Recognizing Thoughts from Bioelectric Patterns? A Brain-Computer Interface with Deep Learning

Jaakko Laiho

School of Science

Thesis submitted for examination for the degree of Master of Science in Technology.

Utrecht 31.07.2020

Thesis supervisor:

Prof. Alexander Jung

Thesis advisor:

M.Sc. Rosa Siervo

Aalto University
School of Science

AALTO UNIVERSITY
SCHOOL OF SCIENCE

ABSTRACT OF THE
MASTER'S THESIS

Author: Jaakko Laiho

Title: Recognizing Thoughts from Bioelectric Patterns? A Brain-Computer Interface with Deep Learning

Date: 31.07.2020 Language: English Number of pages: 6+61

Program: EIT Digital Master School
Major: Embedded Systems

Supervisor: Prof. Alexander Jung

Advisor: M.Sc. Rosa Siervo

The study aims to implement an inter-person brain-computer interface (BCI) capable of classifying brain activity using a deep learning approach on motor imagery electroencephalography (MI EEG) data. The BCI uses a convolutional neural network (CNN) for feature extraction and classification. Offline training is performed on a new and pre-trained model using filtered EEG signals from an open-source 9-subject 4-class MI EEG data set. Performance is assessed within and between subjects. The new model reaches a mean evaluation accuracy of 58% ± 15% within-subject improved by 2.4% ± 1.3% using a pre-trained model. Between-subjects performance is low with a mean accuracy of 35% ± 11%.

Keywords: Motor Imagery, Electroencephalography, Brain-Computer Interfaces, Convolutional Neural Networks, Transfer Learning

AALTO-YLIOPISTO
PERUSTIETEIDEN KORKEAKOULU

DIPLOMITYÖN
TIIVISTELMÄ

Tekijä: Jaakko Laiho

Työn nimi: Ajatuksen tunnistaminen aivosähkön kuvioista? Aivokäyttöliittymä toteutettuna syväoppimisella

Päivämäärä: 31.07.2020 Kieli: Englanti Sivumäärä: 6+61

Program: EIT Digital Master School
Major: Embedded Systems

Työn valvoja: Prof. Alexander Jung

Työn ohjaaja: M.Sc. Rosa Siervo

Työssä tutkitaan aivokäyttöliittymää syväoppimisen menetelmillä. Tavoitteena on toteuttaa yleistettävä malli luokittelemaan aivosähkökäyrällä kuvattua motorisiin mielikuviin liittyvää sähköistä aivotoimintaa. Aivokäyttöliittymä poimii piirteitä ja luokittelee aivotoimintaa konvoluutioneuroverkkoihin pohjustuvalla mallilla.
Malli optimoidaan käyttäen esisuodatettua avoimen lähdekoodin tiedostoa, joka käsittää yhdeksän henkilön aivotoimintaa ja jossa kukin toistaa neljää motorista mielikuvaa kuvattuna elektroenkefalografialla. Mallin suorituskykyä arvioidaan sekä henkilöidenvälisesti että -sisäisesti.
Malli luokitteli aivotoimintaa 58%±15% tarkkuudella ennalta-alustamattattomilla painoarvoilla henkilöidensisäisesti. Mallin tarkkuus parani 2.4%±1.3% siirtovaikutuksella. Malli saavutti henkilöidenvälisesti matalamman 35%±11% tarkkuuden.

Avainsanat: Motoriset mielikuvat, Elektroenkefalografia, Aivokäyttöliittymä, Konvoluutioneuroverkko, Siirtovaikutus

Preface

When I started my thesis work on this project I was at first unacquainted with the fields related to brain-computer interfaces (e.g. neuroscience, machine learning). To design and implement a system capable of understanding the inner workings of a human mind in real-time, and to convert these insights into actionable signals has been demanding for me. The challenge has led me to the deep end of a whole new set of technologies and fields of science - a journey filled with numerous obstacles, many of which seemed insurmountable at first. However, the pursuit of novel technology with the potential to change fundamentally how we interact with our environment to increase the quality of life has kept energy levels high, despite a tumultuous world.

The work would have been impossible without the abundance of support from my friends and family. I would like to give a special thanks to my advisor Rosa Siervo for her continually shared expertise, guidance, and encouragement, without which I would have not made it far. Also, thanks go to the product owner Youri de Koster for presenting such an interesting challenge and allowing me to work on it. I also want to thank my friends and family who have made the journey enjoyable and memorable, and who have invariably helped me to find ways forward at difficult times. Last but not least I want to thank Professor Alex Jung for supervising this thesis.

Additionally, I would like to mention that Aalto university and the EIT Digital Master School have enabled me to acquire a diverse and high-quality education in ICT innovation and management at multiple European locations, and I look forward with anticipation to what the future may bring.

I hope you will enjoy reading this paper and are as excited by the field of brain-computer interfaces and its potential as I have been when producing this Master's Thesis.

Utrecht, 31.07.2020

Jaakko O. Laiho

Contents

Abbreviations

BCI	Brain-Computer Interface
ERD	Event-Related Desynchronization
ERS	Event-Related Synchronisation
ERP	Event-Related Potentials
EP	Evoked Potentials
EEG	Electroencephalography
EOG	Electrooculography
MEG	Magnetoencephalograpy
MI	Motor Imagery
ML	Machine Learning
DL	Deep Learning
CNN	Convolutional Neural Network
SNR	Signal-to-Noise Ratio
ANN	Artificial Neural Network
AI	Artificial Intelligence
ELU	Exponential Linear Unit

1 Introduction

A brain-computer interface (BCI) connects the human brain directly to an external device by measuring user brain activity and translating patterns in that activity into control signals or messages for such an external device (B.-H. Lee et al. 2020; H. K. Lee and Choi 2018; Kar et al. 2018). The artificial output replaces, restores, enhances, supplements, or improves natural central nervous system output by design (J. Wolpaw and E. W. Wolpaw 2012).

BCI is an impactful field of study where research has been conducted to support the recovery, replacement, and extension of natural movement capabilities for motor disabled patients and healthy users alike. BCI empowers such users by enabling the user to control applications and devices in their surroundings interactively through only brain waves, i.e. without a need for muscle activation. Examples of such applications could be a prosthesis, computer, wheelchair, text input system, a novel gaming input device, stroke rehabilitation device, or exoskeleton (Schwartz et al. 2006). Such a solution can also be used to improve the quality of life for subjects who suffer from neuromuscular diseases that limit conventional communication that would otherwise take place through speech or gestures, but that leave the brain otherwise unimpaired, allowing for the retention of cognitive capabilities. Such a scenario could take place for reasons such as having lost motor control because of amyotrophic lateral sclerosis (ALS) (Wijesekera and Leigh 2009), stroke, or spinal cord injury.

Many current BCI solutions make use of electroencephalography (EEG) as it is a way to measure brain activity in a manner that is low-cost, non-invasive, portable, and has a high temporal resolution or sampling rate (Xu et al. 2019; B.-H. Lee et al. 2020; H. K. Lee and Choi 2018; Ortiz Echeverri et al. 2019; Lu, Yin, and Jing 2019). EEG detects the electrical activity of brain cells by continuously measuring potentials between multiple electrodes placed on the scalp (Gevins et al. 1994). The electrodes might be placed over specific parts of the brain. Different brain activities result in different potential patterns, which can be distinguished to some extent. One of the well known and distinguishable EEG patters is those created by motor imagery (MI). MI is a mental strategy in which the user imagines some physical activity, regardless of that activity being executed. Those imagined activities can for instance include opening or closing the left or right hand (Kar et al. 2018; Xu et al. 2019; Lu, Yin, and Jing 2019; Freer and Yang 2020). The brain activity created by such a task can then be decoded based on changes in brain frequency bands in the EEG recordings. The benefit of MI over other EEG patterns is that does not require the configuration of additional external devices. Other such patters might, for example, require the presentation of visual stimulus to elicit specific brain activity. Additionally, MI can be performed voluntarily at short notice.

However, MI EEG control applications for daily assistance can require as much as 20 hours of training split over as many as 50 experimental sessions to get started (Freer and Yang 2020), which is less than ideal for assistive technology. Despite progress made, there is still much room for improvement regarding the accuracy and usability of an MI-EEG-based BCI.

Brain activity decoding methods for BCI perform three operations: The removal of noise from the signal, the reduction of raw data to a more compact set of features, and the classification of those features. Until recently, performing these methods would require significant expertise in EEG signal processing. However, deep learning methods are a popular choice in MI classification research in recent years (Lotte et al. 2018; Zhang et al. 2019) due to having the ability to self-learn needed features from raw data without the need for manual noise removal or feature extraction. This makes them well suited for end-to-end learning in multi-class classification tasks (Schirrmeister et al. 2017; Uktveris and Jusas 2017). A convolutional neural network (CNN) is a type of artificial neural network (ANN) that can learn local patterns in a signal by combining local low-level features from raw input to increasingly high-level features (Y. Zhang et al. 2019).

CNNs draw inspiration from the animal visual cortex (Uktveris and Jusas 2017) and are used with great success in image recognition (Krizhevsky, Sutskever, and Geoffrey Hinton 2012) and acoustic modeling (G. Hinton et al. 2012). However, a static two-dimensional image on which CNNs are known to be successful is a different type of data than the EEG signal, which is a dynamic time series with multiple channels obtained from the scalp surface with electrodes (Gevins et al. 1994). In the simplest case, the EEG signal is one-dimensional. However, to use two-dimensional image classification, the signal needs to be transformed using some method. The mapping of the signal to a time-space representation generates a 2D representation of EEG (Tang, Li, and Sun 2016), in which case the classifying model is interested in the variation of electrical activity across various places on the scalp over some duration of time. Alternatively, the mapping of EEG to a time-frequency representation also generates a 2D EEG signal (Lawhern et al. 2018), which decomposes an EEG signal from a small number of electrodes to their frequency components over some time.

However, despite many examples of impressive progress, there is still room for considerable improvement concerning several important aspects of information extraction from the EEG, including its accuracy, generalizability, and usability for online applications. Further, Zhang et al. 2019 recognize that person-independent classification with high performance is a prerequisite for a widespread diffusion of BCI technology.

This paper proposes a CNN-based model to perform feature extraction and classification for MI EEG signals. It attempts answers the following research questions:

- How accurately can the proposed model classify motor imagery-based brain activity in a brain-computer interface process *within subjects*?
- How accurately can the proposed model classify motor imagery-based brain activity in a brain-computer interface process *between subjects*?

The study conducts three experiments using the proposed model to assess classification accuracy performance to answer those research questions. Firstly, an experiment is conducted using the CNN-based model to assess classification accuracy within-subjects. Next, an experiment examines whether a between-subjects approach reaches comparable accuracies to the within-subject procedure. Lastly, an

experiment assesses whether accuracy can be increased by using transfer learning. The model incorporates a CNN with 4 convolutional blocks and a fully connected output. The model is trained using trials from the public BCI competition IV 2a data set, which consists of 4-class MI trials across 9 subjects (Brunner et al. 2008) to classify minimally processed raw EEG data containing MI brain activity patterns. The model takes as input a 2-dimensional time-space representation of segments of MI EEG trials, which resembles a buffered real-time EEG data stream from a commercially available BCI headset. Every trial contains the performance of exactly one motor imagery task. The model is optimized by cross-entropy loss and performs multi-class classification by assigning a probability for each class of motor imagery task that the trial corresponds to that task.

The remainder of the paper is organized followingly: Section 2 provides background information on BCI and CNN. Section 3 describes the experiments. Thereafter, section 4 presents the contribution and results of the work. Lastly, section 5 provides a discussion and summary of the work.

2 Background

The human brain is an organ that processes information in a highly complex, nonlinear, and parallel way. Not only is it a focal point of this study, but it also inspires the tools themselves used to analyze the brain. The background first touches upon the biology of the relevant components of the human head. Next, it elaborates on electroencephalography, which is a relevant signal acquisition technique of brain activity. The elaboration includes a description of motor imagery, which is a strategy a user can employ to generate specific brain activity. An abstract brain-computer interface process explains the eventual classification of brain activity. Next, the chapter introduces the methods of machine learning used in the implementation of such a process and presents the public data set that contains data relevant to the study.

2.1 Biology

Brain

The nervous system is composed of two parts - the central nervous system CNS, and the peripheral nervous system. The main organ of the central nervous system is the brain, which contains approximately 100 billion nerve cells or neurons. For a model of the brain, see the subsection 'Artificial Neural Networks'. The brain itself consists of three parts - the cerebrum, whose surface is known as the cerebral cortex, the cerebellum, or the little brain highlighted in figure 2, and the brain stem.

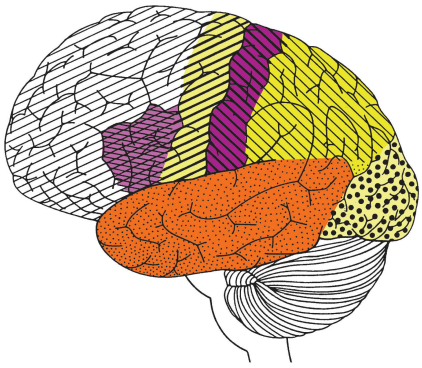

Figure 1: Cerebrum

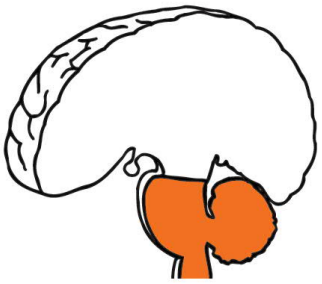

Figure 2: Cerebellum

Further, the cerebrum has four main sections highlighted in figure 1; the frontal, parietal, temporal, occipital lobes. The portion of the frontal lobes that is furthest to the back is the motor area, which is responsible for the control of voluntary movement. The forward areas of the parietal lobes contain the sensory areas which process sensory information, excluding sight or sound. The furthest back areas of the brain are known as the occipital lobes, and they handle visual information. The temporal lobe, which lies underneath the frontal and parietal lobes, processes sound.

The nerve cell

Nerve cells respond to stimuli and transmit information. They comprise of one large branch or axon, multiple short branches or dendrites, and cell bodies or soma. See figure 3 for a visualization. Axons deliver electrical impulses over long distances as they can be over 1 meter long. Dendrites are connected to axons or other dendrites and receive impulses from other nerves. Cell bodies contain most of the metabolism of the nerve cells. In the human brain, a nerve connects to approximately 10 000 other nerves through these dendritic connections. Communication between nerve cells takes place along the dendrites, and the end of the axon converts an electrical signal into a chemical signal and back. Furthermore, Fries 2005 find that to communicate effectively, neuron groups need to achieve neuronal coherence. This means that the activation of neuron groups takes place rhythmically. The rhythmic excitation of neurons results in oscillatory bioelectrical brain activity, which emerges in the form of brain waves. Importantly, the simultaneous activation of clusters of thousands of neurons firing generates current and change in electrical potential potent enough, such that the event can be observed on the person's scalp.

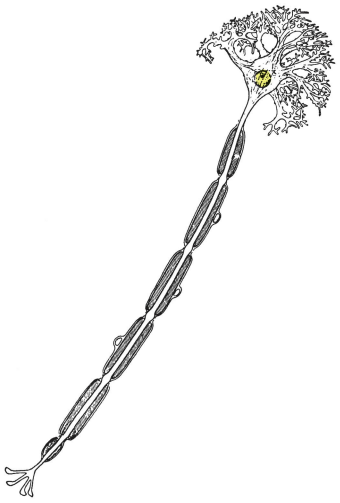

Figure 3: Neuron Architecture

Skin

The electrical properties on the surface of the skin affect the detectability of electrical activity on the scalp, which is important when brain activity is monitored non-invasively through electrical activity on the scalp. Exposed skin has an impedance of around 200 kOhms and skin covered by the hair has an impedance of 120 kOhms in the frequencies relevant to BCI (Rosell et al. 1988). However, the hair itself obstructs physical access to the surface of the skin even though the impedance of hair-covered skin is lower than uncovered skin. In any case, the use of natural skin grease and abrasion of the skin can reduce skin impedances to below 10 kOhms.

2.2 Electroencephalography

Electroencephalography (EEG) devices record the changes in the electrical activity of the brain from the surface of the scalp. By placing multiple electrodes onto the scalp, the EEG device reads the voltage differences generated by brain activity between those electrodes on the surface of the head. However, brain activity is not the only source of signals to the scalp; extra-cerebral source signals come in two forms. Physiological activities such as blinking and the movement of eyes, as well as heart activity, generate electrical activity over the skin, as do the electrical properties of the measurement devices. A group of neurons roughly the size of a diameter of 2.5cm firing together generates enough signal to be measured from the scalp. In other words, clusters of thousands of neurons firing simultaneously produce the raw data output that is an EEG signal. Such an event generates enough current and change in electrical potential that the event can be recorded on the person's scalp non-invasively. (Crosson et al. 2010)

EEG electrode placement employs a standard international system known as the 10/20 EEG system. Image 4 describes sensor placements by using that radial coordinate system. The top of the skull is the origin.

