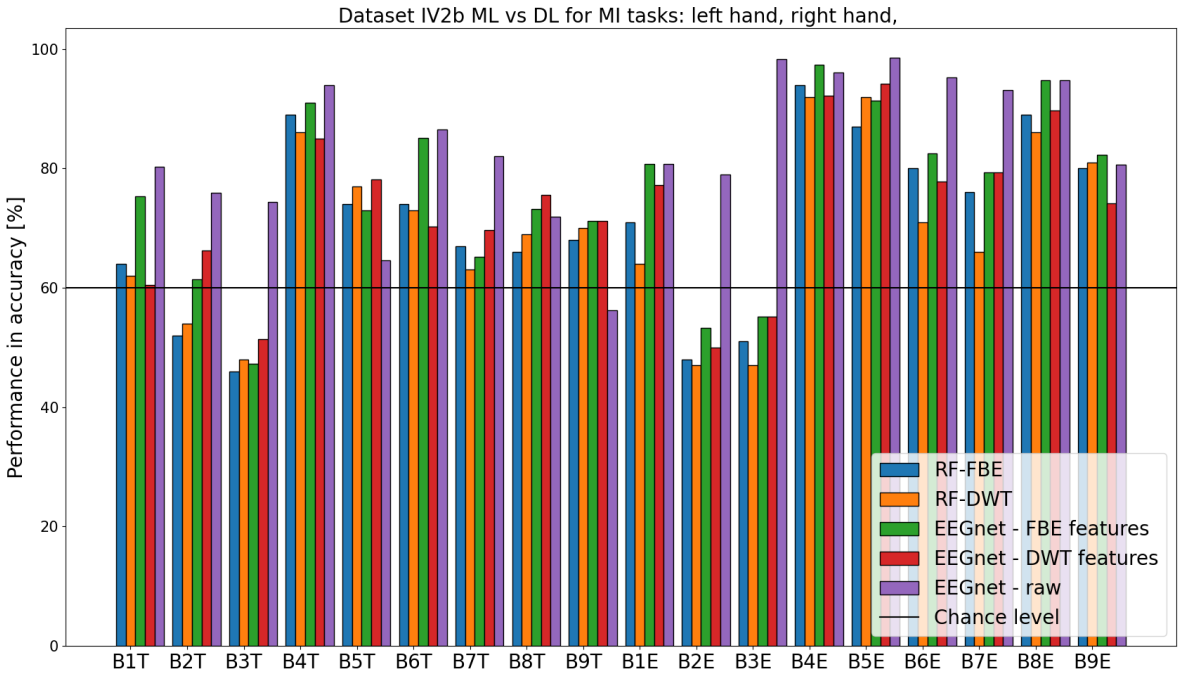


Figure 5.31: Binary classification of left and right-hand for each subject in the IV2b dataset.

Dataset IV2a: Left-Hand and Right-Hand

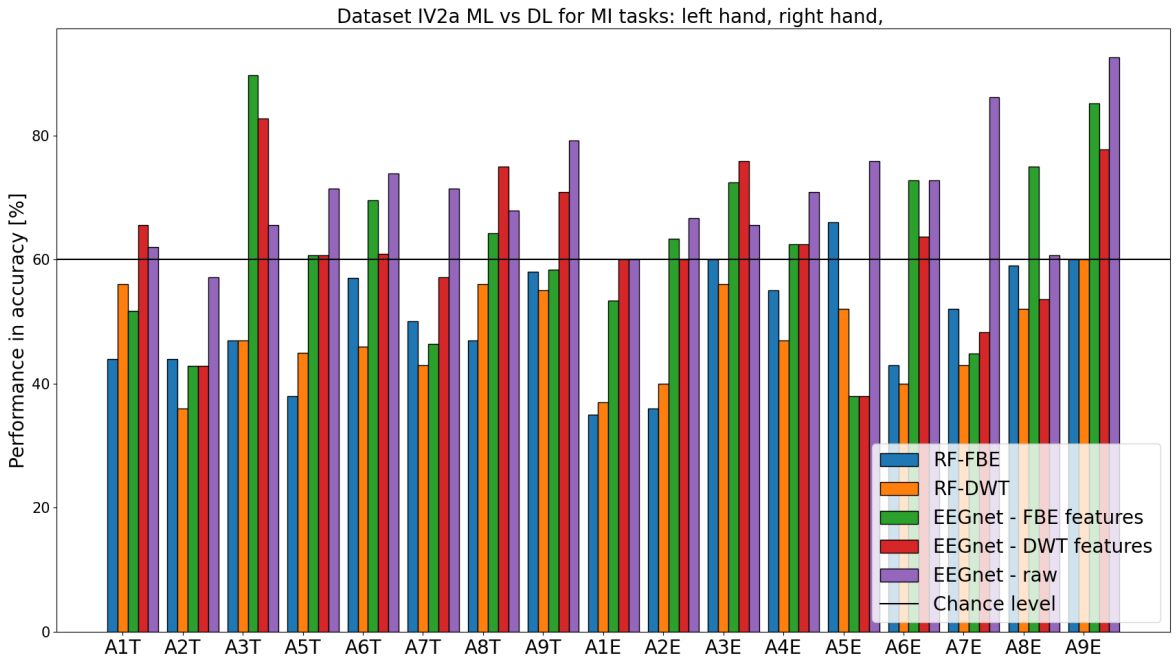


Figure 5.32: Binary classification of left and right-hand for each subject in the IV2a dataset.

Dataset NTNU: Left-Hand and Right-Hand

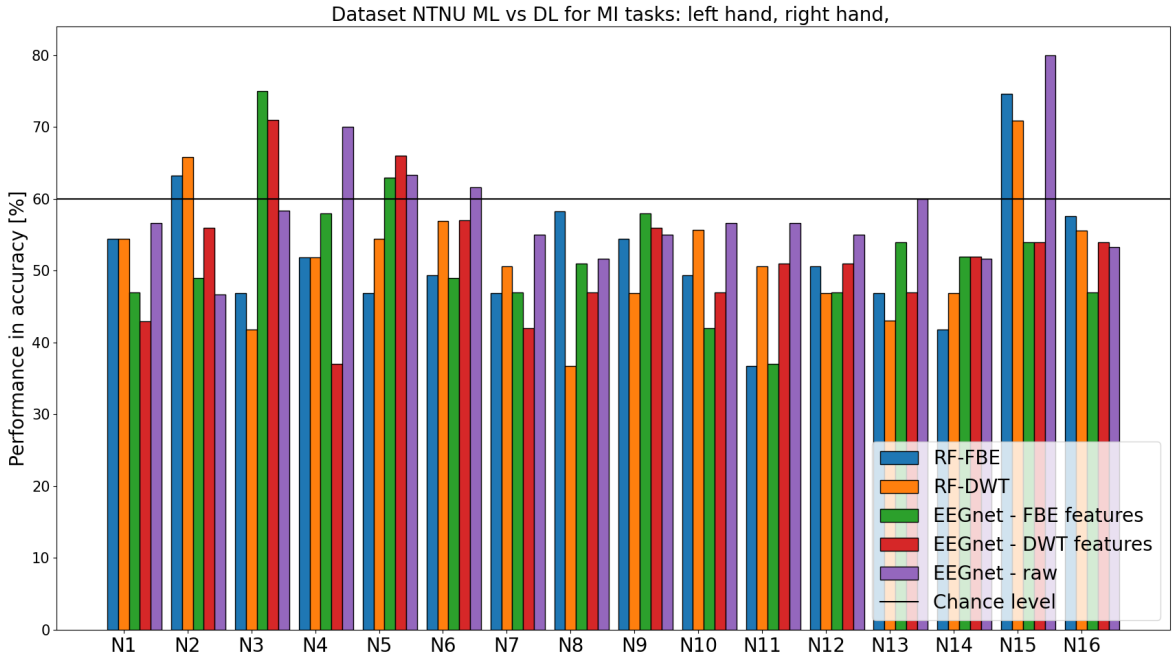


Figure 5.33: Binary classification of left and right-hand for each subject in the NTNU dataset.

Dataset SMR: Right-Hand and Feet

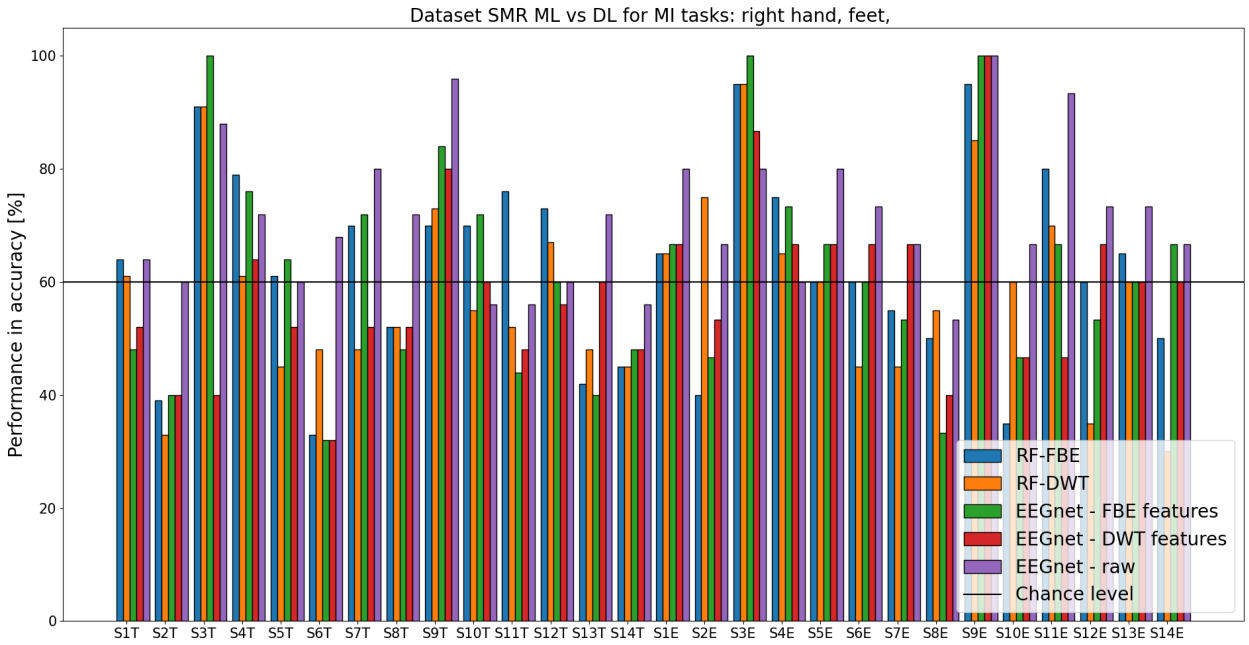


Figure 5.34: Binary classification of right-hand and feet for each subject in the SMR dataset.

