Explorations in Emoji-based P300 BCI-Spellers and Convolutional Neural Network Optimization for SSVEP-based Bio-Signals

PODMORE, JOSHUA,JAMES

Explorations in Emoji-based P300 BCI-Spellers and Convolutional Neural Network Optimization for SSVEP-based Bio-Signals

Joshua J Podmore

Department of Psychology

University of Durham

This dissertation is submitted for the degree of

Doctor of Philosophy

Trevelyan College

2023

Abstract

The past decade has seen significant enhancements in the development of communication-based Brain-Computer Interface (BCI) spellers. These devices often harness brain-based bio-signals via Electroencephalography (EEG) for speller control. To increase the scope of speller viability and functionality we first developed a simplistic emoji-based visual speller paradigm using the P300 bio-signal. The inclusion of emojis over traditional letters, numbers and characters is predicted to enable richer emotional communication capabilities to end-point users with the most severe forms of paralysis. Here is presented a staggered exploration of stimulus design formats ranging between 3, 5 and 7 target emoji arrays positioned from agreeable to disagreeable. In the final iteration of the experimental procedure, a closed-loop system is assessed using 3 neurotypical subjects. This necessitated the real-time capture, pre-processing, classification, and prediction presentation of subject dry-EEG data. The highest-performing single-subject achieved 83% offline classification accuracy for an analysis variant utilizing SMOTE oversampling data augmentation. The final chapter of the thesis focuses on the optimization of pre-processing frequency filters for SSVEP-based bio-signal classification using a range of convolutional neural networks (ShallowConvNet, DeepConvNet, EEGNet & EEGNetSSVEP). All analyses were computed utilizing the open-source 12-target, 10-subject Nakanishi SSVEP Numpad repository. These investigations revealed a positive trend in optimized low-pass filter cutoffs for networks presenting with a greater number of trainable parameters, or a higher model layer count. These results align with current cutting-edge CNN SSVEP classifier research and suggest the effective extraction of SSVEP harmonics is dependent on network complexity. Further, the optimization of aggregated, cross-subject data pre-processing frequency filter cutoffs is shown to enhance subject-level classification performance for both high and low-complexity networks. These methods provide a guideline for research into the optimization of cross-subject dataset pre-processing stages and outline a paradigm for the optimal comparison of CNNs for SSVEP classification.

Declaration

The work in this thesis is based on research carried out at the Department of Psychology, Durham University, United Kingdom. No part of this thesis has been submitted elsewhere for any other degree or qualification and it is all my own work unless referenced to the contrary in the text.

This work was supported by the Centre for Visual Arts and Culture (CVAC) with funding from the Leverhulme Doctoral Scholarships Scheme.

Joshua J Podmore
2023

Acknowledgements

I would first like to extend sincere thanks to my main supervisor, Dr Ulrik Beierholm, who demonstrated limitless patience, kindness and encouragement throughout my PhD journey. Further, a special thanks is owed to Prof. Toby Breckon, his unwavering support and understanding have guided me soundly over the course of both my MRes and PhD. Additionally, I must highlight the contributions made by Prof. Jane Macnoughton, especially in regard to traversing the obstacles posed by the COVID-19 pandemic. I have drawn many times from the deep well of compassion and wisdom offered by this inter-disciplinary supervision team and I am humbled by their vast provision of time and effort in aiding the completion of this work.

I would like to acknowledge the Leverhulme Trust and all of the CVAC scholars, especially my own cohort, for their assistance in refining my academic interests beyond the bounds of science and defining my time spent on campus as truly unforgettable. Moreover, the selfless tenacity with which Prof. Ludmilla Jordanova pursued the task of assisting our completion and the gentleness of care she showed towards the disquieted hearts of so many early career scholars must be applauded.

I must also give thanks to all the staff who assisted me throughout my placement at the James Cook University Hospital. Your openness, affection and stalwart professionalism contributed immensely to the crystallisation of my own career goals and continue to set an incredible example to all with a passion for clinical care. Importantly, I would like to extend a special mention to Geoff whose laughter and warmth as a ward volunteer made me and countless others feel welcome.

My choice to endeavour on these academic pursuits is inextricably linked to the inspiring conversations I shared with Dr Jason Connolly during the first few months of my undergraduate and for this, I will be eternally grateful. Finally, the biggest thanks I have to offer are to my mother, Louise and brothers, Oliver and Luke Podmore. Their unconditional support and kindness have helped me navigate this often challenging process and I will do all in my power to one day repay these debts.

I would like to dedicate this thesis to the memory of Dr Cristiana Cavina-Pratesi who greatly influenced my life and countless others.
The brilliance of her star is unrivalled against the inky midnight of our yesterdays.

Contents

List of Tables

List of Figures

Abbreviations

ACP Advance Care Planning

ALS Amyotrophic Lateral Sclerosis

AoC Accuracy of Classification

AP Average Precision

BCI Brain-Computer Interface

bpm Bits per Minute

bps Bits per Second

CLIS Complete Locked In Syndrome

CNN Convolutional Neural Network

CPU Computational Processing Unit

DCNN Deep Convolutional Neural Network

EEG Electroencephalography

EMG Electromyography

ERP Event-Related Potential

FBCCA Filter-Bank Canonical Correlation Analysis

FDA Fishers Discriminant Analysis

fMRI Functional Magnetic Resonance Imaging

GPU Graphics Processing Unit

GUI Graphic User Interface

HCI Human-Computer Interface

ICA Independent Component Analysis

ILIS Incomplete-Locked In Syndrome

ITR Information Transfer Rate

JPFM Joint Phase-Frequency Modulation

LDA Linear Discriminant Analysis

lsqr Least Squares

MA Mean Accuracy

MEG Magnetoencephalography

MI Motor Imagery

MLP Multi-Layer Perceptron

MOAB Mother of all Benchmarks

NCI Neural-Computer Interface

PCA Principal Component Analysis

QoL Quality of Life

RPT Random Performance Thresholds

SGD Stochastic Gradient Descent

SHP Successive-Halving Pruner

SMOTE Synthetic Minority Over-sampling Technique

SMR Sensorimotor Rhythms

SSVEP Steady State Visual Evoked Potentials

SVD Single Value Decomposition

SWLDA Step-Wise Linear Discriminant Analysis

TPE Tree-structured Parzen Estimation

TRCA Task-Related Component Analysis

Chapter 1

Introduction

1.1 Positioning Statement

The work conducted herein was performed under the challenging conditions arising from the COVID-19 pandemic. The circumstances surrounding the PhD ultimately had a dramatic influence on the research conducted herein. Chapters 2, 3 and 4 of the thesis comprise the totality of all 1st-hand experimentation undertaken. The progress of this research was hampered significantly by the onset of the COVID-19 virus and its associated complications. Specifically, in-person research was restricted, preventing the collection of additional data. This led to an adaptation of the overall thesis, focusing on the methods of optimising the numerous variables associated with EEG pre-processing and neural network-based classification methods. The pandemic also introduced many non-research-related issues, these included disruption to in-person experimentation and data collection, isolation from the university community, compromised supervisory contact (migrated to Zoom), moving home twice in the space of 9 months as well as the exacerbation of PhD-related stresses and anxieties. The author's intent in outlining these factors is to illustrate the context in which this work was undertaken and provide clarity on the somewhat disconnected nature of the two independent research projects defined herein.

1.2 Motivation

The motivation for this thesis, as explored across the studies detailed in Chapters 3, 4, and 5, is rooted in a commitment to advancing the field of alternative and augmentative communication systems, particularly for a specific subset of the patient population whose ability to communicate through volitional means is severely compromised. These individuals, often facing the sudden or gradual loss of the ability to vocalize, are typically treated with a series of communication aid interventions. The range of interventions varies widely, from speech therapy, pictographic boards and eye-code systems to platforms that leverage more advanced technologies such as eye-tracker messaging and browser platforms, custom-residual movement systems, and Brain-Computer Interface (BCI) communication devices. The maintenance of communication between the patient, clinicians, caregivers, friends and family is crucial. The real and perceived reduction in autonomy and agency experienced by patients can have a dramatic negative influence on clinical outcomes and Quality of Life (QoL).

1.2.1 Patient-Centered Treatment and Issues of Consent

In the case of patients with progressive disorders, questions of consent and guardianship are addressed before the onset of a complete-locked-in state via the implementation of Advance Care Planning (ACP) strategies. This involves regular collaborative decision-making on treatment and goals between clinicians, caregivers, and patients before the loss of any communicative ability [1]. Alternatively, for individuals without any means of discernible effective communication, the corresponding legal guardians are consulted and any initial attempts to assist a patient with for example, a BCI-based speller, are performed in the absence of direct consent [2]. Once some level of communication is established the clinicians can of course at this point follow the appropriate guidelines. This is not an ideal set of circumstances and is typically reserved for patients with traumatic brain injury-related aetiological pathways. The means by which researchers justify the use of so-called in-direct consent via legal guardians are informed by the study of patient populations with progressive disorders. This involves conducting repeated surveys with patients as their ability to functionally communicate degrades [3, 4]. These findings provide clinicians with data to identify the core needs of these patients before the attainment of a complete-locked-in state.

1.2.2 Quality of Life

As noted above, communication aids are crucial for individuals with severe dysarthria or anarthria (complete loss of volitional speech) due to several key considerations relating to consent and treatment outcomes. The inability to communicate as has a dramtic impact on the quality of life for both patients and their caregivers. There is a well-documented relationship between communication difficulties and significant psychological distress, including heightened feelings of depression, loneliness, and a sense of being a burden. In [5], researchers revealed that the inability to communicate effectively exacerbates the psychological suffering of ALS patients. These individuals often reported feeling isolated and overwhelmed by their condition, which not only diminishes their overall well-being but also was shown to increase their desire to end their lives prematurely. The study underscores the critical importance of addressing communication challenges in ALS patients as a means of improving their mental health and overall quality of life. Among the several variants of ALS, Bulbar onset is the most severe. This involves the degeneration of the brain stem, particularly the medulla oblongata [6] and can cause dramatic loss in speech and swallowing functions early in the disease, leading to rapid deterioration in communication and quality of life.

Notably, Bulbar onset ALS patients, as opposed to Limb, Respiratory or Axial onset patients, often present with significantly poorer quality of life outcomes. As detailed in [7], this decline is primarily due to the drastically lower degree of independence experienced by these patients, as the early loss of speech and swallowing functions severely impacts their daily lives and ability to communicate. Research involving self-reports of ALS patient attitudes demonstrate that loss of speech and dysarthria more generally has a dramatic negative impact on QoL rating scales. Along these very same lines, the same paper [7] found significant differences in reported depression between ALS Bulbar onset patients using communication aids and those receiving only speech therapy. Specifically, communication aids were associated with lower levels of depression and a better overall mood, highlighting their crucial role in improving QoL for individuals with severe speech impairments. Along these very same lines, [8] reported higher functional independence, a greater ability to convey basic needs and a higher perceived ability to participate in social activities for near-locked-in patients using eye-tracker systems, as opposed to ETRAN letter boards.

Additionally, case study reports have been detailed noting the successful transition from eye-tracker-based systems to P300-based Brain-Computer interface systems [9]. As will be noted and elaborated on in the subsequent Literature Review Chapter, these systems are operated via brain-based bio-signals and can enable the operation of assistive communication devices for

individuals with late-stage ALS or severe tetraplegia following a decline in ocular dexterity. Further, research undertaken to gauge the attitudes and preferences of prospective patients to utilize such advanced technologies indicates a strong willingness to use these devices [10]. This interest in the use of such methods is however qualified by certain operational demands relating to speed of communication, system complexity, selection prediction accuracy, cost, maintenance and comfort [3]. The principle aim of this thesis is to contribute to the corpus of knowledge surrounding the development of communication-based Brain-Computer Interfaces in the ultimate pursuit of improving the QoL and clinical outcomes of individuals with the severest forms of tetraplegia.

1.3 Target Population and Aetiology

The following research is aimed at the development and improvement of BCI-speller technologies for target patient populations presenting with quadriplegia, anarthria (incapable of vocalization) and retained functional vision and movement in one or both eyes. In this instance, the diagnostic definition of quadriplegia follows UK medical standards as the absence of volitional control of the head and all four limbs. More specifically, this blend of conditions is termed Incomplete-Locked-In-Syndrome (ILIS). Crucially, patients presenting with ILIS demonstrate, at minimum, the retention of voluntary motor control for blinking and vertical eye movements. This differs from the so-called Complete Locked-In Syndrome, defined as quadriplegia with residual motor control. Due to these restrictions, the aforementioned target populations can not utilize eye-tracking technologies for communication. Further, the increased severity of the condition typically introduces additional complications to the successful deployment of communication devices owing to a higher incidence of fatigue. Note, that this is a broad generalization and each patient's clinical circumstances vary substantially from case to case.

The most typical aetiology of ILIS involves traumatic injury to the brainstem or ventral pons via haemorrhage[11] or ischemia [12]. Further, progressive disorders characterized by the degradation of the central nervous system can also lead to this unique blend of conditions. These include but are not restricted to, Central Pontine Myelinolysis (CPM), cancerous tumour formation (thoracic level and above), Amyotrophic Lateral Sclerosis (ALS) and multiple sclerosis [13]. Notably, individuals presenting with severe degenerative motor disorders can progress to Complete-Locked-In Syndrome (CLIS). This is characterized by the absence of all volitional motor control in tandem with a state of wakefulness [14]. For instances in which wakefulness can not be established, the patient would likely be re-diagnosed as in a so-called

'vegetative state'.

In the initial stages of such degenerative disorders, patients may retain the capacity for vocalized speech, with even the ability to embellish any utterances with emotive inflexions to provide additional communicative context. As the conditions progress, the standard vocal-based communications can be complemented with pictorial or alphabet-based boards featuring images or characters. In the final stages, highly simplistic spelling devices can be used either independently or with assistance from a clinician to answer simple binary questions. As stated earlier, numerous patients can not use these highly simplistic communication systems due to the eventual loss of reliable control over any muscle groups, even ocular-based movement including blinks and smooth saccades [15].

Crucially, the rate of degeneration varies significantly across individuals and is highly linked to the aetiological pathway of the patient's condition [12]. The most severe outcomes and lowest incidence for recovery are described for individuals with a traumatic injury-based ILIS and CLIS relating to damage of the pons or brainstem. Further, previous research demonstrates that just 40-60% of all patients do not survive past the first 4 months post-injury [16]. These data suggest that BCI researchers must advance the capabilities of BCI-speller devices to enhance life quality for end-point patient users with both highly limited long-term mortality rates and for those individuals with degenerative disorders. Note, that at no point in this thesis is data collected from the aforementioned patient populations. To clarify, studies relating to the thesis herein are purely developmental and all research was conducted in lab settings with typical-healthy subjects. It is the author's intention for this work to function as a foundation for potential future implementations with real end-point user patient populations.

1.4 Emoji-based P300 Speller Rationale

The development of Brain-Computer Interfaces (BCIs) for individuals with severe communication impairments often centres around optimizing alphanumeric speller systems. These systems typically feature a grid matrix, usually 6 x 6 in dimension, containing letters and numbers. To facilitate communication, the system employs a time-locked flashing sequence of these targets. Concurrently, an Electroencephalograph (EEG) records brain activity, capturing changes in micro-voltage across the scalp in response to these visual stimuli. In a P300 speller system, the user focuses on a specific target letter or number, and approximately 300 milliseconds after the target is highlighted, a positive deflection in the EEG signal is detected, propagating from the frontal (Fz) to the central (Cz) and anterior (Pz) electrode locations. De-

spite their potential, these speller systems often lead to user complaints of fatigue, eye strain, and high cognitive load, as highlighted by [17], who identified these issues as significant barriers to user acceptance and effective use of BCIs.

The use of dense target arrays in BCIs, which display many letters and numbers, is problematic for individuals with severe communication impairments due to high rates of visual impairment in this population. Studies, such as [18], report that up to 66% of near-locked-in patients have significant visual deficits, making it difficult for them to effectively use eye-tracker based systems. This limitation underscores the need for alternative communication methods better suited to their sensory challenges. Given these obstacles, studies have found that potential patient users often prefer and achieve higher classification accuracies with two-step selection systems in BCIs. In these systems, target letters and numbers are first grouped into clusters on the screen during an initial selection phase. Users then select individual targets from these clusters in a secondary step. Research by [19] and [20]supports this approach, showing that clustering reduces the cognitive load and visual strain, leading to more efficient and accurate communication for ALS patient groups.

For many individuals with severe communication impairments, such as those with congenital disorders like cerebral palsy or stroke patients with aphasia, text-based communication methods are often unfeasible. For instance, cerebral palsy can restrict language acquisition and comprehension, making traditional text communication impractical [21]. Similarly, stroke patients with aphasia may struggle with language comprehension to the extent that text-based methods are unviable [22]. In such cases, alternative low-tech methods such as communication books or customizable software systems that use symbolic representations of objects or actions provide a practical substitute. These enable individuals to communicate more effectively without relying on text or vocalization-based methods. Tools such as these help bridge the communication gap by allowing users to express needs and ideas through images and symbols.

A handful of studies report the use of symbols, in place of traditional letters and numbers, principally to regulate environmental controls. In most instances, these systems involve presenting between 3 and 12 symbols to address basic functional needs such as emergency alerts, light switching, television controls, temperature regulation, and telephone call activation [23–25]. Additionally, more sophisticated systems have been developed that integrate P300s with electromyographic (EMG) signals in a cluster-based tree design [26]. In these advanced systems, users start by selecting options from a main interface that presents a range of operational

sub-classes, including bed controls, television operations, and wheelchair commands. This hierarchical approach allows users to navigate through various levels of control with relative ease. The perceived complexity of using this method was found to be relatively low across both ALS and typical healthy user, suggesting that long-term implementation could significantly enhance the independence of patients and lead to notable improvements in their quality of life.

The International Classification of Functioning, Disability, and Health (ICF) is a framework designed to assess and describe health and disability comprehensively by evaluating body functions, activities, participation, and environmental factors, aiming to enhance understanding and support across various aspects of functioning. Environmental BCI systems, as discussed above, address the participation category by enabling telephone control; however, these systems are often impractical for individuals with severe incomplete locked-in syndrome due to the absence of volitional speech or residual movement needed to operate text-based input systems.

Communicative participation refers to the ability to engage effectively in interactions and exchange information in various life situations. To better quantify this outcome measure, the self-report Communicative Participation Item Bank (CPIB) was developed [27]. The CPIB helps highlight the impact of patient conditions on everyday communication, including conversations with strangers, group discussions, and telephone calls. This tool has been used to assess the relationship between CPIB items and functional capacities associated with volitional communication control, particularly speech severity (self-reported speech ability), speech usage, and swallowing severity (self-reported swallowing ability). Research has demonstrated that speech severity items accounted for the greatest variance in CPIB responses in a sample of 70 patients with ALS, Parkinson's disease, multiple sclerosis, and head/neck cancer [28]. Extrapolating these findings to individuals who rely on simple eye codes as their primary means of communication underscores the profound social isolation experienced by these patient populations.

Further, in one of the most thorough evaluations of potential CLIS patients' device preferences to date, the authors revealed that emotional expressivity also was categorized as a highly attractive attribute of any prospective communication-based BCI device [29]. Arguably one of the most widely implemented means of emotional expressivity in text-based communications is emojis. These ideograms, originally introduced by Japanese telecom companies to cut down on electronic pager-message length, feature humanoid faces typically yellow in

colouration and vary dramatically in valence and content [30]. The first implementation of a compacted pictorial unit in text-based communications is arguably the emoticon [31]. This describes a sequence of pre-existing characters typically arranged to depict an emotional state and often operates as a paralinguistic element to suffix a statement (for example, ' :) '). This describes a method of clarifying the intended meaning of a text communication via the use of a qualifying icon. These language tools can assist in reducing ambiguity [32], enhancing emotional expressivity [33] and crucially, increasing communication efficiency [34]. Note, that the emoji is arguably more effective in terms of information transfer due to the single character length, and in terms of emotional richness, expressivity [35–37] and impact [38] as compared to emoticons (for review see, [39]).

Moreover, the inclusion of emojis into natural language networks has enabled experimenters to demonstrate the significant influence of emojis in text-based messages [30, 40].Further, the generalized application of emoji across large subject populations [37] has been shown, alongside numerous stratified differences in terms of region [41], gender and age [42, 43]. These suggest that the scope for universal understanding is inherent to emoji application, as well as the ability for specified usage, arguably providing emoji with greater flexibility, albeit lower specificity, than purely text-based communication methods. It could be argued that the hybridized integration of emoji and text-based communications has become ubiquitous with individualized expression in the digital age. Along these very same lines, these tools must be offered to BCI speller users to realise these same individualization goals for the ultimate purpose of enhancing patient quality of life. In pursuit of these aims, the author proposes a study investigating the viability of an emoji-based emotional communication experimental protocol utilizing the P300 waveform.

As is discussed at length in the following chapter (see, subsection 2.6 Emoji-based Speller Designs), it is the author's understanding that emoji have only been utilized once in communication-based BCI studies. As described in [44], a series of 4 emoji were positioned within a 3 x 4 matrix of functional operations and environmental control messages, including temperature regulation, hunger, disagreement and confusion. Here, the authors utilized a convolutional neural network (CNN) for the clasification of signal images to predict user target selections with a cross-subject accuracy of 90%. This previous research is highly promising and it is the author's belief that contributing to this area of BCI development could have substantial impacts on the future quality of life for end-point patient users following additional research and development.

Objective The P300 emoji speller system outlined in this thesis is designed as an alternative communication platform aimed at improving the expressive capabilities of individuals who are unable to communicate through traditional means. The system employs an innovative approach by utilizing emojis, which are mapped to a reduced, one-dimensional pleasure valence scale, allowing for more nuanced emotional communication. The objective here is to demonstrate the operational viability of this method in a lab setting with typical healthy subjects, laying the groundwork for future implementations with real patient populations. A thorough literature review located in the subsequent chapter outlines all necessary concepts in detail regarding EEG systems, BCI configurations and the P300 waveform. This leads onto a discussion of similar, yet distinct prior implementations of image-based BCI systems in order to orient the worked defined here (see, Chapters 3, 4 & 5) within the broader literature.

1.5 Systematic Optimization of SSVEP-based CNN Classification Rationale

As noted above, the vast majority of communication-based BCI systems focus on the use of alphanumeric target arrays. In cases where patients retain the linguistic and cognitive capacity to utilize large alphanumeric arrays, the steady-state visual evoked potential (SSVEP) waveform is frequently employed as a control signal [45]. SSVEP is an oscillatory signal primarily propagated over the occipital lobe following the fixation on a stimulus flickering at frequencies between 6 and 15 Hz [46]. These systems are among the highest-performing BCIs developed to date, capable of supporting up to 160 targets [47] with information transfer rates (ITR) exceeding 100 bits per minute (bpm) [48–51], 200bpm [52–54] and in some case over 300bpm [53, 55].

Over the past decade, iterative improvements to the widely implemented classification method, Filter-Bank Canonical Correlation Analysis (FBCCA), along with advancements in stimulus design and pre-processing techniques, have led to substantial performance gains. Traditionally, a subject's data was correlated solely with perfect sinusoids at the stimulus target flickering rates. However, advanced methods now integrate subject data into the reference filter-bank comparisons and modify stimulus presentation methods to increase signal separability, such as through phase offsetting, leading to significant increases in system accuracy and Information Transfer Rates (ITRs).

Arguably, one of the most impactful advancements has been incorporating the signal harmonics of the reference sinusoids into the comparative Filter-Bank stages [56]. This involves applying multiple bandpass filtering stages to the user's EEG data to isolate both the base stimulus frequencies and their harmonics in the upper-frequency ranges, thereby enhancing classification accuracy [48]. For instance, instead of merely correlating an input signal containing noisy EEG data around 6 Hz against perfect sinusoids at 6, 7, and 8 Hz, the signal is filtered to analyze the expression of all relevant harmonics, such as 6, 12, 18 and 24 Hz, as well as the harmonics of 7 Hz (e.g., 14, 21 and 28 Hz). This significantly boosts classification accuracy as the relative distance between target waveform harmonics increases with each order .

These advanced techniques have also been adopted by researchers using convolutional neural networks (CNNs) to achieve similar outcomes. By expanding the pre-processing filter range to approximately 0.1-80 Hz [54, 57], compared to the original range of around 1-30 Hz [58–61], both the principal and harmonic features of the SSVEP signal can be captured. This broader range has been successful in enhancing the performance of CNN-based BCI systems, however, the relationship between CNN model design and the optimal pre-processing parameters for data filtering has yet to be systematically quantified, indicating an opportunity for further research in this area.

Objective Here the author presents a programmatic system developed in Python for optimizing signal pre-processing filter cutoff frequencies across a range of established CNN classifiers in SSVEP-based BCI spelling applications. The methods outlined here can be broadly adapted for the large-scale optimization of any conceivable pre-processing parameter stage. This work focuses on the critical step of determining data filter ranges, given the significant benefits of including higher frequency ranges to capture harmonic components. The findings are intended not for direct reimplementation but to serve as a guide for conserving resources by providing data-driven estimations for setting pre-processing parameters based on the configuration of a given network. A significant expansion on all topics noted here is positioned in the following literature review chapter.

1.6 PhD Outline and Objectives

In an effort to comprehensively present all related concepts, this thesis begins with a thorough literature review. It then details a series of experiments (Experiments 1, 2, and 3) focused on developing a P300-based emoji speller for emotion communication, exploring various pre-processing methods, stimulus designs, and analytical approaches for P300 evaluation. Following this, the thesis addresses the optimization of convolutional neural network (CNN) classifiers for Steady-State Visual Evoked Potential (SSVEP)-based BCI systems. The final chapter summarizes the research findings and their implications, with additional details and supplementary information provided in the Appendix.

Chapter 2

Literature Review

2.1 Chapter Outline

This literature review comprehensively covers all aspects of the project, starting with detailed discussions of EEG technical hardware and the bio-signals employed, specifically the P300 and SSVEP. The review delves into the foundational concepts and nuances of Brain-Computer Interface (BCI) technology, providing a thorough examination of its technical features and advancements. A critical analysis of existing emoji and pictographic BCI spellers developed to date is also included, offering a historical and technical perspective on how these systems have evolved. Furthermore, this chapter presents an in-depth exploration of Convolutional Neural Networks (CNNs) and models that are directly relevant to the research conducted in this thesis. The literature review is designed to serve as a reference point, offering clarification and insight into the technical and theoretical aspects discussed throughout.

2.2 EEG Signals and Hardware Overview

Before any broader discussions on BCI and more specifically EEG-based BCI research, it is first necessary to outline the technical definitions of EEG and provide a brief history of the hardware and bio-signal acquisition device more generally. The aforementioned EEG systems typically feature arrangements of electrodes positioned over the skull according to the standardized international 10-20 system. These electrodes register differences in micro-voltage (μV) across the lateral surface of brain tissues [62]. At the single-cellular level, neurones undergoing an action potential exhibit a movement of charged particles over the respective axon leading to the generation of a primary electrical current and corresponding magnetic field. Further, the collective action potential of multiple neighbouring neurones can then com-

pound into larger magnetic fields. These so-called 'primary currents' are utilized in Magnetoencephaolgraphy (MEG) applications and include functional connectivity research [63], pre-surgical mapping [64–66] and post-surgical evaluation [67, 68].

Following the initial excitation event, a secondary magnetic field is propagated to balance the variance in electrical potential over the neurone, accordingly, these are termed secondary currents. The interplay between these fields introduces variance in μV amplitudes across the scalp and is registered by EEG-based data acquisition systems. Numerous cognitive and behavioural phenomena have been robustly associated with distinct and replicable changes in μV patterns (for review see, [69]). Typically, these are characterized by an increase in related μV amplitude (Event-Related Potentials) or a change in the prevalence of specific signal frequencies. Invasive EEG methods are relatively prevalent in surrounding literature, this involves the positioning of electrodes directly on (depth electrodes) or near (sub-dural electrodes) target brain regions. Further, so-called sub-dermal corkscrew electrodes can also be embedded directly into the scalp [70]. These methods are utilized nearly exclusively in surgical settings for obvious reasons related to patient well-being, setup times and staff training constraints.

The most common format of EEG-based devices is the non-invasive variant in which electrodes, housed in a flexible cap, are positioned against the scalp. As the electrodes in these systems are significantly further from the brain tissues compared to the invasive methods, numerous guidelines are observed to ensure the collection of high-quality signals. The most common measurement for quality control relating to EEG signal acquisition is the impedance (Ω) level. This describes the ease with which an alternating current can move over a conductive surface. Note, that impedance contrasts with resistance, as this describes an analogous process relating to direct current. In the simplest terms, the lower the level of impedance, the greater the EEG signal quality. Note, that failure to address high impedance levels before data collection has been shown to markedly influence both signal-to-noise ratios [71, 72] and signal coherence [73].

2.2.1 Wet *vs.* Dry EEG

The vast majority of EEG systems available to clinicians, researchers and consumers can be categorized as either wet or dry devices. The so-called wet systems involve the application of a conductive gel to each electrode in the array assembly. These gels (hydrophilic polymers) contain ionic compounds, typically salts, to enhance the signal-to-noise ratio of brain-based bio-signals to the electrodes [74]. The process can be relatively time-consuming and necessitates the provision of subject hair cleansing facilities that can dramatically extend experimental

testing times and place time constraints on the duration of wet EEG data capture. Further, operators must be precise in the application of these gels to avoid electrode coupling. This occurs in instances where conductive gel traces from neighbouring electrode sites meet and the μV amplitude readings from the affected sensors become confounded due to cross-interference. This is especially problematic for ambulatory assessments or experiments involving even moderate physical movement from test subjects. Despite these drawbacks, wet EEG systems are positioned as the gold standard in non-invasive EEG research as these methods achieve the highest resolution signals since impedance values below 5kΩ are easily achievable. Note, that the 5kΩ threshold is a well-established industry standard in non-invasive EEG research [75, 76].

In contrast, dry systems forgo the application of conductive gels and alternatively opt for the use of electrodes with electro-conductive coatings with active noise-shielding properties. Further, dry-EEG sensors are typically engineered to maximise scalp contact via the use of so-called feet-style probes to parse through individual subject hair follicles. Additionally, as these sensors are typically embedded into plastic housings, as opposed to soft caps, spring and foam-based pressure modules are often affixed to the sensor assembly to modify the fit of the device according to the shape of each subject's skull. These devices have been shown to produce EEG signals with extremely high correlations to those collected via wet system setups [77]. Despite this, the ability to achieve sub-5kΩ signals is markedly reduced. This is somewhat offset by significantly higher deployability, further advantages include greater wearability, higher subject recruitment potential, a reduced training curve, enhanced portability and increased accommodation for ambulatory testing.