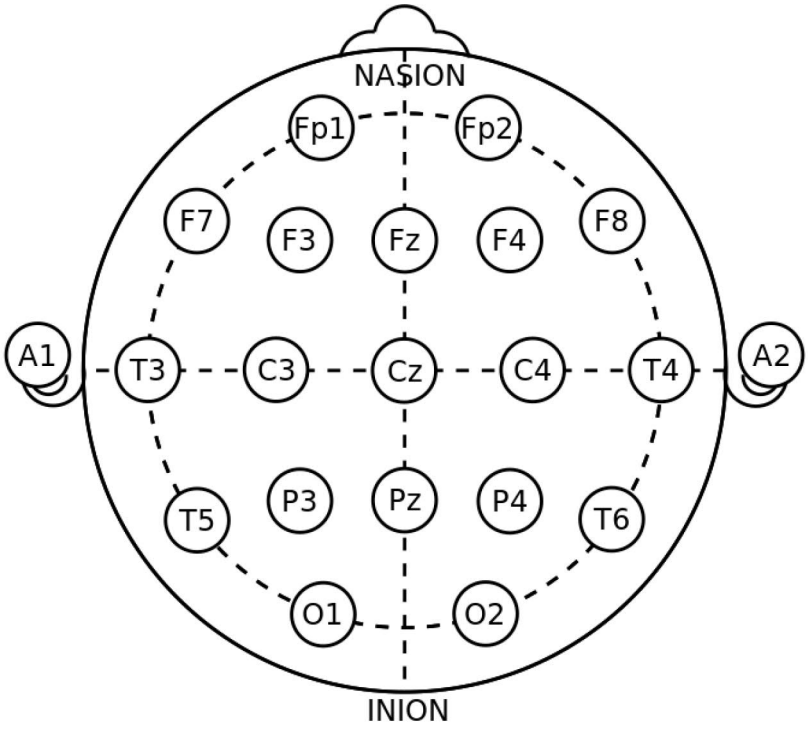

Figure 4: International 10/20 system for EEG sensor placement

In other words, an EEG device measures electrical activity on the scalp. More specifically, EEG records the voltage difference between at least 2 electrodes. The resolution of an EEG device is quantified spatially and temporally. EEG spatial

resolution is in the magnitude of centimeters, because of attenuation or weakening of the signal due to traveling through the low-conductivity bone of the skull; noise signals reduce resolution as well. Spatial resolution can be improved by increasing the number of electrodes used, as well as by using certain filters, or by switching altogether to hybrid BCI devices, which use a combination of different signal sampling modalities. Conversely, the temporal resolution of EEG devices is high - EEG can sample a thousand measurements over the scalp in one second across different sensors. While EEG measures activity in real-time, its ability to measure electrical activity in deeper parts of the brain is limited. (Crosson et al. 2010)

The two types of brain activities known as evoked potentials (EP) and spontaneous EEG referred to in Zhang et al. 2019 are sub-categories neurological phenomena measurable by EEG known to be useful for BCI. The key difference between the two categories of EEG signals is in the source of the stimulus: Spontaneous EEG emerges as a consequence of internally generated stimuli, while EP relies on the presence of external stimuli.

2.2.1 Evoked potential

EP is mainly concerned with the detection of a known waveform in response to external stimuli; these responses are generally robust between individuals and well stereotyped. Event-related potentials (ERP) and steady-state evoked potentials are subcategories of EP signals. They are both further categorized to EP generated through visual, auditory, and somatosensory stimuli, as expanded in table 1. ERPs detect a high-amplitude and low-frequency response to time-locked external stimuli. (Zhang et al. 2019)

	Event-Related Potential	Steady-State Evoked Potential
Visual	(VEP)	(SSVEP)
Auditory	(AEP)	(SSAEP)
Somatosensory	(SEP)	(SSSEP)

Table 1: Evoked potentials

For example, P300 potentials are a well-known visual evoked potential (VEP) signal. Presenting a stimulus in a time-locked event causes the appearance of the classical P300 waveform, which emerges approximately 300 ms after the stimulus. The timing of the signal varies from 250 ms to 900 ms, and its amplitude varies between 5 μV and 20 μV for AEP and VEP. Sutton et al. 1965, who first described it, argue that it may be the most-studied ERP component in selective attention and information processing. They argue that this because P300 has a relatively large amplitude and is effortless to elicit in experiments. (Patel SH 2005)

2.2.2 Spontaneous EEG

Spontaneous or oscillatory BCIs are concerned with the signal powers of EEG frequency bands. These features have a lower signal-to-noise ratio, and more variation between individuals and are thereby more difficult to train for in comparison to EP BCI devices. Examples of spontaneous EEG signals are sleeping EEG, emotional EEG, and motor imagery. Health analysis uses *Sleeping EEG* signals to recognize sleep stages and diagnose sleep disorders. *Emotional EEG* is evaluated in three metrics: valence, arousal, and dominance. The combination of the three metrics forms emotions such as fear, sadness, and anger, and emotion is affected by subjective and environmental factors. Also, gender plays a big part; fear-related emotions are diverse with women, and sadness-related emotions are diverse with males. *Motor imagery* (MI) involves the rehearsal of imaging motion while remaining immobile, which is a counter-intuitive task for most people - getting a feel for MI takes time and practice. An MI action can range from a first-person kinesthetic experience to a third-person view of one's motion, or even the manipulation of an imaginary object (Zhang et al. 2019). Motor imagery has the advantage over the other spontaneous EEG applications in that the user can generate brain activity signals at will with a higher degree of control, at least in comparison to those other spontaneous activities. Further, the user can act without a need for some device by which to generate external stimulus. Overall, MI is particularly applicable to generating control signals in a BCI context due to its voluntary and internal nature.

EEG waves

A BCI observes brain activity through EEG in the form of neural oscillation. This oscillatory signal is caused by the rhythmic firing of neurons that enables their communication (Fries 2005). A mathematical description of neural oscillation consists of frequency, amplitude, and phase, and the different neural frequencies connect to different brain states. A BCI can detect brain states such as deep sleep or unconsciousness, transitioning between deep sleep and wakefulness, inactive wakefulness and relaxation, active wakefulness, and strong concentration and learning from the oscillations. These brain waves are called the delta, theta, alpha, beta, and gamma waves, respectively. Moreover, alpha waves are known usually to occupy the occipital or visual lobes (Y. Zhang et al. 2019). However, the brain waves of the same frequency taking place in the motor cortex instead are known as the mu or sensorimotor waves (Ortiz Echeverri et al. 2019). These are particularly significant in BCI applications. The frequencies to which all aforementioned waves correspond are enumerated in more detail in table 2.

Table 2: Brain waves, descriptions, and frequencies

Brain Activity		
Wave	State	Frequencies
Delta	Deep meditation	0.5 - 3Hz
Theta	Dream	3 - 8Hz
Alpha	Presence	8 - 12Hz
Mu	Voluntary movement	8 - 12Hz
Beta	Active	12 - 38Hz

Motor Imagery and brainwaves

The performance of MI affects brain activity similar to voluntary muscle activation: The signal powers in certain frequency bands change during a voluntary muscle activation or the imagination of voluntary muscle activation. For example, it is known that the imagination of right and left-hand movement results in the attenuation or reduction of signal power in the alpha wave (8-12 Hz) from the contralateral motor cortex area of the brain, also referred to as mu waves, and in a power increase in the beta wave (12-30Hz) from the ipsilateral area of the brain (Kar et al. 2018). These effects are known as event-related desynchronisation (ERD) and event-related synchronisation (ERS), respectively (G. Pfurtscheller and Silva 1999). In other words, the synchronization or desynchronization in the mu and beta bands in the motor cortex distinguishes MI task performance. However, MI exhibits large variation in brain activity between subjects and requires longer training times from users in comparison to EP signals, such as P300 based devices or SSVEP devices. Nonetheless, the lack of external stimuli needed makes it so that MI-based BCI systems can be easier to accept by users (Ortiz Echeverri et al. 2019).

2.3 Brain Computer Interface

Brain-computer interfacing (BCI) is the act of reading brain activity and providing a feedback loop. A more precise definition for a BCI device is such that it meets four criteria: (Gert Pfurtscheller et al. 2010)

1. Direct measurement of brain activity
2. Provides user with feedback
3. Operates online
4. Relies on intentional control: voluntarily choose to perform a mental ask to accomplish some goal

However, while the above better describes active BCIs, there are also passive BCI devices that rely on other forms of control. BCI measures brain activity and generates an output to replace, restore, enhance, supplement, or improve the natural output of brain activity (J. Wolpaw and E. W. Wolpaw 2012). In other words, the BCI changes how the central nervous system interacts with its external or internal environment. Examples of passive BCIs include attention monitoring while driving (Lin et al. 2007) and mood monitoring (Zhang et al. 2019). The development of novel use cases targets the supplementing of natural central nervous system output with applications that control a robotic arm or adjust the indoor lighting with thought alone. Furthermore, BCI devices can utilize one or more categories of brain activity signals; a generic or hybrid BCI uses more than one category of brain activity. Studies compare BCI devices by classification accuracy, average detection time, and information transfer rate. The ratio of correct classifications by all classifications equals classification accuracy. The average detection time takes the difference between a stimulation symbol being presented and a generated command. The information transfer rate describes the amount of information transferred between the human brain and the BCI per minute.

A typical BCI application follows a pattern. Firstly, an offline training phase takes place, where a model is defined and potentially pre-trained. Secondly, the online operational phase recognizes and translates brain activity patterns into commands in real-time.

Figure 5: BCI Application Flow

The second step follows also follows a pattern. The user generates a brain activity pattern using some technique, for example by performing a motor imagery task. These activity patterns are then measured with tools like electroencephalography. The measurements might be pre-processed by filtration with spatial and spectral filters. The BCI obtains features from the pre-processed signals for compact representation purposes, then classifies and consequently translates the features into application commands, and lastly, presents feedback the user to inform the user of the successfulness of the mental command. The process of an online BCI consists of signal acquisition, pre-processing, feature engineering, classification, sending control commands, and feedback. See figure 6 for a visualization.

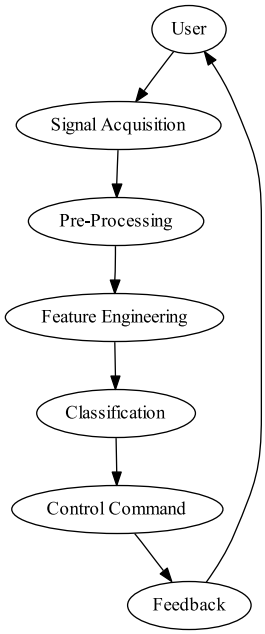

Figure 6: BCI Process

2.3.1 Signal acquisition

Signal acquisition refers to the action of measuring the output of brain activity. Brain activity can be measured to generate signals with electroencephalography (EEG), which is relevant to this study. Alternatively, magnetoencephalography (MEG), functional magnetic resonance imaging (fMRI), or near-infrared spectroscopy (NIRS) are technologies that also measure brain activity. These technologies differ mainly between cost, availability, temporal resolution, and spatial resolution. See table 3 for details. (Crosson et al. 2010)

Table 3: Signal acquisition technology comparison

Technique	Cost	Availability	Temporal Resolution	Spatial Resolution
EEG	low	high	high	low
MEG	high	high	high	low
fMRI	high	low	high	high
NIRS	low	high	low	high

BCI devices come in three varieties, as far as the signal acquisition method invasiveness is concerned. BCI invasiveness is divided followingly:

- Non-invasive methods use sensors placed on the scalp measuring the electrical or magnetic fields generated by the brain,
- semi-invasive methods place electrodes on the exposed surface of the brain, and
- invasive methods place microelectrodes directly into the cortex to measure single-neuron activity.

EEG is a popular method in both science and in commercial use cases because of high usability, low user risk, and the temporal resolution. An EEG device can come in the form factor of an electrode-fitted cap, headset, or even earbuds. Moreover, EEG signals and their subcategories are dominant in recent research according to Zhang et al. 2019. However, EEG comes with several drawbacks. EEG signals are limited by low spatial resolution, which means that, for example, pinpointing brain activity to some single brain cell is not possible (Freer and Yang 2020). Furthermore, the EEG signals have a low signal-to-noise ratio (SNR) which is because that EEG is sensitive to noise from muscle movement such as blinking, and power line noise (Lu, Yin, and Jing 2019). This means that the signal is much affected by non-task-relevant sources, which makes decoding the user's intent fundamentally difficult. Moreover, EEG data is non-stationary (Liyanage et al. 2013), which means that the statistical properties of MI EEG activity change over time, exhibiting wide variance within one subject. In addition, the EEG patterns created by MI vary greatly between subjects as well (B.-H. Lee et al. 2020; Lu, Yin, and Jing 2019).

2.3.2 Pre-processing

Pre-processing filters the signal for noise and removes external signal sources. Raw EEG data are noisy, and the noise can come from three sources, which are the BCI equipment (including leads and electrodes and overall recording system), electrical interference (from the external AC/DC power sources), and the subject's muscle movements (like eye movement and blinking and head or jaw movement and heart activity). Accounting for the aforementioned noise artifacts can require sophisticated tooling, which includes some of the following; According to Lakshmi, Prasad, and Prakash 2014, tools commonly used for this purpose are beyond the use of filters are Independent Component Analysis (ICA), Common Average Referencing (CAR),

Surface Laplacian (SL), Principal Component Analysis (PCA), Common Spatial Patterns (CSP), and Adaptive Filtering (AF). ICA decomposes the EEG data by the characteristics of that data to components to separate artifacts from the signal. CAR removes noise by removing common activity which can be the noise present in the signal. SL estimates the current density that enters or leaves the scalp, which is a way to increase spatial resolution. PCA reduces the dimensionality of data based on the variability of signal properties. CSP transforms an EEG signal into a variance matrix that discriminates between classes maximally by using spatial information to detect patterns. AF can modify signals even if the signal and noise overlap. The advantages and disadvantages of each technique are enumerated in table 2.3.2. Furthermore, it is also possible to control the subject-sourced noise in a research setting by simply discarding any contaminated samples.

Table 4: Comparison of preprocessing methods by Lakshmi, Prasad, and Prakash 2014

Technique	Advantages	Disadvantages
ICA	Computationally efficient. High performance for large data. Decomposes signals into temporal independent and spatial fixed components	Inapplicable for under determined cases Requires more computations for decomposition.
CAR	Outperforms reference methods Yields improved SNR	Finite sample density and incomplete head coverage cause problems in calculating averages
SL	Robust against artefacts generated at regions not covered by electrode cap. Solves electrode referencing	Sensitive to other artefacts Sensitive to spline patterns
PCA	Reduces dimensionality Ranking helps classify data Low loss of information	Underperforms compared to ICA. Assumes data is linear and continuous. Fails to process data for complicated manifold
CSP	Doesn't require knowledge or a priori selection of specific bands	Requires large electrode count Electrode position changes may affect performance.
AF	Modifies signal features according to signals being analyzed Works well for signals with overlapping spectra	

About filters

A BCI system typically pre-processes EEG data by filtering out the frequency ranges beyond those ranges that contain the brain activities of interest. A low-pass filter allows frequency components below some cutoff frequency in some signal to pass while filtering any frequencies above that cutoff frequency. A high-pass filter works oppositely, letting high frequencies pass. A notch filter cuts out all frequencies between some low and high cutoff in a signal, while a bandpass allows the frequencies

in that band between the low and high cutoff pass unaffected. A Butterworth filter is an example of a bandpass filter commonly used due to being maximally flat in the passband, allowing for the full retention of the desired signal while rejecting all frequencies in the stopband (Butterworth 1930). Additionally, all the subsystems in the overall system must be causal for the system to be realizable, which means that the output depends only on past and present input. This must be the case because any online system can only act on such input. Increasing the order of the Butterworth filter increases the steepness of the response roll-off at the higher cutoff resulting in a better filter at the expense of requiring more computation. See figure 7 for a 2-second segment of a raw brain activity signal filtered using a Butterworth bandpass filter to extract theta, alpha, and beta waves.

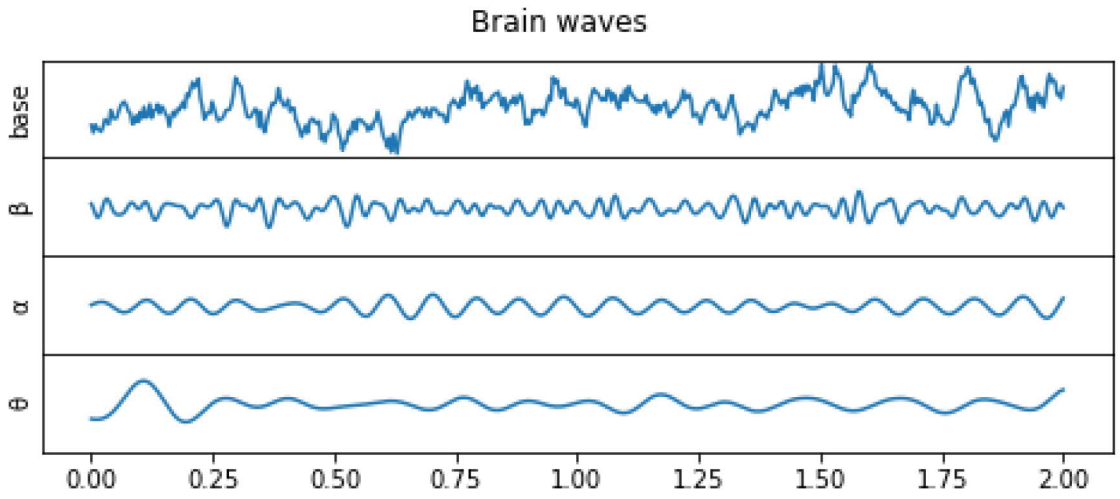

Figure 7: Example 2-second segment Theta, Alpha, and Beta waves

A BCI benefits from signal filtering for at least two reasons: Firstly, it removes the noise component of a direct current (DC) or alternating current (AC) power source. A high-pass filter with a cutoff of around 1Hz removes DC component noise in the signal and a 50Hz or 60Hz notch filter filters out AC power noise. Secondly, BCI classifiers tend to be interested in a subset of all frequencies, and those frequencies outside of that range are not studied.

The use of pre-processing tools in studies similar to this study is briefly summarised. Lu, Yin, and Jing 2019 achieve performance improvements by pre-processing raw EEG data through extracting the alpha, beta, and gamma waves by utilizing a bandpass filter with a range of 7-40Hz, by removing noise artifacts from eye and muscle movement with a 64-component ICA, and by making use of Z-score normalization. Xu et al. 2019 extract the mu and beta waves and utilize a selection of channels, namely the C3, C4, and Cz electrodes. Cecotti and Ries 2016 filter the raw EEG band using an 8-30Hz band to include alpha and beta rhythms, and average EEG signals from C3, C4, and Cz electrodes. Lawhern et al. 2018 instead apply a bandpass filter on the raw EEG to create non-overlapping filter banks by filtering the signal in 4Hz

steps, starting at 4Hz up to 40Hz, which results in having 9 filter banks. Liyanage et al. 2013 apply a 8-30Hz bandpass filter and use CSP. Hartmann, Schirrmeister, and Ball 2018 simply downsample the data to 250Hz down from 5000 Hz to improve training times by reducing the number of training data. Freer and Yang 2020 utilise a 7-30Hz bandpass filter and a Z-normalisation. B.-H. Lee et al. 2020 bandpass filter between 1-60Hz and select the 24 electrodes on the motor cortex. Millan and Mourino 2003 select eight fronto-centro-parietal electrodes, transform them using SL, and extract the 8-30Hz band and perform normalization.