Dataset IV2a: Right-Hand and Feet

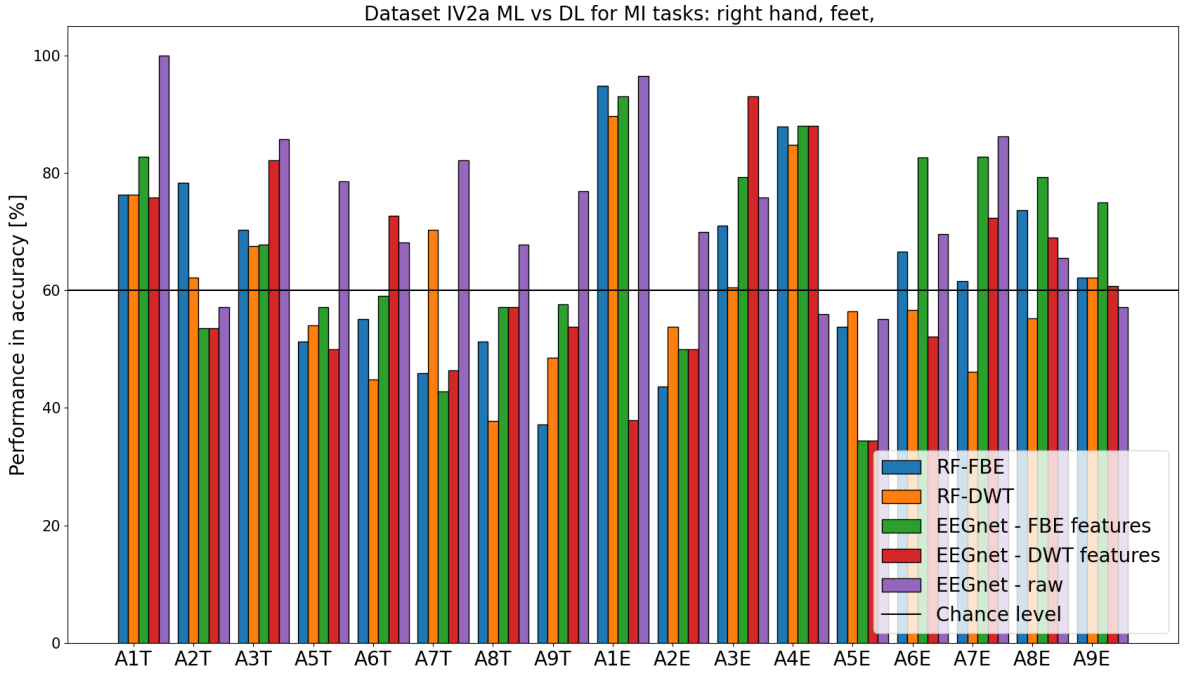


Figure 5.35: Binary classification of right-hand and feet for each subject in the IV2a dataset.

Dataset NTNU: Right hand and Feet

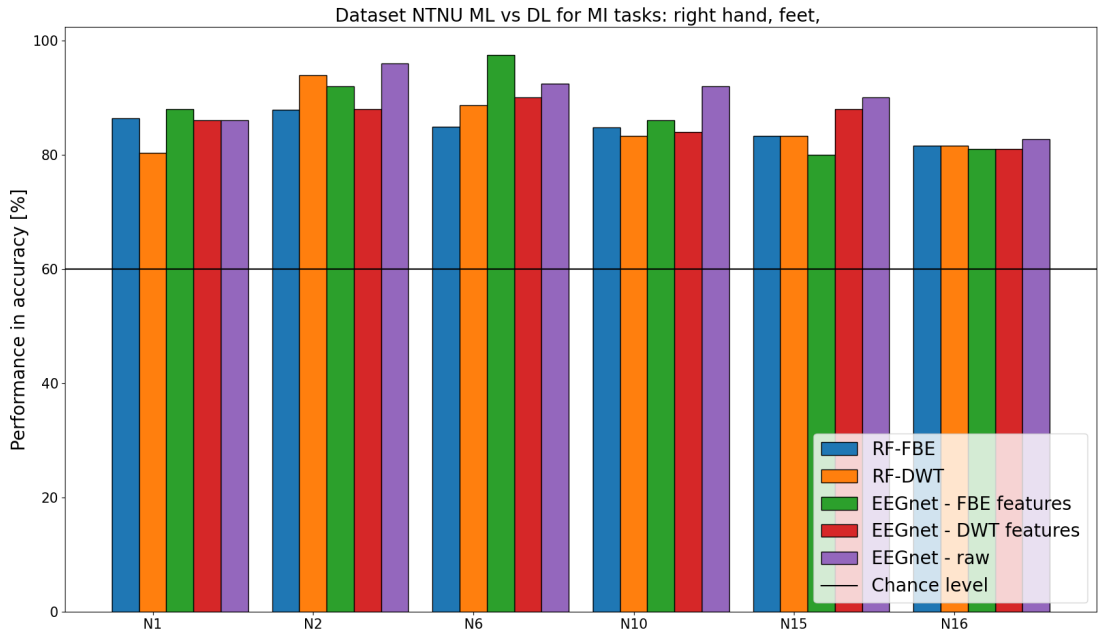


Figure 5.36: Binary classification of right-hand and feet for each subject in the NTNU dataset.

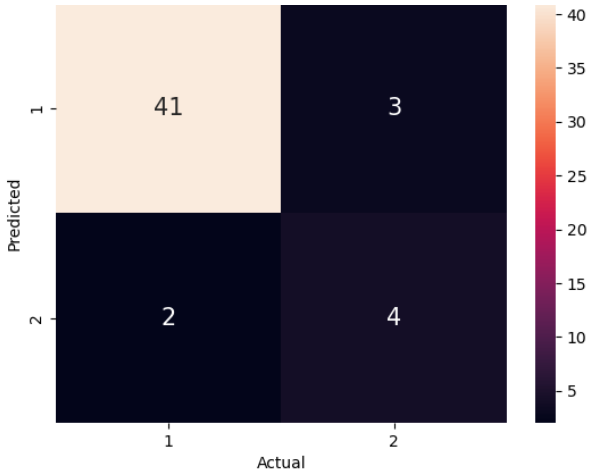


Figure 5.37: Confusion matrix resulting from subject-dependent CNN EEGNet raw data binary classification between right-hand and feet MI, for subject N15 in NTNU dataset. Label 1 refers to right-hand, while 2 refers to feet.

For left-hand and right-hand classification, the performance increased for the majority of the subjects in the IV2b dataset when using EEGNet, compared to RF. The four subjects with a performance below the CL when classifying with RF, namely B2T, B3T, B2E, and B3E, obtain a performance that is significantly over the CL when classifying with EEGNet using raw data, as their performances have increased with over 20%. On the other hand, the performance of the high-performing subjects B5T and B9T has decreased using the same method, as compared to the four other classification methods. For the rest of the subjects, the performance was higher for all EEGNet variations compared to the RF approaches.

For dataset NTNU it can be seen that EEGNet has a different effect on high and low-performing subjects when classifying between left and right-hand MI. For subject N2, which is one of two high-performing subjects when using RF, the performance decrease when classifying with any EEGNet method. For the other high-performing subject, N15, the performance decrease below the CL for EEGNet using decomposition, but it increases when using EEGNet with raw data. For most of the low-performing subjects in dataset NTNU, the performance stays below the CL when classifying with EEGNet. However, subjects N3, N4, and N5 perform below the CL when using ML classifiers, but performs above the CL when using DL. The remainder of the subjects never performs above the CL regarding of classification method. For subjects N6, N8, and N16 the performance decrease when using DL compared to ML, but for the six other low-performers DL increases the performance. Of the three EEGNet variations the highest performance is achieved with EEGNet using raw data, then EEGNet with FBE data.

For every subject in the IV2a dataset, except AT2 and A5E, the performance increase from below the CL to above when using EEGNet. For subject A2T, the performance increase and is highest when classified with EEGNet using raw data, but the performance is still below the CL. Subject A5E had an initial performance above the CL, and the performance increased when using EEGNet with raw data, but decreased below the CL when using EEGNet with either of the decomposition methods.

The second binary classification explored is between right-hand and feet. For dataset SMR, the performance increase for every low-performing subject, except S2T, S6T, S14T, and S8E, when classified with EEGNet. For the majority of the low-performing subjects, the highest performance increase was achieved when using EEGNet with raw data, then EEGNet with FBE. The four subjects who continuously have a performance below the CL, perform best when using EEGNet with raw data. For the high-performing subjects, the performance increase when classifying with EEGNet, except for subjects S11T and S12T where the performance decrease to below the CL. Subject S3T, S3E, and S9E achieve a performance of 100% when classifying with EEGNet using FBE.

For the NTNU dataset, all the subjects achieve a performance above the CL, and DL slightly outperforms ML. However, the confusion matrix shows a large imbalance between the classes, and it can be seen that about half of the *feet MI* trails are misclassified.

For the subjects in the IV2a dataset with a performance below the CL with RF, the performance increase above the CL for every subject except A5E. For the subjects with a prior performance above the CL, the performance increase when using EEGNet, except for A2T, A4E, and A9E where the performance decrease below the CL.

One interesting observation is that each decomposition method affects RF and EEGNet the same, i.e when the performance of the RF with FBE is higher than RF with DWT, the performance of EEGNet with FBE is higher than EEGNet with DWT and vice versa.

5.7.2 Subject-Dependent Multiclass Classification

To explore the performance of CNN as a multiclass classifier, multiclass classification is created for datasets IV2a and NTNU. The three class combinations explored for multiclass classification are *left-hand, right-hand and feet, right-hand, feet and tongue*, and *left-hand, right-hand, feet and tongue*.

The *left-hand, right-hand and feet* classification is conducted for datasets IV2a and NTNU, and the results are presented in Figures 5.38 and 5.40. Due to the class imbalance in the NTNU dataset, a confusion matrix is created, and presented in Figure 5.39. The classification of *right-hand, feet and tongue* and *left-hand, right-hand, feet and tongue* is performed on the subjects in the IV2a dataset, and the results are presented in the Figures 5.41 and 5.42, respectively.

Dataset NTNU: Left-Hand, Right-Hand and Feet

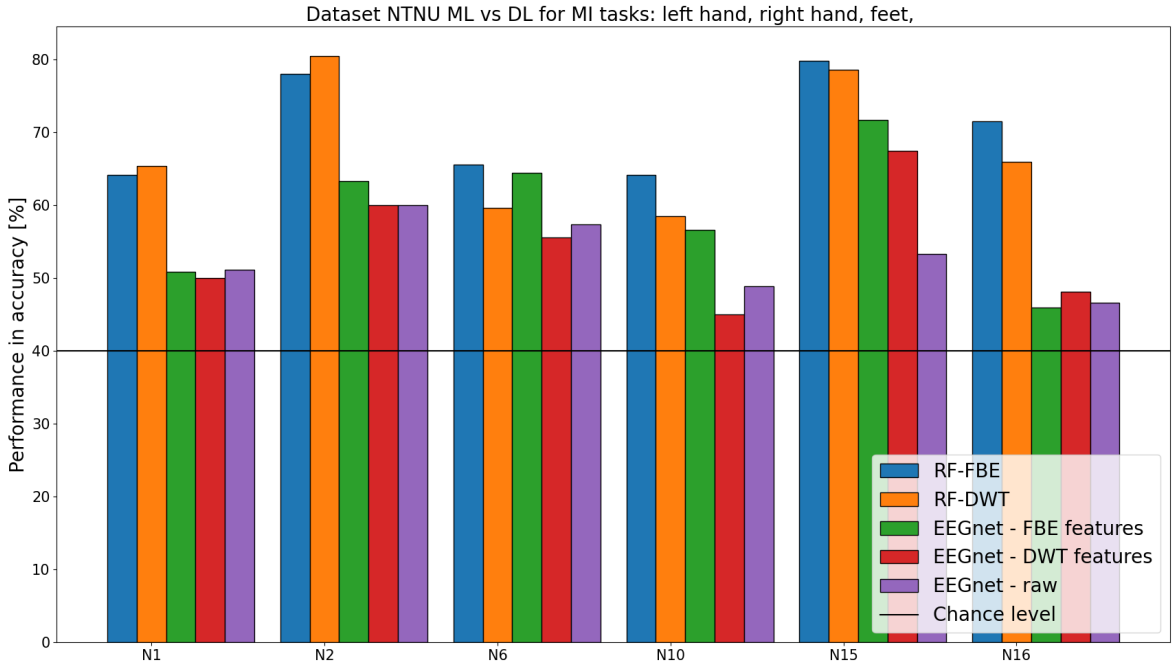


Figure 5.38: Multiclass classification of left-hand, right-hand, and feet MI for each subject in the NTNU dataset.

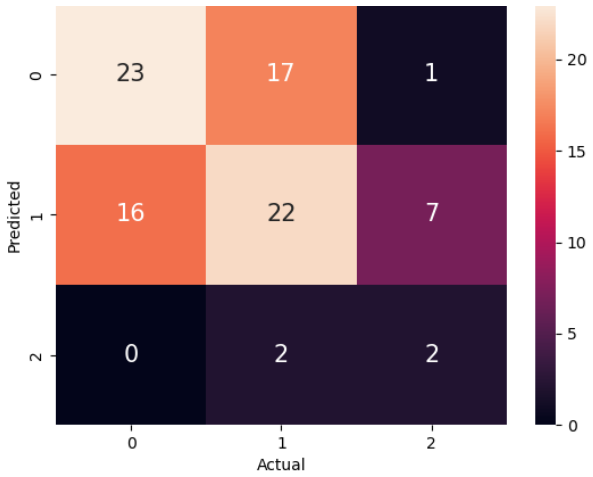


Figure 5.39: Confusion matrix resulting from subject-dependent CNN EEGNet raw data multiclass classification between left-hand, right-hand, and feet MI, for subject N15 in NTNU dataset. Label 0 refers to left-hand, 1 refers to right-hand and 2 refers to feet.