Note, that in the past the benefits of dry-electrode rapid deployability were negated by low user comfort ratings due to the initial commonly adopted pin-style design. Previous research comparing the performance and usability of a range of dry and wet electrode types notes that the active dry gold pin-based electrodes produced by BrainProducts Gmbh were ranked as significantly less comfortable by users [78]. These evaluations were conducted against other user reports featuring the standard flat ring-shaped golden electrodes (EasyCap GmbH) used for wet-EEG, hybrid dry multi-spikes (Quasar Inc.) and passive dry solid gel (BrainProducts Gmbh) electrodes. Crucially, there was no significant difference in the perceived comfort between the latter three electrode variants, with the dry electrode systems retaining the advantage in terms of a lower setup time.

2.3 EEG *vs*. Alternative Brain-based Bio-Signal Acquisition Methods

When comparing non-invasive EEG systems to other methods of brain-based bio-signal acquisition such as functional magnetic resonance imaging (fMRI) it is clear these devices demonstrate significantly reduced spatial resolution. This in combination with the low penetration of the systems (1-2cm) means that signals are poorly localized and research into sub-thalamic structures is unviable. These issues are principally related to the positioning of the sensors on the scalp. Broadly, the interference resulting from both the skull and cerebrospinal fluid ultimately places significant limitations on the spatial resolution of any non-invasive EEG system [79]. Further, the chaotic trajectory that characterizes the EEG signal introduces additional complexity as numerous factors including, electrical noise, subject attentiveness and skin conductivity can lead to minute-by-minute changes in waveform quality. These can obfuscate researcher efforts to consistently replicate past results and divergent outcomes are common within the same subject even for the same experimental session [80–82].

The highly dimensional attributes of the data paired with the expression of numerous non-linear components place significant demands on experimenters and the techniques implemented to extract relevant neural signatures. Despite this, EEG systems feature some of the highest temporal resolution performance statistics of any bio-signal acquisition device class. These qualities arguably provide EEG-based methods with the highest potential for BCI applications owing to the significantly higher upper estimates of resultant information transfer rates. Additionally, EEG-based bio-signals present with the largest number of ultra-low latency waveforms, with the SSVEP requiring just 80-160ms for expression over the visual and associated cortices [83–85]. This low-latency expression paired with high temporal resolution provides a compelling case for the adoption of EEG systems as the primary BCI data acquisition platform.

Further, fMRI systems are typically housed in either hospital or university campuses and require a complement of professionals to operate and maintain, making their deployment in BCI applications unviable. The less invasive EEG methods, as noted above, provide additional obstacles such as clinician costs, surgical risk, added hygiene considerations, lower patient quality of life and dramatically diminished repeatability [86, 87]. This final consideration is key as the long-term positioning of recording equipment in or on target brain tissues invariably leads to scar tissue formation and loss of effective neural signal acquisition. Moreover, previous research involving the consultation of prospective BCI end-point users found only 1 of 17

subjects professed an interest in any form of BCI device requiring a sub-dural implant. This is likely owing to previous trauma associated with post-surgical recovery and the perceived risk of clinical complications worsening or accelerating their current conditions [88]. In summary, these considerations taken together provide a strong case for the pursuit of non-invasive EEG-based BCI research over alternative data acquisition methods.

2.4 Brain-Computer Interface Definition and Terminolgy

Brain-Computer Interfaces (BCIs) are defined as assistive devices utilizing control signals localized to the brain. These systems are typically deployed to aid or replace communication or mobility functions [89–91]. Note, that these do not fall under the same classification as Neural-Computer Interfaces (NCIs), which operate via the collection of control signals from the peripheral nervous system, this includes the use of electromyography (EMG) and eye-tracking [92–94]. Further, so-called Human-Computer Interfaces (HCIs) describe a more generic terminological grouping encompassing both BCIs and NCIs as both involve the connectivity of computer systems with human users. This comprises a wide array of technological considerations such as ergonomic constraints in keyboard design, software system graphic user interface (GUI) layouts, and the positioning of sensors in wearable devices. The broader discipline of HCI has had a significant impact on the development of BCIs by providing the theoretical framework and guiding principles for future technologies based principally on the concept of functionality [95].

As noted above, BCI systems have been developed primarily to restore communication and mobility to individuals with severe paralysis. This covers wheelchair control [96], exo-skeleton commands [97], motor vehicle operation [98] as well as prosthetic hand manipulation in real-world and simulated VR contexts [99, 100]. More broadly, specialized methods for environmental control such as lighting operation in the so-called internet-of-things (IoT) have also been explored yet lie outside the scope of this thesis [101]. Further, systems have been developed for web browsing [102], online/smartphone messaging [103, 104], video-game participation [105], musical composition [106, 107] and crucially, spelling or text entry [108]. At the base level, all BCI systems require hardware for bio-signal acquisition and a corresponding signal amplification unit.

Note that recently many systems have integrated both of these devices into one unit, especially for ambulatory applications. Further, all BCI systems implement some minimal level of signal pre-processing, followed by feature extraction and target classification. Finally, all

closed-loop methods implement a feedback system. This could be configured as a graphical interface or coded series of tones used to update the user on the command issued by the BCI hardware and software ensemble. This relay of information post-classification allows the user to monitor the performance of the system and crucially, implement changes in response to potential errors. Instances involving the real-time or online feedback of this information are the gold standard in BCI research as these typically represent methods with the highest ecological validity.

The most prominent means of evaluating BCI-based system performance relate to the metrics Accuracy of Classification (AoC) and information transfer rate (ITR). The former defines the hit rate of a given classification system and describes the relative number of correct predictions against the number of errors. This is typically expressed either as a percentage between 0 and 100% or in non-integer format between 0 and 1. The former is used to compare BCI systems in terms of speed and is described in the equation positioned below (see, Equation 1). Initially, this statistic was computed to evaluate telecoms-based systems and was adopted at a later stage to determine BCI system performance [109]. Here N corresponds to the total number of targets, in a BCI speller context this relates to the amount of numbers, letters or characters on screen. Further, P refers to the probability of accurate target classification and T denotes the data capture duration period per target. Note, that this algorithmic configuration relates to the computation of offline BCI performance. The addition of all the time needed to calculate the prediction and adequately reset for the next trial must be accommodated to produce valid online performance approximations. Note, that Equation 2 represents the ITR metric in a per-minute format, this is often transposed into bits per second. The studies discussed herein all utilise a bpm presentation format.

Equation 1.

$$B = log2(N) + P * log2(P) + (1-P) * log2((1-P)/(N-1))$$

Equation 2.

$$bpm = B * (60/T)$$

The analysis of any BCI system must consider both AoC and ITR metrics as a relative imbalance can lead to significant decreases in real-world functionality. For example, high-accuracy and low-ITR BCI systems are impractical owing to latency and responsiveness issues. Further, low-accuracy and high-ITR systems suffer from the need to edit or amend output commands, repeatedly stunting the user during live operation. It is clear, that balance

is needed to instil user confidence in the system and foster its adoption for communication or mobility purposes. In previous research, additional measures such as Average Precision (AP) are used to indicate the relative positioning or pattern of classifications. This is strongly related to the phenomenon of overfitting, which describes the biased selection of a target or specific prediction directionality. In these instances, the classifier could be defined as precise as the groupings of predictions made are highly consistent, irrespective of the actual accuracy metrics. Recently, discussions surrounding classifier precision have been superseded by descriptions of the incidence of confusion, as relating to confusion matrices. This effectively describes the same process from the orthogonal conceptual standpoint (for further information see, Figure 3.7). In sum, researchers aim to produce robust classifiers demonstrating high accuracy and precision in tandem with high ITR (bpm) metrics.

2.4.1 BCI Sub-Classes

The most basic distinctions in BCI classifications relate to the locus of the control signal as originating either exogenously or endogenously, in other words in response to an external stimulus or internally generated commands respectively. More specifically, an exogenous BCI requires the use of an external device to excite the corresponding brain-based bio-signal, such as a computer monitor or series of tones. These bio-signals are propagated reflexively, that is to say, without user intention. The only volitional control for the use of such systems is to direct gaze or attention over the intended target stimulus. Along these very same lines, these systems are characterised via the use of so-called bottom-up dynamics as in the case of for example SSVEP waveforms. Conversely, endogenous BCI functions according to top-down control signals. These involve directed, conscious efforts from the user that require no external stimuli and cover systems that operate via imagined motor movements. In these contexts, selection commands are assigned to user limbs and even individual constituent limb structures, for example, hands and whole arm movement.

Often in mobility-based prostheses, there exists a one-to-one mapping between the limb and the desired movement vector, in other instances these are mapped to more generic mobility devices or communication system commands [94, 110]. The systems utilizing these endogenous control features arguably possess greater scope for application as no functional control over any muscle groups is required to operate the systems. In contrast, these methods suffer from the limited number of mappable limb areas, high latencies in signal propagation, a steep user learning curve and a high rate of fatigue induction as opposed to alternative exogenous-based BCI platforms. Owing to these considerations such methods are typically reserved for patient populations with the most severe forms of paralysis. Note, that many such systems are now

in development implementing hybrid control designs. These typically allocate endogenous systems to engage and terminate more elaborate exogenous interfaces. This integration allows viable user populations to benefit from the volitional control features of endogenous systems and the added functionality of exogenous-based platforms [111].

Additional terminological definitions exist in defining the numerous approaches researchers have undertaken in developing BCI systems concerning so-called Active and Reactive BCI, both of which map closely onto the aforementioned endogenous and exogenous BCI classifications. In this instance, the distinctions relate purely to the presence or absence of user intention, as compared to the previous classifications that focus solely on the location of the casual event linked to a given bio-signal propagation [112]. Along these lines, active BCI utilizes a signal that is the result of direct, volitional control. In comparison, reactive systems, utilize waveforms such as the P300 and SSVEP that are propagated exclusively via exposure to external stimuli and can not be activated dynamically. As these definitions map so closely the terms are often utilized interchangeably. A further classification relates to Passive BCI, this involves removing any component of directed user intention and operates similarly to a signal monitoring system. These applications can include emotional state tracking or attentiveness controls to assist in the optimal operation of certain technological platforms.

Finally, definitions relating to the specifics of the control system are important to note. Some platforms operate according to predesignated time-locked blocks, known as synchronous BCI [113]. This is done to standardize the inputs for the classification system and these rigid parameters allow for greater control over the behaviour of a given BCI system. Conversely, asynchronous systems place the locus of control closer to the user, as the operational durations of the system are more fluid. In real-world terms, a synchronous BCI speller could be programmed to collect a pre-coded number of target letters, numbers and characters over the course of a specified number of trial runs. This operationalization typically comes at the cost of relinquishing the ability to begin each spelling run to clinicians, as opposed to the user. In an asynchronous format, the ability to engage and disengage the spelling protocol would be controlled via the user and arguably provides for greater independence and higher potential quality of life outcomes [114]. Despite this, the ability to keep these systems running continuously in the background during periods in which the user does not wish to operate the system is problematic owing to system setup times and potential signal degradation issues. Ultimately, platforms intended for use as asynchronous BCI have similar functional outcomes and can complicate the process for researchers in terms of development. Again, the synchronous methods typically map alongside reactive or exogenous systems, with asynchronous

methods reserved primarily for bio-signals that can be propagated endogenously [115]. Note, that all methods described in this thesis are classified as exogenous, active and asynchronous, EEG-based BCI spellers.

2.5 BCI Configuration and Control Signals

The EEG data acquisition format was selected for the research herein owing to the relatively economical hardware costs, non-invasive methodology and ultra-high temporal resolution [46]. Past research has demonstrated repeatedly that EEG-based methods can perform well in clinical spelling settings and these systems arguably display the most promise due to the potential advances in signal processing and classification techniques currently being explored [116, 117]. Crucially, the control signal and corresponding acquisition device selections are highly dependent on patient user requirements. The authors herein assert that the EEG platform and the related SSVEP and P300 control signals provide the best solution for the target patient population and corresponding functional task: communication via BCI Speller. The field of BCI Speller development aims to enhance or restore communication-based functions for individuals with severe paralysis. This covers the transfer of information in numerous formats including text, numbers, symbols, characters, pictorial icons and binary selection processes (for example yes *vs.* no paradigms). In the past simple alphabet, boards have been replaced with complex branched hex-layouts [91] and currently fully-complimented keyboards are used, traditionally presented via a computer monitor and visually augmented to generate corresponding control signals (exogneous, reactive). The following subsections outline the two most popular control signals available, P300 (2.5.1) and SSVEP (2.6.1) waveforms.

2.5.1 P300

The P300 waveform has been utilized widely in neuroscience literature for numerous BCI and non-BCI-related applications. These include neurological diagnosis, attention and intention research, facial recognition and expression studies, cognitive workload, treatment viability assessments and many more. The signal is classified as an Event-Related Potential (ERP), this describes a waveform propagated as a positive or negative deflection that is temporally locked to an external stimulating event (see, Figure 2.1). Typically, the P300 is characterized by the expression of a large positive deflection 300ms following the onset of an experimental event [118, 119], resulting in a μV amplitude change of around 5-20μV [120]. Note, that there exist numerous sub-components related to the P300 including the N50 and N100. Additionally, the P300 can be sub-categorized within the component into an earlier, P3a, and later P3b wave-

form sub-class [121].

The initial P3a component is characterized by a larger relative increase in μV amplitude between 250-280ms, as compared to the P3b (300-400ms) [122]. The early P3a has been shown to primarily respond in instances of unanticipated modification to experimental stimuli [123], task conditions involving the subject actively ignoring the experimental stimuli [124] and the unexpected early stopping of a task [121]. In contrast, the P3b expression is task-based and relates primarily to the recognition of an unlikely yet expected modification of target stimuli. The study of these sub-components, associated responsiveness to task conditions and the temporal relationships between the waveforms have both informed and corrected numerous theories of working memory, for example, the context-updating theory [122].

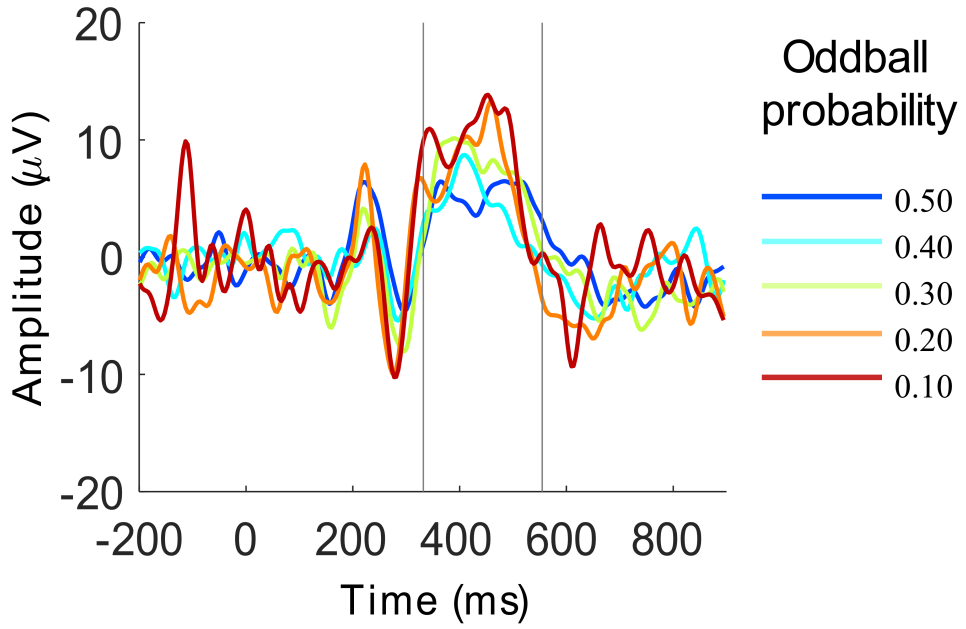

Figure 2.1: The plot shown here is reproduced from [125]. Here, oddball probability is used to explore how memory capacity affects neural responses to infrequent auditory stimuli. This approach involves manipulating learning tasks and oddball probabilities to provide valuable insights into memory function and its variability by analyzing how the brain's response to unexpected events reflects underlying memory processes. This oddball experimental implementation involved subjects listening to a sequence of tones, consisting of frequent low-tone standards and infrequent high-tone oddballs. The subjects form an internal memory of past tones, which they use to predict future tones and respond accordingly. The plot shows the average EEG trace for oddball trials across different blocks, with each block corresponding to a specific oddball probability (OP) indicated in the legend. The traces are colour-coded to reflect these probabilities, demonstrating how the average response to oddball tones varies with the frequency of their occurrence. Note that the image shown here is published under a Creative Commons Attribution 4.0 International License. To view a copy of this license, please visit: https://creativecommons.org/licenses/by/4.0/.

Interestingly, the expression of these waveforms can vary across subjects in terms of precise onset, typically falling within the 250-400ms range [126]. The signals originate from the central Fpz and Cz electrode locations and travel anteriorly towards the parietal regions [127]. Initially, auditory-based stimuli formed the basis of most early investigations into the P300 waveform [118] and have since expanded into the visual sensory modality. Visual methods for inducing P300 responses can include the manipulation of stimuli via flashing, rotation, colour inversion, resizing, flipping and highlighting. Essentially, any means of increasing the relative salience of a given target as compared to previously displayed neighbouring targets are likely viable candidates for the reliable propagation of the P300 waveform. Crucially, the waveform is exogenously generated and also highly dependent on top-down mechanisms including fatigue, attention and stimulus probability estimation. Further, given these top-down effects, the signal presents with a relatively high latency refractory response before returning to baseline. Additionally, the repeated stimulation of the P300 waveform leads to ever-decreasing subsequent peaks, leading to trials in the later experimental stages as demonstrating reduced signal quality [128]. To overcome these issues hybrid methods implementing an array of augmentation techniques into the same experimental session can be utilized to mitigate these issues.

The utilization of the P300 waveform has been encouraged due to the robust, replicable nature of the signal and critically, owing to the gaze independence of the visually-based ERP variant. In other words, the propagation of the P300 waveform for visual stimuli does not require direct fixation of a given target stimulus. In the past, it has been frequently asserted that subjects need only attend the region in which the visual stimuli are positioned to generate the waveform. Note, that some literature casts doubt on the degree of gaze independence expressed by the P300 [129, 130]. Both referenced studies found that P300 amplitude decreased as a function of fixation distance from the attended target.

The reduced expressivity of the waveform translated into lower classification accuracies in a P300-based speller experiment and suggests a classification of semi-gaze independence is more applicable under these conditions [129]. Despite these caveats, P300-based systems arguably possess greater scope for application, especially in patient populations with sub-optimal ocular control for whom eye-tracker or SSVEP-based systems are unviable. Notably, the P300 and associated sub-components (N100) demonstrate reduced expressivity in amyotrophic-lateral sclerosis patients, as compared to healthy age-matched controls. Crucially, these lower amplitudes did not translate into significantly different functional performance in terms of classification accuracy or information transfer rate [131]. Further, the P300 waveform has been successfully implemented in numerous P300-based systems for both

healthy, typical [132, 133] and crucially, real-world, clinical settings [134, 135] (for review see, [136]).

One of the most widely implemented experimental formats utilized to illicit the P300 is the oddball paradigm. In the most simplistic configuration, a singular target is presented to the subjects, often via a computer monitor. The stimulus is coded to present either a common, standard target, for example, a white square and in so-called deviant trials a non-standard stimulus is presented, for example, a black square. Over the course of the trial, the order of presentation for standard or deviant trials can be hard-coded or operated according to a probability function. Note, that the ratio of standard to deviant trials must be biased towards the presentation of the standard target stimulus or ensure the subjects perceive the deviant trial presentations as unexpected events. Following the deviant trial, the P300 waveform propagation is induced and this enables researchers to determine the classification of a given deviant trial independently of the exact stimulus presentation sequence.

The P300 oddball response has been widely utilized in various research contexts, including memory studies, to investigate how the brain processes rare or unexpected stimuli. This research approach involves presenting a frequent standard stimulus interspersed with occasional oddball stimuli that deviate from the norm. Notably, the P300 peak is inversely related to the probability of presentation or the relative degree of surprise associated with a stimulus. This relationship has been documented [137] and highlights how the magnitude of the P300 response decreases as the probability of an event increases, indicating a diminished surprise effect. As seen in [138] a large drop in the P300 area under the curve, a reduced prevalence of associated N200 waveform components and higher component latencies were observed for a checkerboard stimulus layout where target oddball probability increased from 20% to 50% to 80%.

It is crucial to note that the peaking features of the associated signals are reduced and not completely absent from the 50% oddball probability trials, much in the same way these attributes are retained yet diminished in the above Figure (see, Figure 2.1) as shown in [125]. This is likely because the P300 response is not solely dependent on the precise probability of stimulus occurrence but also on the subjective probability perceived by the participant. Previous research emphasises that the P300 response reflects not only the objective rarity of a stimulus but also how surprising it is to the individual, based on their expectations and learning experiences [139]. Further, it has been demonstrated that simply decreasing the probability of the target stimulus by increasing the number of non-target stimuli can have adverse effects

on the P300 response. As discussed in [19, 20, 128, 140–142], task difficulty, probability, and inter-stimulus intervals all impact the P300 response, indicating that an excessive increase in non-target stimuli might negatively affect the quality of the neural response, potentially leading to reduced sensitivity and reliability in detecting the oddball stimuli.

It is important to highlight that the experimental variants can be arranged to manipulate the stimulus intensity or introduce a reward-based feedback mechanism to modify the P300 waveform expression alongside deviant trial frequency ratios to influence the resultant ERP properties (phase and amplitude). Note, that the traditional oddball paradigm format demonstrates a greater incidence of P3b propagation [143]. Along these very same lines, all subsequent P300-based BCI operating according to the oddball design methodology are effectively targeting this later P300 sub-component. Note, that the excessive expression of P3a components could indicate poor subject instruction.

The properties of the P300 noted above led to the use of this waveform in the world's first BCI implementation [144]. Namely, the ability to reproduce the signal in a wide range of the general public [145] and the relative simplicity of experimental design needed for elicitation have lead to the numerous research groups to investigate this signal in a BCI context. Since the publication of this seminal work, the P300 has been widely deployed as a control signal in multiple BCI studies across the globe and innumerable contexts spanning communication and mobility applications. Concerning the BCI speller literature, P300 research greatly influenced the advancement and standardization of graphical user interfaces for the selection of letters, numbers and symbols. These initial systems typically functioned via the presentation of 6 $\times$ 6 target character matrices. Each row and column is pre-coded to generate a relative deviant augmentation event (for example colour inversion) over the course of the trial. At a minimum, each trial involves the augmentation of all stimulus targets twice, once in the corresponding row and once for the corresponding column. Note, that each row and column augmentation is done individually, often featuring a delay between the execution of subsequent sequences.

For a given attended target, one trial should feature two corresponding P3b waveforms. The researchers can cross-reference the pair of temporal deflections and predict the onset of the initial triggering event. Using these two data points it is possible to deduce the target intended for spelling as each character possesses a unique combination of time-locked augmentations. There exist modifications to this base-level paradigm that integrate multiple trials per letter spelt to boost classification accuracy. Note, that the increase in signal quality via averaging leads to additional processing time and a faster rate of P300 waveform amplitude degradation.

This implementation is often restricted via experimental features that reduce the accuracy of the system. The double flash problem describes the common incidence of the same target undergoing individual row and column augmentations in succession. This can significantly reduce the amplitude of the P300 in response to the second augmentation event as the related cerebral regions and corresponding signals have not fully reset to baseline. In other words, the intended re-induction of the P300 waveform occurs before the completion of the refractory period and can significantly attenuate the respective signal expression.

Additionally, the row and column augmentation scheme must be randomised to avoid subject memorisation as the P3b functions via the presence of an unlikely, yet expected change in task-relevant stimuli. Along these lines, any knowledge of the augmentation order before the trial initiation would render the trials intended as deviant to be predictable. This need to consistently modify the order of row and column augmentations can lead to the use of presentation schemes that position stimuli closely in temporal space. Further, after repeated propagation of the P300 waveform, the latency of the signal reduces and can drift significantly. This complicates the process of identifying the correct time-locked deflection and in turn, the intended character for selection. These experimental obstacles are termed adjacency errors and also apply in the spatial domain. This relates to the unintentional direction of attention and or fixation of characters neighbouring the target stimuli. Researchers can mitigate these issues via the use of algorithmic methods to ensure for example that the augmentation of the first row is not immediately followed via the augmentation of the second row. Despite, this only a limited number of row and column augmentation schemes are available under these constraints, in response to these complications alternative methods of stimulus presentation have been explored.

Arguably the most effective means of visual-P300 generation for BCI-spelling developed to date is the asynchronous paradigm [146]. This involves assigning targets with unique augmentation sequences, independent of row or column positioning. For example, a hypothetical trial period could consist of 6 time-locked augmentation events. Distally positioned (spatial domain) targets are grouped, e.g. A, Q and 1, together to augment during the 1st sequence. Further, A, G and 9 are grouped to augment in the 2nd sequence. Essentially, the arrangement of rows and columns is replaced with spatially optimized groupings to maximise the individual target signal profiles over the course of the trial. These techniques dramatically reduce the incidence of both double flash and adjacency errors and also require significantly fewer augmentations per character selected. Further, studies demonstrate that subject-specific attributes can be used to modify the parameters of the stimuli to boost performance. This includes tuning

intra-trial intervals between the onset of subsequent trials, reducing the number of trials per letter spelt, and tailoring the degree of either temporal or spatial separation of augmentations. Moreover, these systems have been integrated into an automated update scheme and change in response to user performance in real-time to further boost speller usability [146].

The utilization of these methods has shown real-world online performance in healthy, typical subjects exceeding 95% classification accuracy at between 94 and 120 bpm for a 72-target (8 $\times$ 9) matrix composed of letters, numbers, symbols and interface controls [146]. Note, that the theoretical maximum throughput of this system is 258 bpm, with some subjects demonstrating zero errors over the course of dozens of characters spelled. These results suggest that the relatively low latency of the P300 waveform (300ms) can be offset via extensive optimization of the stimulus augmentation protocols. These impressive performance metrics are attained via the systematic development of both stimulus design methods and analysis techniques. The bleeding edge of P300-based BCI speller research at present nearly exclusively deploys Step Wise Discriminant Analysis (SWLDA) for the classification of the respective bio-signals. This method is an extension of Fisher's Discriminant Analysis (FDA) and it involves a binary classification, determining if data does or does not contain a P300 by optimizing a discriminant function [147].

In terms of the Row/ Column paradigm, feature vectors are constructed from segments of EEG time series parsed into around 500ms chunks after the initiation of each flash event. Training data are concatenated across channels, to form a temporal image of brain activity across all electrode locations sampled. The discriminant function is optimized with a subset of sample data to identify features relevant to discerning between classes. In other words, data features which contribute to the maximal separation of classes are added to the discriminant function. Further, during this process features calculated as the least significant for the task of class prediction are continuously removed. SWDLA is a constrained and efficient, non-exhaustive search method. These models can over-fit and even demonstrate convergence failure (inability to reach maximal accuracy) if the data used contain significant artefacts, or, do not possess an even distribution of class examples.

The technique has been shown to outperform several alternative analyses including peak picking, covariance and area assessment [144]. Improvements in hardware, operating system efficiency, stimulus display techniques and averaging in pre-processing, have resulted in researchers continually enhancing the performance of P300-based BCI systems [148]. Adaptations to the SWDLA include the addition of a regularization parameter which updates the

discriminant function in response to misclassification. This is intended to reduce the influence of abnormal data and enhance the analyses' generalizability across subjects [149, 150]. SWLDA has also been shown to outperform novel analytical techniques such as support vector machines [147] and multi-layer perceptrons [151, 152].

In consideration of the past research discussed herein, a variant of the SWLDA method is deployed for the classification of the P300 waveform in a simplified emoji-based BCI-emotional communication context (see Chapters 3, 4, & 5). Recent studies demonstrate that the use of face-stimuli overlaying [153], emotional stimuli [154] and the application of colour to stimulus targets [155] can all assist in boosting P300 waveform amplitudes in the short term, ensuring the retention of viable waveforms for detection and classification over the course of protracted recording sessions. Along these very same lines, the efficacy of emoji-based stimuli is assessed for viability to inform future speller applications featuring integrated character-emoji keyboard configurations.

2.6 Emoji-based Speller Designs

At the time of the research conducted herein, emoji-based speller targets had yet to be implemented into a pure emotional expressivity BCI communication platform. The research conducted thus far related specifically to the usage of emoji stimuli as augmentation overlay objects. Note here, that an augmentation overlay refers to the programmatic, time-locked occlusion of a stimulus. The term augmentation relates to the transformation of visual stimulus properties and overlay relates to the quality of this transformation, as characterised by the occlusion of said stimulus. This method, as dictated by its perceptual qualities, is restricted to the visual domain. In the context of a classic visual-P300 BCI speller, the augmentation overlay would be the white square that is overlayed onto the spatial position of given speller targets to induce the evoked potential.

The studies in question probed the efficacy of replacing overlay squares with emoji stimuli in P300 speller contexts and demonstrated significant increases in classifier accuracy [156], with future studies successfully applying these findings in alternative BCI tasks [157]. These results align with studies demonstrating a similar effect utilizing human facial images as overlay objects [153]. It has been suggested that these increases in performance are likely owing to a boost in P300 peaking fostered by the usage of stimulus augmentation objects with a higher level of emotionality and salience [158].

Arguably [44] demonstrates the first use of an emoji integrated array in a P300 speller context. The study made use of a row/column paradigm featuring a 3 $\times$ 4 emoji-icon integrated array composed of 4 face-based emoji and 8 non-face-based icons consisting of pictograms relating to user states, for example cold, warmth, hunger, and thirst. The paper describes the implementation of a deep convolutional neural network (DCNN) (see subsections 2.6.3 & 2.6.4) trained to classify images of the EEG waveforms captured. The methods revealed high classification accuracies for both targets (90%) and non-targets (95%), significantly outperforming both the LDA and Logistic Regression analysis evaluated.

Despite these impressive results, the author of this thesis asserts that the arrangement of these facial emoji stimuli in the emoji-icon integrated array is suboptimal, and the range of emotional expressivity is highly limited. In contrast, previous research into the measurement of human emotions has established the benefits of arranging scales across the continuous dimensions of arousal (low to high) and pleasure (low to high) [159]. Along these very same line, the author of this thesis aims to investigate the efficacy of an emoji-based emotional communication platform utilizing a reduced, one-dimensional pleasure valence scale. Note, that throughout the following text, the term pleasure and affective scales are used interchangeably. Further, it must be stated that the author of this thesis has decided to focus solely on a single continuous dimensional scale, affect, as opposed to both the affective and arousal scales to initially determine the operational viability of this method. It is asserted that even in this reduced format, the implementation of a scale is more conducive to effective emotional communication as compared to the randomised grid format employed in [44]. Future adaptations of this paradigm could feature a hybrid series of scales for increased clarity in emotional expression, this is discussed further in Conclusion chapter, subsection 7.2.1: Future Research.