2.3.3 Feature Extraction

The BCI then converts the pre-processed signals to a more compact set of features using feature extraction tools, before feeding them to the classifier. Feature engineering typically requires significant expertise to be done well, and EEG signal feature extraction commonly uses ICA, PCA, Wavelet Transform (WT), Wavelet Packet Decomposition (WPD), and Fourier Transformation (FT) (Lakshmi, Prasad, and Prakash 2014). ICA and PCA apply to both pre-processing as well as feature extraction. FT transforms a signal from the time domain to the frequency domain. If FT uses one-second segments that overlap by half a second, then the name of the technique in question is power spectral density (PSD). Also commonly taking place is the extraction of beta and alpha or mu frequency bands from raw data using filtering. However, the advancement and success of deep learning techniques in computer vision and speech recognition, and the adoption of those deep learning techniques in the field of EEG-based BCI has reduced the need for manual feature extraction. A comparison of feature extraction methods by Lakshmi, Prasad, and Prakash 2014 is presented in table 2.3.3

Table 5: Comparison of feature extraction methods by Lakshmi, Prasad, and Prakash 2014

Technique	**Advantages**	**Disadvantages**
WT	Analyzes discontinuous signals Analyzes signals in time and frequency domains Extracts energy, distance or clusters	Lacking of WT methodology for noise. Performance limited by Heisenberg uncertainty
WPD	Can analyze the non-stationary signals.	Increased computation time.
FFT	Powerful method of frequency analysis.	Applicable to stationary signals Sensitive to noise Poor time localization

Nonetheless, Xu et al. 2019 compute Short Time FT (STFT) from the electrodes C3, C4, and Cz using window sizes of 64 and overlap to 50 to create spectrum images, from which they extract the mu and beta bands. A similar approach is used by Wang et al. 2018. On the other hand H. K. Lee and Choi 2018 create a spectrum image using Continuous Wavelet Transform (CWT) wherein they use two different so-called mother wavelet known as the Morlet and Bump wavelets; Ortiz Echeverri et al. 2019 also utilise CWT

2.3.4 Classification

Feature extraction is followed by classification. The classifier translates these extracted features into commands that subjects desire to output. Lakshmi, Prasad, and Prakash 2014 found that Linear classifiers and artificial neural networks (ANN) are popular choices. Frequently used linear classifiers are the linear discriminant analysis (LDA), which models the probability density function, and support vector machines (SVM), in which a hyperplane is found where data sets are separated by a gap as wide as possible. Since then ANN based approaches have gained in popularity. The strategy generally used for multiclass BCI is the 'one versus the rest' (OVR) strategy which consists in separating each class from all the others.

Table 6: Comparison of classification methods by Lakshmi, Prasad, and Prakash 2014

Technique	**Advantages**	**Disadvantages**
LDA	Low computational requirements Simple to use Provides good results.	Discriminatory function inapplicable for variance features Preserves complex structures of Gaussian distributions
SVM	Generalizes well Outperforms linear classifiers	Computationally complex
ANN	Robust in practice Computationally simple classifier Versatile to input data	Difficult to build Performance depends hyperparameters

2.4 Machine Learning

Artificial intelligence observes its surroundings (AI) and learns to adapt to it, much like humans. AI is an umbrella term, under which a key idea is machine learning (ML) - the study of computer algorithms that improve over time (Mitchell and Hill 1997). A key problem in ML is determining higher-level insights from low-level perceptions. One can feed in raw data to the ML method to generate predictions. However, it is typically more efficient to feed the model with a compact set of characteristic properties of the data set instead. In a machine learning task, it is necessary only to use a sufficiently large amount of features, even though intuitively it makes sense to use as many features as possible to describe a data point. However, Furthermore, a large number of features increase computational and statistical risk; more computations are required, and the methods will be prone to overfitting. Overfitting happens when the model is fit too closely to some subselection of data instead of generalizing well across all or future data. Dimensionality reduction is about finding such a compression map that transforms a data point to a feature vector, such that from that feature vector it is possible to reconstruct the original data point as accurately as possible using a reconstruction map. However, the choice of features that characterize a data point is critical for overall success, and determining those features can be a difficult task that can cost vast amounts of expert time. A feature vector is the set of features that describes any one instance of data. Features need to be computable easily yet maintain sufficient information about the label. The label represents some high-level facts and is usually difficult to obtain; producing labels can require human expert labor, for example. For this reason, labels are often a scarce resource. Unsupervised learning is an ML method class that can operate in the absence of labeled data - Such methods can extract relevant information from features. However, supervised learning is another important ML method class which reads the features of some data point and predicts some label that approximates the correct label. The production of the mapping from features to labels uses historic data to test various choices and selects one that performs best by some metric. The challenge of such methods is the high amount of data points required and high-dimensionality of those data points, as well as often utilizing non-linear predictor maps, which are computationally demanding.

2.4.1 Artificial Neural Networks

An artificial neural network (ANN) is an ML method that models the way the brain adapts to its environment, usually implemented in hardware or software. For example, one can consider the human brain to be a type of ANN implemented in hardware. Haykin 1994 defines neural networks followingly: “A neural network is a massively parallel distributed processor made up of simple processing units that has a natural propensity for storing experiential knowledge and making it available for use”. A neuron is a fundamental operating unit that processes information in an ANN. A network may employ a large network of neurons or processing units to achieve high performance. Multitudinous neurons connect to other neurons in a network, and each connection between neurons transforms the signal passing through such a network.

Originally discussed in Warren S. McCulloch 1943, Haykin 1994 models a neuron in figure 8. It comprises of three elements, which are the following:

- a set of synapses that multiply an input signal by the weight assigned to that connection
- an adder that sums the weighted input signals, and
- an activation function limits the amplitude range to some finite value

Additionally, neurons may be biased. Bias affects the input of the activation function as though another synapse with a fixed input of +1 were present. This is the same as applying an affine transformation to the activation function, which is a linear mapping that preserves points, straight lines, and planes by keeping parallel lines parallel.

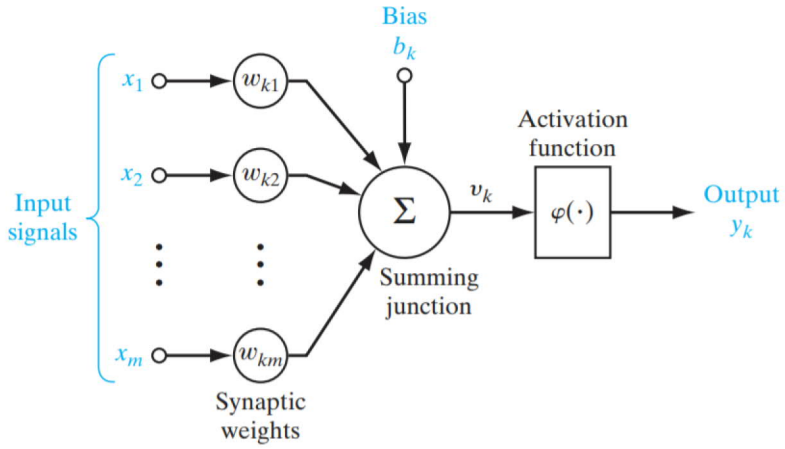

Figure 8: Artificial Neuron (Illus. by Haykin 1994)

See figure 9 for an illustration of four types of activation functions: the threshold function, sigmoid or logistic function, exponential linear unit (ELU) (Clevert, Unterthiner, and Hochreiter 2015), and rectified linear unit (ReLU). The threshold function is almost self-explanatory: If the input is negative, the output is zero, and if the input is positive, the output is one. This is what is known as the McCulloch–Pitts model (Warren S. McCulloch 1943). The sigmoid function is another common form of an activation function, which is a strictly increasing S-curve-shaped that is balanced between linear and non-linear behavior. The sigmoid function has an output range between 0 and 1. Additionally, the hyperbolic tangent function resembles the sigmoid function but differs in that its output ranges from -1 to 1 instead. However, a general problem with the sigmoid and hyperbolic tangent is that they saturate at high and low values. ReLU rectifies the input by outputting 0 for negative values, which has the effect of deactivating some neurons. This results in computational efficiency. ELU is a variant of the ReLU, where a log curve replaces the negative part of the function. Clevert, Unterthiner, and Hochreiter 2015 find that ELUs lead to faster learning

and better generalization performance than ReLUs, in deep networks. Schirrmeister et al. 2017 found a similar effect - ReLU in all layers instead of ELUs reduced the performance of their EEG-classifying deep convolutional neural network.

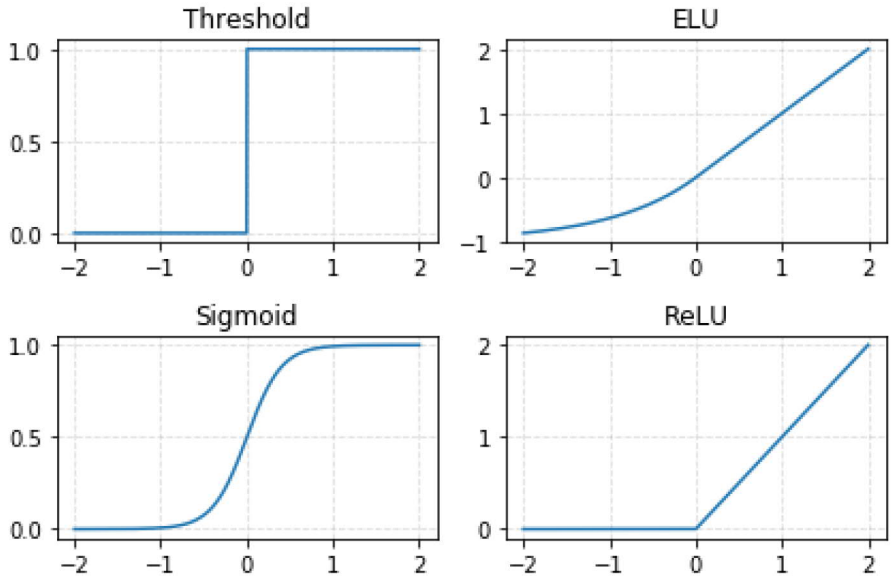

Figure 9: Activation functions

The subset of computationally feasible predictors is called the hypothesis space. If the problem allows linear algebra and the feature space is the euclidian space, the ML methods based on the linear hypothesis space boil down to combinations of basic linear algebra operations, such as vector-matrix multiplications. Standard computing hardware performs matrix operations efficiently, and so it is wise to reformulate any ML problem using vectors as features, if possible. Choosing the features as well as the hypothesis space requires a good understanding of the application. If the label space is finite, a classifier replaces the hypothesis. Alternatively, regression problems replace the hypothesis with a predictor. For example, linear classifiers, which include logistic regression, use classifier maps whose decision boundary is a hyperplane. Determining the best predictor map in the hypothesis space requires the measurement of loss. Squared error loss is widely used in regression problems because it efficiently searches for the optimal predictor within a hypothesis space using gradient descent. Minimizing the squared error loss is equivalent to maximizing likelihood estimation within a linear gaussian model. However, squared error loss is not useful for classification problems involving a discrete label space. A useful loss function for binary classification problems is the logistic loss because those methods can make use of a simple gradient descent method. In the case of a multi-class problem in which only one class label is correct for some feature vector, categorical cross-entropy is a suitable loss function.

The procedure that assigns the synaptic weights of the network is called a learning

algorithm, and examples of common learning algorithms used in supervised learning are the stochastic gradient, and least-mean-square error learning algorithms. The stochastic gradient is an iterative method that approximates the gradient of the network from the whole data set with a random subset of the data to find a set of parameters that ideally results in the fastest learning. The subset selection reduces the amount of data, which in turn reduces computational complexity. This results in faster training but worse convergence rates. The Adam optimizer is an example stochastic gradient descent method by Kingma and Ba 2014 that is well-suited to non-convex optimization problems and is robust to noisy and non-stationary signals. On the other hand, the least-mean-square error minimizes the mean square error, which commonly measures the quality of an estimator.

The choice of good features for any ML application is far from trivial and is perhaps the most difficult task of all. However, deep learning (DL) methods reduce the size of the challenge significantly. DL is the branch of machine learning that learns representations and accomplishes advanced tasks based on multiple neural layers. DL trains a model using extensive data to provide predictions based on new data. A deep ANN differs from ANNs by containing more than one hidden layer. Figure[1] 10 visualizes such a typical example model

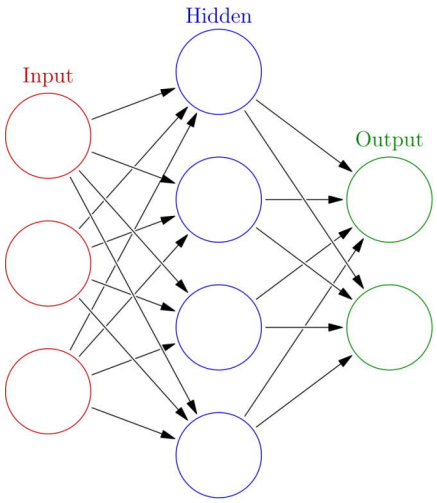

Figure 10: Typical Neural Network

Deep learning models split into three subcategories:

- *Discriminative models* can be used for feature extraction and classification
- *Representative models* only perform feature extraction
- *Generative models* generate data to improve a data set

[1]By Glosser.ca - Own work, Derivative of File:Artificial neural network.svg, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=24913461

Zhang et al. 2019 find that more than 70% of discriminative DL models in BCI research use CNN. They argue that CNN can extract the spatial dependencies, as well as latent discriminative features, powerfully enough for brain activity classification. Because of these reasons, CNNs are commonly used for both feature extraction as well as EEG signal classification, which reduces the need for expert knowledge. However, Ortiz Echeverri et al. 2019 state that CNNs perform better on preprocessed EEG data rather than raw data. Also, they found that kernel size and stride in each convolutional layer have a significant impact on classification accuracy, while the number of convolutional layers is less significant. Similar to images, EEG data have an abundance of features that are difficult to define manually. Furthermore, EEG data points relate to those data that occur close to them in both time and space, albeit that relationship is less obvious than it is with images. An image can represent an EEG sample in a variety of ways. CNNs typically operate on 2D planar surfaces or images. However, a 2D surface fails to represent the spherical geometry of the surface of the scalp where the EEG data originate from without severe distortion. A more generalized form (Cohen et al. 2018) adjusts the CNN operation to preserve spatial proximity. Alternatively, images can use temporal or spectral dimensions in addition to space, which CNNs can represent. MI has promising results with deep learning models. Additionally, Zhang et al. 2019 find that BCI studies frequently use CNN to discriminate between MI EEG based on manual feature extraction for 2-dimensional CNN or temporal CNN or to representatively perform feature engineering to capture latent connections.

2.4.2 Convolutional Neural Networks

Convolutional neural networks (CNN) are a category of discriminative artificial neural networks that utilize temporal and spatial relationships between local data points to determine the useful features of some machine learning problems. Such methods have achieved success in the task of classifying images (Krizhevsky, Sutskever, and Geoffrey Hinton 2012) and acoustic signals (G. Hinton et al. 2012). CNNs reduce the trainable parameters of a fully-connected deep network without compromising performance, which allows the development of more efficient networks. A CNN typically consists of a combination of three different types of layers, which are the following: convolutional layers, pooling layers, and a fully connected output layer. Figure[2] 11 shows a typical CNN architecture that uses such layers.

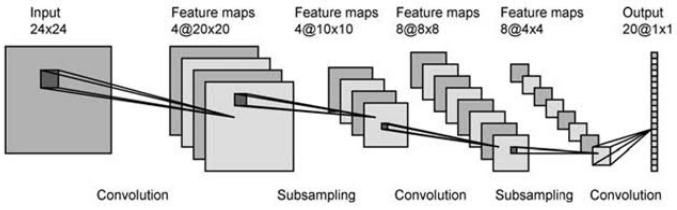

Figure 11: Example CNN Architecture

[2]Source: http://what-when-how.com/wp-content/uploads/2012/07/tmp725d63_thumb.png

A CNN uses convolution and subsampling or pooling to reduce a signal, typically an image, to its essential features. The difference between a dense or fully-connected layer and a convolutional layer is that the latter learns local patterns whereas the former learns relative to a specific point. It does so by processing a possibly padded image using filters of some specified size. These filters pass over the image in user-specified strides and can relate to anything, such as simple edges or complicated patterns. A scalar product applies the filter to sub-images, and that result is passed to an activation layer. Thus, a convolutional network can recognize some similar features in any other part of the signal. Furthermore, convolutional layers can learn increasingly complex concepts with nested convolutional layers. It does so by effectively combining simpler features to increasingly complex features in each layer. A convolutional layer is typically followed by and activation layer that limits the range of the output. Thereafter, a pooling layer performs subsampling, which reduces the resolution of the feature map. It does so by selecting a local average or maximum value typically. Subsampling has the effect of reducing sensitivity to shifts and distortions as well as reducing training time. One CNN can include multiple such convolutional and pooling layers. Finally, a fully connected or dense layer takes as input a one-dimensional vector that represents the previous layers' output. The fully connected layer outputs a list of probabilities for the possible labels, and the highest probability listed is regarded as the classification decision. Since this method typically operates with images, a typical input data has height and width dimensions, as well as a depth value. For example, a typical image might have a three-dimensional depth that represents the color channels red, green, and blue.

2.4.3 Transfer Learning

Zhang et al. 2019 suggest that transfer learning may contribute to the goal of developing a person-independent classification of brain activity. The method works in such a way that the learned parameters of some personalized model are transferred to a similar model to be personalized to a different subject. Transfer learning helps deal with two challenges: A deep learning model requires a large amount of labeled data, and such data is may not be excessively available in the context of EEG signal data. Furthermore, training a deep model is computationally expensive, which can be observed as long training times in model personalization. A transfer learning framework consists of a pre-trained model and a target model with a shared structure, excluding the output layer (Xu et al. 2019). The parameters of the pre-trained model are transferred to the target model and subsequently frozen to some extent. The act of freezing parameters makes them unadjustable in a training phase, which reduces the number of total trainable parameters. This is beneficial due to reducing the computational load of training a model. Thereafter the later layers are trained normally in the target dataset. However, the use of transfer learning assumes that the new task is related to the earlier task for which the pre-trained model parameters were optimized.