Dataset IV2a: Left-Hand, Right-Hand and Feet

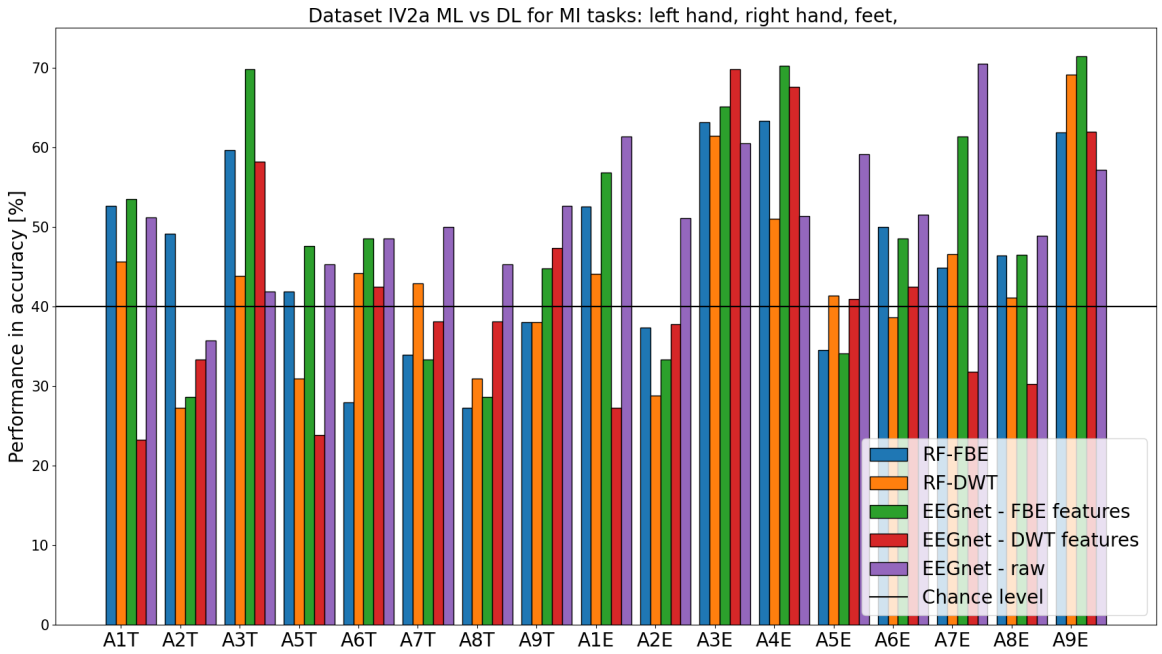


Figure 5.40: Multiclass classification of left-hand, right-hand, and feet for each subject in the IV2a dataset.

Dataset IV2a: Right-Hand, Feet and Tongue

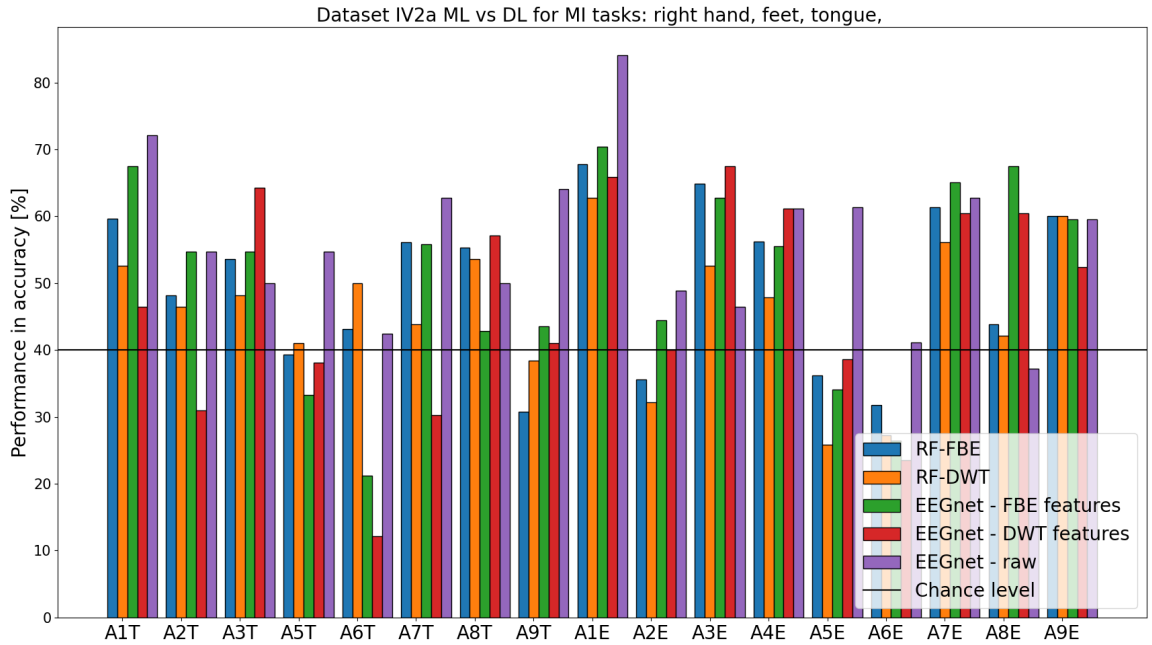


Figure 5.41: Multiclass classification of right-hand, feet, and tongue for each subject in the IV2a dataset.

Dataset IV2a: Left-Hand, Right-Hand, Feet and Tongue

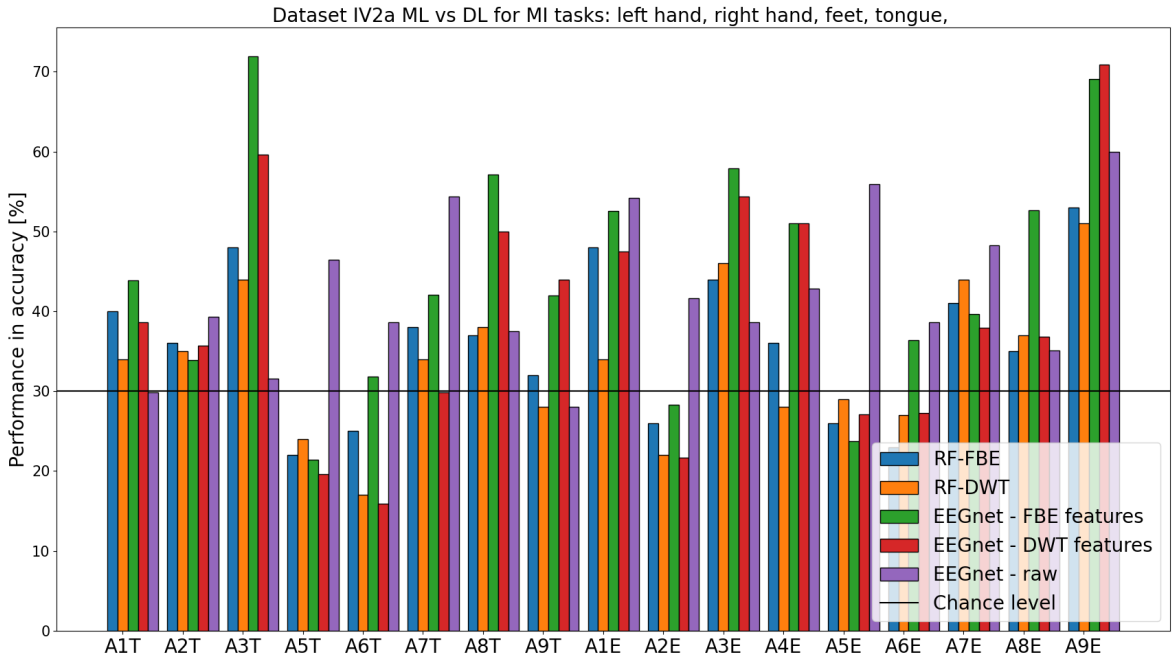


Figure 5.42: Multiclass classification of left-hand right-hand, feet, and tongue for each subject in the IV2a dataset.

For the left-hand, right-hand, and feet classification the performance decreased for every subject in the NTNU dataset when classification with EEGNet, but the performance is still above the CL. Although the confusion matrix shows that the *feet MI* class is four times more misclassified than correctly classified and that the class imbalance is large, there is also some misclassification between the other classes as well.

For dataset IV2a, all of the subjects who have a performance below the CL when using RF, achieve a performance above the CL when using either of the EEGNet variations. Similarly, every subject, except A2T, that performed above the CL when classified with RF, achieves the highest performance when using one of the EEGNet variations. The highest accuracy is achieved by the EEGNet using features from decomposition, especially from FBE.

The second three-class classifier that is explored on the IV2a dataset consists of the classes right-hand, feet, and tongue. All the subjects that have an initial RF performance below the CL, achieve a performance above the CL when applying either of the EEGNet variations. The highest performance for every subject is reached when using either EEGNet with decomposition or EEGNet with raw data. The highest performance is reached for subject A1E with EEGNet using raw data, yielding a performance above 80%.

The only four-class multiclass classifier which is explored is the classification of left-hand, right-hand, feet, and tongue, which was only explored on dataset IV2a. Subjects A5T, A2E, and A5E only achieve a performance above the CL when using EEGNet raw data, and A6T and A6E manage to reach above the CL when using EEGNet with FBE and raw data. For the rest of the subjects, the highest performance is reached when using either EEGNet variations. FBE outperforms DWT when using both RF and EEGNet.

5.7.3 Subject-Independent Classification

The performance of EEGNet as a subject-independent classifier is explored for both binary and multiclass classification. The results can give an indication about whether DL is as affected by the subject variability as ML. The experiments are performed for every dataset and the same class combination that was explored for the subject-dependent classification.

More formally, for the IV2b dataset only binary classification of the classes *left-hand and right-hand* is performed, and the performance for the dataset is presented in Figure 5.43. For the SMR dataset, only binary classification of the classes *right-hand and feet* is performed, and the performance for the dataset is presented in Figure 5.44. For the IV2a dataset, binary classification of *left- and right-hand* and *right-hand and feet*, as well as multiclass classification of the classes *left-hand, right-hand and feet*, *right-hand, feet and tongue* and *left-hand, right-hand, feet and tongue*, is performed. The results for the IV2a dataset are presented in Figure 5.45. Lastly for the NTNU dataset binary classification of the classes *left-hand and right hand* and *right-hand and feet*, and multiclass classification of the classes *left-hand, right-hand and feet*, is performed, with the performance presented in Figure 5.46. Additionally, the confusion matrices for each of the classification combinations were created for the NTNU dataset, to further analyze how the class imbalance affects the performance.

Dataset IV2b

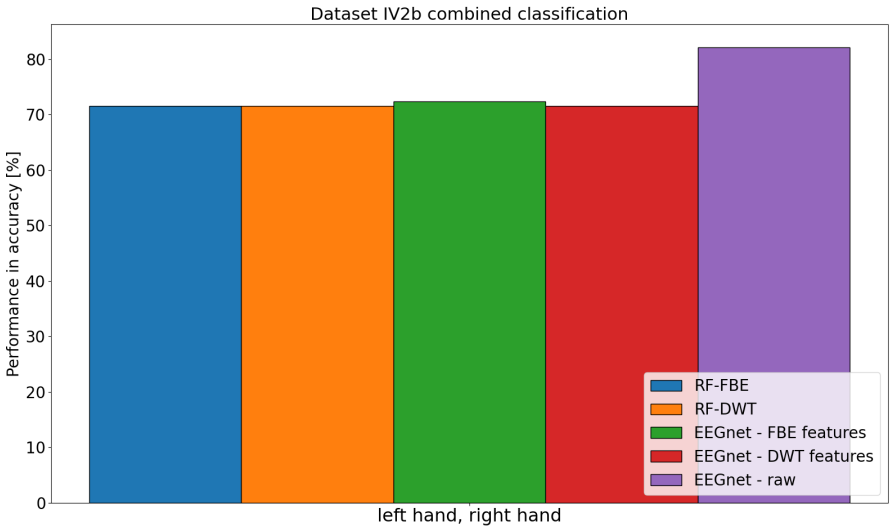


Figure 5.43: Subject-independent binary classification of left and right-hand for the IV2b dataset.