2.6.1 Steady-State Visually Evoked Potentials

The literature surrounding current exogenous-based BCI paradigms also heavily features the application of the Steady State Visual Evoked Potential (SSVEP) as a brain-based control signal. This describes an oscillatory waveform propagated over the surface of occipital and parietal lobes [46, 160]. The signal is detectable as a periodic, phase-locked waveform in response to a 5 Hz stimulus, increasing in amplitude up to 15 Hz and degrading in quality following a non-linear trajectory up to 60 Hz [161]. Owing to these bio-physical constraints, the optimal and commonly used range of SSVEP stimulus frequencies lies between 8 and 15 Hz. These waveforms differ substantively from comparative ERP-based signals such as the P300 waveform, as the SSVEP is purely bottom-up. The signals can be elicited by fixation or attending a flickering visual stimulus. In this instance, flickering refers to the rapid presen-

tation and removal of, for example, a white overlay square or the cycling of stimulus colour inversions. The rate of flicker in the target stimulus is mirrored in the oscillations observed over the aforementioned brain regions. This enables researchers and clinicians to deduce the stimuli fixated or attended by a given subject through the pre-processing and classification of data acquired from relevant target regions.

The latency of the SSVEP components (80-160ms) is extremely low in comparison to alternative brain-based bio-signals [83, 85]. This suggests the theoretical upper information transfer limit of SSVEP-based systems is significantly higher than those utilizing Sensorimotor Rhythms (SMR) (2000-40000 ms) [110] or P300 waveforms (300ms). Further, the signal presents with a short refractory period and robust signal propagation following repeated elicitation. Previous research across a large subject pool (53 participants), shows that the SSVEP can be repeatedly elicited for a 4-target array (8-15 Hz target stimuli) in 95.5% of experimental trials [162]. Along these very same lines, similar investigations revealed just 89% of subjects achieved over 80% classification accuracies for a P300-based experimental setup [145]. These figures dropped significantly for additional evaluations utilizing SMR-based systems that showed just 19% of subjects could reach the same accuracy threshold [163]. In sum, these SSVEP properties, in concert with the primarily bottom-up propagation pathway, ensure that subject fatigue and task learning curves are restricted principally to the stimulus presentation format, as opposed to higher attentional mechanisms. These qualities position the SSVEP waveform as a highly favoured bio-signal for utilization in BCI applications [46].

Note, that it is problematic to determine the supremacy of any of the discussed waveforms from these handful of comparative assessments. This is owing primarily to the differences in experimental design especially relating to the number of targets on screen. The availability of research directly comparing the limits of applicability for SSVEP or P300-based visual BCI spellers is highly limited and primarily restricted to the assessment of hybrid systems. These typically test the performance of sole SSVEP and P300 experimental variants against novel hybrid designs [164, 165]. Both of these papers detail significantly higher classification accuracies and corresponding ITR values at the single-subject level for the hybrid *vs.* single bio-signal based platforms. Despite this, owing to the relatively small number of subjects utilized researchers are restricted in the scope of any conclusions made.

Further, there do exist some significant obstacles unique to the SSVEP waveform that can restrict the scope of this bio-signal in BCI applications. The optimal range of frequencies for signal elicitation (8-15 Hz) possesses extensive overlap with the bandwidth of photic epileptic

seizure induction sensitivity thresholds [166]. The incidence of epilepsy is typically projected at below 1%, further, the photic variant of this condition constitutes between 15 and 20% of individuals affected. In sum, the likelihood of this impacting the end-point user population is minimal. Despite this, considerable efforts must be undertaken in the pre-screening stage of any experimental investigations before SSVEP stimulus exposure. Moreover, recent research suggests that temporally stable SSVEP frequencies could be viable for some subjects and bespoke adjustments to programmed flicker rates could be introduced to avoid low-frequency oscillations with a greater overlap in the photic epilepsy sensitivity ranges [167].

Additionally, the SSVEP has traditionally been classified as a so-called gaze-dependent waveform [168–170]. This limits the number of available BCI end-point users by restricting deployment to populations presenting with dextrous control in at least one eye. Further, it could be argued that the use of eye-tracking speller platforms is more suited to patients with this retained functionality owing to the significantly lower latency of ocular saccades, as compared to the SSVEP signal. Despite this, numerous studies show that the SSVEP would be better defined as a semi-gaze-dependent signal [171, 172]. The majority of studies involve the presentation of multiple flickering targets during concurrent EEG and eye-tracking data acquisition. Typically, subjects are instructed to either fixate and attend to a cued stimulus or gaze at a centrally positioned fixation cross while attending to a cued target. The authors of [173] found that in 71% of the trials assessed viable SSVEP waveforms were propagated during the pure attention condition. These results suggest that some top-down attentional mechanisms influence the expression of the SSVEP waveform, introducing the possibility of widening the established scope of SSVEP-based BCI applications to a larger patient pool. It must be noted that these assessments were largely conducted using ultra-low density target arrays (4 targets on screen) and do not assist researchers in gauging the viability of an exclusively attention-based SSVEP BCI speller for the high-density target matrices deployed in modern speller graphic user interfaces.

The layout of SSVEP-based BCI spellers broadly replicates the same keyboard-style arrangements described in the P300 subsection positioned above (2.5.1). Previously, the number of targets onscreen was highly limited by the narrow bandwidth of the optimal SSVEP elicitation range and the refresh rate of corresponding presentation monitors. Firstly, only 8 integer valued frequencies exist between the 8-15 Hz optimized range. Further, the frequency of a given target must be divisible by the refresh rate of the given presentation screen and produce an integer value. For instances involving non-integer values, the software controlling the respective flickering stimuli will round up or down the corresponding signal. For example, a

monitor with a 60 Hz refresh rate is updated 60 times over the course of one second. In the circumstance of being programmed to present an 11 Hz signal, the monitor is instructed to flicker every 5.45 frames (60 Hz $\div$ 11 Hz). This is not possible, as only a 5th or 6th frame is available for update. In these instances, the hardware simply rounds up, updating every 6 frames.

This results in a decoherence of the intended and actual frequencies presented to the subject. Along these very same lines, only factors of 60 Hz in the optimal range SSVEP range are viable for a 60 Hz monitor (10, 12, 15 Hz), significantly restricting the number of consistently reproducible waveforms. These issues can be somewhat alleviated via the use of ultra-high refresh rate LED-based monitors. Despite this, for a reasonable increase in the number of viable presentation frequencies a significantly higher refresh rate is demanded. For example, the lowest number containing the factors 8, 9 and 10 is 360. Owing to these issues, the cost of acquiring these hardware presentation devices becomes prohibitive and in turn, diminishes the ease of deployability in clinical and research applications.

The limitations imposed by the need for integer-valued target frequencies have since been overcome via the implementation of the approximation method [174]. This involves iteratively alternating between the presentation of two pre-coded frequencies over one refresh cycle. For example, the calibrated interleaving of a 10 Hz ($60 \div 10 = 6$ frames per second) and 12 Hz ($60 \div 12 = 5$ frames per second) signal can lead to the exogenous propagation of an 11 Hz SSVEP waveform. This enables researchers to utilize target frequencies non-divisible by the monitor refresh rate and dramatically increase the density of corresponding speller matrices. Note, that this increase in available target frequencies also leads to a drop in discriminability between the targets. In other words, more robust classifiers are needed to distinguish between 9 and 10 Hz signals, as opposed to 10 and 12 Hz signals.

These complications have been addressed via the so-called Joint Phase Frequency Method (JPFM). This involves applying a graded phase offset to target stimuli positioned proximally in frequency space. In other words, to decrease the correlation of, for example, 9 and 10 Hz flicker signals and corresponding SSVEP waveforms, the 9 Hz signal is programmed with a 0° phase angle and the respective 10 Hz signal is assigned a 180° phase angle. This shifts the initiation point of a given signal from the peak (0°) to the trough (180°) and results in a dramatic decrease in correlation between neighbouring target frequency integer values for the same time point. These stimulus developments have assisted in producing arguably the highest-performing BCI spellers to date in terms of classification accuracy, information trans-

fer rate and stimulus density [175].

2.6.2 Cutting-Edge Classification Methods for SSVEPs: Filter-Bank Canonical Correlation Analysis

Currently, the majority of the highest-performing BCI speller systems utilize variants of the Filter-Bank Canonical Correlation Analysis (FBCCA) method to decode SSVEP target waveforms (for review see, [176]). For the standard CCA method, target classification involves determining the highest correlational coefficient between a suite of reference target signals and all subsampled signals acquired in the corresponding EEG electrode data matrix [177, 178]. This was later expanded into the so-called FBCCA method by increasing the scope of reference signals and input signal pre-processing stages to include the computation of coefficients of target frequency harmonic components. In this instance, both reference sinusoids and subject data are parsed into blocks via frequency filters to isolate 2nd and 3rd-order harmonic sub-components. These adaptations increased cross-subject average ITRs from 95 bpm [179] to over 150 bpm [56]. These harmonic components are essentially reflections of a given target signal in frequency space, for example, the spectrogram of an 8 Hz signal produces relative power deflections at the corresponding 2nd, 3rd and 4th-order harmonic components for 16, 24 and 32 Hz respectively (see, Table 6.1). Note, that the intensity of each subsequent harmonic is reduced proportionally to the noise embedded in the respective multivariate signal. In other words, despite the reduction in overall signal power, the prevalence of associated noise is relative, leaving the signal-to-noise ratio within tolerable levels for effective target signal extraction [176].

Further, the blending of subject-specific signals acquired during pre-screening evaluations with the sinusoidal reference waveforms dramatically enhanced the performance of the corresponding so-called Combined CCA method [180].Briefly, in typical SSVEP stimuli paradigms, the flicker frequencies available for presentation are limited to the factors of the monitor refresh rate. For example, a 60 Hz screen can successfully present flicker frequencies of 60, 30, 20, 15, 12, 10, 6 5, 4, 3, 2 and 1 Hz. As the optimal range of SSVEP is restricted to just 8-15 Hz only 3 SSVEP waveforms would reliably be reproduced using a 60 Hz monitor. Using this method, the only means of increasing the number of available factors in the desired range is the deployment of a monitor with a higher refresh rate.

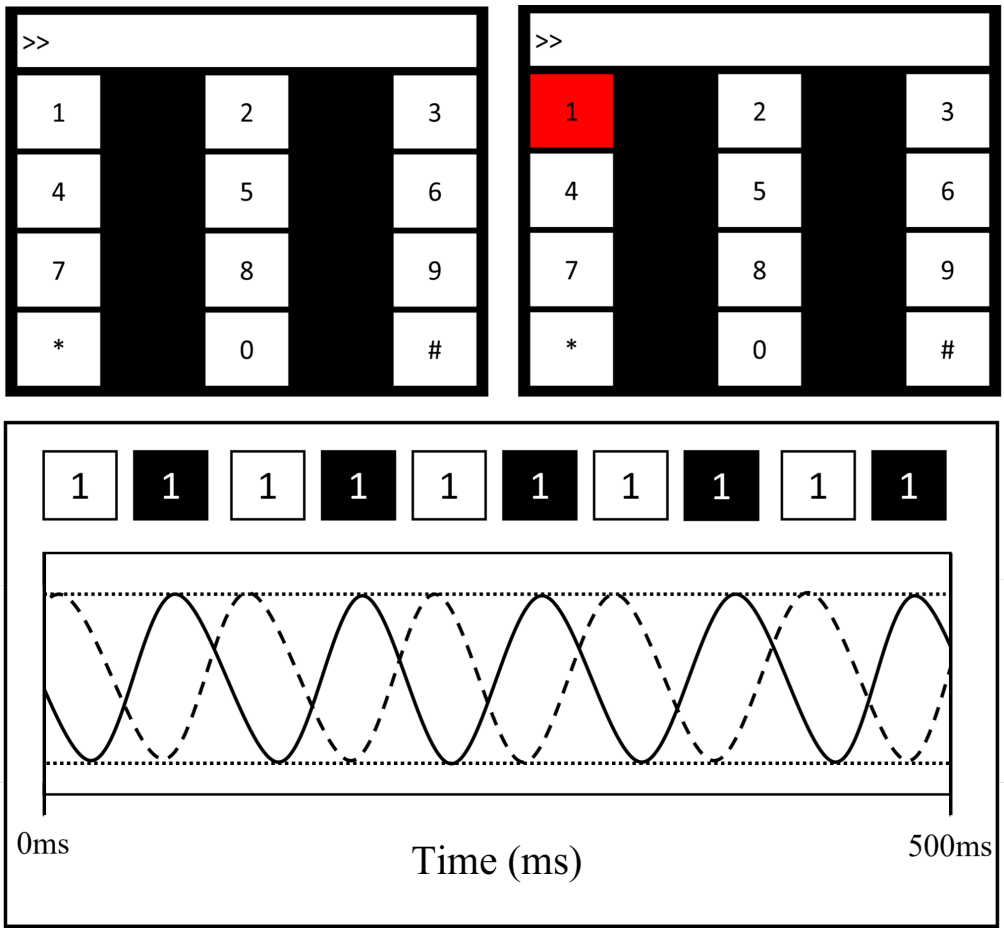

Figure 2.2: Here is presented a series of images depicting the virtual numpad stimuli utilized in the reference experiment conducted by [180]. The upper left quadrant shows the numpad during the rest period and the upper right quadrant illustrates the means of cueing subjects via the overlay of a red square on the target number. Note, that these figures were developed independently of the original article. The lower half of the figure displays a visualisation of a 10 Hz flicker frequency. Positioned above the plot is a representation of a given target number, as the signal oscillates between the amplitude bounds (dotted line) the target number is augmented from standard to an inverted colour representation to generate the flicker. As seen from the plot a total of 5 full cycles are presented over the course of 0.5 seconds indicating that the signal is oscillating at a frequency of 10 Hz. The dashed line is presented alongside the solid line to provide further insight into how modifying the phase angle via the introduction of the temporal offset of a given signal can lead to a dramatic decrease in correlation. This is highly effective for enhancing the discriminability of targets with similar frequencies.

The JPFM dramatically increases the number of available flicker rates using two techniques. Firstly, the phase angle of each target flicker rate profile is modified, meaning the initiation point of targets differs over the given frequency cycle. In other words, some targets are initiated with the cycle at the peak and others at the trough. This significantly decreases the correlation between two signals even for the same frequency (see, Figure 2.2, lower graphic). Further, the method employs a system of rapidly interchanging between two distinct signal profiles for example 10 Hz and 15 Hz to generate an averaging effect that leads to the propagation of a 12.5 Hz SSVEP oscillation (for further info see, original article [180]). At the start of each trial, subjects were cued to attend each target (4 seconds) according to a randomised fixation protocol via the temporary presentation of a red overlay square (1 second). These assessments were conducted over 15 blocks consisting of 12 trials each leading to a total of 180 trials per subject and over the 10 subjects tested a total of 1800 trials were collected. Throughout the stimulus presentation period, concurrent wet EEG data acquisition was performed using the BioSemi ActiveTwo EEG system at 2048 Hz, as per [180], across 8 channels (O1, Oz, O2, PO7, PO3, POz, PO4 and PO8) and referenced against the Cz electrode. Note, all data were later downsampled to 256 Hz offline.

Here, the authors [180] tested numerous Canonical Correlational Analysis (CCA) methods. The authors report impressive classification performance and ITRs for the so-called Combination Method. This is a blended approach incorporating standard CCA reference signal matrices and subject-specific individual template matrices, (see, Table 2.1). This effectively tunes the Combination CCA method at the subject level to provide a bespoke set of reference signals unique to the subject tested [180, 181].

Subjects	Accuracy (%)	ITR
1	78.89	63.33
2	71.67	52.34
3	94.44	92.49
4	99.44	105.47
5	100.00	107.55
6	99.44	105.47
7	98.33	102.14
8	100.00	107.55
9	98.89	103.76
10	86.67	76.72
Mean	92.78	91.68

Table 2.1: Here is presented a table displaying the performance metrics relating to the Combination Method for 1-second data chunks reported in the reference article cited above [180]. Note, that all ITR metrics were computed offline and calculated independently from the original article as these are not stated directly in the main body of the text.

Is should be clear from the discussion above, advances at the cutting edge of SSVEP-based speller performance are related to the symbiotic development of pre-screening, stimulus design and classification method enhancements. Along these very same lines, the inclusion of task-related component analysis (TRCA) has been implemented following a pre-experimental localizer to develop subject-specific spatial filters to parse redundant non-SSVEP-based background noise from subsequent trials [182]. These are then implemented in real-time within the online pre-processing pipeline to enhance the expression of task-relevant SSVEP target waveforms. The aforementioned enhancements allowed for the effective use of data lengths below 400 ms per character spelt and crucially, maintained functional classification accuracies. Note, that typical durations for optimal SSVEP-based speller systems require 500-1000ms for reliable operation.

Further, the use of Bayesian methods to optimize and modify data capture durations in real-time for so-called dynamic stopping protocols has enabled researchers to attain cross-subject average ITR values of 353 bpm [55]. Finally, the use of multiple sequential coding for introducing more than two phase and frequency signal pairs (4) in one monitor refresh cycle [178] to further differentiate target stimuli on screen has been explored in real-time assessments [47]. These build upon the original suggestions outlined in the approximation method [174] and proved effective in developing a functional SSVEP-based speller with 160 targets, operating at an average cross-subject accuracy of 87% with a mean ITR of 78 bpm. In sum, these

impressive results demonstrate that the functional properties of the SSVEP enable researchers to fully explore the depths of scientific creativity and ingenuity. Alone, these results represent the bleeding edge of both SSVEP-based BCI speller research and BCI spellers more broadly as these performance metrics are unrivalled in terms of both ITR and stimulus matrix density.

2.6.3 Neural Network-based Bio-Signal Classification

In many disciplines utilizing volatile time-series data such as weather formation prediction, financial forecasting, and autonomous vehicle tracking, the introduction of neural networks has proven highly profitable (see, [183] for review). The progenitor of these modern methodologies, Multi-Layer Perceptrons (MLPs), operate according to functional and architectural features identified originally in the mammalian visual system. Clusters of nodes, a computational equivalent of neurones, are positioned in operational layers mirroring the striated cortical arrangement found in the occipital lobe [52]. Note that nodes are essentially representations of a specific region in the input data. This could relate to an isolated subsample of pixels in an image or the expression of a specific bandwidth of frequency space for a given length of time-series data. The cascade of inputs and processed signals transverses the network via a series of simulated connections between subsequent model layers and corresponding nodes. The activation of these computational neurones is controlled via a weight function. The responsivity of these weight functions is modified via iterative training procedures adhering to an error minimization protocol. Before the training stage network performance is low as weight values are typically standardized at zero or randomised from a subsample of appropriately scaled values. A loss function calculates the difference between the expected node output and the current node output and updates the relevant weights in the corresponding direction.

MLPs dramatically enhanced the performance of neural networks via the inclusion of a so-called hidden layer, positioned between the input and output layers, providing connections between both. This enables the model to produce a feature map of numerous relevant data components. The simultaneous activations of node clusters tuned for multiple task-critical data representations allow these networks to model highly complex, non-linear relationships [184]. As seen in previous studies, MLPs are not the most effective means of classifying BCI-based bio-signals [185, 186]. This is owing primarily to insufficient network complexity in terms of both node count and operational layer depth. Further, the initial aim to develop biologically inspired computational models to research brain function has now become disentangled with the alternative aim of boosting classification performance for clinical BCI applications. In other words, the goal of researching the operational behaviour of human brain regions has become independent from the task of developing the highest-performing networks

as alternative methods that forgo these constraints, for example, deep neural networks, have shown considerably higher performance in terms of classification accuracy, information transfer rate and bio-signal adaptability.

The advent of GPU-based network training [187] in concert with a renewed focus on raw task-based performance greatly influenced the development of so-called Deep Neural Networks (DNN) for computer vision tasks [188]. This broad area of research has been implemented for numerous highly varied applications including satellite imagery processing [189, 190], medical imaging classification [191–193] and autonomous vehicle control [194]. These models diverge from MLPs principally in terms of layer count, node density, and operational performance, as opposed to, biological validity in architecture design considerations. As noted above, these dramatic increases in computational resource use are afforded via the transition from CPU to GPU hardware. Further, these increased processing demands have been intelligently offset by changing the quality of node weights and connectivity relationships. These modifications include node weight initialization schemes and sparse connectivity between layers [195–199]. The advances in network efficiency have enabled researchers to develop networks with more hidden layers and in turn, allow models to develop ever more complex internal representations in the topmost layers of the network [200].

The convolutional neural network (CNN) is positioned as arguably one of the most popular sub-variants of the deep neural networks currently employed. These models introduce an additional transformation layer during the data output stage within the layer-by-layer, node-by-node connectivity pathway. Note, that in the simplest terms, the convolutional operation, in this context, is a means of reducing or downsampling input data to allow for the amplification and extraction of prevalent data features. The operation is composed of three fundamental components, the input data, filter (or kernel) and output feature map. The convolutional kernel is a data matrix containing values (weights) that represent a randomised data feature. Note, that the kernel must be smaller than the input data, to accommodate for this padding, for example, zero padding, can be applied. The kernel weights and correspondingly sized input data segments undergo an element-wise multiplication. A sliding window traverses the input data, convolving these values with the filter weights to populate an adjacent feature map (output). The quality of the data representations embedded in the kernel weights is updated via an optimization function, for example, stochastic gradient descent (SGD), to enhance the expression of task-relevant data features and ultimately, network performance.

This model design variant stands as the first to outperform trained human experts in tasks concerning visual discrimination [201, 202]. Relevantly, these models have been successfully deployed in multiple bio-signal classifier applications including seizure detection [203, 204], Motor Imagery (MI) classification [61, 205, 206] and robotic limb tracking and instruction [207–209]. These studies demonstrate the adaptability of CNNs for numerous input modalities in the BCI field. Critically, such models have also been effectively implemented for SSVEP classification in assistive speller contexts (see for review, [210, 211]). The developmental trajectory of SSVEP-based CNN classification spans from the implementation of simplistic 1 layer networks, for example, the 1DSCU [212] and ShallowConvNet [61], to more elaborate multi-layer arrangements as seen in PodNet [213], DeepConvNet [61], EEGNet [59] and EEGNetSSVEP [60].

2.6.4 Cutting-Edge Classification Methods for SSVEPs: Convolutional Neural Networks

The highest-performing networks all feature extensive use of online data repositories for training and offline evaluation as well as simulated real-world performance. These vast banks of high-quality data collected from large subject pools are crucial for tuning the high volume of trainable parameters in the associated networks. Most all datasets utilize a variant of the joint, phase frequency method outlined above and feature either numpad, keyboard [49, 52] or keyboard + arrangements [47], increasing the number of targets deployed from, 12, 40 and 160 respectively.

The most significant enhancements in ITR performance reported to date focus either on experimental improvements utilizing current networks, the application of rigorous subject-specific data pre-processing stages or the novel implementation of FBCCA analytical principles in CNN-based solutions [214]. As seen in [50], the performance of the Compact CNN (EEGNet) was boosted in simulated offline trials by 28.2% in terms of cross-subject average performance (147.6 bpm) for a 12-target SSVEP-numpad speller dataset. This was achieved via the implementation of a dynamic windowing protocol and involved using the output of corresponding CNN classifiers to gauge data viability. For instances demonstrating sub-threshold prediction credibility, additional data acquisition was cued reflexively to ensure high classification accuracies. Notably, this system outperformed both the fixed window EEGNet (115.1 bpm) model performance and crucially, an FBCCA (85.2 bpm) implementation also operating with the same dynamic windowing method.

Further, impressive ITR metrics are demonstrated in [54] for the TRCA-DNN (235.21 bpm), a four-layer CNN model, evaluated using two online repositories featuring 35 targets and a combined 105 subjects. The TRCA method, as outlined above [53], generates a bank of subject-specific spatial filters for the extraction of task-relevant data features. These pre-processed signals are then fed into the given CNN model. Notably, the implementation of this pre-processing stage enhanced base DNN performance (167.75 bpm) and demonstrated arguably the highest single-subject real-world online classification performance metric using CNN-based methods for SSVEP decoding to date (318.41 bpm). The highest cross-subject ITR performance in this domain is demonstrated by the CNN model outlined in [57]. Here, only minimal pre-processing of the EEG input data is performed involving the removal of power line noise (50 Hz notch filter), a high-pass filter at 8 Hz and a low-pass filter at 90 Hz. The method of filter bank generation is integrated at the data preparation stage following pre-processing. The original 2D Samples $\times$ Channels input are concatenated across a third axis alongside additional sub-band filtered arrays. This arrangement allows for the expression of the fundamental frequency components generated from the target stimuli in addition to the higher frequency harmonic reflections. As the highest target stimulus frequency of 15 Hz is a factor of the upper limit of the initial pre-processing low-pass filter of 90 Hz harmonics up to at least the 6th order are available for extraction. These 3D filter bank data matrices are fed into a deep 4-layer CNN and produced cross-subject average ITRs of 196.6 bpm and 254.23 bpm for the 70 [49] and 35 subject [52] 40 target keyboard speller datasets respectively.

There still exist numerous potential advances to enhance the performance of the networks defined above. Recent implementations of transfer learning have proven fruitful in boosting classification accuracy and ITRs. Along these same lines, [215] initially trained PodNet, a 4-layer deep CNN, using the aforementioned 70-subject benchmark dataset [176] in a pure cross-subject format as a means of initializing system weights. At this point, a modest AoC and ITR of 73.6% and 93.1 bpm were attained for an isolated 10-subject test set. Following this, subject-specific re-training of exclusively the lowest PodNet convolutional layer was conducted using 50% of each individual subject's data from the 35-subject benchmark dataset [52]. The evaluations using the remaining 25% of single-subject trials demonstrated a cross-subject average AoC and ITR of 95.00 % and 143.13 bpm.

Further, the implementation of so-called Inception modules in SSVEP-based CNN model architectures demonstrates significant potential. Originally developed for traditional computer vision tasks, these operational blocks consist of bottlenecked parallel convolutional filter banks of differing orientations that are later concatenated together alongside a non-bottlenecked

higher-order filter block. This effectively allows for the concurrent extraction of EEG signal data representations with differing temporal dynamics. These analytical properties can be conceived of as a translation of the stratification of EEG signals into the spatially distinct frequency sub-bands employed in FBCCA, to the temporal domain.

With the availability of the aforementioned online repositories, alongside tools for model benchmarking, such as the Mother of all BCI Benchmarks (MOABB), the pathways to validate models in terms of raw performance for different datasets have been firmly established [216, 217]. Despite this, these methods rarely attempt to probe the effects of bespoke dataset pre-processing at the single-subject level for all networks evaluated, as typically the quality of performance is related primarily to AoC, ITR and principally model adaptability, for a range of signals, across multiple subjects. An investigation involving the manipulation of industry-standardized pre-processing parameters relating to core data preparation principles has not been extended beyond the restriction of data acquisition time or arrangement of electrode ensembles. Herein, the author proposes to optimize subject-specific network frequency filter operations (low and high-pass cutoff values) for cross-subject aggregated datasets in a range of publicly available CNN network architectures.

As seen in the recent studies noted above, the ever-increasing implementation of FBCCA principles in CNN model solutions has necessarily increased the relative width of bandpass filters to accommodate for the extraction of higher-order target SSVEP frequency harmonic components. It is crucial to understand the influence of network depth, filter count, complexity and architecture design on this process to inform the development of future models and determine the appropriate tuning techniques for these systems at the single-subject level. Further, the aggregation of cross-subject data for the training of BCI classifiers is a relatively novel phenomenon. Traditionally, methods focused on bespoke baselining and classifier training at the single-subject level due to the perceived overwhelming influence of individual differences and non-stationarity that characterize EEG time-series data. In all notable studies listed above, the aggregation of subject data is fundamental to the results garnered as the flexibility and richness of internal representations developed in aggregate data-trained models are significantly higher as compared to those using exclusively single-subject data. Note, that these methods are only successful in either instance for applications using vast quantities of data. In other words, a highly diverse subject pool is necessary, but not sufficient in the context of successful CNN model training. In response to this conflict in the literature, the optimization methods herein attempt to present and evaluate a subject-specific method for the pre-processing of aggregated datasets.

To summarise, the aim of this research centres around the development of a methodical means of evaluating different CNN architectures for SSVEP-based bio-signal classification. Further, these very same tools can also be applied for the evaluation of optimal hyper-parameters for a given set of data at the cross-subject or single-subject level. In other words, the experiments conducted herein outline a means of creating, optimizing, and evaluating neural network architectures to increase the scope of model applicability and enhance performance to develop more robust and higher accuracy classifiers for end-point patient users. There exist several methods researchers can deploy in the preparation of EEG signals for classification, these include temporal data segmentation via experimental event triggers, time-correction via interpolation, signal referencing via subtraction of either a specific non-relevant cranially positioned electrode or an average of all channels being sampled and active electrode selection in real-time to mitigate noise influence in EEG signals. Historically, these methods and associated parameters have become standardized, despite this in some instances the determination of these values is rather arbitrary.

This includes the selection of low-pass and high-pass EEG filter cutoffs in the EEG pre-processing stages. Some exploratory investigations into the low-pass cutoff values and associated stimulus frequency harmonics have been undertaken [218–220], yet to the author's knowledge there remains no significant work conducted on the optimization of both filter values concerning CNN performance. It is rationalized that setting a high-pass filter bound as close as possible to the target waveform frequencies being captured will result in the lowest amount of redundant hyper-low frequency information from diluting the target signal in EEG data. The same level of theoretical certainty is not present concerning the low-pass filter cutoff in relation to neural network training outcomes. It could be argued that this parameter should be similarly restricted to the upper limit of the target waveform frequency. Despite this, many researchers have successfully utilized SSVEP target waveform harmonic features to boost classification [47, 56, 180] suggesting a higher low-pass cutoff could benefit models depending on network complexity and depth.

Further, there are presently no attempts to characterize the relationship between high and low-pass frequency filter cutoff values. These investigations aim to provide clarity on this crucial pre-processing stage and inform the calibration procedures of subsequent CNN model designs in future research. Moreover, the author asserts that owing to the non-stationary nature of the EEG signal the 'perfect' upper and lower bounds for bandpass filters do not exist, especially when considering the variety of influences on each neural signal across many sub-

jects for many different EEG components. It is predicted that an optimal value for both upper and lower bandpass bounds via automated parameter search is achievable at the single-subject level. To explore these hypotheses the author undertook the task of defining a methodology for the simultaneous automated optimization of signal processing and neural network hyper-parameters to maximize single-subject classification performance when trained on aggregated, cross-subject datasets.

Here, specific relationships between end-point classification accuracy, optimization study duration and the relative number of network training parameters are outlined for a series of commonly utilized models in the domain of SSVEP classification. As will be evidenced in the corresponding results sections, these investigations are highly computationally expensive. Note, that the findings defined herein are intended for use by future researchers in designing optimal search space boundaries by using these results as a guide. To clarify, it is not the author's intention to assert that these methods are viable in any online context, rather they should be utilized in the development of neural network training and development for the enhancement of performance metrics.