2.5 Related work

Other CNN-based MI EEG BCI studies achieve the following results: Uktveris and Jusas 2017 achieve a mean classification accuracy of 68% in the testing set for 4-class MI from BCIC IV 2a problem with less complex feature extraction techniques utilizing such a network, in this case, a fast Fourier transform energy map. B.-H. Lee et al. 2020 produce a mean testing accuracy of 66% in a 7-class MI and 88% in a 3-class MI data set with an end-to-end role assigned CNN. Wang et al. 2018 reach a mean classification accuracy of 92% using a novel 2-class MI data set transformed using a short-time Fourier transform and a scaled exponential linear unit -based 7-layer CNN. Y. Zhang et al. 2019 propose a 3D CNN that gives a mean accuracy of 78% using time-frequency MI data. Lawhern et al. 2018 propose a deep CNN and a shallow CNN with a mean classification accuracy of around 50% and 70%, respectively in the 4-class MI BCI competition IV 2a dataset with features extracted through the use of filter banks and CSP. Tang, Li, and Sun 2016 reach an average accuracy of 86% using a novel CNN, and spatio-temporal EEG from a novel 2-class MI data set. H. K. Lee and Choi 2018 apply their 1-D CNN to the 2-class BCI competition IV 2b data set transformed to a time-frequency representation with CWT and report 78% mean accuracy. Kar et al. 2018 hit a mean accuracy of 70% in the 4-class BCI competition IV 2a data set using their proposed deep CNN and raw EEG data. Schirrmeister et al. 2017 present a set of CNNs that are evaluated against multiple data sets: They propose a deep, shallow, hybrid, and residual CNN that they evaluated against the 4-class MI data sets BCIC IV 2a as well as their larger novel high-gamma dataset. The shallow net has an accuracy of 73% in the smaller data set and 94% accuracy in the larger data set. Ortiz Echeverri et al. 2019 attain 94% accuracy in a 2-class MI dataset (BCI competition III) time-frequency transformed using a derivative CNN. Xu et al. 2019 use the 2-class MI dataset 2b from the BCI competition IV transformed to time-frequency data and attain an average accuracy of 74% using the well-known VGG-16 CNN. Lu, Yin, and Jing 2019 use a novel temporal convolutional network and the large 5-class EEG Motor Movement/Imagery Dataset (EEGMMIDB) from PhysioNet and realize a mean accuracy of 97%.

Table 7: Summary of related work

MI classes	dataset	Accuracy	Author
2-class	BCIC III	94%	Ortiz Echeverri et al. 2019
2-class	BCIC IV 2b	74%	Xu et al. 2019
2-class	BCIC IV 2b	78%	H. K. Lee and Choi 2018
2-class	Novel	92%	Wang et al. 2018
2-class	Novel	86%	Tang, Li, and Sun 2016
3-class	Novel	88%	B.-H. Lee et al. 2020
4-class	BCIC IV 2a	68%	Uktveris and Jusas 2017
4-class	BCIC IV 2a	70%	Kar et al. 2018
4-class	BCIC IV 2a	70%	Lawhern et al. 2018
4-class	BCIC IV 2a	74%	Schirrmeister et al. 2017
4-class	HGD	93%	Schirrmeister et al. 2017
5-class	EEGMMIDB	97%	Lu, Yin, and Jing 2019
7-class	Novel	66%	B.-H. Lee et al. 2020

2.6 Public Dataset

Multiple MI BCI studies (Ortiz Echeverri et al. 2019; Uktveris and Jusas 2017; Kar et al. 2018; Schirrmeister et al. 2017) use the dataset 2A from the BCI competition IV (Brunner et al. 2008) [3]. The dataset consists of a cue-based experiment with 9 subjects performing 4 MI tasks: Imagination of left hand, right hand, both feet, and tongue. The task class labels range between classes 1 through 4, respectively. The 4 classes of MI tasks are distributed across trials in a balanced way. The recording device used 22 active electrodes mapped along with the international 10/20 system, shown in figure 12, to record EEG data. The active electrodes are the following: Fz, FC3, FC1, FCZ, FC2, FC4, C5, C3, C1, Cz, C2, C4, C6, CP3, CP1, CPZ, CP2, CP4, P1, Pz, P2, POZ. Additionally, the data set provides 3 monopolar electrooculogram (EOG) channels that can be utilized for detecting noise from sources like eye movements or blinking (these are called artifacts). The authors marked the collected EEG data for noise artifacts with expert labor in their study.

[3]Dataset available at http://bnci-horizon-2020.eu/database/data-sets . Accessed in Feb 2020. The general data format (GDF) files provided here contain one session per subject each, amounting to 18 files. However, these files are converted to comma-separated value (CSV) tables where each trial in its file, amounting to approximately 5000 files for this study

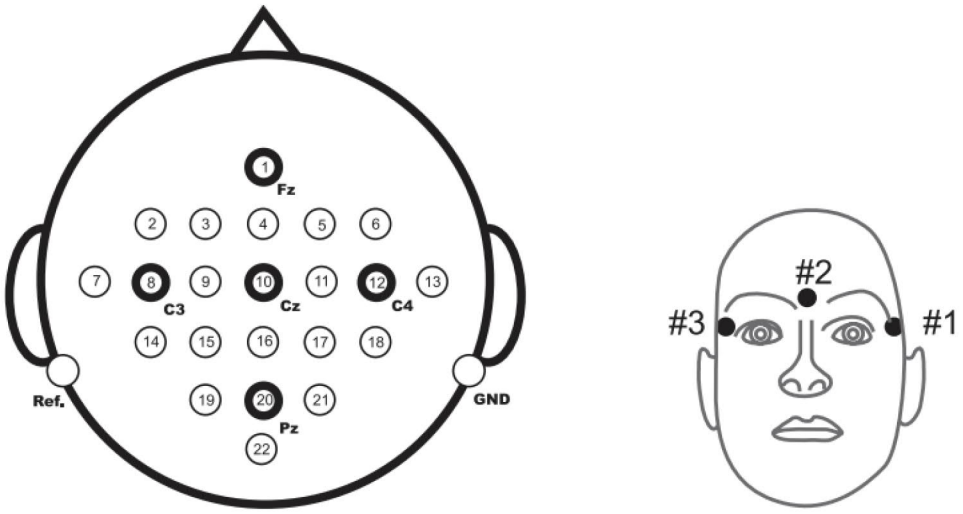

Figure 12: Left: Electrode montage corresponding to the international 10/20 system. Right: Electrode montage of the three monopolar EOG channels. (Brunner et al. 2008)

The data set consists of 9 subjects whom each partook in 2 sessions recorded on separate days. Each session comprised of 6 runs, where one run contained 48 trials. Each trial contained one of the four MI tasks, equally distributed across trials so that there should be 12 trials per class in each run. With measurements being continuously taken across 22 channels at a sampling rate of 250 hertz over a duration of 8-seconds per trial, the total number of data points is around 228 096 000. The total amount of data available is enumerated in table 8. Note that the number of data points is an approximation due to trial duration being approximately 8 seconds.

Subjects	9
Sessions per subject	2
Runs per session	6
Trials per run	48
EEG channels	22
Sampling rate	250Hz
Duration	8s
Total Data Points	228 096 000
Trials	5 184
Trials per subject per class	144
Of which non-contaminated	135

Table 8: Data in BCI Competition IV 2A (Brunner et al. 2008)

The author collected the data followingly: Participants were seated comfortably in front of a computer screen. A dark screen would present a fixation cross accompanied

by an acoustic warning tone to the participant at the beginning of the trial, at time 0 seconds. At time 2 seconds an arrow-shaped cue pointing up, down, left, or right corresponding with a MI task would appear, remaining on-screen for 1.25 seconds. This would prompt the participant to execute the MI task without feedback until time 6 seconds, at which point the fixation cross would disappear. A short pause of around 2 seconds was provided to the participant in the form of a dark screen until the beginning of the next trial. This timing scheme is illustrated in figure 13. The author would pre-process the data with a bandpass filter between 0.5Hz - 100Hz to preserve the relevant information for MI and a 50Hz notch filter to filter out power line noise. The sensitivity of the amplifier was set to 100 microVolts, and so the measurement values range between +-100. (Brunner et al. 2008)

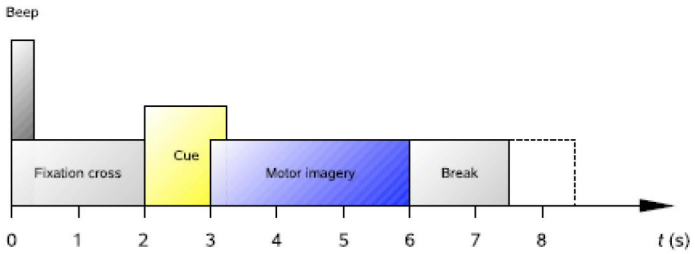

Figure 13: Timing scheme of each trial. (Brunner et al. 2008)

3 Research material and methods

This study focuses on noninvasive EEG data set, in particular MI data, and uses the 4-class BCI competition IV 2A data set (Brunner et al. 2008). Out of the functional neuroimaging technologies, EEG has relatively low cost, high availability, high temporal resolution, and low risk. While the spatial resolution is low, EEG is the target of a substantial majority of published BCI work and is also a logical choice for this study because of its potential for widespread adoption. This study proposes a CNN-based model and trains it using that data set. Data marked for artifacts from those sets are manually excluded, instead of filtering out such noise using the EOG channels. The amount of contaminated samples is small, approximately less than 5% of the data contain contamination, and so the impact of their exclusion on the data set is limited. Also, demographic data and metadata present in the data are further disregarded in this study. The resulting data are considered the raw data or EEG signal in this research.

The preprocessing phase applies temporal filtering to the data. A fifth-order Butterworth bandpass filters the raw EEG signal with a low and high cutoff between 7Hz and 40Hz. This is done for two reasons: 1. To include only those frequency bands that are known to be most important for MI classification, and 2. to exclude noise from eye movements which generates most power in low frequencies, and high-frequency brain activity which is seen not to contain useful information - this approach was used to great success in the winning submission of the original winning solution at the BCI competition IV (Tangermann et al. 2012). Note that this is in addition to the raw EEG data having been bandpass filtered between 0.5Hz and 100Hz and notch filtered at 50Hz upon collection by the author of the dataset. Figure 14 illustrates the effect of the applied pre-processing by visualizing the frequency response of the bandpass filter between frequencies 7 and 40Hz.

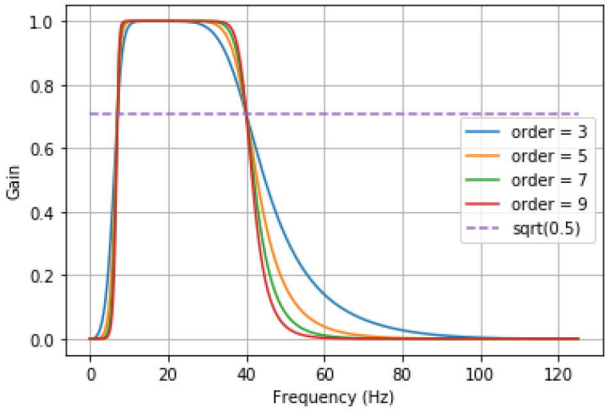

Figure 14: Butterworth bandpass frequency response

Additionally, figures 15 and 16 show the the spectrum of an example trial before and after filtration. On the x-axis are the frequencies, and on the y-axis are the signal powers. Note, the scaling of the y-axis differs in the two figures.

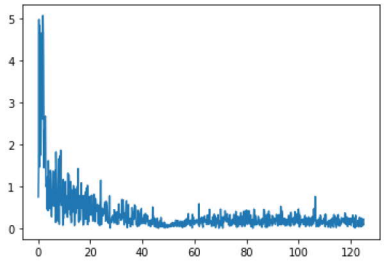

Figure 15: Example trial spectrum before filtering

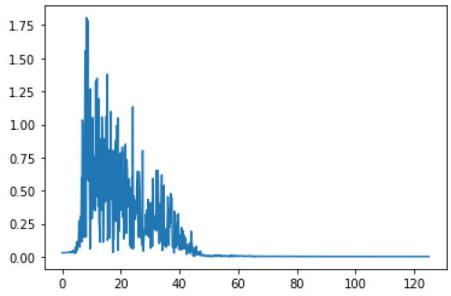

Figure 16: Example trial spectrum after filtering

However, not all parts in the EEG signal in the 8-second trials in the public dataset are equally valuable in terms of being useful for the classification of brain activity. The experimentally found 2-second segment between 2 and 4 seconds in each trial is extracted from the signal, which contains the beginning of the cue presented to subjects and the early parts of the MI task. The reduction of the trial segment size does not adversely affect classification accuracy, while simultaneously enhancing the responsiveness of any online solution employing the model as well as reducing the computational load of the model. A 2-second segment of an example trial at a sampling rate of 250Hz using all 22 EEG channels results in an input data shape of 22×500 pixels. This effectively forms an elongated rectangular grayscale image, visualized here in figure 17 in green.

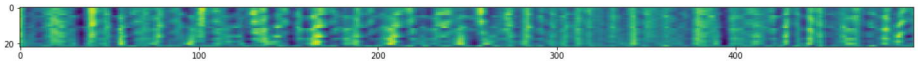

Figure 17: Example 22x500 trial after filtering

3.1 Model

The model can self-learn the relevant features in the input data, as well as perform classification. It draws inspiration from the Deep ConvNet that Schirrmeister et al. 2017 proposed and consists of four convolutional max-pooling blocks and a dense softmax classification block. Blocks 2, 3, and 4, as well as the classification block, are essentially the same as the three standard blocks by the author, while the first convolutional block in this study differs from theirs in that it does not perform the same combination of temporal convolution and spatial filtering, but instead spatially filters each sample. Each convolutional layer uses the exponential linear units (ELU) as activation functions. Convolution over space is performed by each filter in the first block, and the three latter blocks are otherwise standard convolutional blocks. Also, Max pooling with a pool size of (3, 1), a stride of (3, 1), and valid padding is used in each block, and the kernel sizes of the convolutional layers gradually transform the EEG input before it reaches the ultimate classifier layer. Batch normalization is used in each block to facilitate optimization by standardizing the outputs of the layers to a mean of zero, which keeps the layers closer to a normal distribution during training (Ioffe and Szegedy 2015). Note that this is unintentionally left out in the third convolution block. A dropout layer with a probability of 0.5 precedes the three intermediary convolutional blocks, which is a way to reduce overfitting by setting random inputs to zero (Srivastava et al. 2014). These layers are then finally followed by a dense softmax classification layer. The model requires data to be at least 500 samples long and can adapt to longer samples as well. Table 9 summarizes the suggested neural network architecture.

Table 9: Suggested convolutional neural network architecture

Block	Layer	Kernel	Output Shape	Params
Block 1	InputLayer	-	(22, 500, 1)	0
	Reshape	-	(22, 500, 1)	0
25 ELU	Conv2D	22×1	(1, 500, 25)	575
	BatchNormalization	-	(1, 500, 25)	100
	Reshape	-	(500, 25, 1)	0
Stride 3x1	MaxPooling2D	3×1	(166, 25, 1)	0
Block 2	Dropout	-	(166, 25, 1)	0
	Conv2D	10×25	(157, 1, 50)	12550
50 ELU	BatchNormalization	-	(157, 1, 50)	200
Stride 3x1	MaxPooling2D	3×1	(52, 1, 50)	0
	Reshape	-	(52, 50, 1)	0
Block 3	Dropout	-	(52, 50, 1)	0
	Conv2D	10×50	(43, 1, 100)	50100
100 ELU	MaxPooling2D	3×1	(14, 1, 100)	0
Stride 3x1	Reshape	-	(14, 100, 1)	0
Block 4	Dropout	-	(14, 100, 1)	0
	Conv2D	10×100	(5, 1, 200)	200200
200 ELU	BatchNormalization	-	(5, 1, 200)	800
Stride 3x1	MaxPooling2D	3×1	(1, 1, 200)	0
	Reshape	-	(1, 200, 1)	0
Class	Flatten	-	(200)	0
	Dense	-	(4)	804
4 Softmax	Activation	-	(4)	0

This study opts for the Adam optimizer as a learning algorithm (Kingma and Ba 2014). In this study, a subject can only be performing one MI task at a time, and so categorical cross-entropy is the loss function of choice. The data labels are one-hot output encoded, which means that the label is a vector for which each class is represented by one element in such a way that the element for that class is one, and all other elements are zero. The metric of choice used to judge the performance of the model is accuracy, which is to say the number of correct predictions out of all predictions made. Batch size is set to 16, and the number of epochs is set to 75 for training. Table 10 summarizes the other suggested neural network parameters.

Table 10: Suggested model parameters

Parameter	Value
Optimizer	Adam, lr=0.001, $\text{beta}_1 = 0.9, beta_2 = 0.999$
Cost function	Categorical cross-entropy
Metric	Accuracy
Batch size	16
Epochs	75

3.2 Experiment

Firstly, exploratory experiments study the effect of the parameters of the time segment and data set size on classification accuracy. Secondly, the within and between-subject experiments assess model performance. Thirdly, a novel transfer learning experiment measures the improvement in model performance. Overall, the experiments utilize a segmented trial-wise training strategy, i.e. a certain segment of the trial signal is considered as the input, and its label is the training target. Additionally, each experiment preprocesses all used trials as explained earlier. The ordering of the trials is randomized and the data are split to separate sets for training, validation, and evaluation purposes. The models see only training and validation data during the training phase. The used data set consists of 9 subjects' data where each subject produced 600 trials over 2 days or sessions with 4 MI tasks lasting 8 seconds each. The conversion of a trial into input data extracts a time segment of around 2 seconds, and EEG images formed from all channels and those time segments are used for training, validation, and evaluation. All the within-subject experiments use 70% of trials for training, 20% for validation, and 10% for evaluation for each subject. Each experiment reports mean classification accuracy in the evaluation data set, standard deviation, and number of repetitions. The results are visualised using a box plot, which standardly displays the distribution of data using a descriptive five-number summary i.e. minimum, first quartile, median, third quartile, and maximum. A box plot tells about outlier values, helps identify symmetricity of data, determines data grouping tightness, and sees if data is skewed.