Dataset SMR

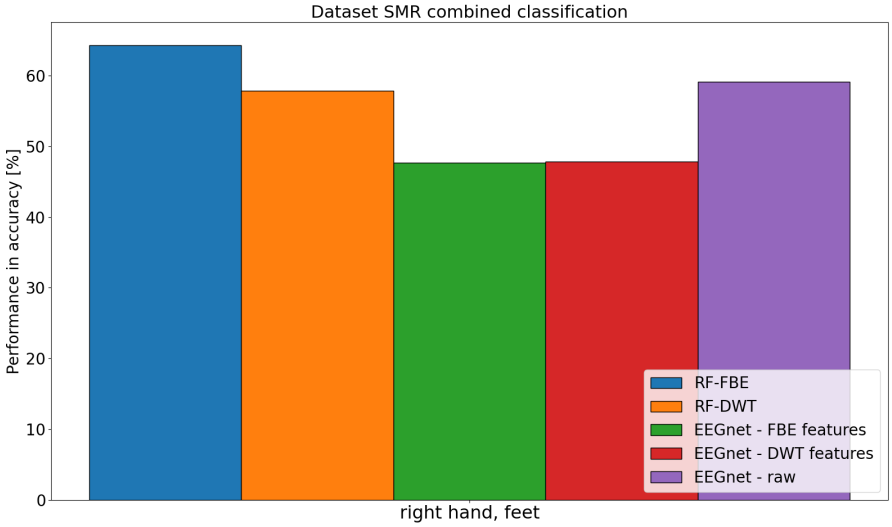


Figure 5.44: Subject-independent binary classification of right-hand and feet for the SMR dataset.

Dataset IV2a

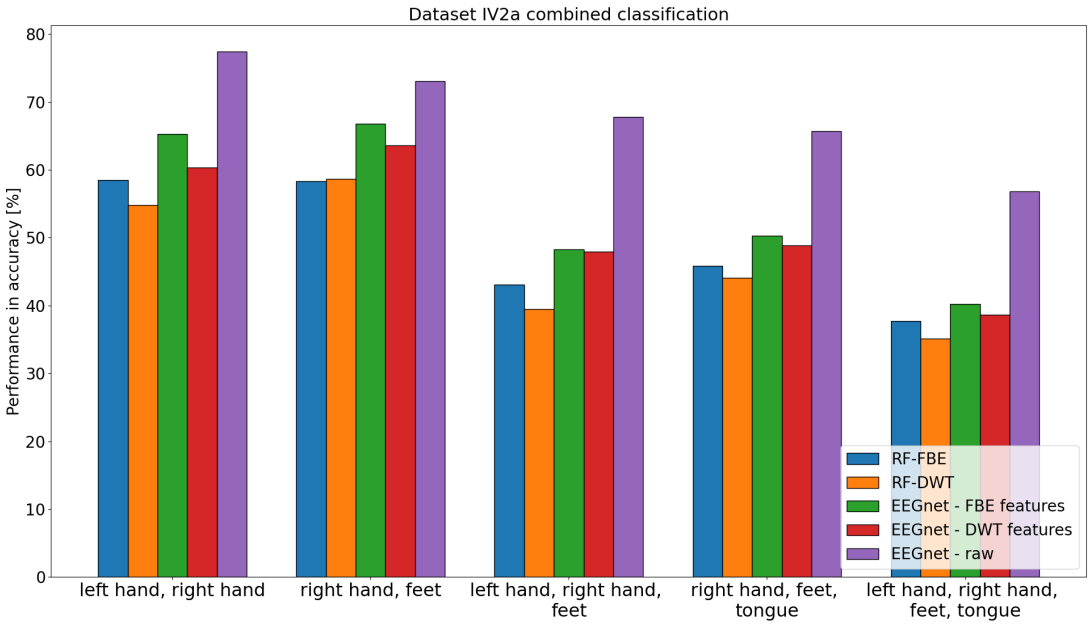


Figure 5.45: Subject-independent classification for the IV2a dataset, using various combinations of classes.

Dataset NTNU

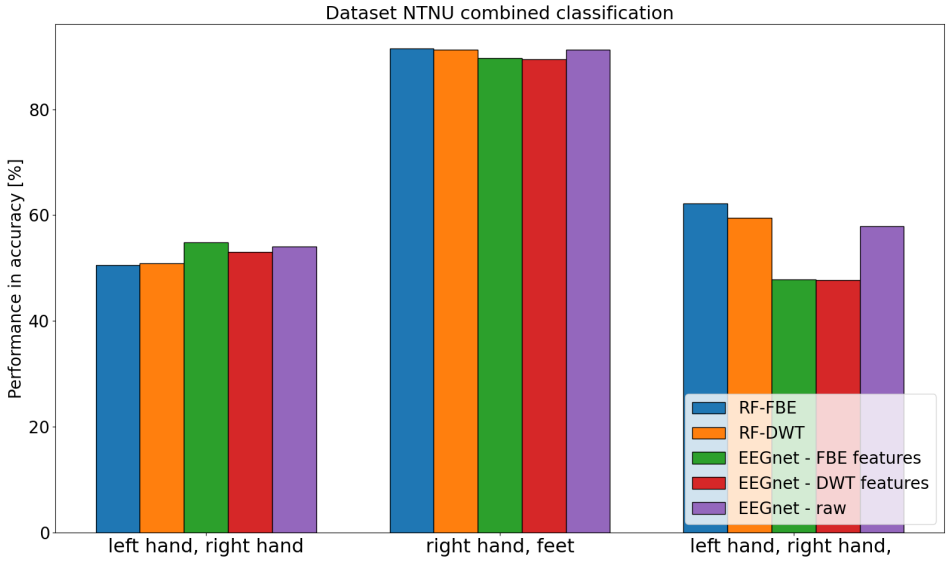
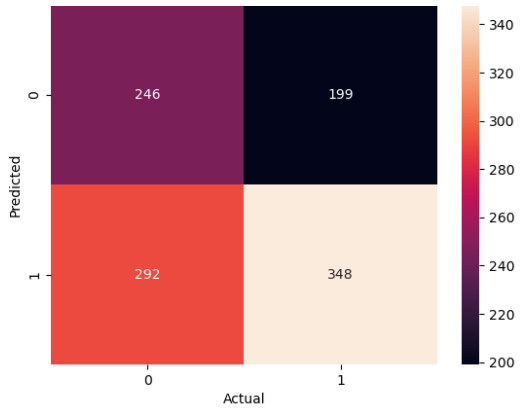
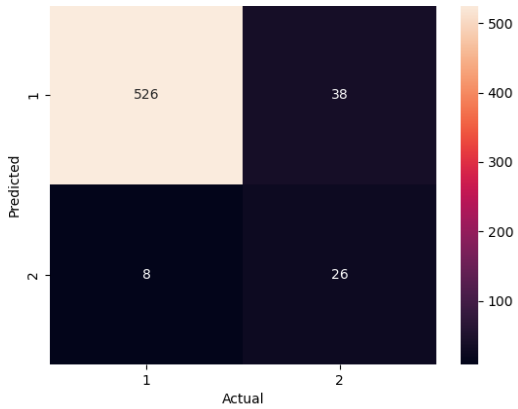


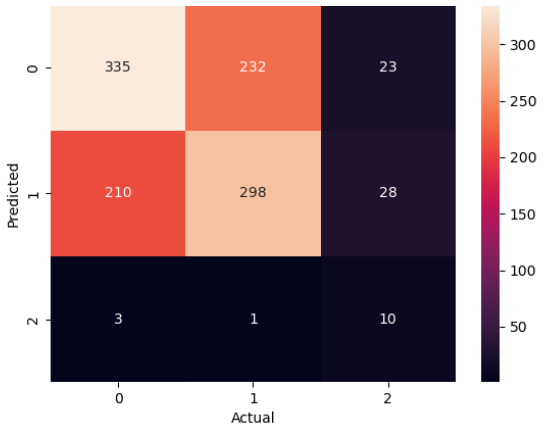
Figure 5.46: Subject-independent classification for the NTNU dataset, using various combinations of classes.



(a) Binary classification of left-hand and right-hand



(b) Binary classification of right-hand and feet



(c) Multiclass classification of left-hand, right-hand, and feet MI

Figure 5.47: Confusion matrices resulting from subject-independent CNN EEGNet raw data classification with NTNU dataset. Label 0 refers to left-hand, 1 refers to right-hand and 2 refers to feet

For binary classification of left- and right-hand MI, the DL classifiers achieve a higher performance than the ML classifiers. For dataset IV2b, the ML classifier already has a high performance of around 70%, and for EEGNet with raw data, the performance increase with over 10%. The two other EEGNet classifiers have similar performance to the two ML classifiers. For the IV2a dataset, the performance is also highest when classifying with EEGNet using raw data, and the performance increases to almost 20% above the CL. For the NTNU dataset, the performance increase by a couple of percentages when using EEGNet, but stays below the CL. The highest performance is achieved with EEGNet using FBE data. The confusion matrix, given in Figure 5.47a, shows that the classifier misclassifies about half of the trials for both classes.

For binary classification of right-hand and feet, the ML classifier has a higher performance than the DL classifiers, for the SMR and NTNU dataset. For the IV2a dataset the DL classifiers, especially EEGNet with raw data, perform higher than ML. The confusion matrix for the NTNU classification, in Figure 5.47b, shows that the majority of the feet MI trials are misclassified as right-hand, as well as that there is a large class imbalance between the two classes. Therefore, the results from NTNU binary classification of right-hand and feet only give an indication of how EEGNet works with an unbalanced dataset, and not how well the classifier classifies the two specific classes. The fact that IV2a

is a low-performing dataset, and SMR is high-performing, is an indication that EEGNet works better with low-performing datasets, compared to ML algorithms. Additionally, for the high-performing datasets, ML performs comparably to DL even though it is simpler.

For multiclass classification of left-hand, right-hand, and feet, DL outperforms ML for the IV2a dataset, and the performance is increased to above 70%. For the NTNU dataset ML performs higher than DL. The confusion matrix, given in Figure 5.47c, shows that there is not only confusion between left-hand and right-hand, but that the majority of the feet trials are misclassified.

The two remaining multiclass classifications are performed using IV2a data. For both of the classifiers, DL performs higher than ML. Of the three DL methods explored, EEGNet with raw data outperforms the others, and EEGNet with FBE slightly outperforms EEGNet with DWT.

Chapter 6

Discussion, Conclusion and Future Work

In this chapter, the acquired experimental results are evaluated and further discussed, in order to get a better understanding of the findings. Following the discussion, a conclusion of the overall contribution to the field of MI-BCIs is given. Lastly, some suggestions for future work and necessary investigations based on our findings are given.

6.1 Summary and Discussion

This section presents a summary and discussion of the individual experimental results obtained in this thesis. Most experiments were conducted to see if some common problems in BCI systems could be solved or reduced. The first experiment was only conducted to determine which features should be used for classification in the subsequent experiments, and thus it is not further discussed. The next experiment tested out the MI detection ability of a proposed classification method, followed by two experiments on inter-subject and inter-dataset TL. Subsequently, a band-window optimization experiment was conducted. Then, channel reduction experiments were performed, where one of the experiments included an ERP and PSD analysis. Lastly, various DL approaches were tested on multiple variations of classes and compared to the ML approaches.

6.1.1 Experiment 2: MI vs No-MI

The proposed classification algorithm using the RF classifier with FBE decomposition showed promising results in the detection of an MI, as shown in Section 5.2. The fact that all metrics, including accuracy, F-score, precision, and recall, achieved a result above 89% for all datasets, indicates that the method is both reliable and robust, which are valuable properties for use in a BCI system.