Chapter 3

Experiment 1: P300-Based BCI-Speller Stimulus Evaluation

3.1 Chapter Outline

This chapter will cover three iterations of the visual-P300 Emoji-Speller experimental design explored throughout the PhD research conducted herein. Each iteration of the paradigm attempts to address the shortfalls of the previous version. It must be noted that the distinct lack of subject data for the final version of the experiment (see subsection 5.3) is owing to the restrictions placed on students and staff alike in the wake of the COVID-19 pandemic (see subsection 1.1). The author intends that this chapter will serve as evidence of critical evaluation skills and the research techniques gained throughout the training process. Note, that all the background literature, rationale and justifications for the following experimental series are contained within the Introduction and Literature Review chapter (see subsections 1.4 & 2.6).Further, all subsequent emoji-based P300-speller experiments (see Chapters 3, 4, & 5) feature two approaches to data organisation, pre-processing and analysis. The original implementation is referred to as Pipeline 1, the follow-up version is referred to as Pipeline 2. The adaptations introduced by the Pipeline 2 approach were performed using a subset of 3 subjects from each experimental variant to address the issues concerning Pipeline 1, for more information please refer to subsections 3.3.5 and 3.4.3.

3.2 Aims

In the first of three experimental formats, subjects were presented with an array of 7 Emoji, ranging in valence from disagreeable to agreeable. Crucially, all systems comprising the stimuli presentation, data collection, data pre-processing and analysis were initially run offline and all on the same machine. This was undertaken to develop a one-stop emoji-speller experiment with plug-and-play functionality for clinical, research, instruction and outreach purposes. This is a crucial first step, as it is necessary to establish that emoji-style stimuli reliably produce P300 waveforms for BCI-speller applications before the development of a fully online emoji-integrated-BCI-speller. The first experiment defined herein involved probing for optimal stimuli features in terms of augmentation method by comparing the classification performance of a simple LDA classifier for a white-overlay square format (referred to as the Flash method) and colour inversion method (referred to as the Inversion method) for target stimulus augmentation.

Note here, that the inversion method involves the re-colouration of all black-coloured elements in a given emoji stimulus to a white colouration. In most instances this involves the change of the emoji perimeter silhouette and outer edges of the emoji facial features, such as the mouth and eyes, inverting from black to white. Additional information can be found in subsection 3.3.3: Stimulus Presentation and the related stimulus screenshots displayed in Figure 3.2. The experiment is adapted in the second iteration via the introduction of a staggered emoji array format (3, 5 & 7 Emoji arrangements). These investigations were undertaken to determine the influence of array density and associated adjacency error issues on subject-level performance. In the final stage of this project, a real-time speller was developed utilizing the 7 Emoji stimulus variant featuring online subject data pre-processing and classifier prediction feedback.

3.3 Method

Here are outlined the methods employed in the investigations relating to Experiment 1. Broadly, this features the use of a 7 Emoji stimulus array design for BCI speller applications and the associated evaluations of the Flash and Inversion augmentation methods described (see, Figure 3.1).

3.3.1 Participants

A total of 10 neuro-typical subjects were recruited from the Durham Psychology Department student population and did not receive payment for participation (5 males, mean age of 23 years, with an age range of 20-27 years). All subjects were screened before experimentation to ensure they presented with normal or corrected to normal vision, had no history of clinical mental illness or epilepsy and were not currently experiencing a skin-based ailment of the scalp. Ethical approval and oversight were granted by the Durham University Psychology Department's Ethics Sub Committee.

3.3.2 Equipment

All EEG time-series data were acquired using the Cognionics Quick-20 headset (Cognionics, San Diego, USA) via wireless Bluetooth connection. All data streams were controlled via LabStreamingLayer and sampled at 500 Hz. Graphical rendering of the stimuli was handled via a dedicated NVIDIA GTX 750ti GPU (2GB VRAM). Note, that before testing, the headset and corresponding electrodes were sanitized with anti-bacterial gel.

3.3.3 Stimulus Presentation

All stimulus presentation software was designed and implemented using the Psychopy Python library [221] and presented using a 68.5cm Samsung LED S27A350H (refresh rate 60 Hz) at a fixed distance from the subject (0.8m). The target emoji stimuli utilized in all studies conducted herein were collected from the open-source OpenMoji repository [222]. The colour of each emoji stimulus was altered from the original yellow colouring to enhance distinguishability [154]. This was used to address the common P300 experimental obstacle known as adjacency error.

This involves the augmentation of stimuli neighbouring the target stimulus triggering a P300 response and can cause a temporal or spatial bleed-over effect (see subsection 2.5.1). This can

delay the refractory period of the P300 waveform and significantly depress signal peak amplitudes. The emoji stimuli are simplistic circular visual targets (diameter: 18mm) evenly spaced across the array at 91mm intervals (see, Figure 3.1). Each of the 7 emoji is positioned centrally across the horizontal axis of the computer monitor. The target emojis represent different levels of emotional valence from disagreeable to agreeable (left to right).

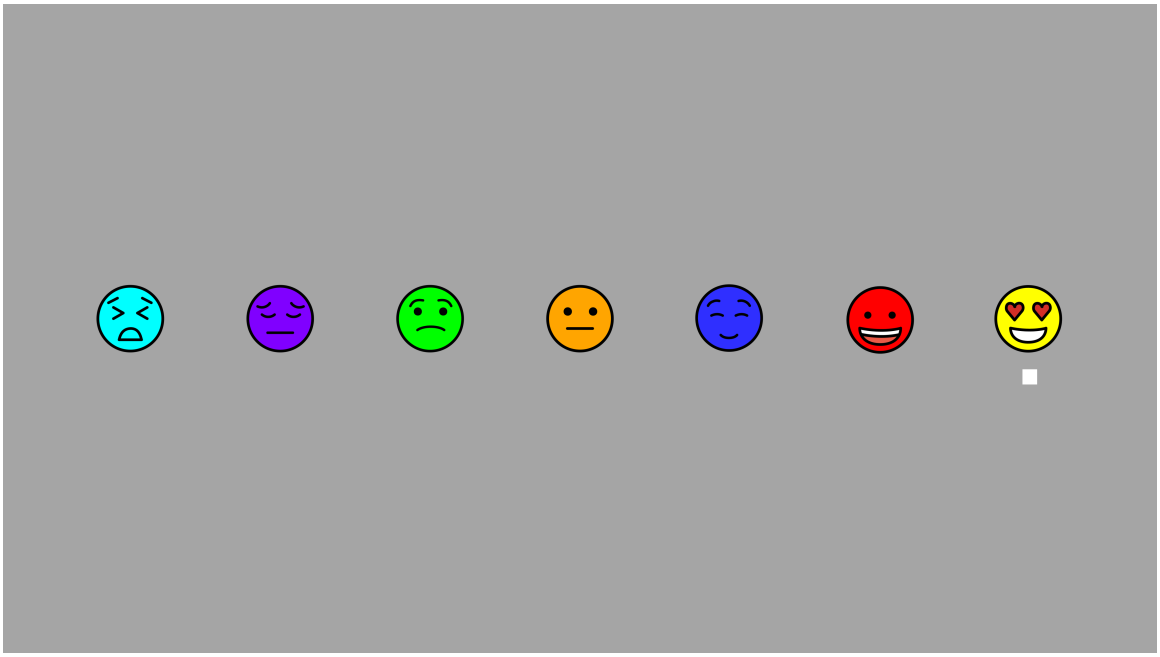

Figure 3.1: The above image is a screenshot of the experimental visual array as seen by the subjects at the start of each experimental trial. The emoji are arranged from left to right in a valance gradient moving from disagreeable to agreeable. The white cue square is positioned under the target emoji throughout the entirety of the experimental trial. From left to right each emoji descriptor and tag used include the 'Persevering Face' (1F623), 'Pensive Face' (1F614), Worried Face' (1F61F), 'Neutral Face' (1F610), 'Smiling Face' (263A), 'Grinning Face' (1F600) and 'Smiling Face with Heart Eyes' (1F60D).

During the testing session, EEG time-series were collected using either a 'flash' or 'inversion' augmentation method (see, Figure 3.2). The flash method involves overlaying a white square onto each respective stimulus according to a time-locked randomisation procedure. The inversion method functions by swapping all black-coloured elements in each respective emoji target to white-coloured elements. All other features of the stimulus design were kept consistent throughout testing, including the duration of stimulus augmentations (fixed at 125ms).

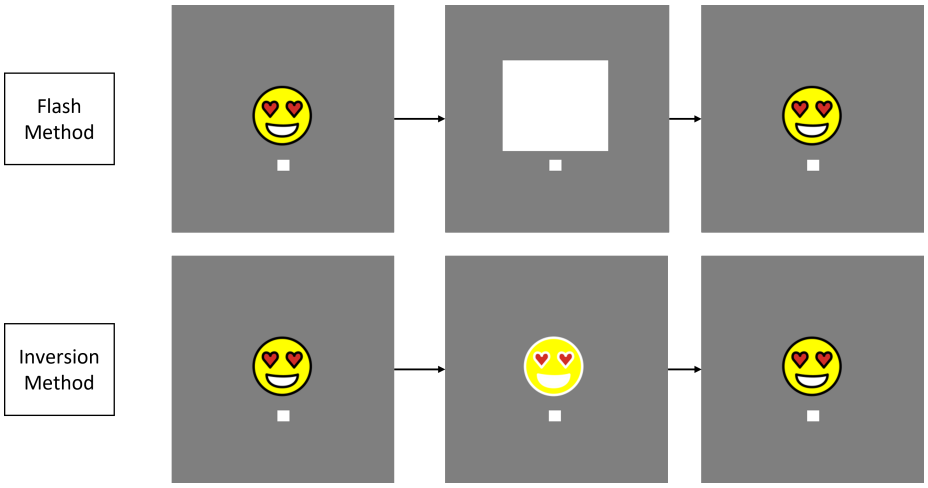

Figure 3.2: The above image illustrates the differences in the augmentation methods employed. The Flash method is displayed in the top row, depicting a white overlay square as the P300-inducing visual emoji augmentation (40mm diameter). The Inversion method is depicted in the second row, illustrating how every black element in each non-augmented emoji is inverted to white.

Each emoji in the array was augmented according to a non-consecutive randomisation program, ensuring that neighbouring targets were not augmented in succession to avoid adjacency error and the double flash problem (see subsection 2.5.1). This principle was enforced within sequences, as well as between sequences such that the selection of the following sequence's 1st augmentation was not the same or a spatial neighbour of the last augmented emoji. The target stimuli were cued to subjects via the presentation of a white cueing square (positioned below the emoji). Each trial was comprised of 5 sequences, with one sequence completing after all emojis in the array were augmented once (see, Figure 3.3).

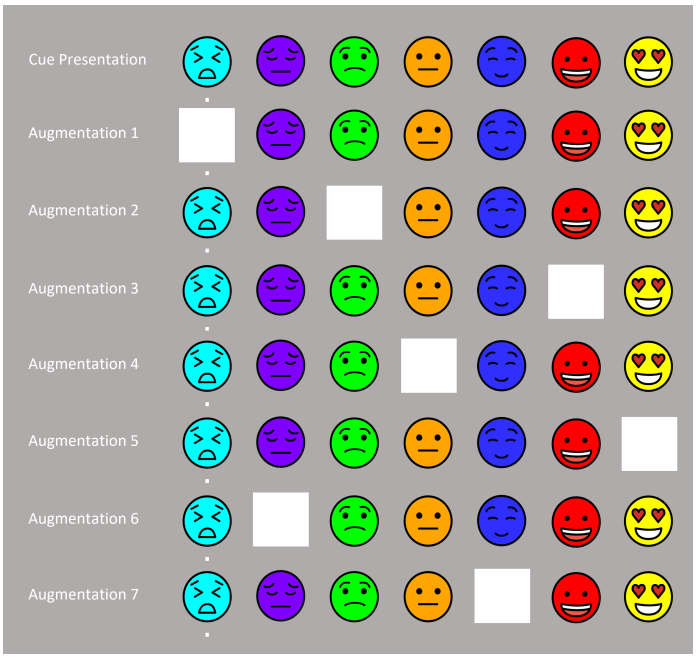

Figure 3.3: Here is presented a figure illustrating one sequence of a flash augmentation trial. The top row depicts the initial stimulus array presented to subjects before the onset of the trial. The subsequent rows demonstrate a pattern of stimulus augmentation adhering to the randomized non-consecutive format. As can be seen, at no point in the stimulus schedule is the proceeding emoji selected for augmentation adjacent to the previously augmented emoji. The pattern of stimulus augmentation in this flash variant example is the same utilized for the presentation of all inverse augmentation method trials.

The inter-sequence intervals were maintained at 375ms, with inter-trial intervals fixed at 1000ms, after which the white square target cue is re-positioned, and the successive trial is initiated. The noted parameters were implemented to ensure that any refractory effects induced by a P300 elicited for the previous time-locked augmentation event would not carry over to the subsequent augmentation or trial event. These intervals were introduced to the detriment of final information transfer rate (ITR) values, and for the benefit of the resulting data quality, reducing the prevalence of temporal bleed-over artefacts.

The inter-stimulus interval (ISI), also commonly referred to as the flash rate, describes the time in milliseconds between the augmentation of successive targets, in this instance emojis. A flash rate of 125ms was chosen based on previous research that aimed to determine the optimal inter-stimulus interval between visual-P300 targets using an analogous 8 x 9 item alphanumeric speller matrix [223]. Here the authors revealed that for a 10-flash run (total number of augmentations for the target per sequence), as the inter-stimulus interval decreased from 250ms to 31.25ms, the group and single-subject Pz average plots showed a dramatic drop in relative P300 peak amplitudes.

For the assessments in this thesis, the array is far more simplistic, consisting of a horizontal spectrum of just 7 targets. Initially, this suggested to the author that using the maximal inter-stimulus interval tested (250ms) was unnecessary. Further, additional benefits for a lower inter-stimulus interval include a shorter experimental period. This was predicted to have a positive influence on data quality by reducing the incidence of subject fatigue in the latter half of the data collection period. Moreover, a lower ISI also presents obvious benefits in terms of a higher theoretically achievable information transfer rate. Given these factors, an ISI of 125ms was implemented.

In sum, each trial comprised all 7 Emoji targets consecutively augmented for a duration of 125ms, followed by an inter-sequence interval of 375ms. Upon the completion of all 5 sequences, a 1000ms inter-trial interval is enforced. In total each trial consists of (5 × ((7 × 125ms) + 375 ms)) + 1000ms, totalling 7250ms. Each subject partook in a total of 4 experimental blocks, with one block consisting of 49 trials. Each stimulus augmentation style was tested twice (2 Inversion blocks and 2 Flash blocks) with a 5-minute break between successive blocks. During these breaks, impedance monitoring and subject comfort were assessed to ensure the acquisition of high-quality data and reduce the incidence of subject fatigue.

Note, that the selection of just 5 augmentations per trial deviates dramatically from the typically utilized 10 augmentation standard, as was implemented in [223]. This decision was primarily made to increase the theoretical upper ITR limit by lowering the total time per trial. Further, a posthoc re-analysis aggregating signals across trials was implemented to simulate a boost in the number of augmentations per sequence into the signal averages, this analysis variant is termed the Collapsed method. The preparations and results of these analyses, as well as the merits of all stimulus parameter settings utilized are discussed in subsections 3.3.5.3 Data Pre-Processing: Pipeline 2 and 3.7.5 Conclusion: Pipeline 2.

3.3.4 Data Acquisition

A Quick-20 Dry EEG Headset (Cognionics) was used for data acquisition at a rate of 500 samples/ per second (500 Hz). Electrodes: Fz, Cz, Pz, P4, P3, O1, O2, A1 and A2 were sampled from the headset concurrently during stimulus presentation (see, Figure 3.4). All EEG time-series data were handled via the LabStreamingLayer (LSL) Python package variant [224]. To improve the signal-to-noise ratio of electrical signals on the surface of the scalp, participant preparation includes gentle and careful rubbing of the scalp underneath each sensor to push aside non-conductive dead skin cells and hair. This procedure is painless but can feel slightly uncomfortable as experimenters rub the blunt, metal-tipped dry sensors against the scalp to ensure optimal seating to the skull.

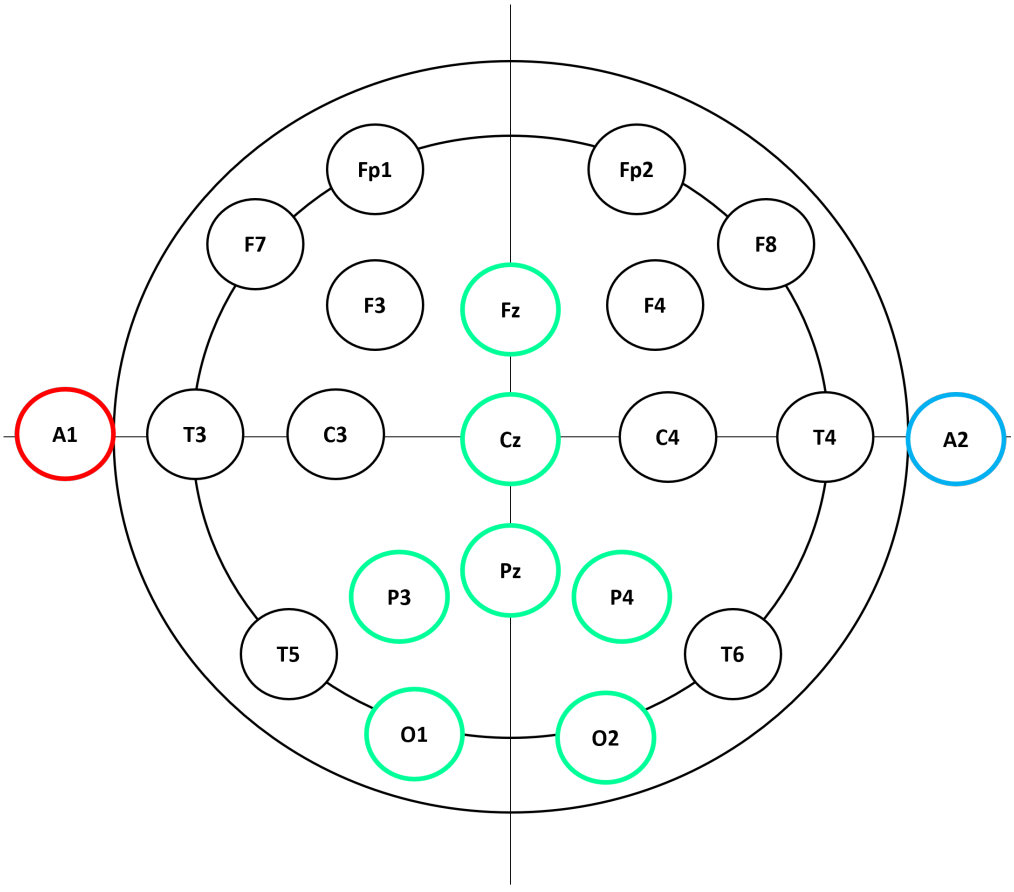

Figure 3.4: Here is presented the electrode arrangement for the Cognionics Quick-20 dry-EEG headset. The sensors are positioned according to the standardized 10-20 EEG data acquisition system. All scalp locations highlighted in green comprise the signal locations used to populate the EEG data array. The reference location, seen in blue (A2) and ground location, seen in red (A1), were utilized to reduce redundant noise in the associated data samples. Note, that the O1 and O2 locations were utilized in place of the traditional Oz location given the reduced density of the electrode array in this headset.

Before the onset of the experiment, researchers attempted to minimize impedance (Ω) values via headset reseating and electrode agitation (as per standardized hardware setup procedures). In typical EEG-based research, impedance values are maintained around 1-10kΩ [75, 76] to maximise the probability of acquiring high-quality time-series data. It must be noted that this was a significant challenge, with headset setup times often exceeding the duration of time subjects spent partaking in the experiment. The full experimental period, including subject preparation (5-10 minutes), testing (30 minutes) and breaks (5 minutes each) required approximately 45 minutes to 1 hour.It is important to state that the EEG capture for all experiments detailed herein did not follow the traditional continuous data collection method in which samples are stored prior to, during and following the experimental period to be later epoched via distinct marker triggers. Instead, the incoming samples were held in a buffer that was periodically sampled during the ongoing trial periods. Here the data was pulled from this stream at the start of each sequence and terminated following the triggering of an internal function monitoring the difference in start and current time. This was originally intended to streamline the process of aggregating and analyzing data in real-time however this produced the unintended consequence of failing to gather samples prior to the onset of the trial for crucial baselining purposes. A baselining method was implemented in the Pipeline 2 approach utilizing the first 50ms of samples as a partial solution to these issues, for further information please refer to subsection 3.3.5.3 Data Pre-Processing: Pipeline 2.

3.3.5 Data Organisation: Pipeline 1

Once the data collection period was completed (all 10 subjects), the individual time series and corresponding labels were collapsed into multiple databanks. These included large aggregate databanks consisting of all trials collected across subjects for each of the stimulus augmentation variants tested. Individual databanks were also generated to allow for single-subject data quality and artefact rejection as well as subject-specific analyses. A significant imbalance in the number of P300 events *vs.* Non-P300 (1:6) events was present in the data, as per standard oddball-based P300-speller experimental designs. The author anticipated that this could introduce the possibility of overfitting into the analysis and addressed this by creating a stratified subsample of Non-P300 events totalling the same number of P300 events. These subsampled data groupings are termed 'class-balanced', as opposed to the 'non-class-balanced' groupings. Further information on the exact arrangement of these data is provided in the Results subsection 3.5.2

As is noted in the title of this subsection, all details herein relate to Pipeline 1. This describes the author's initial data organisation, pre-processing and analyses of these experimental data as well as the corresponding results derived from these methods. As will be evidenced in the following chapters, several key alterations to each stage could have been implemented to improve the procedures detailed in this initial attempt. To address these issues a secondary pipeline, Pipeline 2 was developed and the data was reanalysed following the steps outlined in subsections 3.3.5.3 and 3.4.3. All areas of the thesis relating to either Pipeline 1 or Pipeline 2 will be signposted to ensure clarity.

Given substantial time constraints related to this project the re-analysis defined in Pipeline 2 is performed exclusively with the 3 most promising subjects from Experiment 1. These comprise Subjects 3, 5 and 8. The same approach was taken with the following two chapters relating to Experiments 2 and 3. The selection of subjects differs across these studies and detail is provided throughout to guide the reader at all stages. In sum, Pipeline 2 is characterized as a targetted re-analysis of the most viable data available based on the findings of Pipeline 1 and is designed to address the shortfalls in the aforementioned methodology for all experimental variants of the emoji-based P300-speller described herein.

3.3.5.1 Data Pre-Processing: Pipeline 1

At the sequence level, the data are parsed into 7 Emoji chunks indexing 375ms of EEG time series from the initiation of each time-locked emoji augmentation. All chunks are first pre-processed using the pipeline as described below (see, Figure 3.5).

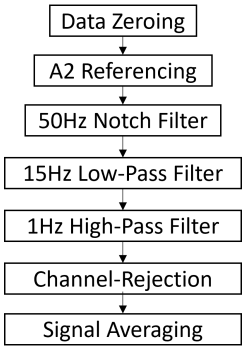

Figure 3.5: Here is presented a graphical illustration of the EEG time-series pre-processing pipeline implemented for the experimental series. Any changes to this base pre-processing format throughout this chapter are strictly additive. Any enhancements to this method are discussed explicitly concerning each respective experiment. That is to say, all data analyzed in the following experiments underwent these pre-processing stages and potentially some additional minor steps.

Initially, the data chunks are zeroed. This involves calculating the mean μV of a given EEG electrode sampled and then subtracting this value from all data points in the single-electrode time-series array. This effectively centres the data at zero, enabling effective visual appraisal of the signal for inspection and presentation. Further, this prevents differences in signal scaling from disrupting the accurate P300 and Non-P300 signal representations. These zeroed data signals are then referenced using the A2 electrode which is positioned over the right ear. Microvoltage signals collected at the A2 electrode are known to be representative of non-salient (Non-P300 relevant) electrical signals generated across the entire head.

The referencing involves the subtraction of corresponding A2 channel samples from all respective electrode sites sampled. This process is intended to remove redundant information from the electrode time series thereby enhancing the signal-to-noise ratio. Note, that the ground and reference compound signal used to remove the common mode signal is handled via the Cognioincs active grounding system [225]. This involves assigning a dedicated ground electrode, in this instance A1, and subtracting this value from the corresponding reference electrode, A2, to produce the compound reference signal.

The powerline noise inherent in all EEG data acquisition was removed via a 50 Hz notch filter. All filters described herein were designed and implemented using the Scipy Python library [226]. The same package was utilized to design and implement two Butterworth filters for data pre-processing. A high-pass filter with a 1 Hz cut-off was used to remove all frequencies below this boundary and a low-pass filter set with a 15 Hz cut-off was used to exclude all frequencies above this level.

3.3.5.2 Channel-Amplitude Rejection: Pipeline 1

Following these signal processing steps, each channel at the sequence level was evaluated in terms of amplitude (μV) range. Initially, the maximum and minimum μV values for each electrode were computed. These were then passed through a Boolean threshold function. In the event, that a channel presented with any values outside the following bounds: $+/-35\mu V$ the channel was identified as abnormal and was subsequently removed from all sequence data matrices. In the event, the channel rejection protocol led to the retention of 2 or fewer channels the entire trial and associated sequences were removed from the analysis.

Previous research has indicated that the standard P300 positive upward deflection ranges around 20μV [120]. This information, alongside the understanding that there are significant individual differences in the propagation of these waveforms, informed the decision to set these

μV bounding thresholds at $+/-35$μV. This level seemed to ensure that all instances of P300 waveform deflection would be captured, with the majority of the far larger movement artefact components being removed. Undoubtedly, some additional confounding signals would bypass these protocols and the researchers intended to parse these out using the aforementioned pre-processing pipeline.

Typically, EEG data rejection involves the thresholding of samples at the trial level. Here the removal of a trial due to the presence of any relevant electrodes leads to the removal of the entire trial. Given the relatively low number of trials here and the tight amplitude threshold (+/- 35 μV), the author instead performed the channel rejection at the sequence level. This dramatically increased the total number of samples retained, while removing noisy channels. The incidence of channel retention was highest for electrodes in which the P300 is maximally expressed (see, Figure 3.6). This was a conscious effort on the part of the experimenters during the EEG setup, whereby impedance levels at these sensor locations were prioritized over distal cranial positions.

As can be seen in Figure 3.6, the Pz sensor demonstrates a considerably lower incidence of channel retention. Admittedly, this could be due to the poor positioning of the EEG kit. Despite this, even after undergoing extensive training and troubleshooting, it was found that the medium-sized Cognionics headset was poorly designed to ensure consistent seating of the Pz sensor against the scalp due to its short length and lack of inward inflexion towards the skull. This issue was present in the vast majority of subjects and primarily the kit was most successful in conforming to the heads of young women, as opposed to large male skulls. Had these steps to perform channel rejection not been taken an inadequate amount of data would have been available for the training and evaluation of the models described herein.

Following these rejection protocols, all signals were averaged across trial sequences to amplify P300 waveform features. The pre-processed data is stored in an aggregate array until all 5 sequences of an experimental trial have been acquired, organised, assessed for artefacts and cleaned. At this point, 5 waveforms (across the 5 trial sequences) for each respective emoji are averaged to generate 7 individual emoji data chunks. Each chunk is then labelled in terms of spatial position and target status (i.e. cued or non-cued targets).

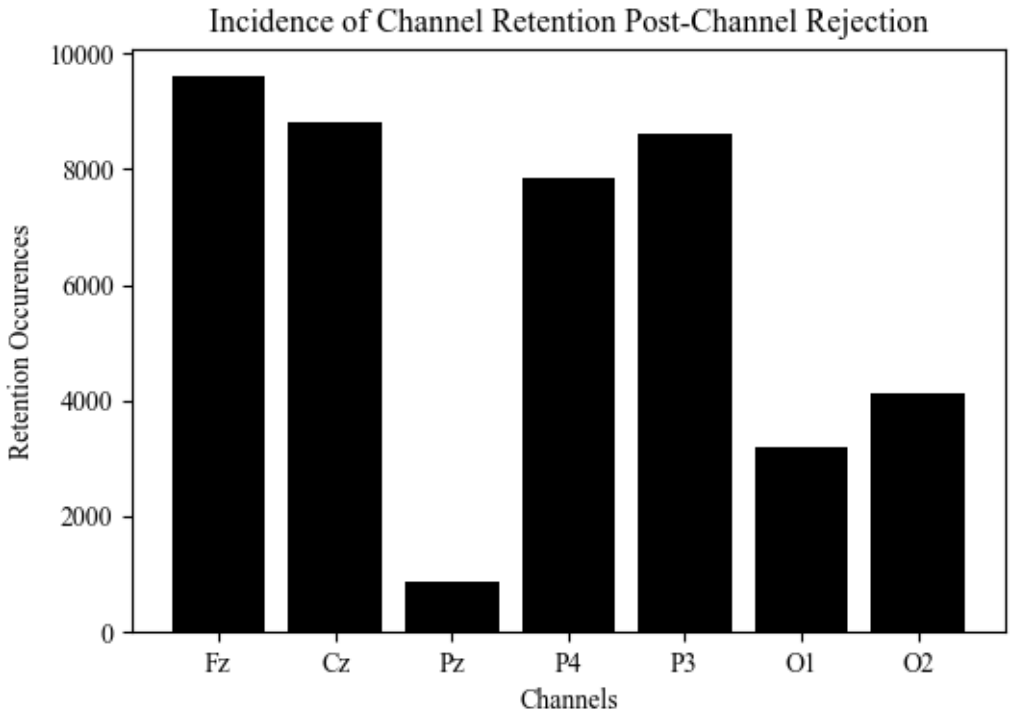

Figure 3.6: The above graph describes the incidence of EEG channel retention following the amplitude-based channel rejection method for all samples collected during the Experiment 1 data acquisition stage. These data are aggregated across all subjects for both the 'Flash' and 'Inversion' augmentation methods and form the Combined dataset (see subsection 3.5.2). The total possible number of events retained per channel amounts to 9800 (5 Sequences $\times$ 49 Trials $\times$ 4 Blocks $\times$ 10 Subjects $=$ 9800). If all channels were included for every sequence this would amount to 9800 events $\times$ 7 channels $=$ 68,600 channel retention events. The total number of times a channel was included in the analysis dataset is represented on the y-axis. The channels sampled from the headset are listed on the x-axis.As is clear from the plot there is a dramatically reduced number of Pz channel retention instances as compared to other sensor locations. The author believes that the Pz sensor mounting arm for the Cognionics Quick-20 dry headset (size: Medium), was poorly designed to seat with the appropriate tension against the skull for the majority of the subjects tested. This greatly increased the number of noise artefacts, manifesting in abnormally high impedances and amplitudes for this sensor location that ultimately led to a high degree of channel rejections. For further information please see subsection 3.3.5.2 Channel-Amplitude Rejection.