Segment size and timing: The experiment uses the larger data set with data from both days 1 and 2 to explore a variety of different time segment parameters. It examines critically the assumption that the 3.5-5.5 second cropping or segment produces the best results, motivated by the fact that it is that segment that contains the time just after the beginning of the cue until just before the end of the cue. At first, the trial is cropped to segments of 2 seconds, such that every slicing into a 2-second segment is considered with a certain degree of granularity: The sampling rate in the data set is 250Hz, which means that a 2-second segment results in 500 samples. The full trial is split to segments (0, 500), (100, 600), (200, 700), etc, which correspond with times 0-2s, 0.4-2.4s, 0.8-2.8s, etc. Additionally, the experiment considers a selection of other parameter sets: Trial samples of the first half (0, 900), the second half (900, 1800), the differently overlapping second half (700, 1600), and the full trial (0, 1800), as well as the default segment of (875, 1375). The experiment evaluates performance over at least 5 repetitions using a larger data set from subject 3. This helps get a sense of the segments that perform better, which are then further examined with other subjects.

Data set size: Next, experiments assess the data set size effect on model performance. The first data set variant referred to as the smaller data set includes all subject data from the day or session 1. The larger data set variant includes data from all subjects from both days 1 and 2 and is approximately twice as large. While it is known that there is intersubject variance in MI EEG data and that data collected on separate days are expected to have large differences, the increase in data set

size should result in a better grouping of prediction accuracy, nonetheless. However, the use of the combined data from both sessions in the data set renders this study uncomparable to the solution submissions in the original competition; The original competition provided the participants with access only to the data from the first session when designing their submissions. The experiments repeat the training and evaluation 15 times for the smaller and larger data sets for each subject.

The within-subject experiment uses the insights on data input parameters found in the exploratory experiments and generates a result for each subject. The experiment selects a configuration based on the parameter set of data set size and trial segmentation found to have the highest mean accuracy across subjects. Using these parameters, the model is repeatedly trained and evaluated using the 70:20:10 ratio for each subject. Additionally, the within-subjects experiment includes the results of the highest-performing data configuration found in the exploratory experiment.

The between-subjects experiment explores model classification performance generalizability. It uses a 70:30 split of one subject for training and 100% of the data of some other subject for evaluation and goes through all subject pairings. The choice of input data parameters considers the insights from the previous experiments. The between-subjects experiment is repeated 15 times for each combination of training subject and evaluation subject.

Lastly, *transfer learning* is explored. The experiment consists of first generating and storing a well-performing classifier model is for each subject using an optimal within-subject experiment design. No-freeze transfer learning indicatively explores every subject pair with a small number of repetitions in the hopes to find some model that should work particularly well in serving as a base model whose use increases classification accuracy for all other subjects maximally. The transfer learning method transfers the learned weights in these models, which the training then uses as a starting point for further fitting to other subjects' data. The method connects the weights of the four convolutional layers of the pre-trained model to a re-initialized fully connected classifier layer. Additionally, different combinations of freezing the convolutional blocks affect the performance. These combinations are the following: Freezing all 4 blocks, freezing only the first block, and freezing none of the blocks. After undergoing such a process, the model training and evaluation proceed normally. The intuition for why this approach should have a positive effect is that it should help the model converge to some local optimum quicker in some sense by moving the starting point for the search be much closer to the eventual goal.

4 Results

In addition to the research conducted, the project contributes a closed source software that contains the tools to load the BCI Competition IV 2a data set, select some desired subset, apply filtration to those data, import the proposed model, train and evaluate that model, save and load any trained models, and apply transfer learning with such models. Implemented of the solution is done in Python[4] using tools such as NumPy[5] and Keras[6]. All software execution takes place in a Jupyter Notebook[7] environment. The results of numerical experiments are stored locally in comma-separated value files, and some are minorly edited using Microsoft Excel[8]

The following paragraphs tabulate, visualise, and elaborate on the results of the experiments. The notation (a, b) refers to a trial segment where 'a' is the segment begin sample and 'b' the segment end sample. As an example, the segment (500, 1500) refers to a trial segment 1000 samples long between 500 and 1500; at a sampling rate of 250 samples per second, such a segment refers to a time slot between 2 seconds and 6 seconds. The results of each experiment for each subject or segment show mean classification accuracies in the evaluation data set, the standard deviation of that accuracy, and the number of repetitions, as well as summary total values, where meaningful. A box plot, bar plot, or matrix represents each result visually. There are some cases where the number of repetitions among subjects or segments differs minorly. This is because some experiment computation configurations overlap with other experiments; the experiments store each result in a centralized manner.

Exploratory work

Running the experiment with default parameters produced the following results shown in table A1 and figure 18. The default parameters were the following: Trial segment ranged between (875, 1375), the data set consisted of both testing and evaluation sessions, and the model used was the proposed within-subjects model. Mean accuracy across subjects was 44.8% ± 15.7%. Subject 3 achieved mean accuracies of 62.9% and the lowest standard deviation of 4.0%.

[4]https://www.python.org/
[5]https://numpy.org/
[6]https://keras.io/
[7]https://jupyter.org/
[8]https://www.microsoft.com/en/microsoft-365/excel

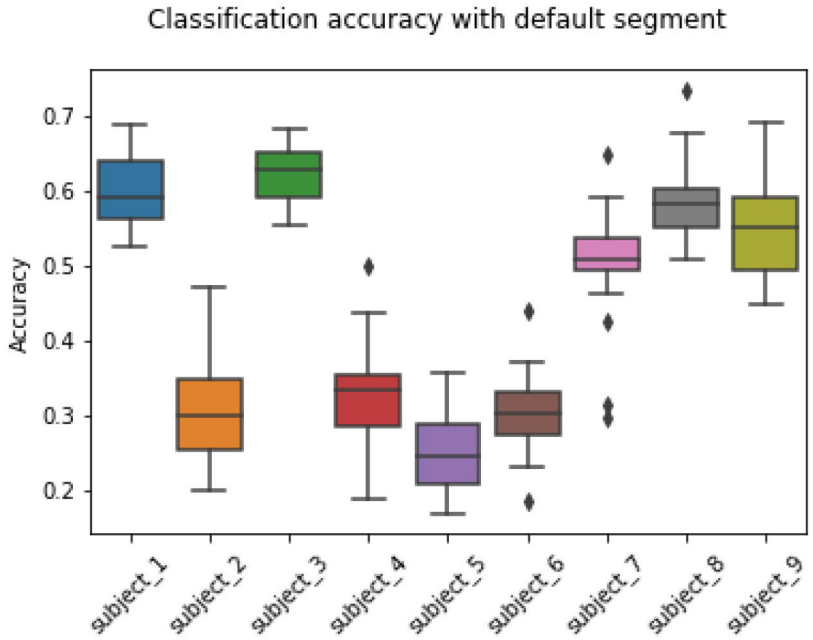

Figure 18: Default experiment within subjects

Varying segments

Subject 3 achieved mean classification accuracies of 69.9% ± 6.5% for a 2-second segment with samples between (700, 1200), which corresponded to times 2.4 - 4.4 seconds. Close behind in performance in the 2-second segment category were the segments between (500, 1000), and (600, 1100). The differences between these three are insignificant. Using a longer segment that closest overlaps with the full motor imagery task execution at (700, 1600) resulted in similar performance with mean accuracies at 68,7% ± 3,9%. The highest mean accuracy achieved is 74.5% ± 5.6% with a segment of (500, 1200), which corresponds to times 2 - 4.8 seconds. Tables A2, A3 and figures 19, 20 show in detail the effects of adjusting segment sizes.

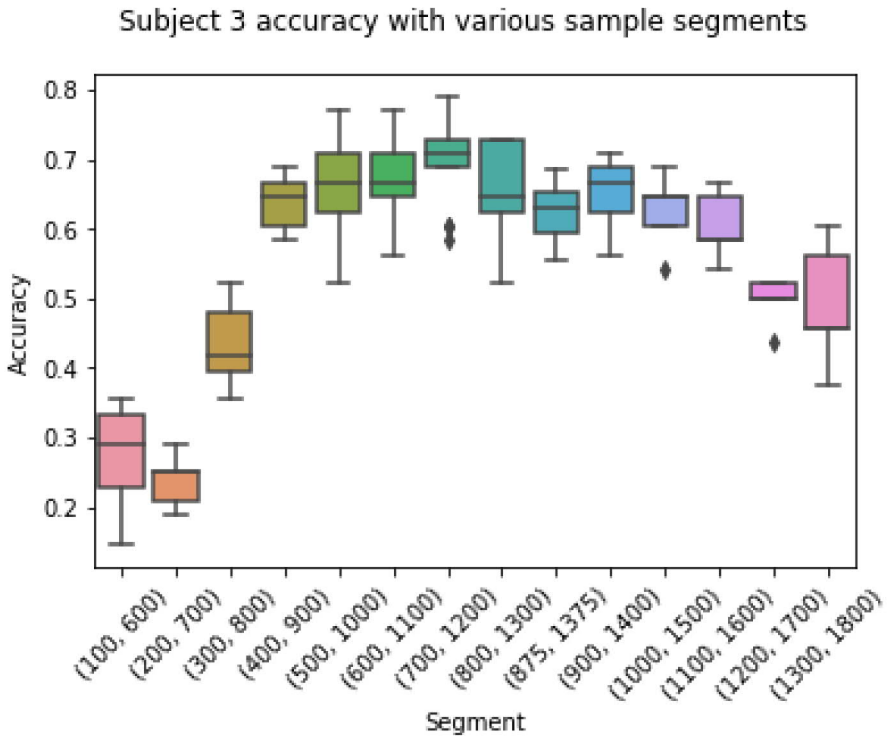

Figure 19: Subject 3 accuracy with varying 500-sample segments

Using a larger segment for MI classification had a small effect on classification accuracy. For example subject 3 reached a mean classification accuracy of 69,9% with a 2-second segment, while a 3.6-second and 7.2-second segment performed worse and better at 62.1% and 67,1%, respectively. A 2.8-second segment was found where accuracy is 74.5%.

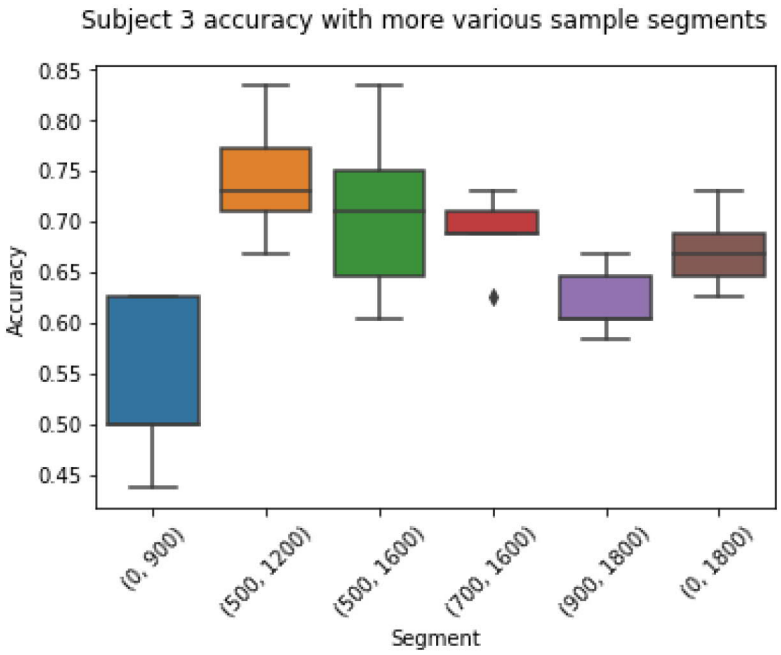

Figure 20: Subject 3 accuracy with varying segments at varying times

Figure 21 and table A4 compare the 3 best performing 2-second segments across all subjects. For many subjects, the effect size of different segments on classification accuracy is small, though for subjects 5 and 6 the effect is particularly large. Mean accuracy for subject 5 is 24.2% for the (700, 1200) segment, and 54.2% for the (500, 1000) segment, differing by 30% points. Similarly, for subject 6, mean accuracy is increased by 21.7% points between the same segments.

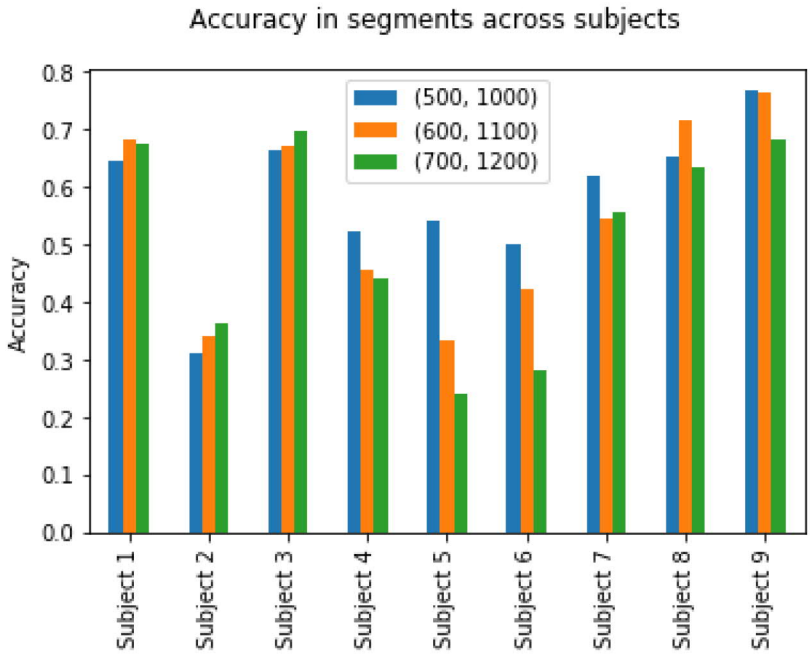

Figure 21: Various segments for each subject

Smaller data set

Mean accuracy for a within-subjects experiment using session 1 data further referred to as the smaller data set reached a mean classification accuracy of 49.3% ± 14.7%. More detailed results can be seen in table A5 and figure 22.

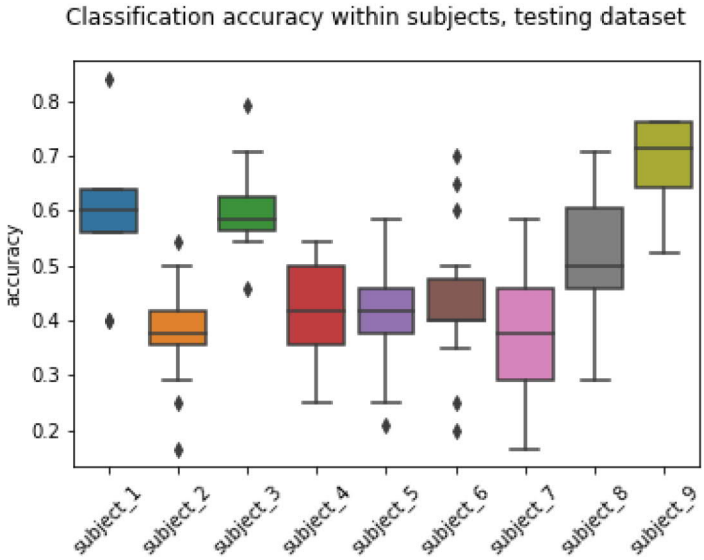

Figure 22: Accuracy with smaller dataset

4.1 Within and Between subjects experiment

Within subject

The model achieved a mean classification accuracy of 58% ± 14.9 within-subject with the larger data set. Mean accuracy within-subjects ranged between 30.8% ± 6% in the worst case by subject 2 and 76.8% ± 6.1% in the best case by subject 9. The procedure was repeated 30 times for each subject. These results, shown in A6, were significant improvements in comparison to results obtained with a smaller data set. The increase in mean accuracy was 8.7% when increasing the data set size by a factor around 2. Figure 23 and table A6 show the overall performance in the within-subject experiment.

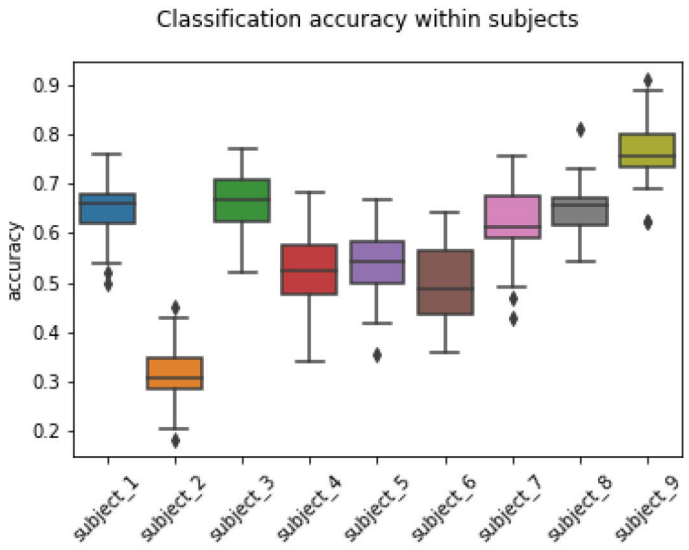

Figure 23: Within subjects classification accuracy

An example training and evaluation in the within subject experiment by subject 1 resulted in the an evaluation accuracy of 72%. The training history is shown here for illustratices purposes, see figures 24 and 25.

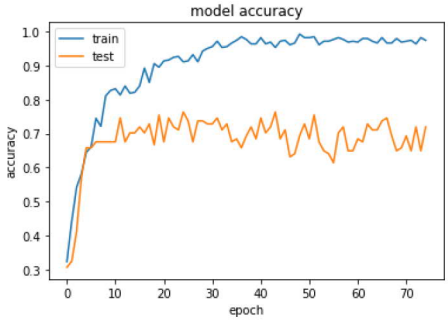

Figure 24: Subject 1 model accuracy history

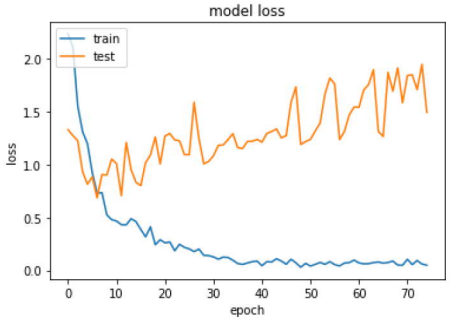

Figure 25: Subject 1 model loss history

Between subject

The model achieved mean classification accuracies of 35% ± 11.1% across all pairs of training and evaluation subjects, including the identical pairs where those two are the same. The diagonal identity line in the matrix is the same as the within-subject experiment from earlier. Figure 26 and table A11 show the mean classification accuracy results for every combination of training and evaluation subject, which totals 81 pairs.