Another interesting finding was that the worst performing datasets in regular classification, dataset NTNU, and dataset IV2a, achieve the best performance for MI detection. This may be due to the fact that the objective of detecting the occurrence of an MI is much simpler than classifying the type of MI, and the data needed for classification will also be different. Thus, the worst-performing datasets may contain information that is more distinguishable in terms of MI detection compared to MI classification. This also indicates that the algorithm is better at capturing the features relevant for MI detection in these datasets, which leads to better performance. However, further investigation should be done to fully understand why this occurs.

6.1.2 Experiment 3: Transfer Learning

TL was explored through an inter-subject and an inter-dataset technique, testing both LOO and LSO approaches.

Using inter-subject TL, in Section 5.3.1, is the better approach to the subject-dependent model in over half, more specifically 54.17%, of the subjects, indicating its effectiveness. It also became evident that TL is a more effective and robust approach for the low-performing subjects as compared to the high-performing, which can be explained by the fact that high-performing subjects produce features that are more MI-specific, while low-performers produce more features which are more relevant to other brain activity than MI, due to lack of MI ability. To then use a pre-trained model on these low-performers, the classification performance may be improved due to the model recognizing the actual MI features in the subject, instead of the irrelevant features. Additionally, it was seen that the LSO approach outperforms the LOO approach for high performers, which may again be due to their ability to produce distinguishable MI patterns, thus obtaining powerful features, which are necessary to get an accurate classification. Therefore, adding their features to the source model may increase the performance, as more relevant features are added to the model. The opposite was seen for low-performers, as adding their poor features to the source model may decrease the performance as features irrelevant for MI are added.

The other TL approach was based on an inter-dataset approach, in Section 5.3.2. Here, the results indicated that the TL approach used between datasets should be based on an LSO approach, where the source model is solely built on data from high-performing subjects, as this method achieved the highest performance. This is a reasonable result as the high-performers are able to produce recognizable MI patterns, and the source model is thus trained on relevant MI features, and adding a few samples from the target subject, in an LSO approach, additionally trains the model on some features specific to the subject's characteristics, which again increases the performance. Furthermore, this method has generalizability as it performed well for all datasets.

6.1.3 Experiment 4: Sliding Window and Frequency Band Optimization

An optimization process exploring various combinations of the time window and frequency bands for data extraction was conducted both subject-dependently and subject-independently with the aim of capturing the relevant ERD/ERS patterns for MI. The optimal band-window combination was obtained by evaluating their performance using the RF classifier.

The subject-dependent optimization technique was proven successful, as all subjects achieved an increased performance after optimization, compared to using the standard method, as given in Section 5.4. This increase was particularly evident in low-performers, where the majority of them exhibited a performance above the CL after optimization. This resulted in them becoming high performers and shows a great indication that an optimization process may assist in resolving the problem of BCI illiteracy.

It is also worth noting that the subject variability was also reflected in the determined optimal fre-

quency band and time window across the subjects. This indicates that this type of optimization process should be performed for each individual subject, to consider their individual differences, in order to get the best effect and performance increase. This conclusion is also supported by the fact that the subject-independent optimization was less successful than the subject-dependent optimization.

Even though most datasets had a gain in performance after the subject-independent optimization, some still remained under the CL. Dataset IV2b and SMR even decreased in performance after optimization, which again may be due to the great subject variability within the datasets. Additionally, the optimization process only used one frequency band for decomposition, thus less data and information was obtained compared to the regular method, which may have led to a decrease in performance. However, as dataset IV2b and SMR are high-performing datasets, this decrease may also indicate that optimization is more efficient for low-performing subjects and datasets.

To get an overview of the most significant bands for classification, a statistical analysis was presented in Section 5.4.3. This analysis indicated that optimal frequencies around the upper mu band, mu_{high} and Beta_1, and the upper Beta band, Beta_4 and Beta_5, were frequently determined. This confirms the theory that the upper mu band is task-specific and can be used for MI task determination, as stated in [15]. The upper Beta band is also a reasonable result, as this band is close to the frequencies for concentration, which is a key element in producing a good mental image of the task [86]. There was also a noticeable difference in the frequency band distribution between high-performers and low-performers. A total of 88% of the high-performing subjects had their optimal frequencies in the upper mu and high Beta bands, whereas the low-performing subjects had a more scattered distribution. This finding may indicate that the high and low performers can be predicted based on their optimal bands, as an optimal band outside upper mu and high Beta may indicate the lack of a clear MI in the subject's data, and the classification may be performed using irrelevant information, causing poor performance.

6.1.4 Experiment 5: Channel Reduction using Greedy Backward Elimination Algorithm

A channel reduction experiment was conducted on datasets NTNU, IV2a, and SMR. The recurring observation for each of the three datasets explored, section 5.5, is that the performance increased for the majority of the subjects when a single channel was removed. These results indicate that some channels contain redundant information, which negatively affects the classification.

For the majority of the low-performing subjects, the highest performance was achieved after removing several channels. The increase in performance was also highest for the low-performing subjects, where some managed to perform above the CL. In contrast, for the already high-performing subjects, the growth in performance was not as significant compared to the low-performing subjects, and the majority reached their highest performance after removing only one or a few channels. For some high-performing subjects, especially in the SMR dataset, the performance decrease when removing the first channel and then increases later. This decrease might be due to the high-performing subjects creating more high-quality, MI relevant data, and thus removing channels does not just remove subject-dependent data, but also some MI-specific data.

The ideal subset of channels for each subject was inconsistent, with low-performing subjects weighing the channels differently than the high-performing subjects. This indicates that the selection of channels is subject-dependent, making it challenging to determine a general channel subset for the entire dataset. The subject-independent classification also benefitted from reducing the number of channels by one, but the performance only increased by a small percentage. The small improvement is likely due to the fact that the ideal channel subset is subject-dependent.

The three datasets contain different MI classes, which impacts the amount and what channels are selected, as this selection is influenced by each MI class. All the datasets contain the right-hand MI, but as the SMR dataset has unlabeled channels it is difficult to exploit its results in future recording sessions. Both the NTNU and IV2a dataset contains the left-hand and right-hand MI classes, and both give high weighting to channel C4 and C3, which also corresponds to the theory of which areas in the brain are activated during these MI tasks [15].

Although the results indicate that subject MI enhancement through channel reduction using the greedy elimination algorithm is significantly subject-dependent, the method has shown improvement for every subject, and can still be used to reduce the processing for individual subjects.

The topographic map analysis showed that the maps gave some insight into which regions in the brain are affected and activated by each MI class, as well as the differences between low and high-performing subjects. A comparison of which regions had the highest and lowest intensity and which channels the corresponding subject weighted with high and low importance, showed that the high-intensity regions and channels had the lowest importance. This might be due to the signal varying in time, and channels with a lot of activity will have a small mean compared to signals that are more stationary in time. Although topographic maps visualize the spatial distribution of the signal, it does not show the complexity of the signal, which needs to be extracted with decomposition and feature extraction. The distribution of the measurement changes over time, and the topographic map can only show a single time instance or the average over a measurement.

6.1.5 Experiment 6: ERP and PSD Analysis

The channel selection for dataset IV2b, as described in Section 5.6, suggests that C4 is the most important channel for MI classification. Therefore, an ERP analysis was performed specifically on this channel. The analysis suggests that there is a difference in the ERP for a high-performing versus a low-performing subject. The high-performing subject exhibits a consistent and distinct ERP, while the low-performing subject's ERP shows unpredictability and fluctuations, indicating a possible lack of concentration or inability to produce a clear MI [86]. Additionally, the power of the ERP for the high-performing subject is approximately twice as strong as that for the low-performing subject, indicating a higher level of task engagement and stronger cognitive abilities [87]. Furthermore, the stronger and more consistent ERP in the high-performing subject may also indicate superior MI skills compared to the low-performing subject [88].

There was also a clear difference in the PSDs between the high-performing and low-performing subjects. The high-performer had the most variations in the theta, mu, and Beta frequency bands, which

corresponds to the established theory for MI [15]. In contrast, the low-performer also had increased variations in frequencies around 60-125 Hz, in an oscillatory pattern.

By exploiting this finding, it could be possible to identify and predict high-performing subjects prior to classification, by looking at their ERP signal's strength and structure, as well as their PSD. This could potentially save time in processes such as subject selection for the creation of TL models and training programs can be tailored to the specific subject. However this experiment was only conducted on two subjects, and further exploration is necessary to validate and confirm these findings.

6.1.6 Experiment 7: Classification with CNN

The experiments with DL explored binary and multiclass classification of unprocessed and processed data for both subject-dependent and subject-independent classification, in section 5.7. For the binary subject-dependent classification the performance increase significantly for the low-performing subjects when classified with DL compared to ML. For the higher-performing subjects, the performance of DL classifiers does not have such a significant increase or effect. By increasing the performance of the low-performing subjects, the overall performance for each dataset increases as well. The results indicate that EEGNet is less affected by the intra subject variability than RF.

Considering that the results from the multiclass subject-dependent classifiers are produced for the NTNU and IV2a datasets, which have overall low performances, it is uncertain how the performance and effect of EEGNet are on a high-performing dataset. The results from NTNU as a multiclass classifier should also be disregarded, due to the unbalanced nature of the dataset which caused the accuracy metric and actual classification performance to contradict each other. The results from the classification of the IV2a datasets indicate that for the majority of the subjects, the highest performance is achieved with EEGNet using FBE. This indicates that some form of additional signal decomposition and feature extraction is needed for EEGNet to be able to successfully perform multiclass classification. The change in choice of preferred classifier between binary and multiclass, i.e. from using raw data to FBE processed data, could also be due to the *tongue MI* class, which is only explored for multiclass classification. Thus, further multiclass subject-dependent analysis should be performed for a higher-performing dataset, and analysis of the tongue MI classification.

Subject-independent classification performs significantly differently than subject-dependent. First, comparing the different EEGNet variations, EEG using raw data perform better than EEGNet using either of the decomposed data, for all the datasets. For the low-performing dataset, IV2a, the DL classifiers outperform the ML classifiers. On the other hand, for the higher-performing dataset, SMR, the DL classifiers perform poorer than the ML classifiers. This is an indication that the CNN as a feature extraction method, manages to remove some inter-subject variability, which increases the low-performing dataset performance.

The results for the NTNU datasets classification with *feet MI* indicate how well the DL classifier performs with unbalanced data. Especially the multiclass classification with NTNU subjects shows that RF performs better than EEGNet for unbalanced datasets, which is assumed from theory [58]. This is evident from both the subject-dependent and subject-independent classification.

Although EEGNet with unprocessed data has an overall higher performance than RF, the complexity of the classifier, in both computational power and time, made it less useful. For EEGNet with processed data, which of course takes time to process, also had a higher computational cost than using RF. There is also a lot more variability with using CNN since there is randomness in some layers, such as dropout. EEGNet works well for subjects and datasets with low performance but has comparable performance to RF for high-performing subjects.

6.1.7 Comparison with Results from Literature

To give a summary of the results obtained in this thesis, and see how they compare to previously obtained results in the literature, Table 6.1 is provided. The Table includes the highest achieved accuracy performance for the explored techniques, together with some results from the literature introduced in Section 3.4. This is given for all four datasets, namely IV2b, IV2a, SMR, and NTNU, and for multiple class combinations. It is worth noting that the highest achieved performance from our thesis is for the best subject in each dataset, while literature is given as an average of the dataset.

In general, it can be seen that the explored methods achieve comparable results to previous results from the literature. Common for all the datasets is that the CNN-EEGNet method gives the highest score for binary classification, but for multiclass other proposed methods from the literature have shown higher performance.

In [1],[2], and [69], the NTNU dataset was concluded to be a poor dataset in relation to MI. However, the proposed optimization process in this thesis highly increased the overall performance of the dataset and is thus proven effective, even for datasets with initially unsatisfying performance. For the NTNU dataset, our methods worked better than the previously tested methods from the literature.

However, as the results from the thesis were taken from the best subject, they do not show the diversity in performance from the worst to the best subject, and are therefore not directly comparable to literature, as most of the literature is presented as an average for the dataset. Nevertheless, considering the impressive performances achieved for some subjects, it is an indication that the explored methods show great promise and effectiveness for use in the field of MI-BCI.