3.3.5.3 Data Pre-Processing: Pipeline 2

As noted above, a secondary data preparation methodology, Pipeline 2, was implemented for these same data to address the shortcomings of Pipeline 1. To avoid excessive repetition of information, only the differences in these two methods are discussed here, for instances where the methods are shared please refer back to the previous Pipeline 1 subsection. As seen in Table 3.1, the differences between the methods are fairly extensive.Firstly, data zeroing is not

applied in this series of pre-processing steps. The application of a high-pass finite-impulse response filter to remove ultra-low frequencies from the EEG signal effectively reduces the signal offset from zero caused by large drifting components in this frequency range. Additionally, the author has implemented a dedicated baselining method. As stated in subsection 3.3.4 Data Acquisition, samples were only collected from the initiation of the trial. On reflection, the author thought it necessary to utilize data points from the first 50ms of each trial for baselining to remove DC drift, normalize the data around a common reference point and improve the overall signal-to-noise ratio.

This involves computing the average of the first 50ms in each data sequence for every EEG channel separately. These average values are then subtracted from all samples in each respective EEG channel. In tandem, these two steps effectively remove the majority of large drift components and therefore make the application of zeroing redundant.This should not influence the quality of the resulting P300 averages given that the most important waveform components for oddball paradigm-derived Event-Related Potentials are the characteristic negative drop in micro-voltage around 200ms and a positive deflection around 300ms.

Method	Pipeline 1	Pipeline 2
Zeroing	*X*	
Grounding	*X*	*X*
Notch Filtering: 50Hz (Powerline)	*X*	*X*
Notch Filtering: SSVEP Removal		*X*
Filtering Method: IIR zero-phase	*X*	
Filtering Method: FIR zero-phase		*X*
High-Pass Filter: 0.1Hz		*X*
High-Pass Filter: 1Hz	*X*	
Low-Pass Filter: 15Hz	*X*	*X*
Baselining: 1st 50ms Avg.		*X*
Amplitude-based Channel Rejection	*X*	*X*
Impedance-based Channel Rejection	*X*	
Cross-Channel Averaging	*X*	*X*
Num Sequences per Trial	*5*	*5 & 10*
Downsampling	*X*	
Oversampling		*X*
Pooled-Subject Classifier Training	*X*	
10-fold Cross-Validation		*X*
Localizer Pre-Training	*X*	

Table 3.1: Here is a table displaying all the pre-processing techniques implemented for the Pipeline 1 and 2 methods. This is to be used for comparative purposes to aid in the understanding of how these two approaches differ. Principally, this secondary methodology (Pipeline 1) was undertaken to address the shortfalls of Pipeline 1, especially with regards to the baselining methodology, new high-pass filter cutoff value, implementation of 10-fold cross-validation (see subsection 3.4.3.1 Cross-Validation), oversampling via SMOTE data interpolation (see subsection 3.4.3.3 Oversampling via SMOTE), and cross-trial collapsing to artificially boost the number of augmentations per trial.

Further, I have transitioned from implementing an Infinite Impulse Response (IIR) filter to a Finite Impulse Response (FIR) filter, utilizing a zero-phase filtering method [227]. This FIR filtering technique is less susceptible to inducing unwanted reflection artefacts and other edge effects. Additionally, the high-pass filter implemented to remove ultra-low frequency drift was dropped from increased from 0.1 to 1Hz to more effectively capture all waveform components related to the P300 bio-signal.Moreover, the continuous and regular presentation of visual stimulus augmentations at intervals of 125ms (see subsection 3.3.3: Stimulus Presentation, Figure 3.3) will likely lead to the elicitation of a corresponding SSVEP waveform.

This should be around 8Hz, given $1second \,/\, 8peaks = 125ms$. I have addressed this via the application of an additional notch filter centred in this area of the frequency space. Note, that the stimulus variants detailed in Experiments 2 and 3 (see subsections 4.2 & 5.4.3) feature modifications to the augmentation onset intervals, here the SSVEP-targetted notch filters are adjusted to accommodate for this please see the appropriate subsections for further information (4.3.7 & 5.4.10). Further, these signals were removed in order to prevent the occurrence of SSVEPs in both the Target and Non-Target class samples from presenting with strong similar oscillatory characteristics. In the event that both these classes feature a high incidence of 8 Hz waveforms it is likely that the corresponding classifiers will group these signals together. In order to maximise the separability of these signals, the notch filter was applied to all samples.

As is noted in the final paragraph of subsection 3.3.3: Stimulus Presentation for the Pipeline 1 method, 5 sequences were collected per trial, aggregated and transformed into trial-level averages. The erroneous decision to limit the number of sequences to 5 per trial was made in pursuit of maximising the information transfer rate of the system, to the detriment of the target and non-target averages, class separability and in turn, the classification accuracies. To clarify again, the use of this number of sequences per trial is not in line with standard P300-speller design principles. Typically 10 or 15 augmentations are used initially with stimulus parameters such as the inter-trial intervals and number of augmentations per stimuli being modified following sustained maximal performance by the user[135].

Here, to address the relatively low number of augmentation events per emoji for each trial (5) I have collapsed the signals of each neighbouring trial. This involves first computing the average of all target and non-target events corresponding to the respective emoji stimuli augmentations within a single trial. This is then repeated for the next trial. Following this, the two trial-level averages for the target samples are averaged together. This process is then repeated for the remaining non-target trials. The pairing of non-target segments, corresponding to the non-cued emoji across the two trials was implemented based on the spatial positioning of the emoji onscreen. This method of aggregating and averaging was performed in an attempt to more closely mirror the current methods implemented in alternative widely used P300-speller applications [228].

Further, this should increase the relative quality of the individual trial averages, boosting the embedded P300 features and reducing noise artefacts. Both the Inversion and Flash methods were tested over 2 blocks, comprising 49 trials each. Hence, for each subject, following this collapsing procedure, just 49 trials remained for training and evaluation purposes. The rela-

tive influence of this procedure is compared against a Non-Collapsed variant for each set of experimental data evaluated. Note, all instances of the simulated increase in augmentations per emoji are referred to as following the 'Collapsed Approach', the original implementation using the standard 5 sequences per emoji average is referred to as the 'Non-Collapsed' approach.

3.4 Analysis: Pipeline 1

Here is provided all information relating to the analyses of data collected during Experiment 1 following the original Pipeline 1 approach. These data were all analysed using the Linear Discriminant Analysis (LDA) classification method. This involves building a discriminant function from data features that are capable of parsing the two target classes (P300 *vs.* Non-P300 waveforms) with the highest degree of separability [229]. Typically, the most optimal outcomes are achieved for classes with high covariance and feature vectors with normal distributions. The volatile characteristics of EEG time series and associated target waveforms (non-stationarity) in combination with cross-channel noise often lead to these analysis prerequisites being unfulfilled and a given LDA classifier can easily overfit [230]. To accommodate for these issues spatial filters can be applied to address the balance in variance within the underlying class groupings identified. The implementation of these methods alongside artefact removal protocols can in some instances enhance class separability and ultimately improve classification accuracies.

This can be accomplished by standardized bandpass filtering or via the use of dimensionality reduction methods such as Principal Components Analysis (PCA) and Independent Components Analysis (ICA). These techniques involve the decomposition of numerous variables into distinct data groupings that characterize the original data in a condensed format to increase the efficacy of subsequent feature selection or extraction tasks. The PCA method assumes zero correlation between underlying data components in the spatial or temporal domains. This differs from ICA which operates under the principle that any given subset of components demonstrates maximal statistical independence in just one domain, for example, time or space exclusively [231].

The outcome of ICA is to decompose for example an input signal into a subset of independent source signals where these sub-signals are maximally separated [232]. Note, that the improper application of these methods can reduce the prevalence of target waveform expression in the respective input signals and potentially lead to model overfitting [233]. In all subsequent ex-

perimental implementations defined herein, no dimensionality reduction method was applied in the final pre-processing pipeline as the inclusion of these steps either had negligible or detrimental effects on the final performance metrics produced.

The LDA method is currently positioned as the gold standard in P300 analysis (see subsection 2.5.1) and for these reasons, this method was re-implemented here using the Python Scikit-learn library [227]. The data evaluated herein could be broadly categorised along two key parameters, augmentation style (Flash *vs.* Inversion *vs.* Combined) and subject inclusion (Pooled / Single-Subejct). A grid search method was implemented to determine the optimal LDA configuration of solver type and shrinkage factor to maximise the average performance accuracy. Specifically, this process was introduced to tune both the solver type and the shrinkage factor. Of all solver types evaluated, the Least Squares (lsqr) method was found to be optimal in all test instances.

This involves the computation of a line of best fit in which the error, the distance of each sample point from the separating line, is maximally reduced. In the wider context of the LDA method, the categorization of a given emoji-level data chunk as either a P300 or Non-P300 waveform is dependent on its relative positioning above or below this line (for further information see, [234]). In this context, shrinkage is a regularization method used to enhance the covariance matrix estimations on which the given LDA solver operates to parse the target classes. This is especially useful for instances in which there is a low number of data samples and can assist in boosting classification accuracies as well as model generalizability. The application of the shrinkage factor is scaled between 0, no shrinkage, and 1.0, maximal shrinkage.

3.4.1 Downsampling Class Balancing Considerations: Pipeline 1

Due to numerous factors that are discussed at length in the corresponding conclusions and reflections subsections (see, 3.7.1 & 3.8.3), the aforementioned non-class balanced data partitions produced highly overfit LDA models. The initial results reported herein relate to the analyses of these non-class balanced data and are restricted solely to the Flash augmentation method. This is owing to the highly similar performance of all models for all augmentation variants and both the cross and single-subject levels relating to this data configuration. To see more information, figures and discussion relating to the Inversion and Combined non-class balanced datasets see, Appendix, Figures A.1-11.

In response to the aforementioned findings, the author redirected this study towards inves-

tigating the impact of class balancing on LDA classifier performance. To these ends, each dataset was reprocessed to ensure the number of P300 and Non-P300 waveforms were equal. The class-balanced Non-P300 events were selected according to a ranking method. One non-target emoji was chosen based on an aggregate value assigned to each of the stimuli. For each emoji, a value was computed to define its distance from the target emoji in both time and space. For instance, the P300 target emoji may have been on the extreme left of the array (spatial position 0) (see, Figure 3.1). Following this example, if the Non-P300 target emoji was positioned at the extreme right of the array it would be assigned a spatial distance value of 6, as it is six emoji from the location of the target emoji.

Further, the temporal distance values were computed according to the presentation order of the augmentations. For instance, the example trial sequence may involve the P300 target emoji being augmented first in the sequence. The Non-P300 target emoji, spatially positioned at the extreme right, may have been augmented 2nd. This means it would be assigned a temporal position value of 1, as it is just a single time unit's distance from the P300 target emoji. The temporal position value for each non-target emoji changes over the course of each trial, as every trial consists of 5 sequences, each operating according to unique non-consecutive randomised augmentation instructions. Following this, each non-target emoji possesses 5 temporal positioning values which are then averaged. The spatial positioning values and averaged temporal positioning values are then summed to create the aggregate value. The emoji demonstrating the highest compound positioning value is selected to populate the Non-P300 data matrix.

This was done to maximise the temporal and spatial distance between the P300 and Non-P300 datasets to reduce the incidence of adjacency error or double-flash artefacts (as mentioned above) from introducing undesirable elements into the class-balanced dataset. These dataset partitioning methods also influence the presentation of data throughout the chapter. It must be noted that regarding the non-class-balanced data results all Non-P300 events are used to compute the signal averages shown (see, Figure 3.2). The class-balanced data averages utilize Non-P300 data events that correspond exclusively to one non-cued emoji stimuli (see, Figure 3.3).

3.4.2 Random Performance Thresholds: Pipeline 1

For the experiment in question, all original datasets possess a P300:Non-P300 ratio of 1:6, as the total array of emojis presented is 7, with one stimulus randomly selected as the target before testing. Despite this, the offline analysis defined herein does not adhere to a 14.3% random performance threshold, as per $100 \div 7$ targets. This is owing to the organisation of the

data and the classifier method implemented. Each trial leads to the collection of 6 non-target emoji averages and one corresponding cued emoji average signal containing the target P300 waveforms. All these samples are utilized in the training and assessment of the classifiers discussed herein. As the method selected for prediction is an LDA model, parsing P300 and Non-P300 trials, the performance is always evaluated in terms of a 50% random performance threshold, as effectively the 7-target array has been reduced to a binary classification task. Along these very same lines, it is for these reasons that the random performance threshold does not vary following the implementation of the class-balancing protocols.

Note, that this applies to all variants of the offline analyses discussed throughout this thesis and across all emoji array density variants, including the staggered 3, 5 and 7 Emoji stimulus assessments conducted in Experiment 2: Variable Array Density Assessments (see subsection 4.3.5). The only instance in which the random performance thresholds correspond directly to the number of targets on screen relates to the real-time assessments performed in Experiment 3: Real-Time Feedback Implementation (see subsection 5.5.3). This is because only one of the 7-emoji average signals is selected as a target emoji to inform the visual feedback function. In other words, the offline analyses can be viewed as operating at the sample level, P300 *vs.* Non-P300, in contrast to the LOCRT classifier and associated functions operate at the trial level. These differences are restated in all relevant subsections to aid results interpretation.

3.4.3 Analysis: Pipeline 2

Here are positioned all the details relating to the analysis for Pipeline 2 in relation to Experiment 1. Note, in order to avoid excessive repetition I will briefly outline the differences between the methodology implemented here as compared to Pipeline 1, for further details please refer back to subsection 3.4. For the Pipeline 2 analysis, the same LDA classifier was implemented, however, given the uniformity observed in the selection of the least squares method solver type following the grid search protocol for all instances in the Pipeline 1 analysis, this process was bypassed. Further, the absence of any meaningful relationship between the shrinkage factor values and corresponding model performance metrics led the authors to refrain from reporting these values in any of the Pipeline 2 results. In sum, the LDA classifier, using a least squares solver method was implemented in all instances noted here, for additional information on the adaptations made regarding the cross-validation, oversampling method and argumentation against a sequence-labelling approach please see below.

3.4.3.1 Cross-Validation

To more effectively evaluate the performance of all models assessed via the Pipeline 2 method a 10-fold cross-validation procedure was implemented via the scit-kit learn KFold library [227]. This involved first defining a train and test data split of 9:1, here 10% of all samples within any given fold were positioned in an isolated subset exclusively for evaluation purposes. The *k*-folding procedure implemented here initially involves a stratified shuffling of all samples and corresponding labels, into 10 class-balanced subsets. The cross-validation protocol involves isolating one subset for evaluation purposes and training on the remaining 9. This iterates across all 10 subsets and accumulates performance metrics such as classification accuracy in the process. At the end of the evaluation cycle, a mean classification accuracy is produced, along with a corresponding standard deviation to assist in assessing the stability of this metric across the 10 subdivisions tested. Here, as both of the 49 trial Flash or Inversion method blocks were collapsed during the analysis, a total of 98 target emoji samples and 588 non-target emoji samples were collected per subject per augmentation method. Here each test set *k*-fold comprised 10/98 target samples and 59/588 non-target samples, with the remaining 88 target samples and 529 non-target samples making up the remaining 9 training folds.

3.4.3.2 Statistical Tests of Significance

This section describes all statistical tests of significance used in the Pipeline 2 analysis. Specifically, the Shapiro-Wilk test, one-sample t-test, and Permutation Tests were all implemented using the Python SciPy stats library [226]. These methods are implemented for all BCI emoji-speller variants detailed throughout Chapters 2, 3 and 4 relating to the Pipeline 2 approach (for further information see subsection 3.3.5.3)

Single-Subject Assessment: One-Samples T-Test All results concerning single-subject classification accuracies in the Pipeline 2 analysis are evaluated using a one-tailed One Sample t-test [235] based on the mean accuracy calculated from 10-fold cross-validation and the associated standard deviation. Three mean accuracy metrics are assessed: Overall accuracy, Target accuracy, and Non-Target accuracy. Overall accuracy reflects the performance of the single-subject trained LDA model across all test set samples. Additionally, Target (P300) and Non-Target (Non-P300) accuracies are separately evaluated to identify potential model-specific class bias. The accuracies from each fold, which contribute to the mean values, are first tested for normality using the Shapiro-Wilk test ($p > 0.05$) [236]. The mean accuracy is then compared against a 50% random performance threshold, appropriate for this binary target *vs.* non-target classification task. The one-tailed t-test is used to determine whether the

mean accuracy significantly exceeds this threshold. A p-value significance of 0.05 is applied. All assessments described here were implemented using the Python Scipy stats t_dist library.

Between Subject Assessments: Permutation Test Here, a comparison of the mean accuracy metrics (Overall, Target, Non-Target) across subjects against the random performance level of 50% is conducted using the Permutation Test. The permutation test is a non-parametric statistical method used to assess the significance of observed differences between groups by evaluating how these differences would distribute under the null hypothesis [237]. In this context, the groups consist of the accuracy metrics (such as Overall, Target, and Non-Target) and an adjacent array representing the 50% chance threshold. Permutation refers to the process of systematically rearranging the accuracy values and their associated thresholds to generate a distribution of the test statistic under the null hypothesis, thereby determining the probability of observing the actual difference by chance [237]. Given that the data consist of mean accuracies from only three subjects (3, 5, and 8), this approach is particularly suitable because it does not rely on assumptions of normality or homogeneity of variances, which are often violated with small sample sizes. The primary assumptions of the permutation test are that the observations within each group are exchangeable under the null hypothesis and that the test statistic used is calculated appropriately. These assumptions are met as the mean accuracies are computed from independent subjects, and the permutation test inherently accounts for the exchangeability of the data under the null hypothesis. Although the permutation test is robust and does not require normality, the test's statistical power is limited by the small sample size. To address this limitation, a larger number of permutations (10,000) was employed to enhance the precision and reliability of the p-value estimation. The significance level for the test is set at 0.05.

Non-Collapsed *vs.* Collapsed Assessments: Permutation Test To analyze the paired differences between mean accuracy values for the Non-Collapsed (5 sequences per averaged sample) and Collapsed (10 sequences per averaged sample) data partitions, the Permutation Test is also employed here. Unlike the Permutation Test described above, involving the comparison of mean accuracy metrics against a 50% chance threshold across all subjects, this paired permutation test focuses on the differences between conditions for the same subjects. Initially, the observed differences in mean accuracy between the Non-Collapsed and Collapsed conditions are calculated for each subject. The same permutation process is employed by rearranging the paired condition labels for each subject and recalculating each respective test statistic to produce a distribution of test statistic values. A comparison of the observed test statistic to this distribution produces a p-value showing the likelihood of observing such dif-

ferences by chance [237]. Notably, this approach allows for the comparison of conditions without relying on assumptions of normality or homogeneity of variances. Again, given the small sample size, the statistical power of the test is inherently limited. This is mitigated by performing a large number of permutations (10,000). The significance level for this test is set at 0.05.

Flash *vs.* Inversion Augmentation Method: Permutation Test In a similar vein, the performance comparison between the Flash and Inversion augmentation methods employs the same paired permutation test methodology. This comparison examines the overall mean accuracies for the same data partition (Non-Collapsed or Collapsed) across subjects. For each subject, paired accuracy values from the Flash and Inversion stimulus methods are compared, mirroring the approach used in the Non-Collapsed *vs.* Collapsed analysis. Specifically, the permutation test is applied to the mean accuracy values obtained from 10-fold cross-validation for both augmentation methods (see subsection 3.4.3.1). By applying this statistical approach, the goal is to determine whether there is any significant difference in classification performance between the two augmentation techniques. The repeated permutation process evaluates whether the observed differences in mean accuracy could have arisen by chance, with the p-value providing insight into the statistical significance of any observed differences. This methodology ensures a reliable comparison between Flash and Inversion methods without the need for strict distributional assumptions, while still maintaining robust analysis even with limited subject data.

3.4.3.3 Oversampling via SMOTE

As noted above in Pipeline 1, I implemented a downsampling method to address the significant class imbalance (1:6) between the target (cued) emoji samples and the non-target (non-cued) emoji samples. This method effectively mitigates the issues surrounding classifier overfitting, despite this the concurrent drop in the number of samples available to the classifier for training also introduces significant issues as the amount of data used to define the separation of these classes has been dramatically reduced. To try and address the class imbalance without the aforementioned drop in the total number of training samples the author has implemented an over-sampling method. This involves artificially boosting the representation of the minority class [238], in this instance the Target P300 data samples.

The most basic implementation of this technique involves copying existing samples and the corresponding labels and aggregating them into the training set. This highly simplistic application is extremely susceptible to classifier overfitting towards the minority class, as the

variance within the oversampled class is artificially lower than the majority class, it reduces the complexity of defining these samples within the given search space [238]. To address this data-driven methods such as the Synthetic Minority Over-sampling Technique [239] (SMOTE) have been widely adopted. This method generates novel minority class data by interpolating samples within the minority class and in some instances, reducing the prevalence of the majority class.

In this thesis, the technique was implemented via the widely adopted ImLearn Python library [240]. Here the default parameters were utilized in all instances. This involved upsampling the number of target samples such that the quantity of the minority Target class reached parity with the Non-Target samples. Here, the default ImLearn SMOTE settings concerning interpolation were utilized. This involves determining the 5 k-nearest neighbours of the target sample and interpolating this original signal with the neighbouring samples. The degree of linear interpolation between the target sample and each of the 5 nearest neighbours is executed according to a uniform random distribution between 0 and 1. Here, a value of 0 indicates that essentially all of the original target signal is retained and 1 denotes that following the interpolation, the signal would effectively be identical to the nearest neighbour. This method of oversampling introduces diversity into the newly generated samples while also ensuring that the synthetic data contain features represented within the target class.

Note, that these methods were applied exclusively after the segregation of the data into train and test cross-validation k-folds. At no point was any synthetic data positioned inside the test set for any of the cross-validation assessments. Here the 88 target emoji samples assigned for training were oversampled to produce 441 synthetic samples, to reach a total of 529 samples, equally the quantity of data in the non-target class training subset. Following this data augmentation, 83.3% of all samples relating to the target class for any given k-fold were comprised of synthetic data. It could be argued that the author could have implemented majority class undersampling to reduce this high intra-clas ratio of Real:Synthetic data. It is important to state that the functionality of the ImLearn SMOTE library does not extend to the implementation of majority class undersampling.

Further given the significant time constraints placed on this project, the exploration of this additional methodology without a clear parameter threshold positioned this investigation beyond its scope. Moreover, the use of the previously outlined cross-validation method, in tandem with the exclusive testing of models on novel, non-synthetic data should provide some reassurance during the evaluation of the models discussed herein. Note, that in the collapsed anal-

ysis variants involving the averaging of data across neighbouring trials to simulate a relative increase in the number of augmentation sequences per trial from 5 to 10, the ratios between these data divisions all remain consistent.

3.4.3.4 Sequence-Labelling

Given the substantial considerations made in this thesis to address issues relating to class balance, the author believes that it is important to assert their reasoning for not implementing a sequence re-labelling method. This technique forgoes the individual labelling of time-locked sequence segments as belonging to target or non-target emoji stimulus augmentations. Here a binary target *vs.* non-target problem is transformed into a 7-class problem where the onset of the target emoji stimulus augmentation would be used directly to label the data. In other words, for a given sequence where the cued target emoji is augmented 1st, this would be assigned a label of 1, if augmented 7th, this would be given a label of 7.

The label would apply to the entire non-segmented sequence during which all stimuli in the array are augmented once. In the original method, a time window of averaged data around 375ms in duration is passed to the classifier to determine the presence of a time-locked negative component at 200ms and a positive deflection around 300ms. In this alternative sequence-labelled method, the entire sequence period would be passed comprising well over 1 second of data. Here the complexity of the classification problem increases dramatically. In the original iteration, following sequence averaging the classifier is trained to determine the absence or presence of P300-like features, in contrast to the sequence-relabelled method the classifier must distinguish P300-like features at 7 different temporal locations. In other words, the number of trials per class is dramatically lower than the binary classification method while also being a far more complex classification problem.

	Onset 1	Onset 2	Onset 3	Onset 4	Onset 5	Onset 6	Onset 7	Total
Raw Counts	63	73	54	68	75	83	74	490
Multiple Rounded	60	70	50	65	75	80	70	470
5 Sequence Avg.	12	14	10	13	15	16	14	94
Collapsed	6	7	5	6	7	8	7	46

Table 3.2: Here are shown the augmentation onest counts relating to Subject 3 for both Flash augmentation variant blocks, consisting of 49 trials each. The Onset monikers denote the temporal position of each respective sequence target emoji augmentation. As seen in the Raw Counts column for Onset 1, 63 inter-trial sequences had a presentation scheme that featured the target emoji being augmented first out of all 7 emojis onscreen. In other words, as soon as the trial began the target cued emoji, which could be positioned in any of the 7 emoji onscreen spatial positions, was augmented first. As discussed in subsection 3.3.3 Stimulus Presentation, there was a 125ms interval between the onset of the stimulus augmentations, this is the amount of time separating each temporal onset label detailed here. The Multiple Rounded row denotes the closest multiple of the number of sequences involved in the cross-trial averaging and the raw count when rounded down. Here there are 63 raw count samples, the intention here is to demonstrate the process for a 5-sequence average, here 60 is the closest factor as it is only possible to round down. This is because all sample averages must have the same number of sequences. Finally, the 'Collapsed' row relates to the number of samples retained following the averaging between neighbouring samples, this was done to simulate increasing the number of augmentations per sequence to improve the average plots and increase target vs. non-target separability.

Further, instead of 7 separate samples being returned per sequence, just 1 sample is returned using this method. It might be assumed that given 1 label is provided per sequence, as opposed to 1 target and 6 non-target labels, and that 7 emoji are augmented over a multiple of 7, 49 trials, the data would be perfectly class balanced. Here 49 trials with 5 sequences each would theoretically return 245 samples, with 35 samples for each of the 7 classes and 70 samples each over the 2 blocks per augmentation variant. However, as noted in subsection 3.3.3 Stimulus Presentation, the stimuli augmentation protocol followed a non-consecutive randomisation procedure. This ensured that the augmentation of a given emoji was never followed by the augmentation of itself or its direct spatial neighbour. These were the only constraints placed on the presentation scheme, no accommodations were made to enforce the even distribution of temporal onset times for the target emoji augmentations.

This in tandem with the fact that the number of sequences per trial (5) is smaller than the number of emoji onscreen means that there is a high probability there will be an uneven distribution of class samples across the 7 temporal onset labels. As seen above in Table 3.2, this is well evidenced by the Flash augmentation trials collected from Subject 3. The respective

counts demonstrate substantial variability in samples per class. Note, that to effectively average across these samples it would be necessary to find the closest factor of the number of sequences per average and round this down, for instance, Subject 3 Onset 1 has 63 instances. Here, as seen in Table 3.2 if using 5 sequences per average we would need to round this down to 60. Following this, the remaining sequences are averaged to produce 12 cross-sequence averaged samples.

Finally, as stated above, the author intended to assess the relative contribution of artificially increasing the number of augmentation sequences per sample. Here, this would involve collapsing neighbouring sequences together to simulate the collection of 10 augmentation events per emoji. As is seen in the final row (see Table 3.2) the total number of trials in the dataset is just 47 samples. This is without even applying the train-test split, meaning around just 42 samples for training and 4 samples for validation per subject. Notably, following this approach, a lower number of total trials are available here than in the highly restrictive downsampling method implemented in the Pipeline 1 approach (see subsection 3.5.2), and the complexity of the problem has been dramatically increased (binary to 7-class problem).

3.4.3.5 Onset-Labelling

In contrast to the sequence-labelling method described above, if the data were to be cut into segments and labelled with the temporal onsets as a function of distance from the onset of the target augmentation event, a large imbalance in classes would necessarily emerge in addition to significant averaging complications. For example, assume the cued emoji is augmented first, here the first 375ms time chunk would be labelled as 1, indicating that this data chunk occurred immediately after the onset of the Flash or Inversion stimulus augmentation (see, Table 3.3, first row). All subsequent time windows would be assigned the labels 1, 2, 3, 4, 5, 6 and 7. Then assume a trial for which the cued emoji is augmented last, the labels would hence be -7, -6, -5, -4, -3, -2 and -1, as all previous augmentations occurred before the onset of the target augmentation. Feasibly, these could converted to absolute values if all assumptions regarding the expected waveform behaviour before, but most importantly, following the onset of the the cued time chunk are ignored. A table is provided (see, Table 3.4) to illustrate that the labelling of sequence data segments following this method would lead to an accumulation of labels for the data windows neighbouring (here labelled as 2) the onset of the target augmentation event (here labelled as 1).

Further, given the refractory effects associated with the P300 waveform, it is likely that during the aggregation of samples positioned the same distance from the onset, but that occur

either before or after the target augmentation e.g. 2 *vs*. -2, substantial variance would be present within these groups (see Table 3.3, Onset 2). To avoid this, the explicit labelling of time segments as occurring before or after the target augmentation onset would necessarily lead to a total of 13 different labels for this 7-Emoji array. Additionally, it would not be possible to compute label averages as the presentation scheme used did not adhere to perfect randomization. As noted above, the scheme was developed to avoid spatial double flashing and adjacency errors, therefore in many instances there would be a huge imbalance in the number of time-segments needed to construct the corresponding cross-sequence averages.

Onset 1	Onset 2	Onset 3	Onset 4	Onset 5	Onset 6	Onset 7
1	2	3	4	5	6	7
2	1	2	3	4	5	6
3	2	1	2	3	4	5
4	3	2	1	2	3	4
5	4	3	2	1	2	3
6	5	4	3	2	1	2
7	6	5	4	3	2	1

Table 3.3: Here are shown the counts relating to a hypothetical temporal onset-labelling scheme for the 7-Emoji experimental variant. Here in the top row are the Onset fields representing each of the 7 time-locked onsets of the 7 on-screen emoji through the experimental sequence, each spaced 125ms apart. The values in the main portion of the table denote the relative temporal distance of the given onset time to the actual target cued stimulus augmentation event. Here, 1 indicates that the time window corresponds directly to the onset of the target stimulus, and a value of 2 indicates that the data window directly neighbours the onset of the target stimulus augmentation, either immediately before or after the event. Along these lines, a 7 is only achieved if the target stimulus is augmented at the very start of a sequence or the very end. Here all possible permutations of the relative temporal position of the target augmentation onset are presented, note that these do not represent actual subject data and are only provided to explain the class imbalance inherent to this method.

	Onset 1	Onset 2	Onset 3	Onset 4	Onset 5	Onset 6	Onset 7	Total
Counts	7	12	10	8	6	4	2	49

Table 3.4: Here are shown the counts of the labels accumulated via the hypothetical onset-labelling example positioned above. The counts refer to the number of times the given onset label appears in Table 3.3.