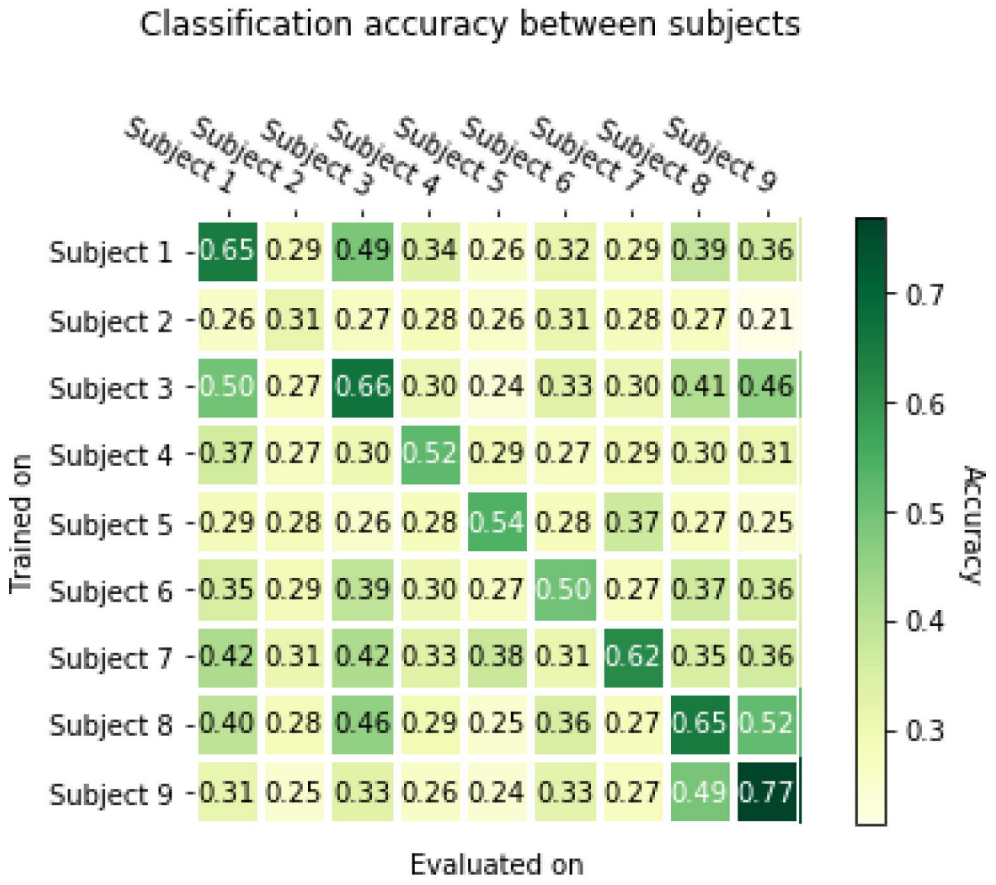

Figure 26: Between subjects classification accuracy matrix

The between-subject matrix consists of approximately 1000 samples, with a mean classification accuracy of 32% ± 7% for non-identical training and evaluation subject pairs. Training and evaluation were repeated for those pairs around 15 times which resulted in the accuracies the matrix shows. With an enrollment ratio of 0.01, an alpha of 0.05, and an 80% power, any pair with an accuracy above 38% has a significant effect. For example, this phenomena can be seen in subject pairs like (1, 3), (1, 8), (3, 1), (3, 8), (3, 9), (6, 3), (7, 1), (7, 3), (8, 1), (8, 3), (8, 9), (9, 8), which accounts for 1 per 6 or 16.7% of all non-identical pairs.

4.2 Transfer learning experiment

The results of the combinations of using a model pre-trained on one subject and then further trained with other subjects can be seen in table A13 and figure 27 without any frozen blocks. The procedure was repeated for pair twice. The results indicate that the model classified subject 2's data poorly and subject 9's data well regardless of which pre-trained model is used as the starting point.

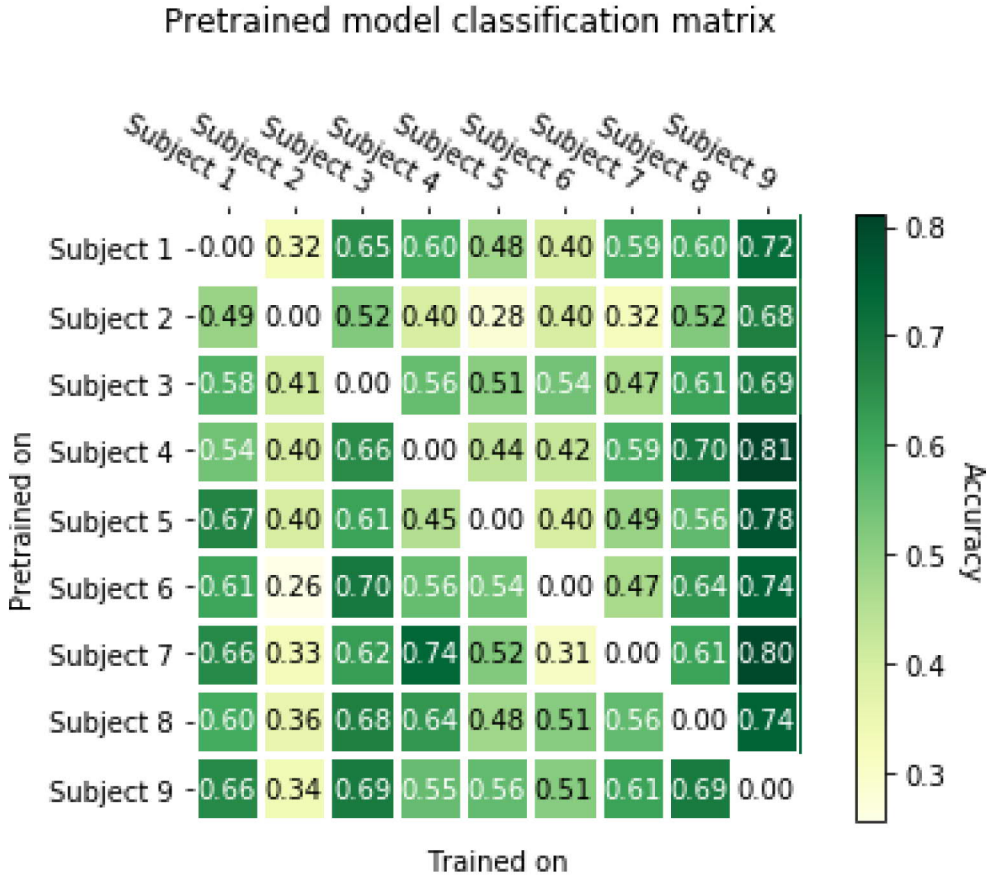

Figure 27: Transfer learning indicative matrix

Transfer learning experiments with freezing used the pre-trained well-performing within-subjects model by subject 9. The model reached mean classification accuracies of 32.3% ±11.3 with such a setup. Subject 8 achieved the maximal mean accuracy of 54.7% ± 6.0%. The experiment was repeated 4 times per subject, at which point the results did not further improve when repeating the experiment. See table A7 and figure 28 for detailed results.

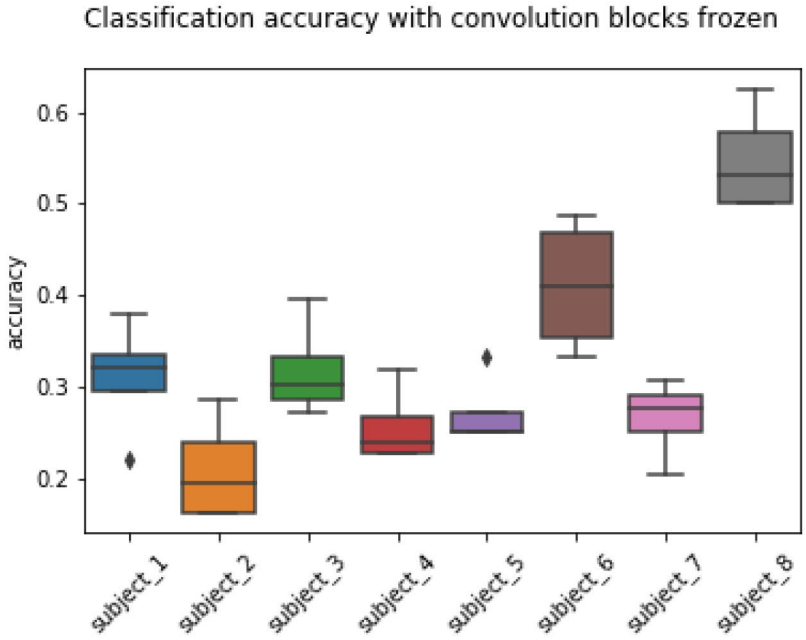

Figure 28: Transfer learning full freeze accuracy

Freezing only the first convolution block resulted in somewhat better performance than a full freeze, reaching a mean accuracy of 37.3% ± 12.3%. Transferring and trainig was repeated 10 times for each subject. See table A8 and figure 29 for detailed results of transfer learning with the first convolutional block weights frozen.

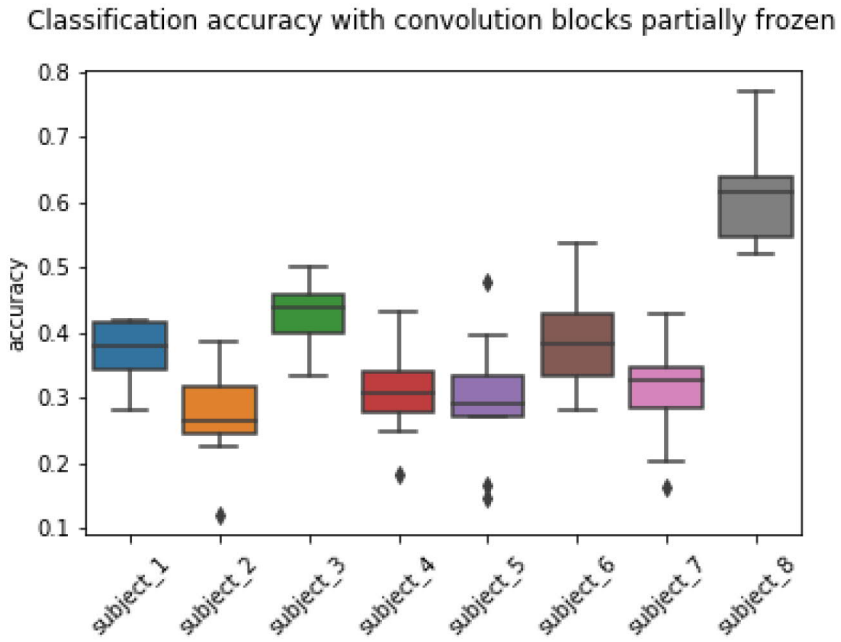

Figure 29: Transfer learning partial freeze accuracy

Table A9 and figure 30 report that using subject 9's model as a pre-trained model for subjects 1 to 8 resulted in the mean classification accuracies of 58.2% ± 13.1%. The procedure was repated 60 times per subject.

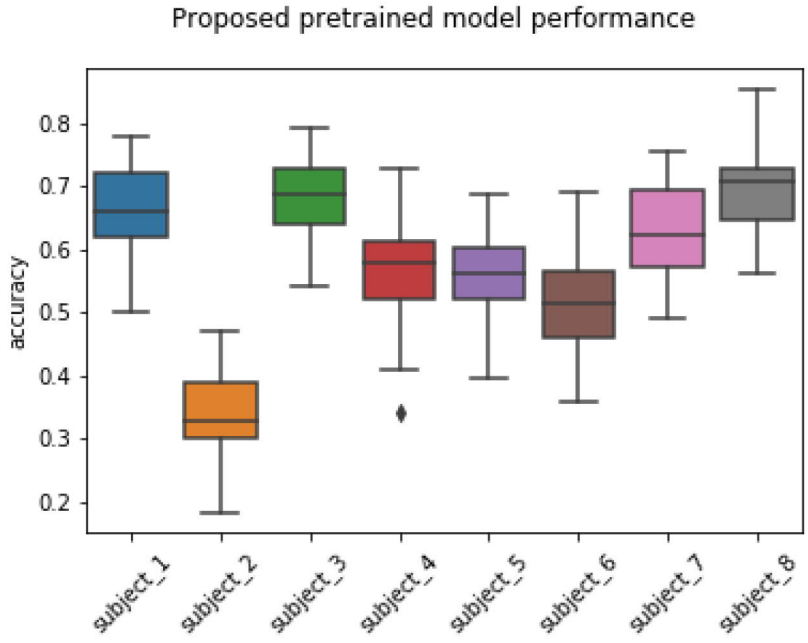

Figure 30: No-freeze transfer learning performance

On average, no-freeze transfer learning increased mean classification accuracies by 2.4% points for subjects 1 to 8. Figure 31 illustrates the effect.

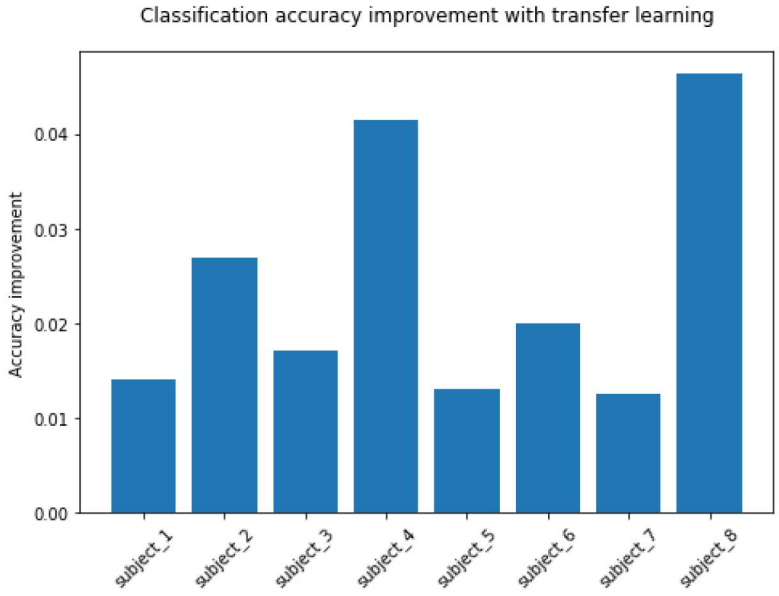

Figure 31: Accuracy improvement per subject

A side-by-side of the within-subjects and transfer learning results can be seen in figure 32.

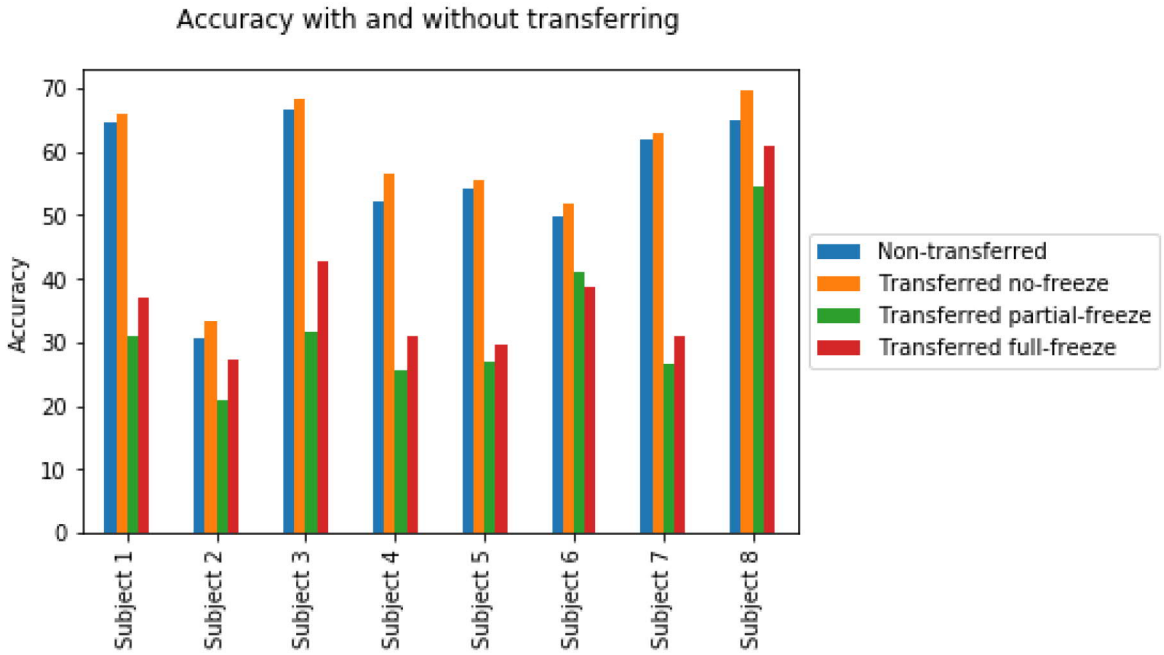

Figure 32: Accuracy improvement with transfer learning

Smaller data set after transfer

Mean accuracy for a within-subjects experiment using a no-freeze pre-trained model and the smaller data set reached a mean classification accuracy of 49.9% ± 14.2%, again omitting subject 9. See table A12 and figure 33 for the results.