Dataset	Classes	Method	Channels	Performance	Reference
IV2b	LH, RH	RF+FBE+Inter-subject TL-LSO	3	98.33%	5.3.1
	LH, RH	RF+FBE+Inter-dataset TL-LSO	3	89.82%	5.3.2
	LH, RH	RF+FBE+BW Optimization	3	98.02%	5.4.1
	LH, RH	CNN-EEGNet+raw data	3	98.55%	5.7.1
	LH, RH	SVM+BW Optimization+FBCSP	3	84.50%	[30]
	LH, RH	RF+FBE	3	78.07%	[1, 2]
	LH, RH	CNN+SAE	3	78.00%	[76]
	MI, No-MI	RF+FBE	3	89.89%	5.2
IV2a	LH, RH	RF+FBE+Inter-subject TL-LSO	22	58.33%	5.3.1
	LH, RH	RF+FBE+Inter-dataset TL-LSO>CL	3	77.03%	5.3.2
	LH, RH	RF+FBE+BW Optimization	22	81.82%	5.4.1
	LH, RH	CNN-EEGNet+ raw data	22	92.59%	5.7.1
	LH, RH	SVM+BW Optimization+CSP	22	82.13%	[30]
	RH, F	CNN-EEGNet+ raw data	22	100.00%	5.7.1
	RH, F	SVM+BW Optimization+CSP	22	85.70%	[30]
	LH, RH, F,	CNN-EEGNet+FBE	22	71.14%	5.7.2
	RH, F, T	CNN-EEGNet+raw data	22	84.09%	5.7.2
	LH, RH, F, T	RF+FBE+Inter-subject TL-LSO	22	50.00%	5.3.1
	LH, RH, F, T	RF+FBE+BW Optimization	22	57.47%	5.4.1
	LH, RH, F, T	RF+FBE+Channel reduction	15	64.38%	5.5.2
	LH, RH, F, T	CNN-EEGNet+FBE	22	71.92%	5.7.2
	LH, RH, F, T	LDA+CSP+ITL	22	77.00%	[77]
	MI, No-MI	RF+FBE	22	94.87%	5.2
NTNU	LH, RH	RF+FBE+Inter-subject TL-LSO	7	68.33%	5.3.1
	LH, RH	RF+FBE+Inter-dataset TL-LSO	3	60.45%	5.3.2
	LH, RH	RF+FBE+BW Optimization	7	78.48%	5.4.1
	LH, RH	RF+FBE-Channel reduction	6	76.25%	5.5.1
	LH, RH	CNN-EEGNet+ raw data	7	79.82%	5.7.1
	LH, RH	CNN	7	60.63%	[69]
	LH, RH	SVM+CSP	7	63.02%	[69]
	LH, RH	RF+FBE	7	53.69%	[1, 2]
	MI, No-MI	RF+FBE	7	92.12%	5.2
SMR	RH, F	RF+FBE+Inter-subject TL-LSO	15	82.76%	5.3.1
	RH, F	RF+FBE+BW Optimization	15	100.00%	5.4.1
	RH, F	RF+FBE-Channel reduction	14	100.00%	5.5.3
	RH, F	CNN-EEGNet+FBE	15	100.00%	5.7.1
	RH, F	RF+CSP	15	79.30%	[74]
	RH, F	sLDA+CSP	15	77.15%	[74]
	MI, No-MI	RF+FBE	15	90.74%	5.2

Table 6.1: Overview of the best results obtained in this thesis compared to results found in previous literature. With classes left-hand (LH), right-hand (RH), Feet (F), Tongue (T), MI, and No-MI. *BW* stands for Band-Window. The *Performance* is given in accuracy, and the *Reference* column indicates where the result can be found, either in one of our sections or in the literature.

6.2 Conclusion

The main goal of this thesis was to look into some common challenges and problems related to MI-BCI systems, with the aim of finding solutions and improvements.

One of the main challenges in BCI systems is the BCI illiteracy of the users. The findings in this thesis suggest that an optimization process, channel reduction or the use of a CNN classifier could improve the illiteracy in a subject. This can be concluded due to the great increase in classification performance after these methods were applied. These methods were especially effective for low-performing subjects, where the optimization methods had an average increase of 19.23% for these subjects, suggesting that all subjects can achieve an accuracy above the CL if provided with the appropriate frequency band and time window during processing and classification. The CNN approach also showed promise in reducing the intra- and cross-subject variability and even got most subjects above the point of BCI illiteracy. Thus, these methods are robust in improving the BCI illiteracy within a subject.

Another objective in relation to BCI systems is to create a reliable system with the ability to accurately interpret the user's intentions. Firstly, MI detection is a crucial aspect of a BCI system. In this thesis, the use of an RF classifier combined with an FBE decomposition obtained a high accuracy and good performance metrics for this type of classification. This result is valuable for the development and implementation of a reliable BCI system. Secondly, it is important to choose a good classifier for the specific type of MI task being performed. In this thesis, the CNN classifier, especially the one using raw data as input, showed promising results in the classification of both binary and multiclass MI, and thus it is the preferred classifier to obtain a BCI system which accurately executes the intended user-command. Thus, a BCI system may benefit from using RF with FBE for MI detection, and CNN for MI classification, in order to become more reliable.

As mentioned, there is a time-consuming calibration process related to the customization of a BCI system to a user. The results in this thesis imply that the use of TL could reduce or remove this process, thus saving time and resources. This is valid as TL do not need to initialize and train a new model on each subject, but uses pre-trained models and adapts them to the new subject, while still retaining a good performance, as proven in this thesis. This approach also solves the problem of limited data for training, as previously acquired data can be used on the new subject. The findings also suggest that the TL should follow a LSO approach, especially for high-performers with valuable features for MI classification, while for some low-performing users, the TL approach should be LOO, as their features may interfere with the MI classification process. Additionally, it is particularly important that pre-trained models utilize well-performing subjects, so it would work for the generally good subjects and be built on MI-specific features.

Additional challenges connected to the use of BCI systems are the wearability, usability, and availability of the BCI systems to the intended user. If the system should be used continuously, it is important that it is wearable and comfortable, thus a channel selection process that reduces and optimizes the channels is favorable. The results from channel reduction suggested that decreasing the amount of

channels increases the accuracy, but the favorable number and placements of the channels was subject dependent. The reduction of channels will also reduce the related cost and thus make the system more accessible for the users which requires such systems [5].

Another possibility to reduce the cost related to the BCI systems should be to detect if the users will perform well or not, prior to excessive processing, data acquisition, and tests, thus reducing the resources needed for each subject. The findings in this thesis indicate that an ERP and PSD analysis on the subject's raw EEG data may help predict the subject's performance even before any processing or classification is performed. The findings also indicated that there is a relation between the optimal frequency band and the performance of the subject, so performing a simple optimization process and analysis on a few samples may predict the performance, as high-performers tend to have an optimal band in the upper mu or high Beta frequencies. This may help to identify BCI illiteracy or lack of MI ability before extensive resources or time is used on the subject, so that relevant measures, such as training adaptation and tailoring, can be taken in early stages. Lastly, using pre-trained models in a TL approach may also assist in reducing the cost related BCI as less time for data acquisition is needed, thus making the whole process of attaining a tailored BCI system more accessible for the users.

6.3 Recommendations for Future Work

Even though the findings in this thesis are promising, there are still several techniques that should be tested to further validate and improve the utilized approaches.

As the findings in this thesis suggest that a binary classifier performs better than a multiclass classifier, a future experiment should be to use multiple binary classifiers to build a multiclass classifier. For example, if one should classify the four MI classes of right-hand, left-hand, feet, and tongue, one could start by continuously using an MI detector to determine if an MI has occurred. Once an MI is detected, a binary classifier is used to check if the MI is a hand or non-hand task. For hand tasks a left-hand versus right-hand classifier can be used, while for non-hand, a feet versus tongue classifier can be used. The output of the final classifier will then be the determined MI task. Using this kind of step-wise classification may help with BCI illiteracy, as an optimization may be used on each step and features may be adapted to the individual classifiers.

Further experiments should also be conducted in predicting if a subject will be well-performing or BCI illiterate based on their raw signal or before extensive processing, in order to save time and cost. This can be done by looking more in-depth into the ERPs and PSDs for more subjects to validate the findings in this thesis, in addition to verifying the hypothesis that there is a connection between performance and optimal frequency band [15].

Additionally, in this thesis, the only hyperparameter that was changed for CNN was the dropout probability, which changed for subject-dependent and subject-independent classification. The kernel length was also changed based on the sampling frequency. Since the CNN works as a feature extractor as well as a classifier, the hyperparameters should be explored and tuned for each dataset. The depth, choice of activation functions, and model optimizer can be explored to get a deeper understanding

of what work and does not work with EEG data. The performance of EEGNet as a multiclass classifier should also be explored with an MI dataset that performs well with other classifiers.

Also, as CNN (EEGNet) was proven to be the superior classifier for all the datasets, a TL experiment using CNN as the classifier should be conducted. Such an experiment could be performed by changing and adapting the hyperparameters of CNN to the new dataset. This could be especially useful as CNN has shown great results and improvements in the weaker subjects, and could thus help conquer the BCI illiteracy problem. The computational cost of using CNN can also be reduced by using TL to build and change the model.

Furthermore, the channel reduction, and therefore also selection, performed in this thesis was at most performed using 22 channels. To get more valid and generalizable results, even more, channels should be tested, as 22 channels are already a small subset of available channels. Other, more complex, approaches to channel selection should be explored, such as the different versions of the non-dominated sorting genetic algorithm (NSGA) [89].

Finally, the algorithms and optimization techniques should be tested online in a real-time BCI-system which is connected to a computer. This requires a well-designed data acquisition system and a well-performing MI detector, i.e. classifier. If the online tests are successful, the system may be connected to a device, like a drone, for real-time control using the MI tasks as commands.

References

[1] Victoria Kenworthy. Motor imagery classification for bci using random forest and gradient boosting techniques. Technical report, NTNU, 12 2022. URL http://dx.doi.org/10.13140/RG.2.2.13430.68164.

[2] Karoline Nylænder. Motor imagery classification for bci using random forest and gradient boosting techniques. Technical report, NTNU, 01 2023. URL http://dx.doi.org/10.13140/RG.2.2.12306.73923.

[3] Bernhard Graimann, Brendan Z Allison, and Gert Pfurtscheller. *Brain-computer interfaces: Revolutionizing human-computer interaction*. Springer Science & Business Media, 2010.

[4] Alejandro Antonio Torres-García and Marta Molinas. Analyzing the recognition of color exposure and imagined color from eeg signals. In *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 386–391. IEEE, 2019.

[5] Piotr Wierzgała, Dariusz Zapała, Grzegorz M. Wojcik, and Jolanta Masiak. Most popular signal processing methods in motor-imagery bci: A review and meta-analysis. *Frontiers in Neuroinformatics*, 12, 2018. ISSN 1662-5196. doi: 10.3389/fninf.2018.00078.

[6] Eiichi Naito, Per Roland, and H.Henrik Ehrsson. I feel my hand moving: A new role of the primary motor cortex in somatic perception of limb movement. *Neuron*, 36:979–88, 01 2003. doi: 10.1016/S0896-6273(02)00980-7.

[7] Dale Purves, George J Augustine, David Fitzpatrick, William C Hall, Anthony-Samuel LaMantia, James O McNamara, and S Mark Williams. *Neuroscience*. Sinauer Associates, Inc., 3 edition, 2004.

[8] Simone Dall'Orso, Taylor Hamstreet, and Silvia Muceli. The "little person" in the brain who helps to direct our movements. *Frontiers in Young Minds*, 11:750301, 2023. doi: 10.3389/frym.2022.750301.

[9] Qiong Gui, Maria V Ruiz-Blondet, Sarah Laszlo, and Zhanpeng Jin. A survey on brain biometrics. *ACM Computing Surveys (CSUR)*, 51(6):1–38, 2019.

[10] Saeid Sanei and Jonathon A Chambers. *EEG signal processing*. John Wiley & Sons, 2013.

[11] Priyanka A Abhang, Bharti W Gawali, and Suresh C Mehrotra. Technological basics of eeg recording and operation of apparatus. In *Introduction to EEG-and speech-based emotion recognition. 2nd national conference on innovative paradigms in engineering and technology*, 2016.