The author hopes that through these explanations it is clear the sequence and onset-labelling approaches to data labelling are not viable given the inherent class-imbalance and complica-

tions relating to sequence averaging. Further, the original intent of these implementations, class-balancing, is not achievable without additional synthetic data generation and this would add additional complexity to an already highly involved data preparation process. Moreover, both techniques involve training a classifier for either a 7 or potentially 13 class problem, as opposed to the original binary, Target (oddball) *vs*. Non-Target task. In sum, the decision has been made not to pursue these investigations as it is extremely likely that the resultant classifiers will not have sufficient data to effectively train relative to the complexity of the class separation task at hand.

3.5 Results: Pipeline 1

Here are presented all results relating to Experiment 1. This comprises an investigation into the efficacy of the Flash and Inversion augmentation methods (see, Figure 3.2) for eliciting the propagation of visual P300 waveforms in a 7-target emoji-based BCI speller format.

3.5.1 Post-Processing Data Info: Pipeline 1

If all subject trials are included (see subsection 3.5.2) the dataset would constitute 13720 events, comprising a total of 1960 P300 events and 11760 Non-P300 events. A train-test data partition was performed to reserve 10% of the total number of events for the evaluation of each respective data partition. Similarly, if all trials for one subject are retained, the test dataset constitutes 196 P300 test events and 1176 Non-P300 test events. Again, within each data partition, the samples are further broken down into pooled-subject (aggregated data across all subjects) and single-subject sets to probe for the effects of data aggregation on LDA model performance. All data used in the evaluation stage of the pooled-subject subset is composed of 10% of the samples from each subject tested, hence all associated classifiers are termed pooled-subjects classifiers.

3.5.2 Data Partitions: Pipeline 1

The following subsection is intended to outline the reasoning behind the use of multiple data subsets. Firstly, the principle aim of this experiment was to explore if there were significant differences in classification performance between the two augmentation methods: Flash and Inversion (see subsection 3.3.3). Additionally, the author intended to observe any significant changes in classification performance if the data collected across the two augmentation methods were combined.

This is not typically performed with P300 data, as there are numerous individual differences in the propagation of the waveform. These differences can even pose obstacles to data aggregation for the same subject during the same session, as the P300 is an attentional top-down component, meaning small changes in fatigue or concentration can have dramatic influences on waveform expression. In contrast, previous studies have explored this process of pooled-subject data aggregation and observed significant improvements in classifier performance [58].

Dataset	Class Balancing	Inversion Data	Flash Data	Total Num Events	Num Test Events	Test Events Post-Rejection
Flash No Balance	FALSE	FALSE	TRUE	6860	686	685
Inversion No Balance	FALSE	TRUE	FALSE	6860	686	679
Combined No Balance	FALSE	TRUE	TRUE	13720	1372	1364
Flash Balance	TRUE	FALSE	TRUE	1960	196	198
Inversion Balance	TRUE	TRUE	FALSE	1960	196	198
Combined Balance	TRUE	TRUE	TRUE	3920	392	392

Table 3.5: This is a table key for denoting the data partitions used to investigate the classification across and within subjects for the individual Flash and Inversion augmentation methods in addition to the Combined dataset that is comprised of signals across both methods. The first column provides a moniker for each data partition. The second column refers to the presence (TRUE) or absence (FALSE) of class balancing (see subsection 3.4.1). The following Inversion and Flash columns are presented to indicate the augmentation method used to drive the signals comprising the data partition in question. The Num Events column details the total number of events held in the data partitions listed. Note that the Inversion No Balance and Combined No Balance datasets are not discussed in the main body of the thesis. This is because the associated results are highly similar to those collected for the Flash No Balance dataset. To avoid excessive repetition, these results are positioned in the Appendix A.1 & A.2.

It must be noted that in some instances (e.g., Flash Balanced and Inversion Balanced), datasets (see, Table 3.5) present with a higher number of test events post-rejection than the previously outlined maximum number of test events. This is shown in the first row for the Flash Balanced dataset which highlights the maximum number of test events possible if 10% of all trials are subsampled for evaluation purposes, at 196. In the immediate right-hand column, the actual number of test events included post-channel rejection is 198. This is because the data partitioning was performed at the single-subject level. For each subject, a total of

980 trials were collected for the P300 and selected Non-P300 event samples constituting the class-balanced sample. In sum, this comprises 1960 data events. After this data aggregation step, partitions between training and test samples are implemented. This involved isolating 10% of the total dataset for evaluation purposes. As 10% of 196 is a non-integer value (19.6) the functions utilized to perform this operation default to rounding this pre-determined value up to 20.

As will be seen in subsequent sections (Table 3.7), some subjects did not return 20 events for the 10% test set data partition. This is due to certain subjects providing data with a higher incidence of channel rejection. As described above, if a single trial presented with 5 or more channels demonstrating μV amplitude volatility outside the defined thresholds (see subsection 3.3.5.2) the entire trial was not included in the final dataset. This led to some subjects presenting with a total number of test events below the 196 test event maximum. In these instances, if the subjects returned less than 195 events (e.g., 194 events) then the class-balancing protocol would retrieve 10% of this trial data and round down the non-integer (e.g. 19.4) to 19. This should clarify the presence of additional samples in the post-channel rejection column. Note, that these slight imbalances in the exact proportions of class events per data partition do not influence the random-performance thresholds discussed.

3.5.3 Flash Method Results: No Class Balancing: Pipeline 1

The results reported in this subsection refer to analyses undertaken on the Flash No Balance data partition (refer to Table 3.5). All data organisation, pre-processing and analysis was conducted using the Pipeline 1 approach (see subsection 3.3.5). These data contain all subject time-series gathered during the concurrent visual presentation of the Flash method of stimulus augmentation. The ratio of P300 events:Non-P300 events is 1:6. These analyses represent an ecologically valid means of assessing the functionality and robustness of the experimental design. Note, that all analyses were undertaken offline. It must be highlighted that these are the only non-class balanced results reported in the main body of the text. As noted in the above table (Table 3.5), this was done to avoid the excessive repetition of highly similar results and associated interpretations. All analyses relating to the Inversion No Balance and Combined No Balanced datasets can be found in Appendix: A.1 and A.2.

3.5.3.1 Pooled-Subject

As can be seen in Table 3.6, the LDA classifier trained using the pooled-subject data achieved greater than random performance (>50%). To clarify, these LDA models were trained using

samples pooled across all subjects tested and were evaluated using isolated test data (10%) from all subjects. Mean accuracy is high (83.60%), with a clear imbalance in performance metrics for P300 events (MA=0%) as compared to Non-P300 events (MA = 100%). When applying the grid search technique, the optimal combination of solver and shrinkage value was shown to be the lsqr method at 0.01, respectively.

	Mean Accuracy (%)	P300 Accuracy (%)	Non-P300 Accuracy (%)	Solver	Shrinkage	Num Test Events
Pooled Subjects	83.6	0	100	lsqr	0.01	685

Table 3.6: A table of classification performance and optimization results for the Flash No Balance dataset (refer to, Table 3.5 for data partition info). Here, this classifier was trained on data pooled from all 10 subjects. The training was comprised of the first 90% of all samples from each subject, with the remaining 10% of test data being the final samples collected for each respective subject. The Mean Accuracy column refers to the overall classification performance of the trained LDA models across the respective test dataset. The P300 Accuracy column denotes the classification performance exclusively for P300 event predictions (otherwise known as 'hits'). The Non-P300 Accuracy column details the LDA classifier performance only for the Non-P300 events comprised in the test dataset. The Solver column states the solver method selected via the grid search optimization method (see subsection 3.4). The term lsqr refers to the Least Squares (lsqr) solver method. The Shrinkage column denotes the shrinkage factor determined via the grid search method to demonstrate the highest classification performance attained. Finally, the Num Test Events column refers to the total number of events comprising the test dataset for each data partition listed.

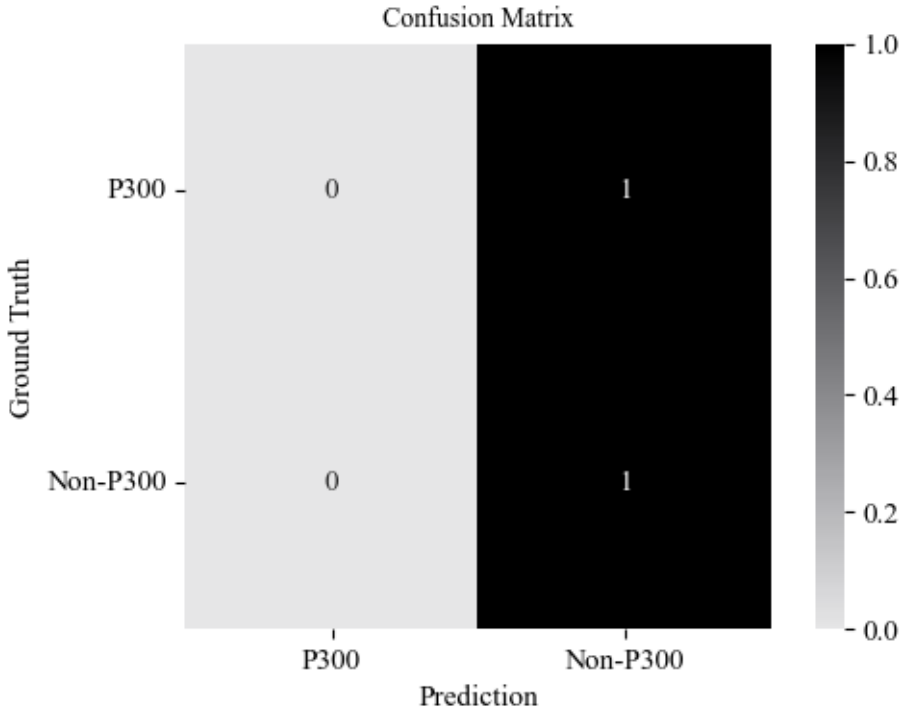

Figure 3.7: Here is displayed a confusion matrix illustrating the normalized classification performance of trained LDA models for the Flash Method No Balance pooled-subject data partition (see, Table 3.5). The normalization of classification performance percentages involves rescaling (normalizing) these values between 0 and 1. The key to the right of the matrix is presented to aid in the interpretation of classification performances, with darker shadings indicating a high incidence of prediction selection and lighter shadings indicating a lower incidence of selection. Optimal classification performance would be represented by dark squares across both the top-left and bottom-right squares (diagonal), with light-grey shadings for the opposing squares. This would represent high incidences of correctly identifying P300 events as P300 events (hits) and likewise for Non-P300 events. Further, such a plot would indicate that there is a very low incidence of confusing P300 for Non-P300 targets or vice-versa. Note, that all values in the confusion matrix sum to 2.

In the above figure (see, Figure 3.7) the confusion matrix indicates a selective bias towards the prediction of Non-P300 events in every instance evaluated. All test events were assessed by the LDA classifier as being Non-P300 events, leading to a hit score of 1 for the Non-P300 class. The hit score for P300 events was found to be zero, with significant confusion present in the classification of these target waveforms.

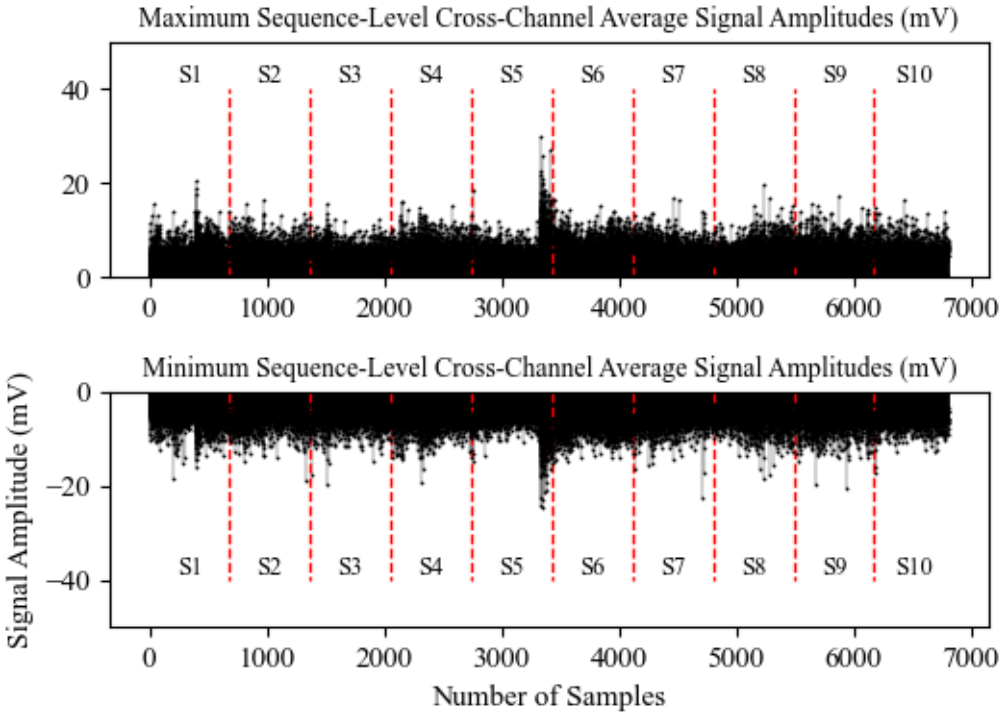

Figure 3.8: This is a stem plot of all the μV ranges (pooled-subject) of each event in the Flash No Balance data partition (see, Table 3.5). All samples are listed across the x-axis and the y-axis notes the maximum (upper plot) and minimum (lower plot) μV values for all pre-processed samples. These plots are used for data-quality assessments to eliminate the possibility that the variances in performance across subjects (S) can be attributed to significant variances in μV amplitudes. Any large variances in these values could indicate the presence of movement artefacts, improper seating of the sensors to the skull or poor subject scalp conductivity. The total number of samples collected, 6860, is the product of 7 emoji $\times$ 49 trials $\times$ 2 blocks $\times$ 10 subjects. As seen in both subplots, the relative positioning of the individual subject data inside the aggregate data array is denoted via the corresponding bounds indicated via red dotted lines and associated subject moniker. Note, that the trials comprising each data subset are organized chronologically according to the randomised presentation schedule.

Inspection of Figure 3.8 reveals a relatively uniform distribution of high and low EEG amplitude (μV) maximum and minimum values for each respective emoji sample. The vast majority of events are positioned inside a band comprising $+/-15$ μV, with a significant minority of events falling beyond these bounds. It is clear that some subject data expresses significantly higher EEG μV amplitude values with a clear correlation in the expression of extreme values across both negative and positive subsets. The large peaking and negative deflections present at the centre of the amplitudes plot are derived exclusively from trials relating to Subject 5. These results will be discussed further in the subsequent within-subject analyses subsection.

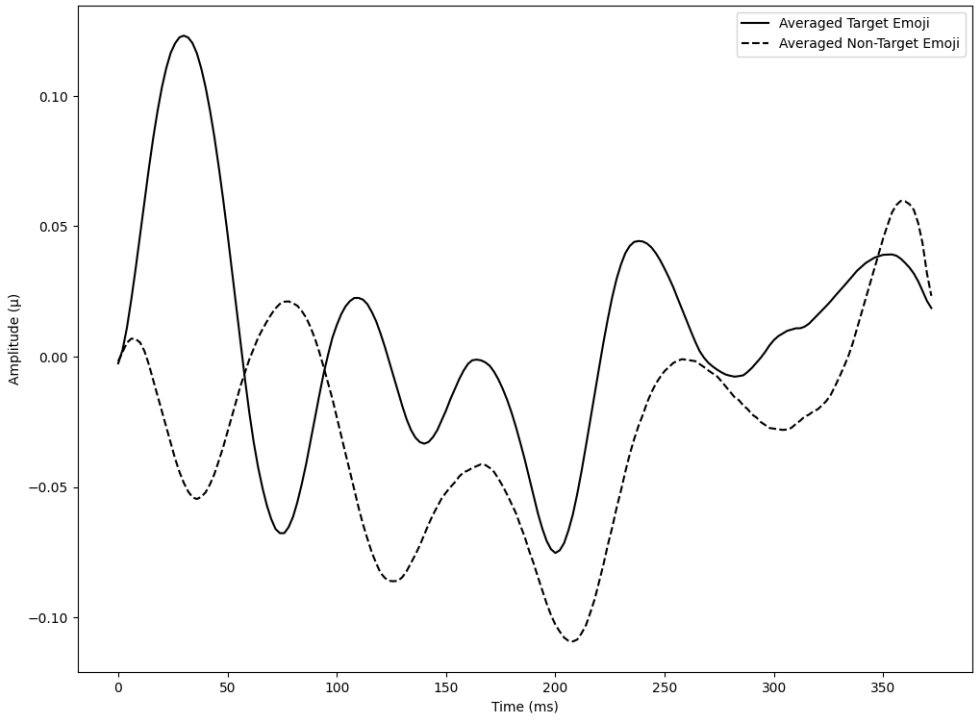

Figure 3.9: Here is shown a Cz grand average plot for all trial P300 (solid line) and Non-P300 (dashed line) events respectively collected during the Flash No Balance data partition (refer to, Table 3.5 for data partition info). Each augmentation event generated a data stream marker used to window the data into 375ms chunks. Given the onset and offset times of the augmentations through each trial sequence, some data is shared across data chunks for different emoji targets. These waveforms are computed by averaging across all P300 (relating to cued target emoji instances) or Non-P300 events (non-cued target emoji instances) and isolated exclusively to the central Cz channel. The averages generated across these classes amplify underlying EEG waveform patterns embedded in the signals. Further, all plots were baselined by computing the average of the first 50ms of the samples collected. Note, that these baselining measures were implemented exclusively for presentation purposes and were not applied during the Pipeline 1 approach to data pre-processing as stated in subsection 3.3.5.3, see Table 3.1. [241].

As shown in Figure 3.9, the P300 signal exhibits more variance than the Non-P300 average signal. There is a strong negative component at 50ms, leading to a mild positive deflection at 300-350ms. Notably, a strong initial peak and oscillatory component is visible. Further, the end-point peaking component of the Non-P300 waveform is ultimately higher. This suggests that subject's attentiveness to the target cue was insufficient for clear visual differentiation between the waveforms. Additionally, an oscillation with a periodicity of around 125ms (8Hz) is observable for the P300 averaged signal, which is likely an SSVEP resulting from the 125ms stimulus augmentation interval used in this experimental variant (see subsection 3.3.3 Stimulus Presentation).

3.5.3.2 Within-Subject

This section features the results of LDA models trained and evaluated exclusively using single-subject data (refer to, Table 3.7). There is minimal variance in classification performance across the subject-level models (86.56 +/-2.23\%). The classification performance reflects many of the same properties as the Pooled-Subject performance. The P300 accuracy metrics are all fixed at 0\%, with a selection bias towards Non-P300 class accuracy at around 100\%. All single-subject LDA classifier-optimized grid searches demonstrated a preference for the lsqr solver method. Further, the average shrinkage value (0.15) varied quite substantially (+/-0.12), ranging from 0.01-0.25 units.

	Mean Accuracy (%)	P300 Accuracy (%)	Non-P300 Accuracy (%)	Solver	Shrinkage	Num Test Events
Subject 1	86.96	0	100	lsqr	0.1	69
Subject 2	86.96	0	100	lsqr	0.24	69
Subject 3	86.96	0	100	lsqr	0.24	69
Subject 4	86.96	0	100	lsqr	0.25	69
Subject 5	88.06	0	100	lsqr	0.02	67
Subject 6	85.51	0	98.33	lsqr	0.01	69
Subject 7	85.51	0	98.33	lsqr	0.09	69
Subject 8	88.06	0	100	lsqr	0.15	67
Subject 9	85.29	0	100	lsqr	0.18	68
Subject 10	85.51	0	100	lsqr	0.22	69
Single Subject Avg.	86.58	0	99.67	n/a	0.15	68.5
Single Subject Var.	1.39	0	0.84	n/a	0.12	1

Table 3.7: A table of classification performance metrics and optimization results for the Flash No Balance dataset (refer to, Table 3.1 for data partition info) for which all model acuracy metrics were computed using single-subject data during training and evaluation. The left-hand column denotes each subject assessed, along side the Single-Subject Avg. This is a mean of all single-subject performance metrics and does not indicate the presence of any pooled data evaluations. Similarly, the Single-Subject Var. row denotes the variance in these given single-subject metrics. For further information regarding the peformance metric and parameter fields please refer back to Table 3.6.

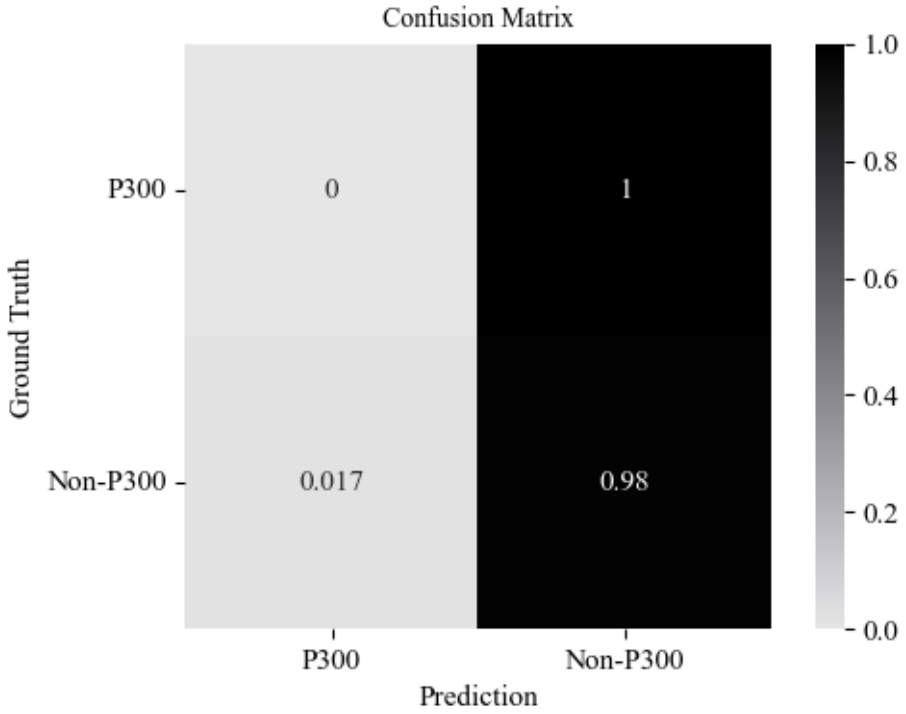

Figure 3.10: Displayed here is a confusion matrix generated via the LDA classifier results for Subject 6 relating to the Flash No Balance data partition (refer to, Table 3.5). This illustrates the incidence of Non-P300 and P300 selections via the trained model. All values above zero positioned in the top-right to bottom-left diagonal orientation indicate misclassification, otherwise referred to as confusion. In the orthogonal direction, values greater than 0 indicate the incidence of accurate classification.

Two subjects (Subjects 6 & 7) demonstrated sub-100% Non-P300 event classification accuracies (98.33%). As seen above in Figure 3.10, the confusion matrix of Subject 6 illustrates that one Non-P300 event was erroneously evaluated as a P300 event. The variance in total samples retained post-pre-processing is highly uniform (average=68.5 events, $+/-1$). As seen in Figure 3.11, for the subject with the lowest incidence of events retained (Subject 5), the EEG waveform μV amplitude ranges are mostly reflective of the pooled-subject dataset (see, Figure 3.8). Any significant and prolonged movements during the experimental period by the test subject would reveal substantially higher variances than those seen here. This data is presented here to illustrate that these results can not be ascribed to errors in data collection or noise artefacts. For additional information on the interpretation of this figure see, Figure 3.8.

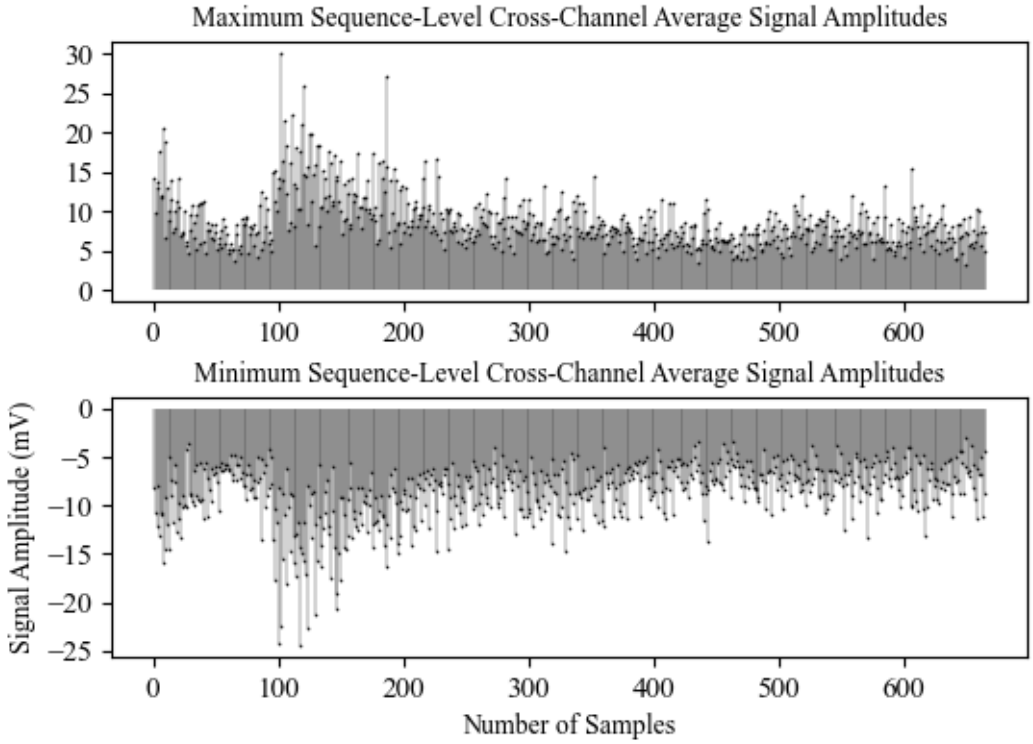

Figure 3.11: This shows the distribution of the maximum positive and negative values of all respective samples collected from Subject 5 (refer to, Table 3.7) retained post-processing. For further info on interpretation see, Figure 3.8.

3.5.4 Flash Method Results: Class-Balanced: Pipeline 1

This subsection discusses results regarding the Flash Balanced dataset (refer to, Table 3.5) computed via the Pipleine 1 approach (see subsection 3.3.5.1). These data are a class-balanced (downsampled) subset of the subject EEG time-series data collected using the Flash method of stimulus augmentation. The ratio of P300 events: Non-P300 events is 1:1. These analyses were undertaken in an attempt to mitigate the confounding influence of overfitting observed in the results described above (see, corresponding conclusion subsection 3.7.1, for further information).

The same protocol is used for the selection of comparator emojis in the average plots is used for the class balancing method as above. This involves selecting a comparator emoji with the maximal spatial and temporal separability from the cued target emoji to enhance the separability of respective features and reduce the presence of bleed-over effects in the evaluation of these data.

3.5.4.1 Pooled-Subject

	Mean Accuracy (%)	P300 Accuracy (%)	Non-P300 Accuracy (%)	Solver	Shrinkage	Num Test Events
Pooled Subject	43.26	38.98	48.45	lsqr	0.6	198

Table 3.8: A table of classification performance metrics and optimization results from the pooled-subject Flash Balanced dataset (refer to Table 3.5 for data partition info). For further info on table-field headings refer to, Tables 3.6 & 3.7.

As can be seen in Table 3.8, the LDA classifier trained using the pooled-subject data achieved lower than random performance (<50%). Mean accuracy (MA) is low (43.26%) with a clear balance in performance metrics for Non-P300 events (MA=48.45%) as compared to P300 events (MA=38.98%). When applying the grid search technique the optimal combination of solver and shrinkage values was found to be the Least Squares (lsqr) method at 0.6, respectively.

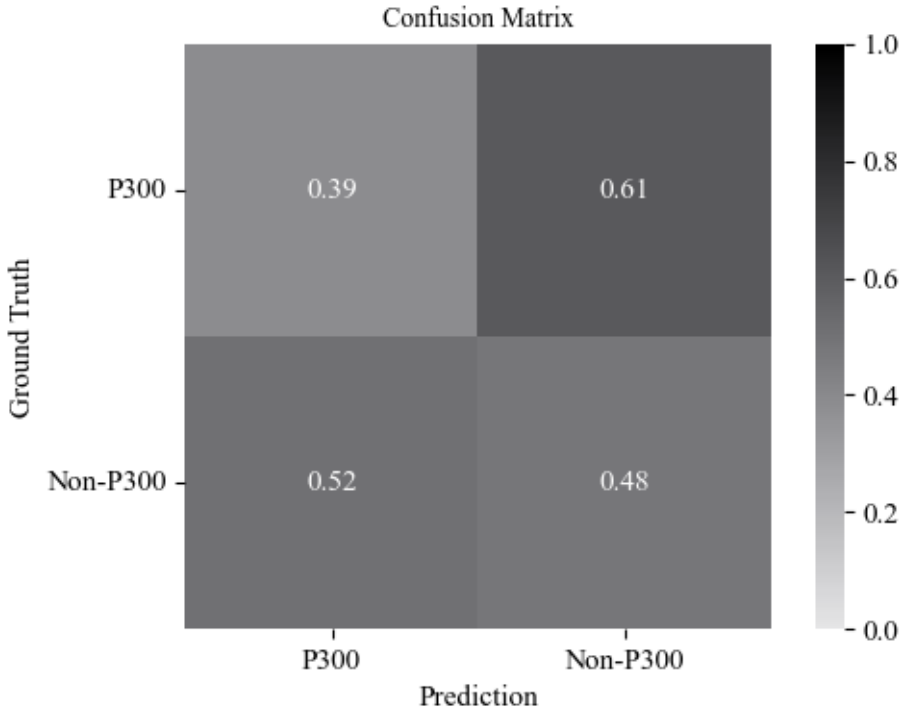

Figure 3.12: Displayed here is a confusion matrix generated via the LDA classifier results for the Pooled-Subject Flash Balanced data partition (refer to, Table 3.5). For more info on interpretation see, Figure 3.7.

In the confusion matrix above (see, Figure 3.12), the prediction preferences of the LDA model trained using the pooled-subject data are displayed. The classification of both P300 and Non-P300 targets is below the random performance threshold. Additionally, the figure illustrates that there exists a slight selection bias for the Non-P300 class resulting in the misclassification of many P300 events as belonging to this class.