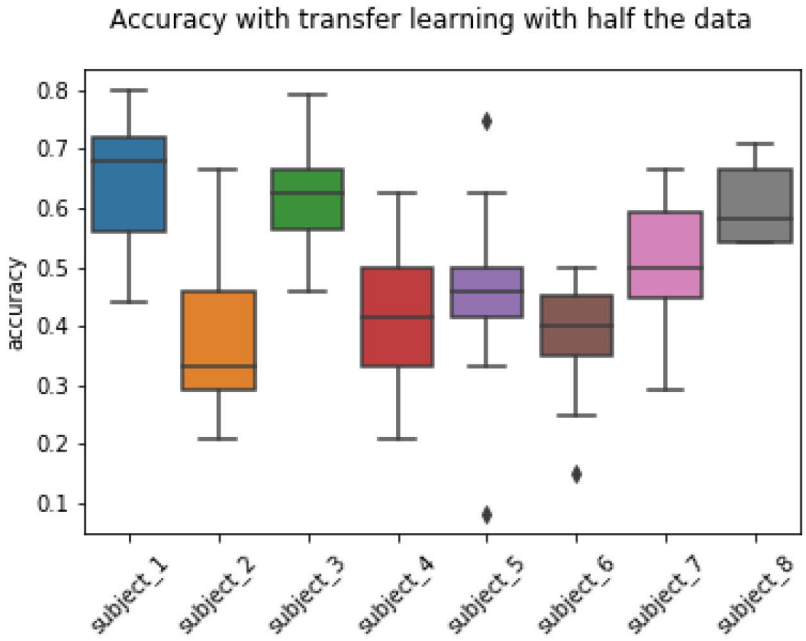

Figure 33: Accuracy with transfer learning with a smaller dataset

5 Summary

This study explored the performance of a CNN-based 4-class MI EEG classifier model operating on a minimally pre-processed BCI competition IV 2A dataset. The results indicate answers to the research questions repeated here:

- How accurately can the proposed model classify motor imagery-based brain activity in a brain-computer interface process *within subjects*?
- How accurately can the proposed model classify motor imagery-based brain activity in a brain-computer interface process *between subjects*?

The proposed CNN model classified 4-class MI EEG with a mean accuracy of 58% ± 14.9%, and employing transfer learning improved mean classification accuracies by around 2.4% points. This is around 10% points worse than other studies using similar data found, see table 7. Nonetheless, this means that the same model can classify brain activity with an accuracy high enough to be useful for most subjects (but not all). Between-subject accuracy reached a mean accuracy of 35.1% ± 11.1%. In practice, the model will hardly outperform random chance in all cases without first performing personalization, which means that any BCI application utilizing the model still needs to collect data from each end-user and personalize the model using that data first. Curiously, there were certain subject pairs for which the model classified brain activity between subjects with a significant and sizeable effect.

Within-subject performance is found dependent on the size of the data set, as made clear by the significant difference in the accuracies of the within-subjects model when using a small and large data set. Therefore, it seems plausible that further increasing the number of trials per subject should increase accuracy with a high likelihood. A segment that encapsulated the change from focus/rest to MI task had the best potential for classifying MI tasks based on the classification accuracies in the differently segmented trials - as opposed to a segment between the beginning and end of a MI task, or outside of the task altogether. However, the optimal timing of the segment varied across the subjects, which may indicate a delay of varying size in the change to the discriminative brain activity of MI task performance. Overall, the segment from (500, 1000) is considered to have the best performance. Increasing the segment size lead to improved accuracy in some instances, but the relation between the two is not clear. Moreover, the length of the segment will directly affect the overall response time of a BCI solution linearly, which is undesirable. Overall, accuracy seemed to depend on segment timing than length.

Transfer learning increased classification accuracy in every case when used optimally as found in this study, indicating that the method allows the model to reach a similar performance with fewer trials. Similarly, the accuracies seem slightly higher when training a model post-transfer with a smaller data set when using no-freeze transfer learning compared with training a model from scratch. However, the results do not make evident some model to be best for use as a pre-trained model for other subjects. It seems as though a model that showed poor performance in the earlier within-subject experiment should also be inadmissible to work well as a pre-trained

model. Similarly, a better performance within-subject might be more suitable as a pre-trained model. However, freeze-based transfer learning approaches proved unsuccessful, paling in comparison to the choice of utilizing a non-freezing-based transfer learning approach. Freezing more blocks reduced the ability of the model to generalize within-subjects as made evident by the results of the full freeze. In other words, even minimally freezing model weights has a large adverse effect on classification accuracy. Overall performance is still low, and so the time gained due to the reduction of the computational load in the training phase servers little purpose. Such methods work better with very large amounts of data, which is not the case here. Also, MI EEG signals vary greatly between subjects as well as within-subjects. However, even low-level features extracted by the first convolutional block did not seem to be directly reusable, surprisingly, which the first-block-only freeze variant demonstrated. When a pre-trained model is utilized, most beneficial is to leave all blocks unfrozen to enable the further trainability of the model weights, which adaption to new subjects requires.

Validity

However, this study produces limited evidence to support a high external validity for the model. The data set used was produced in a controlled laboratory setting, whereas the solution should be applicable in real-life scenarios. In such scenarios, data collection and device use may take place in an environment filled with external stimuli with limited controllability, which may affect model performance. Furthermore, the number of subjects in the study was small and the author provided limited demographic information on the partaking subjects. More experimentation is needed to say with confidence whether the model generalizes for a larger group of subjects. Moreover, the results fail to elucidate the exact relation between the number of trials used in training and the accuracy the model would be expected to achieve. In other words, the number of trials needed to guarantee robust performance for any subject remains unclear. Additionally, the classification model has not been applied to any real-life use case, and so online performance and usability remain unclear.

Implications

The proposed model using pretraining has its accuracy improved by around 10% points when doubling the data set size, which is in the ballpark of results by Schirrmeister et al. 2017 who found a performance increase with their similar deep ConvNet of around 15% points when switching to a data set with around thrice the number of trials. However, in a real-life use case, more trials result in longer training times, which then adversely affects the adoptability of the eventual BCI device. The small increase in classification accuracy that is achieved through transfer learning in this study is in line with the results from Xu et al. 2019, who also use transfer learning CNN for MI EEG classification. However, they choose a time-spectrum to represent input data and different model architecture. Their focus was on reducing training times, and interestingly, their framework features a partial freeze of layers

5.1 Future research

Future research can focus on other data sets. Among the known public data sets, the EEGMMIDB by PhysioNet (Schalk et al. 2004) is a larger MI EEG data set with more subjects, more channels, and more classes. Alternatively, further data can be collected by future research in a more life-like setting, to gain more external validity for a well-functioning BCI solution. Furthermore, it may be worthwhile to explore the use of a smaller segment to discover if response rates can be improved with a minimal adverse effect on accuracy. While increasing segment sizes did not seem to increase accuracy overall, it would be surprising to find if there were some significantly shorter segment that achieves similar performance as the 2-second segment identified here. What is more, pretraining slightly improves accuracy. However, it does not directly make the model generalize between subjects, thereby failing to bypass the need to personalize the model altogether. What remains unexplored by this study is the effect the use of a pre-trained model may have on reducing the amount of training data required by other subjects to reach similar performance, thereby potentially reducing personalization times by some detectible extent. For example, it could be the case that generating a pre-trained model with a larger data set should result in better model generalizability. Additionally, the usability of a BCI headset may be increased by reducing the number of electrodes utilized. Adapting the model to a smaller electrode count can be studied to see what effect it may have on performance. However, the lower electrode count requires an adaptation in input representation - time-frequency-based representations show promise. A BCI application that requires minimal to no time for personalization may be interested in opting for evoked potential brain activity altogether because it is known to be more robust between subjects. However, such an application would require an external method to generate stimuli. Another way to reduce the personalization time for new subjects would be to reduce the number of motor imagery classes. Additionally, it may be worthwhile to investigate why in the between-subjects experiment certain subject pairs can directly reuse each others' models. Perhaps there is some deep similarity in brain activity of subjects, or perhaps this effect is created by the model used. However, CNNs are notoriously difficult to interpret (Schirrmeister et al. 2017), and so why this happens is unclear. An experiment with a larger number of sbujects combined with demographic information may provide insights.

References

Brunner, Clemens et al. (2008). "BCI Competition 2008 – Graz data set A". In: URL: http://www.bbci.de/competition/iv/desc_2a.pdf.

Butterworth, S. (1930). "On the Theory of Filter Amplifiers". In: *Wireless Engineer* 7, pp. 536–541. URL: https://www.changpuak.ch/electronics/downloads/On_the_Theory_of_Filter_Amplifiers.pdf.

Cecotti, Hubert and Anthony Ries (July 2016). "Best practice for single-trial detection of event-related potentials: Application to brain-computer interfaces". In: *International Journal of Psychophysiology* 111. DOI: 10.1016/j.ijpsycho.2016.07.500.

Clevert, Djork-Arné, Thomas Unterthiner, and Sepp Hochreiter (2015). *Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)*. arXiv: 1511.07289 [cs.LG].

Cohen, Taco S. et al. (2018). *Spherical CNNs*. arXiv: 1801.10130 [cs.LG].

Crosson, Bruce et al. (2010). "Functional imaging and related techniques: an introduction for rehabilitation researchers". In: *Journal of rehabilitation research and development* 47 (2). ISSN: 0748-7711. DOI: 10.1682/jrrd.2010.02.0017.

Freer, Daniel and Guang-Zhong Yang (Jan. 2020). "Data augmentation for self-paced motor imagery classification with C-LSTM". In: *Journal of Neural Engineering* 17.1, p. 016041. DOI: 10.1088/1741-2552/ab57c0. URL: https://doi.org/10.1088%2F1741-2552%2Fab57c0.

Fries, Pascal (Oct. 2005). "A mechanism for cognitive dynamics: neuronal communication through neuronal coherence". In: *Trends in cognitive sciences* 10, pp. 474–80. ISSN: 1364-6613. DOI: 10.1016/j.tics.2005.08.011.

Gevins, Alan et al. (May 1994). "High resolution EEG 124-channel recording, spatial deblurring and MRI integration methods". In: *Electroencephalography and Clinical Neurophysiology* 90 (5), pp. 337–358. DOI: 10.1016/0013-4694(94)90050-7. URL: https://www.sciencedirect.com/science/article/pii/0013469494900507.

Hartmann, Kay Gregor, Robin Tibor Schirrmeister, and Tonio Ball (Jan. 2018). "Hierarchical internal representation of spectral features in deep convolutional networks trained for EEG decoding". In: *2018 6th International Conference on Brain-Computer Interface (BCI)*. DOI: 10.1109/iww-bci.2018.8311493. URL: http://dx.doi.org/10.1109/IWW-BCI.2018.8311493.

Haykin, Simon (1994). *Neural Networks: A Comprehensive Foundation*. 1st. USA: Prentice Hall PTR. ISBN: 0023527617.

Hinton, G. et al. (2012). "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups". In: *IEEE Signal Processing Magazine* 29.6, pp. 82–97.

Ioffe, Sergey and Christian Szegedy (2015). *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*. arXiv: 1502.03167 [cs.LG].

Kar, Aupendu et al. (Nov. 2018). "A Deep Convolutional Neural Network Based Classification Of Multi-Class Motor Imagery With Improved Generalization". In: vol. 2018. DOI: 10.1109/EMBC.2018.8513451.

Kingma, Diederik P. and Jimmy Ba (2014). *Adam: A Method for Stochastic Optimization*. arXiv: 1412.6980 [cs.LG].

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey Hinton (Jan. 2012). "ImageNet Classification with Deep Convolutional Neural Networks". In: *Neural Information Processing Systems* 25. DOI: 10.1145/3065386.

Lakshmi, M. Rajya, T. Srinivas Prasad, and Rawool Anupkumar Prakash (2014). "Survey on EEG Signal Processing Methods". In: URL: https://www.researchgate.net/publication/328419840_Survey_on_EEG_Signal_Processing_Methods.

Lawhern, Vernon J et al. (July 2018). "EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces". In: *Journal of Neural Engineering* 15.5, p. 056013. DOI: 10.1088/1741-2552/aace8c. URL: https://doi.org/10.1088%2F1741-2552%2Faace8c.

Lee, Byeong-Hoo et al. (2020). "Classification of high-dimensional motor imagery tasks based on an end-to-end role assigned convolutional neural network". In: arXiv: 2002.00210. URL: https://arxiv.org/abs/2002.00210.

Lee, H. K. and Y. Choi (2018). "A convolution neural networks scheme for classification of motor imagery EEG based on wavelet time-frequecy image". In: *2018 International Conference on Information Networking (ICOIN)*, pp. 906–909. URL: https://ieeexplore.ieee.org/document/8343254.

Lin, Chin-Teng et al. (2007). *Development of a Wireless Embedded Brain - Computer Interface and Its Application on Drowsiness Detection and Warning*. DOI: https://doi.org/10.1007/978-3-540-73331-7_61.

Liyanage, Sidath et al. (June 2013). "Dynamically weighted ensemble classification for non-stationary EEG processing". In: *Journal of neural engineering* 10, p. 036007. DOI: 10.1088/1741-2560/10/3/036007.

Lotte, F et al. (Apr. 2018). "A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update". In: *Journal of Neural Engineering* 15.3, p. 031005. DOI: 10.1088/1741-2552/aab2f2. URL: https://doi.org/10.1088%2F1741-2552%2Faab2f2.

Lu, N., T. Yin, and X. Jing (2019). "A Temporal Convolution Network Solution for EEG Motor Imagery Classification". In: *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)*, pp. 796–799. URL: https://ieeexplore.ieee.org/document/8941681.

Millan, J. R. and J. Mourino (June 2003). "Asynchronous BCI and local neural classifiers: an overview of the adaptive brain interface project". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 11.2. ISSN: 1558-0210. DOI: 10.1109/TNSRE.2003.814435.

Mitchell, Tom and McGraw Hill (1997). *Machine learning*. ISBN: 0070428077.

Ortiz Echeverri, Cesar et al. (Oct. 2019). "A New Approach for Motor Imagery Classification Based on Sorted Blind Source Separation, Continuous Wavelet Transform, and Convolutional Neural Network". In: *Sensors* 19, p. 4541. DOI: 10.3390/s19204541.

Patel SH, Azzam PN (2005). "Characterization of N200 and P300: Selected Studies of the Event-Related Potential." In: *Int J Med Sci*, 2(4):147–154. DOI: 10.7150/ijms.2.147. URL: http://www.medsci.org/v02p0147.htm.

Pfurtscheller, G. and F.H. Lopes da Silva (1999). "Event-related EEG/MEG synchronization and desynchronization: basic principles". In: *Clinical Neurophysiology*. DOI: 10.1016/S1388-2457(99)00141-8. URL: https://doi.org/10.1016/S1388-2457(99)00141-8.

Pfurtscheller, Gert et al. (2010). "The hybrid BCI". In: *Frontiers in Neuroscience* 4, p. 3. ISSN: 1662-453X. DOI: 10.3389/fnpro.2010.00003. URL: https://www.frontiersin.org/article/10.3389/fnpro.2010.00003.

Rosell, Xavier et al. (Sept. 1988). "Skin impedance from 1 Hz to 1 MHz". In: *IEEE transactions on bio-medical engineering* 35, pp. 649–51. DOI: 10.1109/10.4599.

Schalk, G. et al. (2004). *BCI2000: A General-Purpose Brain-Computer Interface (BCI) System*.

Schirrmeister, Robin Tibor et al. (Aug. 2017). "Deep learning with convolutional neural networks for EEG decoding and visualization". In: *Human Brain Mapping* 38.11, pp. 5391–5420. ISSN: 1065-9471. DOI: 10.1002/hbm.23730. URL: http://dx.doi.org/10.1002/hbm.23730.

Schwartz, Andrew B et al. (Oct. 2006). "Brain-controlled interfaces: movement restoration with neural prosthetics." In: *Neuron* 52. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2006.09.019.

Srivastava, Nitish et al. (2014). "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". In: *Journal of Machine Learning Research* 15.56, pp. 1929–1958. URL: http://jmlr.org/papers/v15/srivastava14a.html.

Sutton, Samuel et al. (1965). "Evoked-potential correlates of stimulus uncertainty". In: *Science*, 150(3700):1187–1188. DOI: 10.1126/science.150.3700.1187. URL: https://science.sciencemag.org/content/150/3700/1187/tab-pdf.

Tang, Zhichuan, Chao Li, and Shouqian Sun (2016). "Single-trial EEG classification of motor imagery using deep convolutional neural networks". In: *Optik*. DOI: 10.1016/j.ijleo.2016.10.117. URL: https://doi.org/10.1016/j.ijleo.2016.10.117.

Tangermann, Michael et al. (2012). "Review of the BCI Competition IV". In: *Frontiers in Neuroscience* 6, p. 55. ISSN: 1662-453X. DOI: 10.3389/fnins.2012.00055. URL: https://www.frontiersin.org/article/10.3389/fnins.2012.00055.

Uktveris, Tomas and Vacius Jusas (2017). "Application of Convolutional Neural Networks to Four-Class Motor Imagery Classification Problem". In: ISSN: 2335-884X. DOI: http://dx.doi.org/10.5755/j01.itc.46.2.17528.

Wang, Zijian et al. (Jan. 2018). "Short time Fourier transformation and deep neural networks for motor imagery brain computer interface recognition". In: URL: https://doi.org/10.1002/cpe.4413.

Warren S. McCulloch, Walter Pitts (Dec. 1943). "A logical calculus of the ideas immanent in nervous activity". In: *The bulletin of mathematical biophysics* 5.4, pp. 115–133. ISSN: 1522-9602. DOI: 10.1007/BF02478259. URL: https://doi.org/10.1007/BF02478259.

Wijesekera, Lokesh and Nigel Leigh (2009). "Amyotrophic lateral sclerosis". In: *Orphanet Journal of Rare Diseases*. DOI: 10.1186/1750-1172-4-3. URL: https://doi.org/10.1186/1750-1172-4-3.

Wolpaw, Jonathan and Elizabeth Winters Wolpaw (2012). *Brain-Computer Interfaces: Principles and Practice*, pp. 3–12.

Xu, G. et al. (2019). "A Deep Transfer Convolutional Neural Network Framework for EEG Signal Classification". In: *IEEE Access* 7, pp. 112767–112776. URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8772136.

Zhang, Y. et al. (2019). "Portable brain-computer interface based on novel convolutional neural network". In: *Computers in Biology and Medicine* 107, pp. 248–256. ISSN: 0010-4825. DOI: https://doi.org/10.1016/j.compbiomed.2019.02.023. URL: http://www.sciencedirect.com/science/article/pii/S0010482519300708.