[12] Kai Zhang, Guanghua Xu, Xiaowei Zheng, Huanzhong Li, Sicong Zhang, Yunhui Yu, and Renghao Liang. Application of transfer learning in eeg decoding based on brain-computer interfaces: a review. *Sensors*, 20(21):6321, 2020.

[13] Dennis Mcfarland, Laurie Miner, Theresa Vaughan, and Jonathan Wolpaw. Mu and beta rhythm topographies during motor imagery and actual movements. *Brain topography*, 12:177–86, 02 2000. doi: 10.1023/A:1023437823106.

[14] Navneet Tibrewal, Nikki Leeuwis, and Maryam Alimardani. Classification of motor imagery eeg using deep learning increases performance in inefficient bci users. *PLOS ONE*, 2022. doi: 10.1371/journal.pone.0268880.

[15] Gert Pfurtscheller and Christa Neuper. *Dynamics of Sensorimotor Oscillations in a Motor Task*, pages 47–64. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010. ISBN 978-3-642-02091-9. doi: 10.1007/978-3-642-02091-9_3.

[16] Priyanka A Abhang, Bharti W Gawali, and Suresh C Mehrotra. Technological basics of eeg recording and operation of apparatus. In *Introduction to EEG-and speech-based emotion recognition. 2nd national conference on innovative paradigms in engineering and technology*, 2016.

[17] Trans Cranial Technologies. 10/20 system positioning manual. `https://trans-cranial.com/docs/10_20_pos_man_v1_0_pdf.pdf`, (2012). Last accessed 9-01-2023.

[18] A. Runnova, Anton Selskii, Anton Kiselev, Rail Shamionov, Ruzanna Parsamyan, and Maksim Zhuravlev. Changes in eeg alpha activity during attention control in patients: Association with sleep disorders. *Journal of Personalized Medicine*, 11:601, 06 2021. doi: 10.3390/jpm11070601.

[19] Myrto Stylianou, Nik Murphy, Luis Peraza, Sara Graziadio, Ruth Cromarty, Alison Killen, John O'Brien, Alan Thomas, Fiona Lebeau, and John-Paul Taylor. Quantitative electroencephalography as a marker of cognitive fluctuations in dementia with lewy bodies and aid to differential diagnosis. *Clinical Neurophysiology*, 129, 03 2018. doi: 10.1016/j.clinph.2018.03.013.

[20] Teodoro Solis-Escalante, Gernot Müller-Putz, and Gert Pfurtscheller. Overt foot movement detection in one single laplacian eeg derivation. *Journal of neuroscience methods*, 175:148–53, 09 2008. doi: 10.1016/j.jneumeth.2008.07.019.

[21] Li Hu and Zhiguo Zhang. *EEG Signal Processing and Feature Extraction*, page 13. Springer Nature Singapore Pte Ltd, 2019.

[22] Christa Neuper, Reinhold Scherer, Miriam Reiner, and Gert Pfurtscheller. Imagery of motor actions: Differential effects of kinesthetic and visual–motor mode of imagery in single-trial eeg. *Cognitive Brain Research*, 25(3):668–677, 2005. ISSN 0926-6410. doi: 10.1016/j.cogbrainres.2005.08.014.

[23] Erik K. St Louis, Lauren C. Frey, Jeffrey W. Britton, Jennifer L. Hopp, Pearce J. Korb, Mohamad Z. Koubeissi, William E. Lievens, and Elia Pestana-Knight. *Electroencephalography (EEG): An Introductory Text and Atlas of Normal and Abnormal Findings in Adults, Children, and Infants*, pages 8–20. American Epilepsy Society, 2016. ISBN 978-0-9979756-0-4. doi: 10.5698/978-0-9979756-0-4.

[24] Shobiha Premkumar. Subject identification using eeg signals and supervised learning. Master's thesis, NTNU, 2020.

[25] M.M. Bradley and A. Keil. Event-related potentials (erps). In V.S. Ramachandran, editor, *Encyclopedia of Human Behavior (Second Edition)*, pages 79–85. Academic Press, San Diego, second edition edition, 2012. ISBN 978-0-08-096180-4. doi: 10.1016/B978-0-12-375000-6.00154-3.

[26] Shravani Sur and Vinod Sinha. Event-related potential: An overview. *Industrial psychiatry journal*, 18:70–3, 01 2009. doi: 10.4103/0972-6748.57865.

[27] John Dempster. Chapter six - signal analysis and measurement. In *The Laboratory Computer*, Biological Techniques Series, pages 136–171. Academic Press, London, 2001. doi: 10.1016/B978-012209551-1/50039-8.

[28] Ognen A. Petroff, Dennis D. Spencer, Irina I. Goncharova, and Hitten P. Zaveri. A comparison of the power spectral density of scalp eeg and subjacent electrocorticograms. *Clinical Neurophysiology*, 127(2):1108–1112, 2016. ISSN 1388-2457. doi: 10.1016/j.clinph.2015.08.004.

[29] Aurora Saibene, Mirko Caglioni, Silvia Corchs, and Francesca Gasparini. Eeg-based bcis on motor imagery paradigm using wearable technologies: A systematic review. *Sensors*, 23:2798, 03 2023. doi: 10.3390/s23052798.

[30] Nitesh Singh Malan and Shiru Sharma. Time window and frequency band optimization using regularized neighbourhood component analysis for multi-view motor imagery eeg classification. *Biomedical Signal Processing and Control*, 67, 03 2021. doi: 10.1016/j.bspc.2021.102550.

[31] Simanto Saha and Mathias Baumert. Intra-and inter-subject variability in eeg-based sensorimotor brain computer interface: a review. *Frontiers in computational neuroscience*, 13: 87, 2020.

[32] Simanto Saha, Khondaker A. Mamun, Khawza Ahmed, Raqibul Mostafa, Ganesh R. Naik, Sam Darvishi, Ahsan H. Khandoker, and Mathias Baumert. Progress in brain computer interface: Challenges and opportunities. *Frontiers in Systems Neuroscience*, 15, 2021. ISSN 1662-5137. doi: 10.3389/fnsys.2021.578875.

[33] Gerwin Schalk, Dennis Mcfarland, Thilo Hinterberger, NR Birbaumer, and Jonathan Wolpaw. Bci2000: a general-purpose brain-computer interface (bci) system. *IEEE Trans. Biomed. Eng.*, 51:1034–, 07 2004. doi: 10.1109/TBME.2004.827072.

[34] Dimitris G Manolakis and John G Proakis. *Digital signal processing*. Pearson Education, 2014.

[35] Chul Hwan Kim and Raj Aggarwal. Wavelet transforms in power systems. part 1: General introduction to the wavelet transforms. *Power Engineering Journal*, 14(2):81–87, 2000.

[36] Shui-Hua Wang, Yu-Dong Zhang, Zhengchao Dong, and Preetha Phillips. *Wavelet Families and Variants*, pages 85–104. Springer Singapore, Singapore, 2018. ISBN 978-981-10-4026-9. doi: 10.1007/978-981-10-4026-9_6.

[37] Amara Graps. An introduction to wavelets. *IEEE computational science and engineering*, 2(2): 50–61, 1995.

[38] Amara Graps. An introduction to wavelets. *IEEE Computer Society*, 2(2):1, 8–10, 1995.

[39] Samina Khalid, Tehmina Shehryar, and Shamila Nasreen. A survey of feature selection and feature extraction techniques in machine learning. *Proceedings of 2014 Science and Information Conference, SAI 2014*, pages 372–378, 10 2014. doi: 10.1109/SAI.2014.6918213.

[40] Luis Alfredo Moctezuma and Marta Molinas. Classification of low-density eeg for epileptic seizures by energy and fractal features based on emd. *Journal of biomedical research*, 34(3):180, 2020.

[41] Emmanuel Didiot, I. Illina, Dominique Fohr, and Odile Mella. A wavelet-based parameterization for speech/music discrimination. *Computer Speech & Language*, 24:341–357, April 2010. doi: 10.1016/j.csl.2009.05.003.

[42] Firas Jabloun and A. Cetin. The teager energy based feature parameters for robust speech recognition in car noise (pdf). *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, 1:273–276, 04 1999. ISSN 0-7803-5041-3. doi: 10.1109/ICASSP.1999.758115.

[43] Theodore H Bullock, Matthew C McClune, John T Enright, and et al. Use of the fractal dimension for the analysis of electroencephalographic time series. *Electroencephalography and clinical neurophysiology*, 95(2):111–117, 1995. doi: 10.1007/s004220050394.

[44] Chang-Ting Shi. Signal pattern recognition based on fractal features and machine learning. *Applied Science*, pages 3–5, July 2018. doi: 10.3390/app8081327.

[45] Nicole N. Nickens. Measures of central tendency. In *The SAGE Encyclopedia of Educational Research, Measurement*, volume 4, pages 1038–1039. SAGE Publications, Inc., 2018. doi: 10.4135/9781506326139.

[46] Dorothy J. Musselwhite and Brian C. Wesolowski. Variance. In *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation*, volume 4, pages 1790–1793. SAGE Publications, Inc., 2018. doi: 10.4135/9781506326139.

[47] Gonzalo Safont, Addisson Salazar, Antonio Soriano, and Luis Vergara. *Combination of multiple detectors for EEG based biometric identification/authentication*, pages 230–236. IEEE, 10 2012. doi: 10.1109/CCST.2012.6393564.

[48] Hae-Young Kim. Skewness. In *The SAGE Encyclopedia of Educational Research, Measurement*, volume 4, pages 1529–1530. SAGE Publications, Inc., 2018. doi: 10.4135/9781506326139.

[49] John T. E. Richardson. Kurtosis. In *The SAGE Encyclopedia of Educational Research, Measurement*, volume 4, pages 940–942. SAGE Publications, Inc., 2018. doi: 10.4135/9781506326139.

[50] Marco Cocconcelli, Matteo Strozzi, Jacopo Cavalaglio Camargo Molano, and Riccardo Rubini. Detectivity: A combination of hjorth's parameters for condition monitoring of ball bearings. *Mechanical Systems and Signal Processing*, 164(108247):150–155, 2021. ISSN 0888-3270. doi: 10.1016/j.ymssp.2021.108247.

[51] Bo Hjorth. Eeg analysis based on time domain properties. *Electroencephalography and Clinical Neurophysiology*, 29(3):306–310, 1970. ISSN 0013-4694. doi: doi.org/10.1016/0013-4694(70)90143-4.

[52] Qiong Liu and Ying Wu. *Supervised Learning*, pages 3243–3245. Springer US, Boston, MA, 2012. ISBN 978-1-4419-1428-6. doi: 10.1007/978-1-4419-1428-6_451.

[53] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[54] Afek Ilay Adler and Amichai Painsky. Feature importance in gradient boosting trees with cross-validation feature selection. *Entropy*, 24(5):687, 2022. doi: 10.3390/e24050687.

[55] Arash Khoda Bakhshi. Real-time crash prediction for a long low-traffic volume corridor using corrected-impurity importance and semi-parametric generalized additive model. *Journal of Transportation Safety & Security*, 03 2021. doi: 10.1080/19439962.2021.1898069.

[56] Fernando Martinez-Taboada and José Redondo. The siesta (seaav integrated evaluation sedation tool for anaesthesia) project: Initial development of a multifactorial sedation assessment tool for dogs. *PLOS ONE*, 15:e0230799, 04 2020. doi: 10.1371/journal.pone.0230799.

[57] L Breiman. Random forests. *Machine Learning*, 45:5–32, 10 2001. doi: 10.1023/A:1010950718922.

[58] Jafar Tanha, Yousef Abdi, Negin Samadi, Nazila Razzaghi, and Mohammad Asadpour. Boosting methods for multi-class imbalanced data classification: an experimental review. *Journal of Big Data*, 7(1):1–47, 2020. doi: 10.1016/j.eswa.2011.09.033.

[59] Muhammad Tariq Sadiq, Xiaojun Yu, Zhaohui Yuan, Muhammad Zulkifal Aziz, Naveed ur Rehman, Weiping Ding, and Gaoxi Xiao. Motor imagery bci classification based on multivariate variational mode decomposition. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(5):1177–1189, 2022. doi: 10.1109/TETCI.2022.3147030.