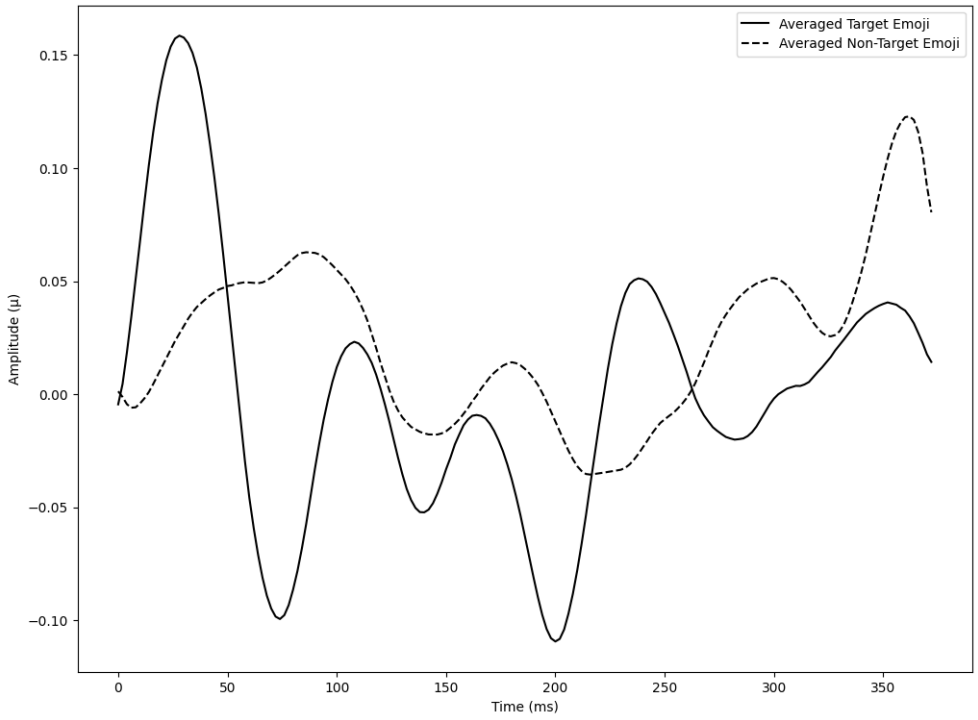

Figure 3.13: Here is shown a Cz grand average plot for all trial P300 (solid line) and Non-P300 (dashed line) events respectively collected during the computed for the cross-trial P300 (solid line) and Non-P300 (dashed line) events respectively. These waveforms were generated using the Flash Balanced data partition (refer to, Table 3.5). Note: the signals used to create this Non-P300 average all correspond to all non-cued emoji segments collected.

The cross-sample average signals (see, Figure 3.13) do not represent typical P300 and Non-P300 waveforms. The P300 average (solid line) displays a relatively large negative component at around 100ms and 200ms. Despite this, the tail end of the waveform does not contain the necessary feature of a large 300ms positive deflection. The Non-P300 average waveform presents with a relatively high range of amplitude values, alongside a mild positive deflection around 300ms.

3.5.4.2 Within-Subject

The within-subject performance averages peaked above the random performance threshold (56% $+/-$16.45%) (see, Table 3.9). Classification performance for P300 events (50.13%) is non-significantly different to the random performance threshold of 50%. Crucially, it must be noted that Subjects 3, 6 and 9 all performed well above the random classification threshold, with Subject 3 achieving a mean accuracy of 75%. The solver method tuned for each subject individually returned the lsqr method as the most optimal in all instances. In contrast, the shrinkage values differ significantly across subjects, diverging from the trend of sub 0.2

shrinkage values observed in the Flash No Balance data partition results (see Table 3.6). The variance in shrinkage is considerable across subjects (average=0.63, $+/-0.5$), spanning the entire 0-1 value scale.

	Mean Accuracy (%)	P300 Accuracy (%)	Non-P300 Accuracy (%)	Solver	Shrinkage	Num Test Events
Subject 1	55	44.44	63.64	lsqr	0.96	20
Subject 2	50	33.33	63.64	lsqr	0.99	20
Subject 3	75	77.78	72.73	lsqr	0.6	20
Subject 4	65	44.44	81.82	lsqr	0.01	20
Subject 5	57.89	42.86	66.67	lsqr	0.99	19
Subject 6	65	66.67	63.64	lsqr	0.65	20
Subject 7	45	44.44	45.45	lsqr	0.08	20
Subject 8	42.11	42.86	41.67	lsqr	0.99	19
Subject 9	60	60	60	lsqr	0	20
Subject 10	45	44.44	45.45	lsqr	0.98	20
Single Subject Avg.	56	50.13	60.47	n/a	0.63	19.8
Single Subject Var.	16.45	22.23	20.08	n/a	0.50	0.5

Table 3.9: A table of classification performance metrics and optimization results for single-subject trained and evaluated LDA models using the Flash Balanced dataset (refer to Table 3.5 for data partition info). For further info on table-field headings refer to, Tables 3.6 & 3.7.

As can be seen in Table 3.9, the variance in total samples retained post-pre-processing is highly uniform (average=19.8 events, $+/-0.5$). This is further reinforced by the plot positioned above (see, Figure 3.14), as a significant majority of samples present are well below a $+/-20$ μV boundary, despite the $+/-35$ μV amplitude threshold employed. Some increased volatility in samples can be observed for Subjects 3 and 8, these will both be discussed further in the corresponding conclusion section (see subsection 3.7.1).

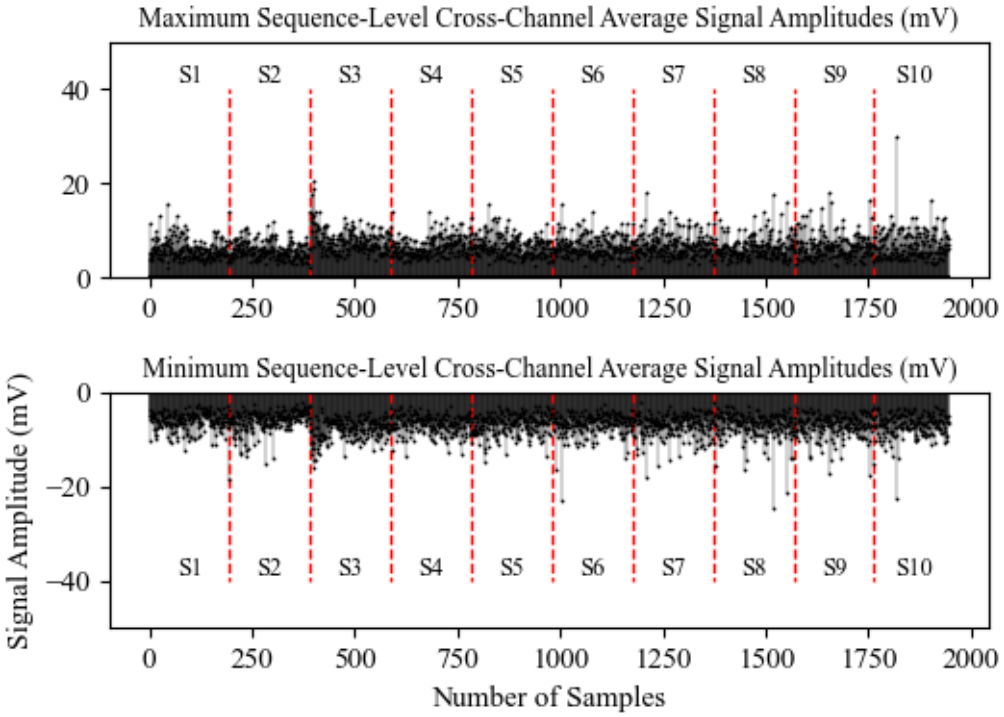

Figure 3.14: This shows the distribution of the maximum positive and negative values for every respective sample across all subjects for the Flash Balanced dataset (refer to, Table 3.5) retained post-processing. For further information on interpretation see, Figure 3.8. The total number of samples attainable for this data partition is 2 (emoji) $\times$ 49 (trials) $\times$ 10 (subjects) $\times$ 2 (blocks) $=$ 1960.

The consistency of class predictions is presented below for two subjects at either end of the performance scale in Subject 3 (high-performing) and Subject 8 (low-performing). As seen in Figure 3.15, Subject 3 demonstrates minimal confusion across all prediction states and no clear bias towards either class.

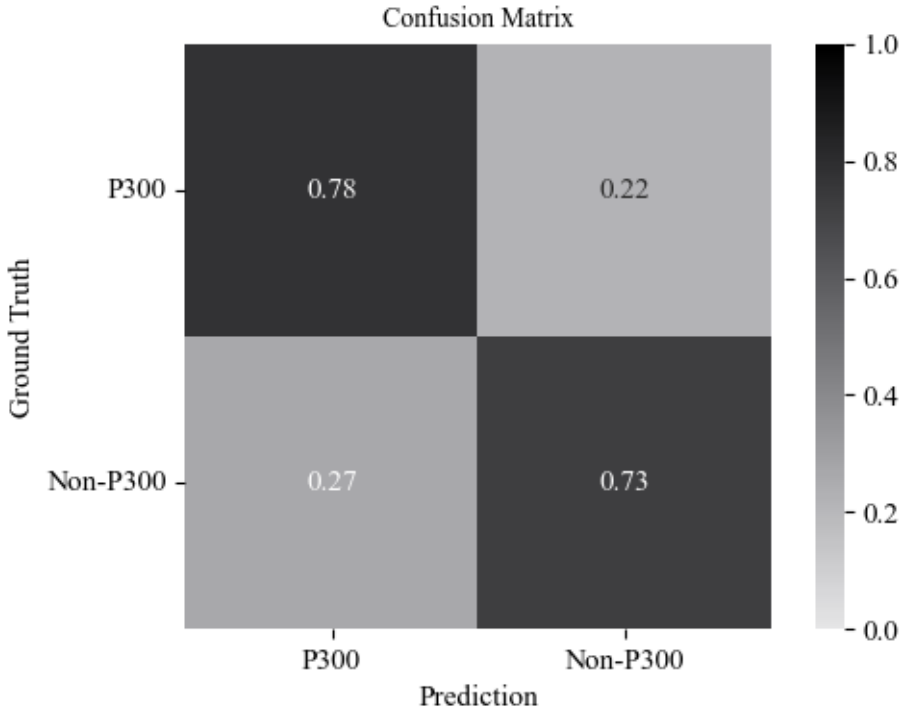

Figure 3.15: Displayed here is a confusion matrix generated via the LDA classifier results for Subject 3 in the Flash Balanced data partition (refer to, Table 3.5). For more information on interpretation see, Figure 3.7.

Figure 3.16, illustrates that for Subject 8 the LDA discriminant function failed to develop an accurate representation of either class, resulting in numerous misclassifications and sub-random threshold performance $<$50%).

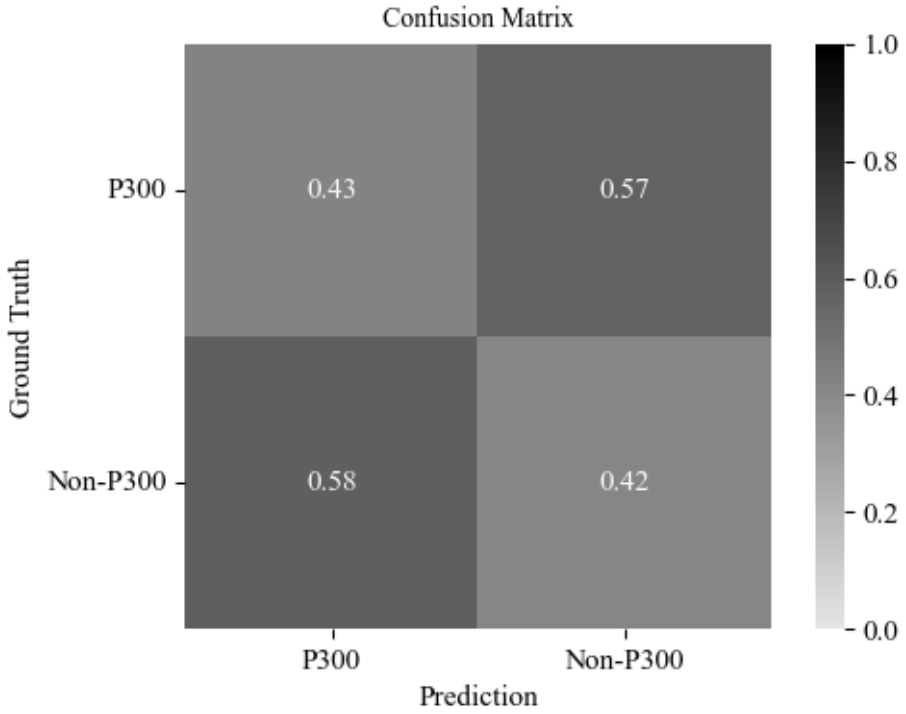

Figure 3.16: Displayed here is a confusion matrix generated via the LDA classifier results for Subject 8 in the Flash Balanced data partition (refer to, Table 3.5).

As shown below, there is no evidence of significant μV amplitude range differences between the two subjects (Subjects 3 & 8). Both demonstrate positive and negative maximal deflections between $+/-15$ μV and provide a representative example of the recordings collected across all subjects sampled.

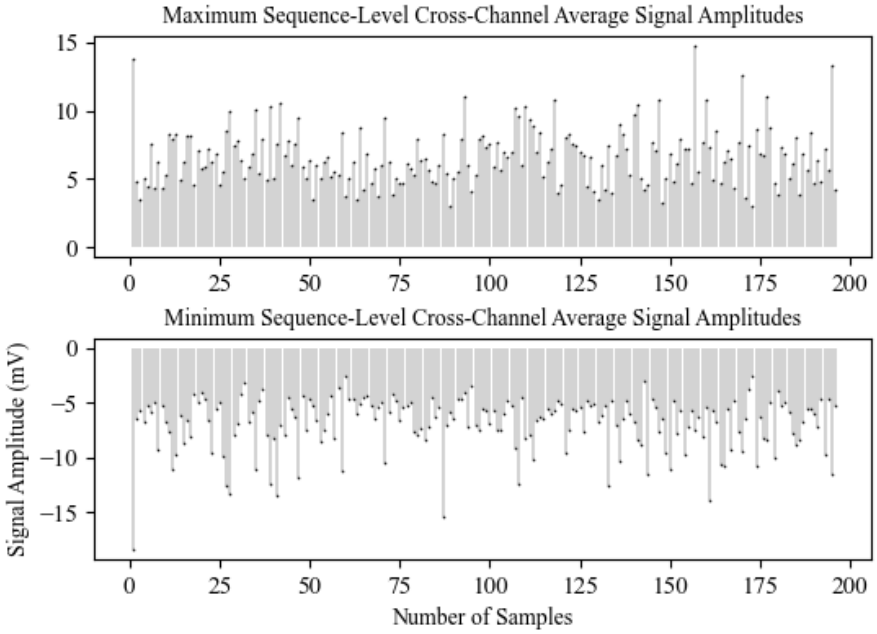

Figure 3.17: This shows the distribution of the maximum positive and negative values for every respective sample for Subject 3 in the Flash Balanced dataset (refer to, Table 3.9) retained post-processing. For further information on interpretation see, Figure 3.8. The total number of samples attainable for this within-subject data partition is 2 (emoji) $\times$ 49 (trials) $\times$ 1 (subject) $\times$ 2 (blocks) $=$ 196.

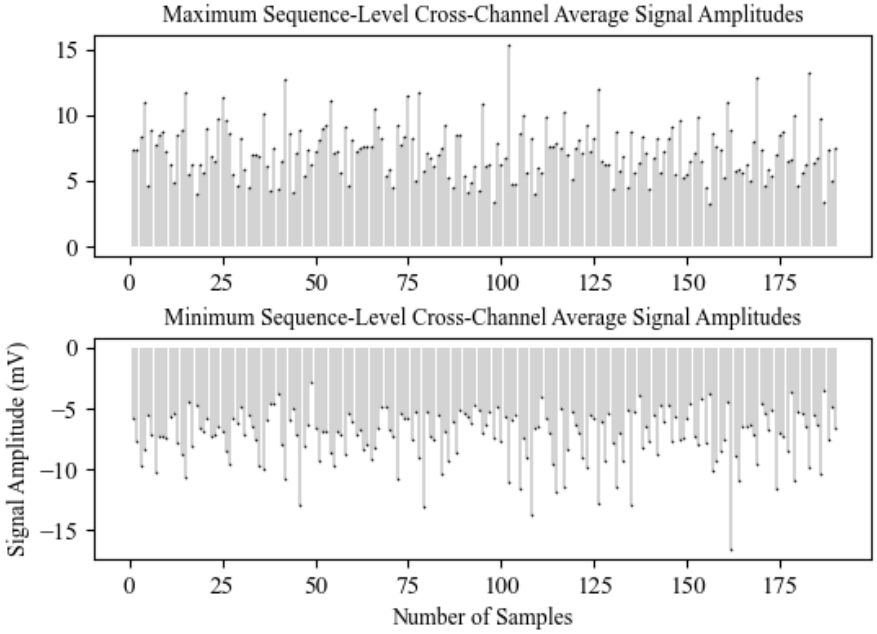

Figure 3.18: This shows the distribution of the maximum positive and negative values for every respective sample for Subject 8 in the Flash Balanced dataset (refer to, Table 3.5) retained post-processing.

3.5.5 Inversion Method Results: Class-Balanced: Pipeline 1

This subsection concerns all results generated for the Inversion Balanced dataset (refer to, Table 3.5). These data were collected utilizing the Inversion method of stimulus augmentation (see, Figure 3.2, lower panel). This involves reversing all emoji stimulus elements from a black to a white colouration to elicit the propagation of a visual P300 waveform. All samples are class-balanced (downsampled) to enforce a strict 1:1 ratio of P300 and Non-P300 events.

3.5.5.1 Pooled-Subject

	Mean Accuracy (%)	P300 Accuracy (%)	Non-P300 Accuracy (%)	Solver	Shrinkage	Num Test Events
Pooled Subject	48.83	52	46.02	lsqr	0	198

Table 3.10: The table contains classification performance metrics and optimization results from the LDA models trained using the pooled-subject Inversion Balanced data partition (refer to, Table 3.5).

As can be seen in the table above (Table, 3.10) the results of the pooled-subject analyses revealed a sub-random (<50%) classification accuracy. The grid search optimisation determined the optimal solver method as lsqr with a shrinkage value of 0.

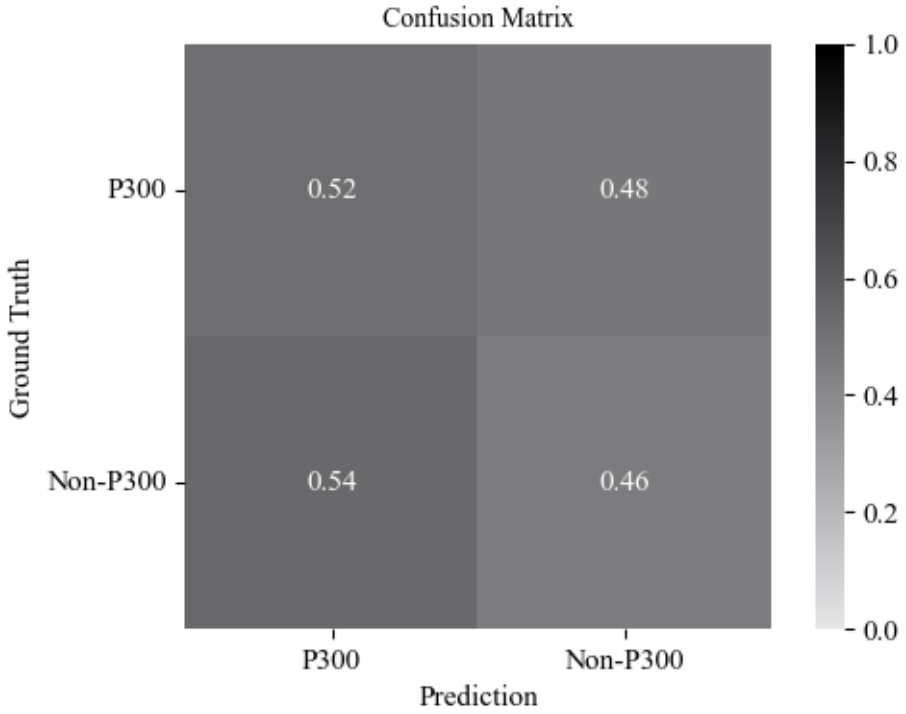

Figure 3.19: Presented is a confusion matrix generated via the LDA classifier results for the Pooled-Subject Inversion Balanced data partition (refer to, Table 3.5).

The above confusion matrix (see, Figure 3.19) illustrates the predictions generated via the LDA model trained using the Inversion Balanced pooled-subject data. This figure shows that the LDA classifiers demonstrated significant difficulty in accurately identifying P300 and Non-P300 events as belonging to the corresponding classes.

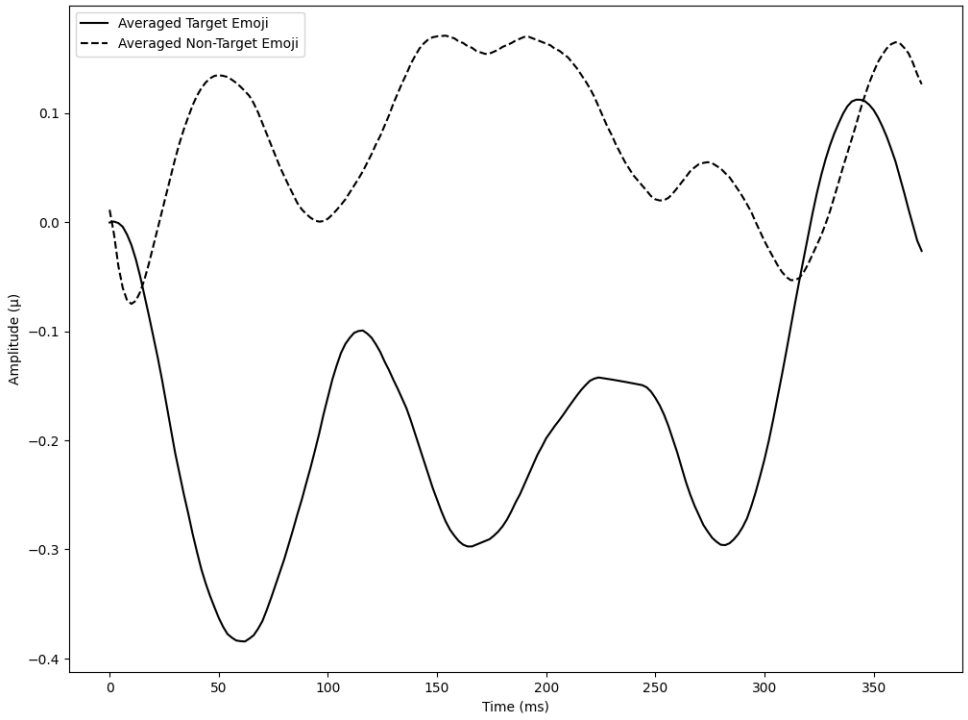

Figure 3.20: The plot herein displays the Cz grand averaged signals from all subjects for the Inversion Balanced data partition. The solid line shows an averaged signal for all P300 events and the dashed line shows an averaged signal for the Non-P300 events.

The figure above (see Figure 3.20) displays the grand-averaged Cz waveforms for both P300 (solid line) and Non-P300 (dashed line) events in the Inversion Balanced data partition. The P300 waveform is characterized by a strong oscillatory component with a periodicity of approximately 125ms (8Hz), likely due to the 125ms stimulus augmentation interval. This suggests that adjacent emoji, in addition to the cued target, were at least partially attended to during the task. The P300 signal exhibits a significant positive deflection crest at 300-350ms, which is typical of a P300 waveform. The Non-P300 waveform, while displaying some P300-like features such as a negative component around 100ms and a smaller positive peak at 300-350ms, has a much lower amplitude in terms of μV. However, due to its relatively stable baseline with less drift, the Non-P300 signal has a larger area under the curve. This is likely owing to the author's decision to perform non-continuous data collection, preventing the use of a more standard baselining method for example, using the previous 500ms, as opposed to the first 50ms.

3.5.5.2 Within-Subject

The classification results table for the Inversion Balanced data partition reveals only a single subject (Subject 3) achieved above random performance for both the P300 (66.67%) and Non-P300 (63.34%) event classification accuracies (see, Table 3.11). The single-subject average of 50.50%, $+/-12.5\%$ is broadly representative of the poor overall performance of the subjects sampled. The grid search optimisation revealed an exclusive preference for the lsqr solver method and an average shrinkage value of $0.42 +/-0.49$.

	Mean Accuracy (%)	P300 Accuracy (%)	Non-P300 Accuracy (%)	Solver	Shrinkage	Num Test Events
Subject 1	40	30	50	lsqr	0.97	20
Subject 2	40	60	20	lsqr	0.06	20
Subject 3	65	66.67	63.34	lsqr	0.11	20
Subject 4	50	55.56	45.45	lsqr	0.07	20
Subject 5	60	50	70	lsqr	0.81	20
Subject 6	45	44.44	45.45	lsqr	0	20
Subject 7	55	33.33	72.73	lsqr	0.62	20
Subject 8	50	66.67	41.67	lsqr	0.11	18
Subject 9	55	50	60	lsqr	0.93	20
Subject 10	45	50	40	lsqr	0.49	20
Single Subject Avg.	50.5	50.67	50.86	n/a	0.42	19.8
Single Subject Var.	12.5	18.34	26.37	n/a	0.49	1

Table 3.11: The table contains classification performance metrics and optimization results for single-subject LDA models from the Inversion Balanced data partition (refer to, Table 3.5).

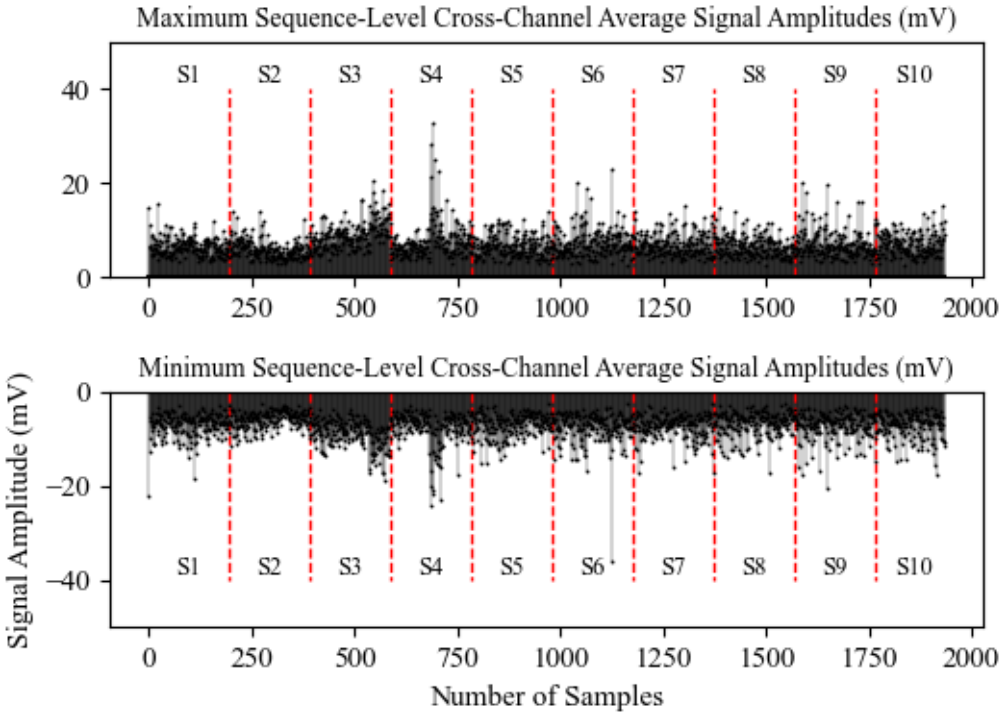

Figure 3.21: The figure illustrates the distribution of the maximum positive and negative values for every respective event across all subjects for the Inversion Balanced dataset (refer to, Table 3.5) retained post-processing. For further info on interpretation see, Figure 3.8

The μV amplitude plots (refer to, Figure 3.21) demonstrate a relatively normal distribution of maximum and minimum amplitude values for every respective event in the Inversion Balanced data partition. The variance in samples retained post-processing can be found in the classification Table 3.11, showing the only subject to have samples removed is identified as Subject 8.

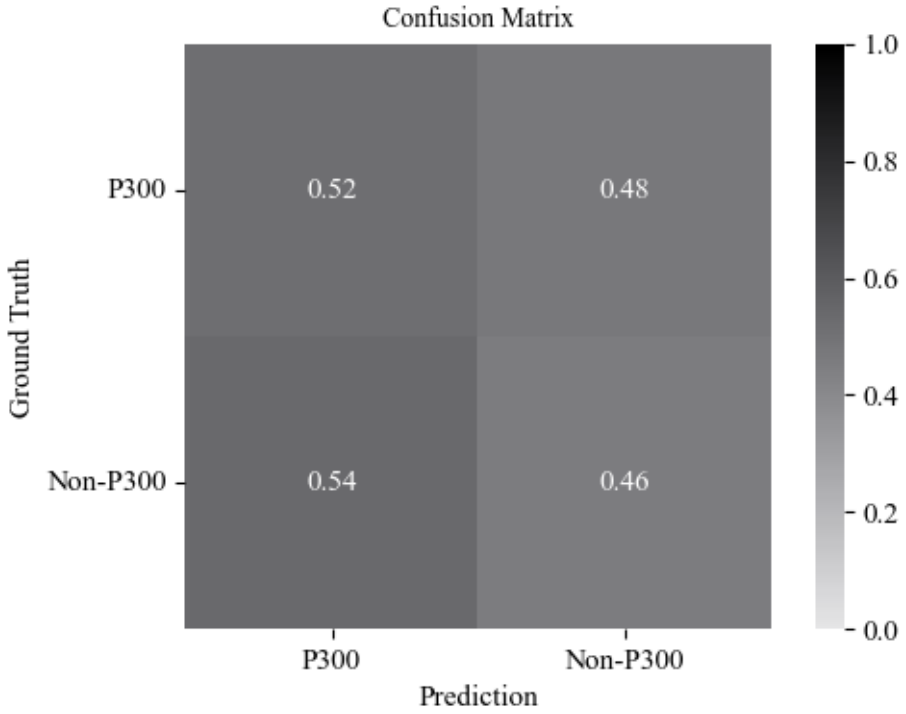

Figure 3.22: Here is presented a figure displaying a confusion matrix generated via the LDA classifier results for Subjet 3 in the Inversion Balanced data partition (refer to, Table 3.5).

As seen in the figure above (see, Figure 3.22), the LDA classifier did not effectively discriminate the classes based on the characteristics of the data relating to each class. These data features only allowed for a highly marginal increase in classification accuracy above random performance ($>$50%).