Zhang et al. (2019). *A Survey on Deep Learning based Brain Computer Interface: Recent Advances and New Frontiers*. arXiv: 1905.04149 [cs.HC].

A Tables

Table A1: First trial within subject with default parameters

Subject	**Mean**	**N**
Subject 1	60.4 ± 4.9	20
Subject 2	31.0 ± 7.1	20
Subject 3	62.9 ± 4.0	20
Subject 4	32.5 ± 7.4	20
Subject 5	24.9 ± 5.3	20
Subject 6	30.1 ± 5.7	20
Subject 7	50.0 ± 8.1	20
Subject 8	59.0 ± 6.3	18
Subject 9	54.6 ± 6.7	18
Total	44.8 ± 15.7	176

Table A2: Classification accuracies with 500-sample segments with Subject 3

Segment	**Mean**	**N**
(100, 600)	27.1 ± 8.5	5
(200, 700)	23.7 ± 4.1	5
(300, 800)	43.3 ± 6.7	5
(400, 900)	63.7 ± 4.3	5
(500, 1000)	66.5 ± 5.6	33
(600, 1100)	67.3 ± 5.2	20
(700, 1200)	69.9 ± 6.5	13
(800, 1300)	65.0 ± 8.6	5
(875, 1375)	62.9 ± 4.0	20
(900, 1400)	65.0 ± 5.8	5
(1000, 1500)	62.5 ± 5.5	5
(1100, 1600)	60.4 ± 5.1	5
(1200, 1700)	49.6 ± 3.4	5
(1300, 1800)	49.2 ± 9.2	5

Table A3: Classification accuracies with sporadically varying windows with Subject 3

Segment	**Mean**	**N**
(0, 900)	53.8 ± 8.4	5
(500, 1200)	74.5 ± 5.6	13
(500, 1600)	70.7 ± 7.0	17
(700, 1600)	68.7 ± 3.9	5
(900, 1800)	62.1 ± 3.4	5
(0, 1800)	67.1 ± 4.0	5

Table A4: Three best segments for each subject

Subject	**Segment**	**Mean**	**N**
Subject 1	(500, 1000)	64.7 ± 6.4	31
Subject 1	(600, 1100)	68.1 ± 6.3	16
Subject 1	(700, 1200)	67.5 ± 6.1	8
Subject 2	(500, 1000)	30.8 ± 6.0	36
Subject 2	(600, 1100)	34.6 ± 6.2	18
Subject 2	(700, 1200)	35.3 ± 6.5	11
Subject 3	(500, 1000)	66.5 ± 5.6	33
Subject 3	(600, 1100)	67.3 ± 5.2	20
Subject 3	(700, 1200)	69.9 ± 6.5	13
Subject 4	(500, 1000)	52.3 ± 8.4	28
Subject 4	(600, 1100)	45.5 ± 8.5	15
Subject 4	(700, 1200)	44.3 ± 10.1	8
Subject 5	(500, 1000)	54.2 ± 7.3	28
Subject 5	(600, 1100)	33.5 ± 7.9	15
Subject 5	(700, 1200)	24.2 ± 5.1	8
Subject 6	(500, 1000)	49.9 ± 7.8	28
Subject 6	(600, 1100)	42.2 ± 8.8	15
Subject 6	(700, 1200)	28.2 ± 6.1	8
Subject 7	(500, 1000)	61.8 ± 7.5	26
Subject 7	(600, 1100)	54.6 ± 5.7	15
Subject 7	(700, 1200)	55.6 ± 11.0	8
Subject 8	(500, 1000)	65.1 ± 6.2	28
Subject 8	(600, 1100)	71.6 ± 7.8	14
Subject 8	(700, 1200)	63.3 ± 9.3	8
Subject 9	(500, 1000)	76.8 ± 6.1	36
Subject 9	(600, 1100)	76.3 ± 6.4	14
Subject 9	(700, 1200)	68.3 ± 6.9	8

Table A5: Prediction accuracies within subjects, smaller data set

Subject	**Mean**	**N**
Subject 1	58.7 ± 10.3	15
Subject 2	37.8 ± 9.4	15
Subject 3	60.0 ± 7.7	15
Subject 4	42.2 ± 9.3	15
Subject 5	41.4 ± 10.3	15
Subject 6	44.0 ± 13.4	15
Subject 7	38.3 ± 13.0	15
Subject 8	51.4 ± 12.2	15
Subject 9	69.5 ± 7.4	15
Total	49.3 ± 14.7	135

Table A6: Within subjects classification accuracies

Subject	**Mean**	**N**
Subject 1	64.7 ± 6.4	31
Subject 2	30.8 ± 6.0	36
Subject 3	66.5 ± 5.6	33
Subject 4	52.3 ± 8.4	28
Subject 5	54.2 ± 7.3	28
Subject 6	49.9 ± 7.8	28
Subject 7	61.8 ± 7.5	26
Subject 8	65.1 ± 6.2	28
Subject 9	76.8 ± 6.1	36
Total	58.0 ± 14.9	274

Table A7: Transfer learning full freeze results

Subject	**Mean**	**N**
Subject 1	31.0 ± 6.6	4
Subject 2	20.9 ± 5.9	4
Subject 3	31.8 ± 5.5	4
Subject 4	25.6 ± 4.3	4
Subject 5	27.1 ± 4.2	4
Subject 6	41.0 ± 7.5	4
Subject 7	26.5 ± 4.4	4
Subject 8	54.7 ± 6.0	4
Total	32.3 ± 11.3	32

Table A8: Transfer learning partial freeze results

Subject	**Mean**	**N**
Subject 1	37.2 ± 4.7	10
Subject 2	27.3 ± 7.6	10
Subject 3	42.9 ± 5.0	10
Subject 4	30.9 ± 6.8	10
Subject 5	29.8 ± 9.8	10
Subject 6	38.7 ± 7.9	10
Subject 7	31.0 ± 8.0	10
Subject 8	60.8 ± 7.7	10
Total	37.3 ± 12.3	80

Table A9: No-freeze transfer learning classification accuracies

Subject	**Mean**	**N**
Subject 1	66.1 ± 6.3	71
Subject 2	33.5 ± 6.7	64
Subject 3	68.2 ± 5.8	64
Subject 4	56.4 ± 8.1	62
Subject 5	55.5 ± 7.2	62
Subject 6	51.9 ± 7.4	62
Subject 7	63.0 ± 7.1	62
Subject 8	69.7 ± 7.1	63
Total	58.2 ± 13.1	510

Table A10: Improvement of no-freeze transfer

Subject	**Mean**
Subject 1	1.4
Subject 2	2.7
Subject 3	1.7
Subject 4	4.1
Subject 5	1.3
Subject 6	2.0
Subject 7	1.3
Subject 8	4.6
Total	2.4 ± 1.3

Table A11: Between subjects classification accuracies

Trained on	**Evaluated on**	**Mean**	**N**
Subject 1	Subject 1	64.7 ± 6.4	31
Subject 1	Subject 2	29.1 ± 1.6	15
Subject 1	Subject 3	48.7 ± 3.1	15
Subject 1	Subject 4	33.6 ± 3.0	15
Subject 1	Subject 5	26.3 ± 2.0	15
Subject 1	Subject 6	31.6 ± 3.3	15
Subject 1	Subject 7	29.3 ± 2.2	15
Subject 1	Subject 8	38.6 ± 3.6	15
Subject 1	Subject 9	36.2 ± 3.8	15
Subject 2	Subject 1	26.1 ± 2.8	15
Subject 2	Subject 2	30.8 ± 6.0	36
Subject 2	Subject 3	26.8 ± 3.2	15
Subject 2	Subject 4	27.7 ± 3.2	15
Subject 2	Subject 5	25.7 ± 2.1	15

Subject 2	Subject 6	31.0 ± 2.5	15
Subject 2	Subject 7	27.5 ± 1.8	15
Subject 2	Subject 8	27.3 ± 2.2	15
Subject 2	Subject 9	21.3 ± 2.7	15
Subject 3	Subject 1	49.8 ± 3.7	15
Subject 3	Subject 2	27.3 ± 1.7	15
Subject 3	Subject 3	66.5 ± 5.6	33
Subject 3	Subject 4	30.1 ± 2.6	15
Subject 3	Subject 5	24.0 ± 2.1	15
Subject 3	Subject 6	33.1 ± 3.2	15
Subject 3	Subject 7	30.0 ± 2.2	15
Subject 3	Subject 8	41.4 ± 3.2	15
Subject 3	Subject 9	45.7 ± 3.5	15
Subject 4	Subject 1	36.6 ± 2.1	15
Subject 4	Subject 2	26.8 ± 1.8	15
Subject 4	Subject 3	30.4 ± 3.1	15
Subject 4	Subject 4	52.3 ± 8.4	28
Subject 4	Subject 5	29.3 ± 2.6	15
Subject 4	Subject 6	27.1 ± 3.2	15
Subject 4	Subject 7	29.3 ± 3.5	15
Subject 4	Subject 8	29.7 ± 2.2	15
Subject 4	Subject 9	30.7 ± 2.8	15
Subject 5	Subject 1	28.8 ± 2.2	15
Subject 5	Subject 2	28.4 ± 2.2	15
Subject 5	Subject 3	25.6 ± 3.1	15
Subject 5	Subject 4	28.1 ± 2.4	15
Subject 5	Subject 5	54.2 ± 7.3	28
Subject 5	Subject 6	27.8 ± 2.8	15
Subject 5	Subject 7	36.9 ± 3.2	15
Subject 5	Subject 8	27.2 ± 2.4	15
Subject 5	Subject 9	25.4 ± 4.2	15
Subject 6	Subject 1	35.4 ± 4.3	15
Subject 6	Subject 2	28.7 ± 1.5	15
Subject 6	Subject 3	39.4 ± 4.0	15
Subject 6	Subject 4	30.0 ± 1.6	15
Subject 6	Subject 5	27.1 ± 2.7	15
Subject 6	Subject 6	49.9 ± 7.8	28
Subject 6	Subject 7	27.0 ± 2.1	15
Subject 6	Subject 8	36.8 ± 3.3	15
Subject 6	Subject 9	35.7 ± 3.7	15
Subject 7	Subject 1	42.3 ± 2.7	15
Subject 7	Subject 2	31.2 ± 1.4	15
Subject 7	Subject 3	41.9 ± 3.2	15

Subject 7	Subject 4	33.0 ± 2.9	15
Subject 7	Subject 5	37.8 ± 2.6	15
Subject 7	Subject 6	31.2 ± 2.2	15
Subject 7	Subject 7	61.8 ± 7.5	26
Subject 7	Subject 8	35.0 ± 3.7	15
Subject 7	Subject 9	36.3 ± 4.3	15
Subject 8	Subject 1	39.6 ± 5.3	15
Subject 8	Subject 2	28.5 ± 3.1	15
Subject 8	Subject 3	45.6 ± 5.7	15
Subject 8	Subject 4	29.5 ± 3.3	15
Subject 8	Subject 5	25.3 ± 2.4	15
Subject 8	Subject 6	35.5 ± 3.0	15
Subject 8	Subject 7	27.5 ± 2.5	15
Subject 8	Subject 8	65.1 ± 6.2	28
Subject 8	Subject 9	51.7 ± 3.0	15
Subject 9	Subject 1	31.5 ± 3.2	15
Subject 9	Subject 2	25.3 ± 1.9	15
Subject 9	Subject 3	33.2 ± 4.2	15
Subject 9	Subject 4	26.0 ± 2.0	15
Subject 9	Subject 5	24.5 ± 1.4	15
Subject 9	Subject 6	32.6 ± 2.4	15
Subject 9	Subject 7	27.3 ± 1.8	16
Subject 9	Subject 8	49.2 ± 4.3	16
Subject 9	Subject 9	76.8 ± 6.1	36
Total	35.1 ± 11.1	81	

Table A12: Accuracy within subjects with transfer learning and a smaller data set

Subject	**Mean**	**N**
Subject 1	63.8 ± 11.0	19
Subject 2	38.6 ± 11.7	19
Subject 3	61.8 ± 8.4	19
Subject 4	41.0 ± 11.9	19
Subject 5	45.8 ± 12.8	20
Subject 6	38.7 ± 9.0	20
Subject 7	50.2 ± 10.8	20
Subject 8	60.3 ± 6.6	19
Total	49.9 ± 14.2	155

Table A13: Transfer learning matrix

Subject	**Pretrained on**	**Segment**	**Mean**	**N**

Subject 2	Subject 1	(500, 1000)	32.0 ± 5.9	3
Subject 3	Subject 1	(500, 1000)	65.3 ± 7.3	3
Subject 4	Subject 1	(500, 1000)	60.2 ± 11.2	2
Subject 5	Subject 1	(500, 1000)	47.9 ± 2.9	2
Subject 6	Subject 1	(500, 1000)	39.7 ± 1.8	2
Subject 7	Subject 1	(500, 1000)	59.2 ± 2.9	2
Subject 8	Subject 1	(500, 1000)	60.4 ± 2.9	2
Subject 9	Subject 1	(500, 1000)	72.2 ± 7.9	2
Subject 1	Subject 2	(500, 1000)	49.0 ± 4.2	2
Subject 3	Subject 2	(500, 1000)	52.1 ± 5.9	2
Subject 4	Subject 2	(500, 1000)	39.8 ± 11.2	2
Subject 5	Subject 2	(500, 1000)	28.1 ± 10.3	2
Subject 6	Subject 2	(500, 1000)	39.7 ± 9.1	2
Subject 7	Subject 2	(500, 1000)	31.6 ± 13.0	2
Subject 8	Subject 2	(500, 1000)	52.1 ± 5.9	2
Subject 9	Subject 2	(500, 1000)	67.8 ± 1.6	2
Subject 1	Subject 3	(500, 1000)	58.0 ± 5.7	2
Subject 2	Subject 3	(500, 1000)	40.8 ± 5.8	2
Subject 4	Subject 3	(500, 1000)	55.7 ± 8.0	2
Subject 5	Subject 3	(500, 1000)	51.0 ± 1.5	2
Subject 6	Subject 3	(500, 1000)	53.8 ± 0.0	2
Subject 7	Subject 3	(500, 1000)	46.9 ± 2.9	2
Subject 8	Subject 3	(500, 1000)	61.5 ± 4.4	2
Subject 9	Subject 3	(500, 1000)	68.9 ± 6.3	2
Subject 1	Subject 4	(500, 1000)	54.0 ± 5.7	2
Subject 2	Subject 4	(500, 1000)	39.8 ± 4.3	2
Subject 3	Subject 4	(500, 1000)	65.6 ± 7.4	2
Subject 5	Subject 4	(500, 1000)	43.7 ± 11.8	2
Subject 6	Subject 4	(500, 1000)	42.3 ± 1.8	2
Subject 7	Subject 4	(500, 1000)	59.2 ± 0.0	2
Subject 8	Subject 4	(500, 1000)	69.8 ± 1.5	2
Subject 9	Subject 4	(500, 1000)	81.1 ± 1.6	2
Subject 1	Subject 5	(500, 1000)	67.0 ± 1.4	2
Subject 2	Subject 5	(500, 1000)	39.8 ± 13.0	2
Subject 3	Subject 5	(500, 1000)	61.5 ± 4.4	2
Subject 4	Subject 5	(500, 1000)	45.5 ± 12.9	2
Subject 6	Subject 5	(500, 1000)	39.7 ± 5.4	2
Subject 7	Subject 5	(500, 1000)	49.0 ± 2.9	2
Subject 8	Subject 5	(500, 1000)	56.2 ± 5.9	2
Subject 9	Subject 5	(500, 1000)	77.8 ± 6.3	2
Subject 1	Subject 6	(500, 1000)	61.0 ± 1.4	2
Subject 2	Subject 6	(500, 1000)	25.5 ± 4.3	2
Subject 3	Subject 6	(500, 1000)	69.8 ± 1.5	2

Subject 4	Subject 6	(500, 1000)	55.7 ± 1.6	2
Subject 5	Subject 6	(500, 1000)	54.2 ± 0.0	2
Subject 7	Subject 6	(500, 1000)	46.9 ± 2.9	2
Subject 8	Subject 6	(500, 1000)	63.5 ± 7.4	2
Subject 9	Subject 6	(500, 1000)	74.4 ± 11.0	2
Subject 1	Subject 7	(500, 1000)	66.0 ± 5.7	2
Subject 2	Subject 7	(500, 1000)	32.7 ± 0.0	2
Subject 3	Subject 7	(500, 1000)	62.5 ± 11.8	2
Subject 4	Subject 7	(500, 1000)	73.9 ± 1.6	2
Subject 5	Subject 7	(500, 1000)	52.1 ± 5.9	2
Subject 6	Subject 7	(500, 1000)	30.8 ± 3.6	2
Subject 8	Subject 7	(500, 1000)	61.5 ± 1.5	2
Subject 9	Subject 7	(500, 1000)	80.0 ± 9.4	2
Subject 1	Subject 8	(500, 1000)	60.0 ± 2.8	2
Subject 2	Subject 8	(500, 1000)	35.7 ± 7.2	2
Subject 3	Subject 8	(500, 1000)	67.7 ± 4.4	2
Subject 4	Subject 8	(500, 1000)	63.6 ± 3.2	2
Subject 5	Subject 8	(500, 1000)	47.9 ± 8.8	2
Subject 6	Subject 8	(500, 1000)	51.3 ± 3.6	2
Subject 7	Subject 8	(500, 1000)	56.1 ± 4.3	2
Subject 9	Subject 8	(500, 1000)	74.4 ± 4.7	2
Subject 1	Subject 9	(500, 1000)	66.3 ± 7.2	32
Subject 2	Subject 9	(500, 1000)	33.9 ± 6.3	27
Subject 3	Subject 9	(500, 1000)	67.1 ± 6.1	27
Subject 4	Subject 9	(500, 1000)	56.9 ± 7.9	27
Subject 5	Subject 9	(500, 1000)	55.4 ± 5.5	27
Subject 6	Subject 9	(500, 1000)	52.9 ± 7.6	27
Subject 7	Subject 9	(500, 1000)	61.9 ± 6.3	27
Subject 8	Subject 9	(500, 1000)	69.0 ± 6.0	28