[60] Keiron O'Shea and Ryan Nash. An introduction to convolutional neural networks. *CoRR*, abs/1511.08458, 2015.

[61] Rikiya Yamashita, Mizuho Nishio, Richard Do, and Kaori Togashi. Convolutional neural networks: an overview and application in radiology. *Insights into Imaging*, 9, 06 2018. doi: 10.1007/s13244-018-0639-9.

[62] Vernon Lawhern, Amelia Solon, Nicholas Waytowich, Stephen Gordon, Chou Hung, and Brent Lance. Eegnet: A compact convolutional network for eeg-based brain-computer interfaces. *Journal of Neural Engineering*, 15, 11 2016. doi: 10.1088/1741-2552/aace8c.

[63] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 3(1):1–40, 2016.

[64] Fabien Lotte, Laurent Bougrain, Andrzej Cichocki, Maureen Clerc, Marco Congedo, Alain Rakotomamonjy, and Florian Yger. A review of classification algorithms for EEG-based brain-computer interfaces: a 10 year update. *Journal of neural engineering*, 15(3):031005, 2018.

[65] Jasmin Kevric and Abdulhamit Subasi. Comparison of signal decomposition methods in classification of eeg signals for motor-imagery bci system. *Biomedical Signal Processing and Control*, 31, 01 2017. doi: 10.1016/j.bspc.2016.09.007.

[66] Muhammad Tariq Sadiq, Xiaojun Yu, Zhaohui Yuan, Zeming Fan, Ateeq Rehman, Inam Ullah, and Guoqi Li. Motor imagery eeg signals decoding by multivariate empirical wavelet transform based framework for robust brain-computer interfaces. *IEEE Access*, PP:1–1, 11 2019. doi: 10.1109/ACCESS.2019.2956018.

[67] Kaishuo Zhang, Neethu Robinson, Seong-Whan Lee, and Cuntai Guan. Adaptive transfer learning for eeg motor imagery classification with deep convolutional neural network. *Neural Networks*, 136, 12 2020. doi: 10.1016/j.neunet.2020.12.013.

[68] Amardeep Singh, Ali Abdul Hussain, Sunil Lal, and Hans W. Guesgen. A comprehensive review on critical issues and possible solutions of motor imagery based electroencephalography brain-computer interface. *Sensors*, 21(6), 2021. ISSN 1424-8220. doi: 10.3390/s21062173.

[69] Kjesti Brynestad and Erlend Vatsvåg. An asynchronous motor imagerybased brain-computer interface for two-dimensional drone control. Master's thesis, NTNU, 2021.

[70] Abbas Salami, Mohammad Bagher Khodabakhshi, and Mohammad Hasan Moradi. Fuzzy transfer learning approach for analysing imagery bci tasks. In *2017 Artificial Intelligence and Signal Processing Conference (AISP)*, pages 300–305, 2017. doi: 10.1109/AISP.2017.8324101.

[71] Le Song and Julien Epps. Classifying eeg for brain-computer interfaces: learning optimal filters for dynamical system features. In *Proceedings of the 23rd international conference on Machine learning*, pages 857–864, 2006.

[72] Jaime Riascos Salas. *Do I have a third arm? Towards a Supernumerary Motor Imagery Brain-Computer Interface in Virtual Reality*. PhD thesis, Federal University of Rio Grande do Sul, 03 2019.

[73] Michael Tangermann, Klaus-Robert Müller, Ad Aertsen, Niels Birbaumer, Christoph Braun, Clemens Brunner, Robert Leeb, Carsten Mehring, Kai Miller, Gernot Mueller-Putz, Guido Nolte, Gert Pfurtscheller, Hubert Preissl, Gerwin Schalk, Alois Schlögl, Carmen Vidaurre, Stephan Waldert, and Benjamin Blankertz. Review of the bci competition iv. *Frontiers in Neuroscience*, 6, 2012. ISSN 1662-453X. doi: 10.3389/fnins.2012.00055.

[74] David Steyrl, Reinhold Scherer, Oswin Förstner, and Gernot R Müller-Putz. Motor imagery brain-computer interfaces: random forests vs regularized lda-non-linear beats linear. In *Proceedings of the 6th international brain-computer interface conference*, pages 241–244, 09 2014. doi: 10.3217/978-3-85125-378-8-61.

[75] Kai Keng Ang, Zheng Yang Chin, Chuanchu Wang, Cuntai Guan, and Haihong Zhang. Filter bank common spatial pattern algorithm on bci competition iv datasets 2a and 2b. *Frontiers in Neuroscience*, 6, 2012. ISSN 1662-453X. doi: 10.3389/fnins.2012.00039.

[76] Yousef Rezaeitabar and Ugur Halici. A novel deep learning approach for classification of eeg motor imagery signals. *Journal of Neural Engineering*, 14:016003, 02 2017. doi: 10.1088/1741-2560/14/1/016003.

[77] Joshua Giles, Kai Keng Ang, Lyudmila S. Mihaylova, and Mahnaz Arvaneh. A subject-to-subject transfer learning framework based on jensen-shannon divergence for improving brain-computer interface. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3087–3091, 2019. doi: 10.1109/ICASSP.2019.8683331.

[78] R Leeb, C Brunner, G Müller-Putz, A Schlögl, and G Pfurtscheller. Bci competition 2008–graz data set b. *Graz University of Technology, Austria*, pages 1–6, 2008.

[79] Ander Ramos-Murguialday and Niels Birbaumer. Brain oscillatory signatures of motor tasks. *Journal of neurophysiology*, 113:jn.00467.2013, 03 2015. doi: 10.1152/jn.00467.2013.

[80] Kjersti Brynestad. EEG-based motor imagery classification using dwt-based feature extraction with svm and riemannian geometry-based classifiers. Technical report, NTNU, 2020. DOI: 10.13140/RG.2.2.23210.13761.

[81] Talha-Kedir Bousbia-Salah, Ait-Ameur. Biorthogonal wavelet for eeg signal compression (pdf). *Mechanical Systems and Signal Processing*, 2011. doi: pdf/10.1145/2093698.2093707.

[82] Na Ji, Liang Ma, Hui Dong, and Xuejun Zhang. Eeg signals feature extraction based on dwt and emd combined with approximate entropy. *Brain Sciences*, 9:201, 08 2019. doi: 10.3390/brainsci9080201.

[83] OpenBCI. MNE-Features: Extracting features from eeg data with MNE-Python. `https://mne.tools/mne-features/generated/mne_features.univariate.compute_pow_freq_bands.html`, 2021.

[84] Luis Alfredo Moctezuma and Marta Molinas. Eeg channel-selection method for epileptic-seizure classification based on multi-objective optimization. *Frontiers in neuroscience*, 14:593, 2020.

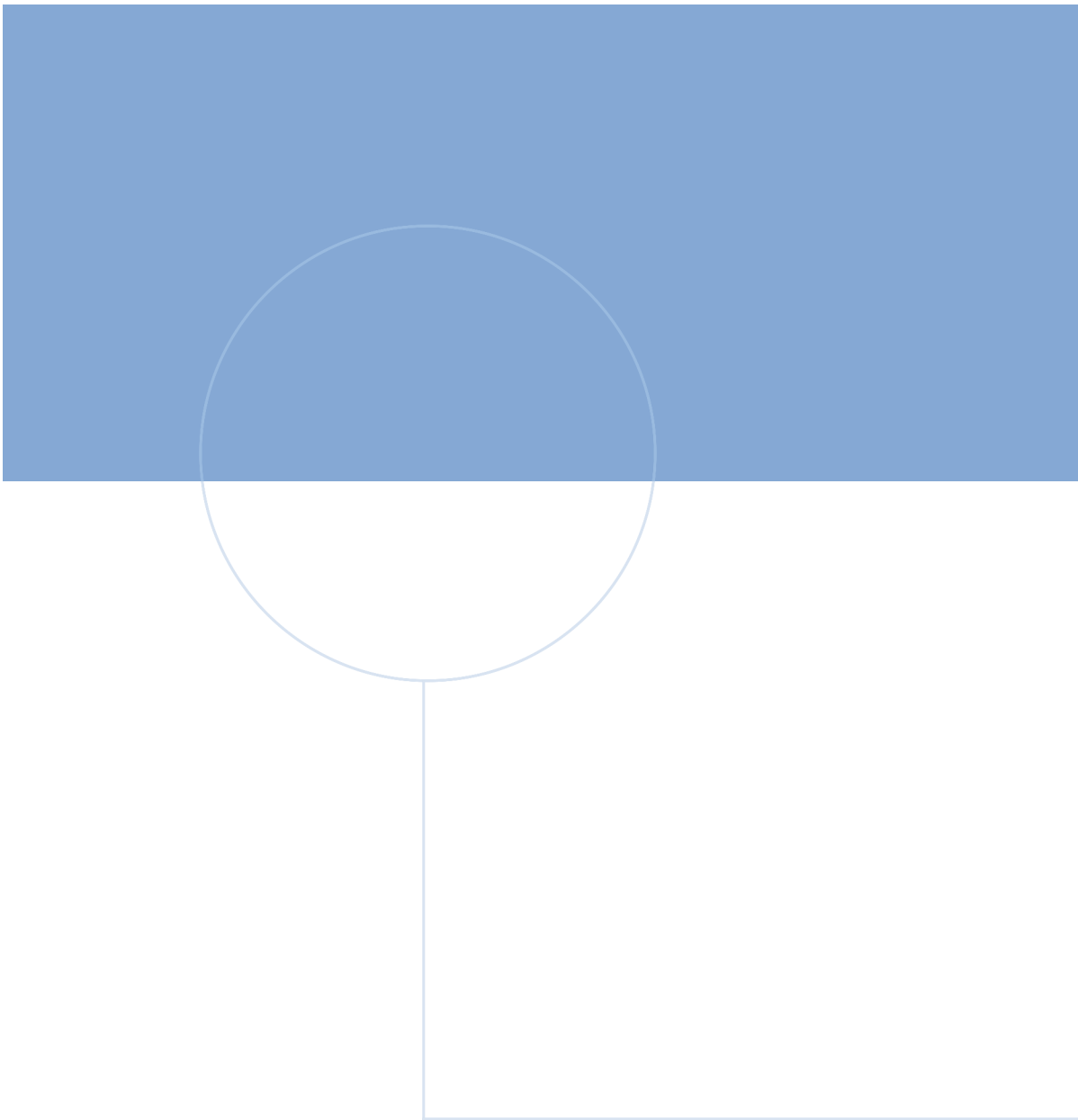
[85] Aleksandra Vuckovic and Bethel Osuagwu. Using a motor imagery questionnaire to estimate the performance of a brain-computer interface based on object oriented motor imagery. *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology*, 124, 03 2013. doi: 10.1016/j.clinph.2013.02.016.

[86] Babak Mahmoudi and Abbas Erfanian. Electro-encephalogram based brain–computer interface: improved performance by mental practice and concentration skills. *International Federation for Medical and Biological Engineering*, 44:959–968, 2006. doi: 10.1007/s11517-006-0111-8.

[87] Tianyu Jia, Chong Li, Xinyu Guan, and Linhong Ji. Enhancing engagement during robot-assisted rehabilitation integrated with motor imagery task. In *Proceedings of the 2019 International Conference on Intelligent Medicine and Health*, ICIMH 2019, page 12–16. Association for Computing Machinery, 2019. ISBN 9781450372862. doi: 10.1145/3348416.3348420.

[88] Diana H. Romero, Michael G. Lacourse, Kristen E. Lawrence, Steven Schandler, and Michael J. Cohen. Event-related potentials as a function of movement parameter variations during motor imagery and isometric action. *Behavioural Brain Research*, 117(1):83–96, 2000. ISSN 0166-4328. doi: 10.1016/S0166-4328(00)00297-7.

[89] Luis Moctezuma and Marta Molinas. Towards a minimal eeg channel array for a biometric system using resting-state and a genetic algorithm for channel selection. *Scientific Reports*, 10, 09 2020. doi: 10.1038/s41598-020-72051-1.



NTNU

Norwegian University of
Science and Technology