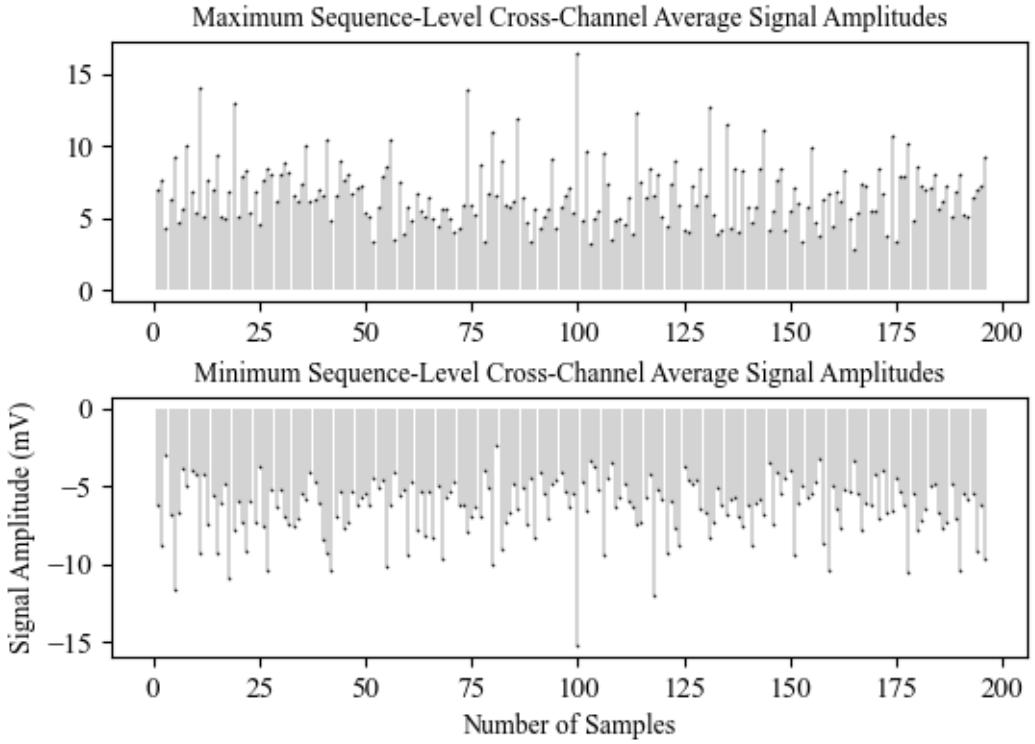

Figure 3.23: The figure illustrates the distribution of the maximum positive and negative values of every respective sample for Subject 3 in the Inversion Balanced dataset (refer to, Table 3.5) retained post-processing. For further info on interpretation see, Figure 3.8.

The μV amplitude range plot for Subject 3 demonstrates a high level of uniformity for both the positive and negative maximum deflections found across events, with all samples presenting well below $+/-20\mu$V (see, Figure 3.23). As seen in Table 3.11, all samples collected from Subject 3 were included in the analyses, with none of these data being removed via the channel-rejection protocols.

3.5.6 Combined Method Results: Class-Balanced: Pipeline 1

The following analysis relates to the data partition Combined Balanced (refer to, Table 3.5). This is an aggregate dataset, combining the data for all 10 subjects across both augmentation methods implemented (Flash and Inversion) utilizing the same data pre-processing and analysis methods outlined in Pipeline 1 (see subsection 3.3.5.1). The total number of samples attainable post-processing for this dataset amounts to 2 (emoji) $\times$ 49 (trials) $\times$ 4 (blocks) $\times$ 10 (subjects) $=$ 3920. An average of 39.2 events were retained per subject ($+/-1.5$).

3.5.6.1 Pooled-Subject

	Mean Accuracy (%)	P300 Accuracy (%)	Non-P300 Accuracy (%)	Solver	Shrinkage	Num Test Events
Pooled Subject	51.64	51.69	51.58	lsqr	0.77	392

Table 3.12: A table of classification performance metrics and optimization results from the pooled-subject Combined Balanced data partition (refer to, Table 3.5).

As can be seen in Table 3.12, the mean accuracy (51.64%) is just above purely random performance (50%). Both the P300 (51.69%) and Non-P300 (51.58%) event classification accuracies are marginally above random performance. The results in this instance do not have much significance, given the only slight (around 1.5%) increase above the random performance threshold level. The grid search optimisation revealed the lsqr method at a shrinkage rate of 0.77 as the most effective combination of training parameters.

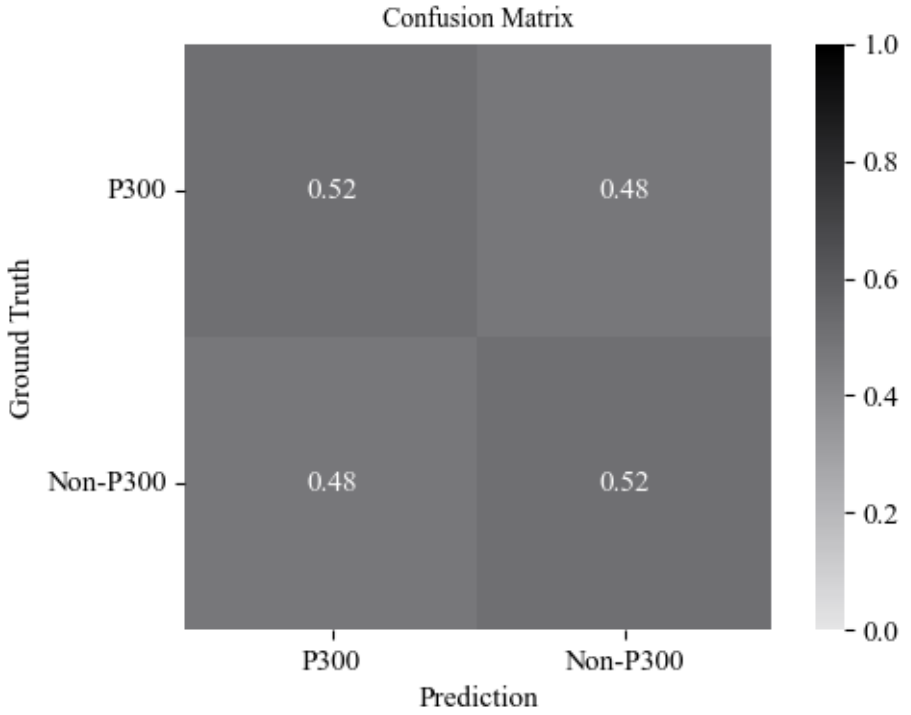

Figure 3.24: Here is presented a confusion matrix generated from the results of the LDA analyses conducted using pooled-subject data for the Combined Balanced data partition (refer to, Table 3.5).

The figure above (see, Figure 3.24) illustrates the LDA classifier predictions in finer detail. The diagram reveals that the model did not learn to accurately distinguish between P300 and Non-P300 waveforms, with an arguably near-random performance pattern, owing to the relatively even distribution of predictions across both classes tested.

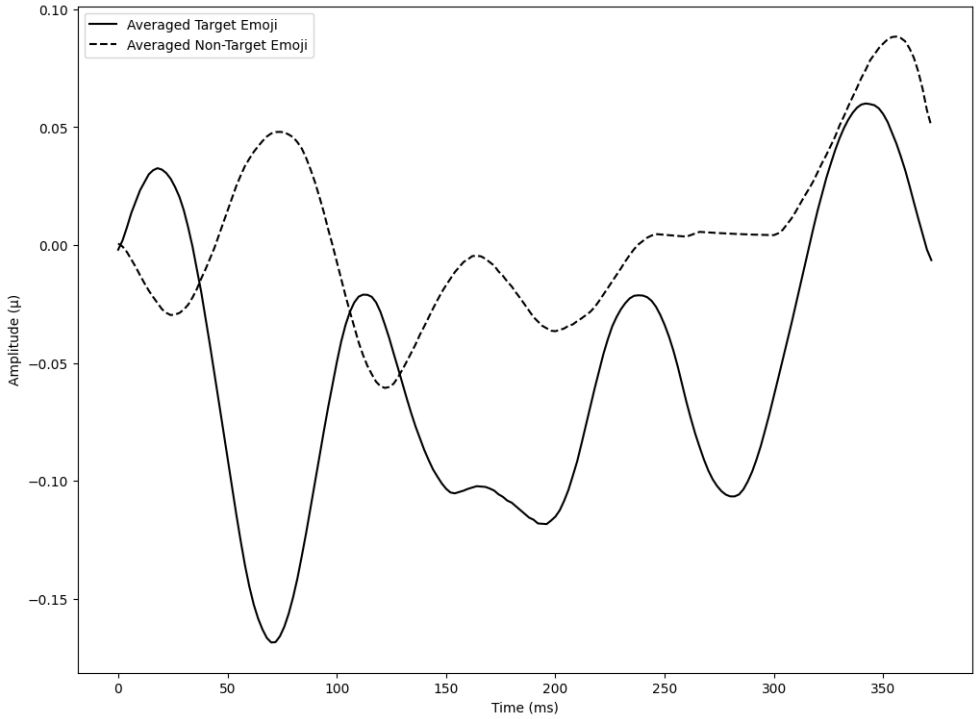

Figure 3.25: This is a Cz grand average plot for the P300 (solid line) and Non-P300 (dashed line) signals across the Combined Balanced data partition for all 10 subjects sampled in this experimental variant (refer to, Table 3.5).

In the above figure (see Figure 3.25), the P300 average signal (solid line) exhibits significant negative drift, complicating direct comparisons with the Non-P300 signal. An oscillation with a periodicity of approximately 125ms is also observed, likely reflecting an 8Hz SSVEP induced by the 125ms stimulus augmentation interval. Although the P300 signal shows much greater variance, the higher relative amplitude of the Non-P300 waveform results in a larger area under the curve. Given these observations, it is unlikely that the P300 peaking component is effectively utilized by the LDA models for separating these data into distinct classes.

3.5.6.2 Within-Subject

As can be seen from the classification table (Table 3.13), the single-subject average mean accuracy (44.59%) broadly represents the single-subject performance across the whole dataset. No subjects from the evaluated data partition achieved a mean accuracy significantly greater

than random performance, with only one marginal subject (Subject 9) approaching the random performance threshold. The grid search optimization undertaken to determine the best training parameters for the LDA classifier led to the exclusive selection of the lsqr solver method and demonstrated significant variance in the selection of the shrinkage metric (average=0.51 $+/-0.49$).

	Mean Accuracy (%)	P300 Accuracy (%)	Non-P300 Accuracy (%)	Solver	Shrinkage	Num Test Events
Subject 1	30.77	31.25	30.43	lsqr	0.00	39
Subject 2	41.03	43.75	39.13	lsqr	0.01	39
Subject 3	52.20	63.64	38.89	lsqr	0.99	40
Subject 4	52.50	45.45	61.11	lsqr	0.38	40
Subject 5	38.46	46.67	33.33	lsqr	0.93	39
Subject 6	50.00	45.45	55.56	lsqr	0.62	40
Subject 7	35.00	27.27	44.44	lsqr	0.15	40
Subject 8	45.95	52.38	37.50	lsqr	0.00	37
Subject 9	56.41	56.25	56.25	lsqr	0.99	39
Subject 10	43.59	31.25	52.17	lsqr	0.98	39
Single Subject Avg.	44.59	44.34	44.88	n/a	0.51	39.2
Single Subject Var.	12.82	18.19	15.34	n/a	0.50	1.5

Table 3.13: A table of classification performance metrics and optimization results from the Combined Balanced data partition (refer to, Table 3.5).

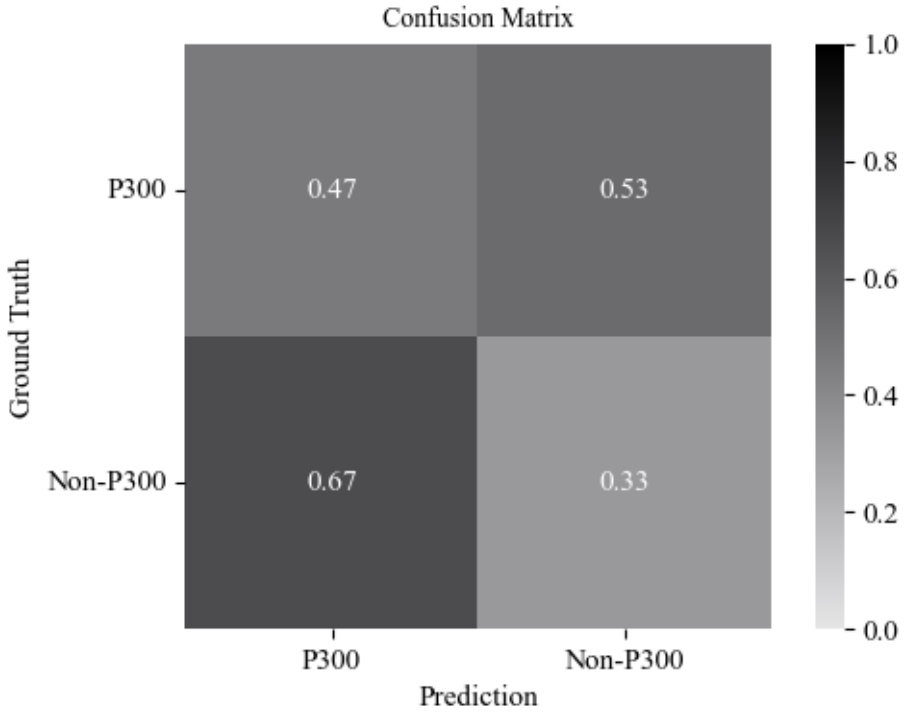

Figure 3.26: Here is shown a confusion matrix generated from the results of the LDA analyses conducted using data from Subject 5 for the Combined Balanced data partition (refer to, Table 3.5).

When inspecting the figure above (see, Figure 3.26), significant confusion can be observed between the prediction of Non-300 and P300 events (and vice versa). Some biasing of predictions is evident in the heightened incidence of misclassifying Non-P300 events as P300 events (lower left quadrant). Further, a significantly reduced tendency to correctly predict Non-P300 events as such is also observed (lower right quadrant).

3.6 Results: Pipeline 2

Here are presented all results relating to Experiment 1 utilizing the Pipeline 2 data organisation, pre-processing and analysis stages, for further details please refer back to subsections 3.3.5.3 and 3.4.3. This comprises an investigation into the efficacy of the Flash and Inversion augmentation methods for the proposed 7-emoji BCI communication platform. Note, as seen in Tables 3.6, 3.8, 3.10 and 3.12 relating to the Pipeline 1 approach, the performance of all models trained using data aggregated across subjects either failed to produce classification accuracies above the 50% random threshold or demonstrated significant overfitting. Further, at the single-subject level, only a handful of subjects demonstrated classification performance

above 50%. The author has selected the 3 highest-performing subjects from the Pipeline 1 assessments, Subjects 3, 5 and 8 to validate the Pipeline 2 approach. This was done due to project completion time constraints in addition to a desire to avoid presenting similar low-quality results. Here, all models are trained using data from a single subject, feature data augmentation via the SMOTE over-sampling technique and report the subject-level performance for the Flash and Inversion method, as well as the Collapsed and Non-Collapsed trial methods.

3.6.1 Data Partitions: Pipeline 2

In the table positioned below (Table 3.14), all data quantities are reported following trial rejection, cross-validation-based test set isolation and training set data augmentation via SMOTE. Here, only the tables relating to the Flash augmentation method are shown, as highly similar distributions of quantities were defined for the Inversion augmentation method (see Tables A.3 & A.4). As noted in subsection 3.4.3.1 Cross-Validation, over the course of both 49 trials blocks for either augmentation method a total of 98 trials were collected for the P300 targets and 588 trials collected for the Non-Target instances. Here, the total number of samples retained following amplitude-based channel rejection is listed, for more information on the details of this method see subsection 3.3.5.2 and Table 3.1. As is expected several trials were excluded due to exceptionally high micro-voltages to retain data quality. Further, the train-test split of 9:1 is executed before data augmentation to ensure no synthetic samples are erroneously positioned within the test set.

Additionally, the proportion of real *vs.* synthetic data within the over-sampled minority P300 class is presented. Further, a secondary table illustrating the same distribution of dataset quantities is shown for the so-called Collapsed dataset variant. As noted in subsection 3.3.5.3 (Data Pre-Processing: Pipeline 2), this involved artificially simulating the presence of 10 sequences per trial, as opposed to the original 5 sequences. This was done by collapsing the samples of P300 and corresponding Non-P300 data segments from two neighbouring trials and averaging across these data. The process was undertaken to gauge the relative performance enhancement of the predicted increase in signal-to-noise ratio. As can be seen from both tables mentioned, there are no significant differences between subjects in the number of trials retained post-channel rejection for either data preparation variant (Non-Collapsed *vs.* Collapsed).

	Flash Non-Collapsed						
	Total Post-Rejection		Test		Train		
Subjects	P300	Non-P300	P300	Non-P300	P300 (Real)	P300 (Synthetic)	Non-P300
Subject 3	95	587	10	59	85	443	528
Subject 5	97	585	10	59	87	439	526
Subject 8	94	581	9	58	85	437	522

Table 3.14: Here is presented a table detailing the distribution of sample quantities for the datasets associated with the Flash Non-Collapsed Pipeline 2 approach. All samples here are composed of signals collected over all 5 sequences of each trial, for more information see subsection 3.3.5.3. The 'Total Post-Rejection' column denotes the number of samples retained for each of the 3 respective subjects following the amplitude-based channel rejection (+/-35μV), for further information see subsection, 3.3.5.2. The 'Test' column defines the number amount of samples assigned per subject for the evaluation of all associated LDA classifier models. Here 10% of the total post-rejection samples for the P300 and Non-P300 samples were isolated for these purposes. This assignment was repeated 10 times according to a 10-fold cross-validation procedure for model performance metric assessment, for more information see subsection 3.4.3.1. The final 'Train' column contains information on the number of Real, non-augmented, P300 samples. Further, the number of Non-P300 samples is noted here on the far right sub-column. Positioned in the central sub-column is the number of Synthetic samples generated via linear interpolation between samples in the Real P300 subset utilizing the SMOTE method (see subsection 3.4.3.3 Oversampling via SMOTE). As can be seen, the sum of the Real and Synthetic P300 samples equals the number of Non-P300 train samples. This was done to ensure a 1:1 ratio of P300 and Non-P300 samples in an attempt to mitigate the confounding influence of overfitting. To avoid excessive repetition of similar results, the equivalent data distribution table for the Non-Collapsed Inversion augmentation method data is positioned in Appendix Table A3.

	Flash Collapsed						
	Total Post-Rejection		Test		Train		
Subjects	P300	Non-P300	P300	Non-P300	P300 (Real)	P300 (Synthetic)	Non-P300
Subject 3	45	292	5	29	40	223	263
Subject 5	46	291	5	29	41	221	262
Subject 8	44	289	4	28	40	222	261

Table 3.15: Here is presented a table detailing the distribution of sample quantities for the datasets associated with the Flash Collapsed Pipeline 2 approach. All samples here are composed of signals from 10 sequences, this was generated by averaging corresponding Target and Non-Target samples for 2 neighbouring trials containing 5 sequences each. For further information on field headings and interpretation please refer to the table above (Table 3.14). Note, that the ratios between Target and Non-Target samples for all datasets listed, including the proportion of Real vs. Synthetic P300 instances mirror those in the Non-Collapsed data preparation variant (see table above). To avoid excessive repetition, the data distribution table related to the Inversion augmentation method is positioned in Appendix Table A.4.

3.6.2 Flash Method Results: Non-Collapsed: Pipeline 2

The results reported in this subsection refer to analyses undertaken on the: Flash Non-Collapsed data partition (refer to, Table 3.14). These samples were collected during the implementation of the Flash augmentation method involving the overlay of a white augmentation square for a given on screen emoji stimulus (for more information see, Figure 3.2). All data organisation, pre-processing and analysis were undertaken using the Pipeline 2 approach (see subsections 3.3.5.3 and 3.4.3.1). As stated above, all samples herein were computed via the averaging of all 5 sequences per trial.

	Overall		Target		Non-Target	
Subjects	Acc Mean	Std Dev	Acc Mean	Std Dev	Acc Mean	Std Dev
3	0.74*	0.05	0.81*	0.06	0.70*	0.06
5	0.73*	0.03	0.79*	0.04	0.68*	0.07
8	0.80*	0.03	0.88*	0.02	0.72*	0.06

Table 3.16: Here is presented a table showing the performance metrics associated with Subjects 3, 5 & 8 for the Flash Non-Collapsed data partition (see, Table 3.5). All results were computed following the stages laid out in the Pipeline 2 data organisation, pre-processing and analysis methodology. Here all individual samples are composed of averages computed across all 5 augmentation sequences within each respective trial (see subsection 3.3.5.3 Data Pre-Processing: Pipeline 2). The 'Overall' column details the mean classification accuracy ('Acc Mean') and associated standard deviation ('Std Dev') of the single-subject trained LDA model for all samples within the corresponding classifier test set (see Table 3.5 for further information). This mean accuracy value was generated by computing the average of all classification accuracies reported for each of the 10-folds of the cross-validation procedure implemented (see subsection 3.4.3.1). The 'Target' and 'Non-Target' columns report the within-class accuracies for the respective cued Target P300 samples and non-cued Non-Target Non-P300 samples. Note, that as the Target samples were augmented via the SMOTE data interpolation method to balance the ratio between the classes, here around 83% of all Target class samples were composed of synthetically generated instances. The final row, 'Avg.', reports a mean of all data points within the respective column. To be clear, this does not represent LDA model results computed on pooled-subject aggregated data. Here, in all one-sample t-tests computed the threshold for significance was set at $p < 0.05$ (denoted via *). This was done given the relatively small sample sizes and associated test set quantities.

In these analysis, the Shapiro-Wilk Test was used to evaluate the normality of accuracy metrics computed at the single-subject level for each 10-fold cross-validation. The test results consistently showed p-values greater than 0.05, indicating that the accuracy metrics for each subject approximately followed a normal distribution (see, Table 3.16). Each individual mean accuracy metric was then compared against a fixed random performance threshold of 50% using a one-tailed, one-sample t-test. This test was oriented to measure whether the mean accuracy significantly exceeds this base value. The degrees of freedom for the t-test were computed as the number of folds in the cross-validation minus one (i.e. $k-1$). All metrics reported corresponding p-values less than 0.05, indicating that these mean accuracy values are significantly higher than the 50% chance level.

Note, that for the Overall classification results all subjects demonstrated mean classification accuracies above the 70% functional performance threshold. Despite this, both Subjects 3 and 5 are highly marginal given the corresponding standard deviation metrics. These indication that for a portion of the cross-validation folds the classification accuracy dipped below the 70% functional performance cutoff. These considerations do not apply Subject 8, with a

mean accuracy of 80% and associated standard deviation 0.03, the results can be considered representative of surpassing both the 50% random chance and 70% functional performance thresholds. It is crucial to note that this relatively high classification accuracy is primarily the result of an exceptionally high Target sample accuracy mean of 88%. Notably, all subjects demonstrate a bias in classification towards the Target samples, as opposed to the Non-Target samples of around 6% across subjects. The efforts undertaken to address the imbalance in samples across these classes via the SMOTE interpolation method could have introduced some bias towards the target class.

Here, comparisons were made between the pooled-subject accuracy mean variants (Overall, Target and Non-Target) and the chance 50% performance level via the permutation method (see subsection 3.4.3.2). Only the Target accuracy variant produced values significantly below the 0.05 p-value threshold (p=0.045). Note, that these comparisons are highly limited given the small number of subjects in the assessment. Despite this, there is a clear difference in the average of the Target means (0.83) and the Non-Target means (0.70). The results could indicate the presence of classifier bias towards the Target class, particularly given the high prevalence of synthetic samples in the training dataset.

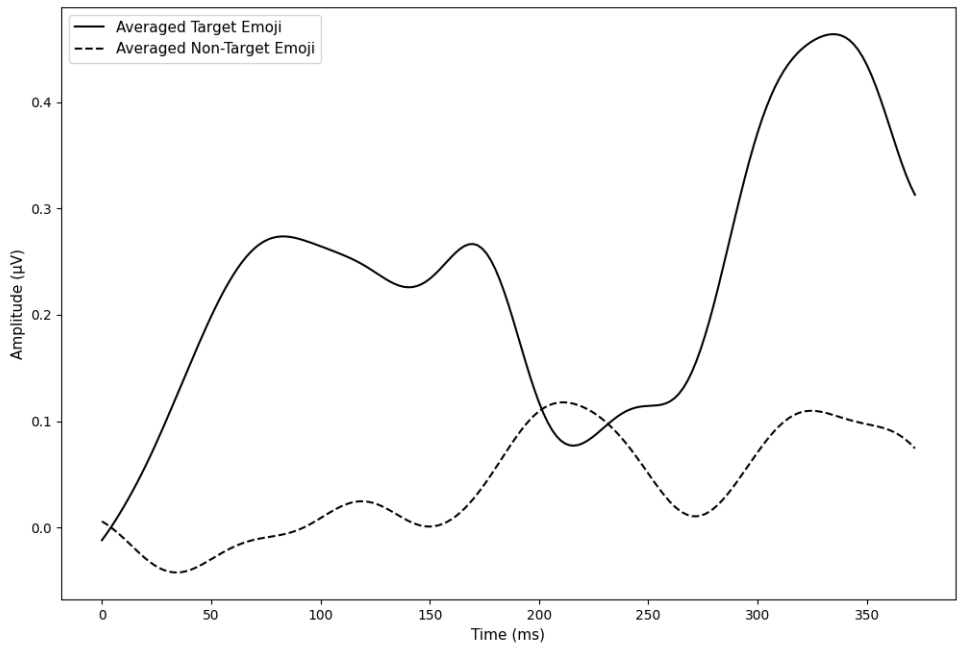

Figure 3.27: This is an average plot constructed exclusively from the Cz electrode for all P300, Averaged Target Emoji (solid line) and Non-P300 Averaged Non-Target Emoji (dashed line) samples collected across Subjects 3, 5 & 8 for the Flash Non-Collapse data partition (see, Table 3.5). As can be seen, the time dimension is positioned on the x-axis (0-375ms) and the micro-voltage range is oriented to the y-axis. Here, all samples were processed using the Pipeline 2 data pre-processing methodology (see subsection 3.3.5.3 Data Pre-Processing: Pipeline 2). Initially, this involves applying the pre-processing stages to all samples for a given trial trial. This features A2 electrode referencing, the application of a standard 50Hz powerline notch filter, the implementation of a pair of 0.1 high-pass and 15Hz low-pass finite-impulse responses filters as well as the application of an 8Hz notch filter to remove potential SSVEP artefacts induced via the 125ms stimulus onset interval. At this stage, the sequence then undergoes amplitude-based-channel rejection to remove any poor-quality samples. Given this average plot is constructed exclusively from Cz electrode samples, the identification of samples outside the +/-35µV range leads to the entire sequence being excluded. Further, each sequence is parsed into 7 segments (375ms windows) via corresponding markers indicating the onset times of the 7 emoji-augmentation events, in this case, Flash events. Following, the segments were baselined using the average of samples collected over the first 50ms of the given data chunk and then labelled according to the respective spatial emoji location. This is repeated for all 5 sequences in the respective trial and segments corresponding to the individual time-locked emoji augmentations are averaged together. These same steps are repeated for all trials in a given subject and then extended to all 3 subjects in this analysis. Finally, every Target P300 and Non-Target P300 averaged sample are aggregated into separate arrays and a grand pooled-subject mean signal is computed to generate the plot seen here. A total of 282 P300 samples and 1747 Non-P300 samples were utilized respectively. At no point, were any synthetic P300 samples included in the construction of these average signals. Note, that the primary difference between the Non-Collapsed and Collapsed data variants relates to the number of sequences assigned to each sample average. As the plot here features all samples it represents a linear combination of all signals within the given augmentation variant meaning the corresponding average plot generated using the Collapsed data is identical to the figure shown here.

As seen in Figure 3.27, the Cz electrode grand average signals generated using all Target and Non-Target samples for all 3 Subjects assessed demonstrate signal quality commensurate with the classification results observed. The Target P300 grand average presents with both characteristic waveform features induced via a visual oddball paradigm, namely, a moderate negative deflection around 200ms and a strong positive component around 300ms. Further, the redundant waveforms and noise components associated with the Non-P300 samples have been effectively diminished via deleterious averaging. As stated in the corresponding caption (see, Figure 3.27), the grand average plot positioned here is effectively identical to the plot generated using the Flash Collapsed data partition as the samples used to construct both Target and Non-Target signal averages are shared. This is due to the fact that the Collapsed signals feature an artificially increased number of sequences per sample (10 sequences), as compared to the Non-Collapsed data partition (5 sequences). In sum, when averaging across all samples within these respective data partitions the same average plots are produced. It is for this reason only one grand average plot is presented for each stimulus augmentation variant.

3.6.3 Flash Method Results: Collapsed: Pipeline 2

The results reported in this subsection refer to analyses undertaken on the: Flash Collapsed data partition (refer to, Table 3.5). All data organisation, pre-processing and analysis were undertaken using the Pipeline 2 approach (see subsections 3.3.5.3 and 3.4.3.1). Note, that all samples herein were computed via the averaging of 10 sequences across two neighbouring trials to artificially increase the number of sequences per trial and gauge the relative influence of the method on the resulting classifier performance metrics.

	Overall		Target		Non-Target	
Subjects	Acc Mean	Std Dev	Acc Mean	Std Dev	Acc Mean	Std Dev
3	0.75*	0.07	0.84*	0.07	0.66*	0.10
5	0.72*	0.05	0.80*	0.05	0.65*	0.10
8	0.76*	0.07	0.87*	0.07	0.66*	0.08

Table 3.17: Here is presented a table showing the performance metrics associated with Subjects 3, 5 & 8 for the Flash Collapsed data partition (see, Table 3.14). All results were computed following the stages laid out in the Pipeline 2 data organisation, pre-processing and analysis methodology. Here all individual samples are composed of averages computed across 10 augmentation sequences consisting of samples spanning two neighbouring trials (see subsection 3.3.5.3 Data Pre-Processing: Pipeline 2). For additional information on table field headings and interpretation please refer to Table 3.16.

As detailed for the Non-Collapsed Flash data partition above (see, Table 3.1), here all subjects relating to all respective accuracy variants produced mean, cross-validated accura-

cies significantly above the chance level of 50% ($p<0.05$). Again, the same trend emerged with corresponding LDA classifiers trained at the single subject level reporting a bias in the classification accuracy for Target samples. The permutation tests performed within the mean Target classification accuracy grouping were highly significantly different ($p=0.016$) as compared to chance (50%). In contrast, the Overall ($p=0.053$) and Non-Target (0.064) were both extremely marginal. The inclusion of additional subjects would likely lead to the reporting of significant results for both these groupings. Despite this, the low number of subjects for these assessments precludes the author from making broad generalizations regarding pooled-subject performance and system generalizability.

Finally, each set of accuracy metrics across both the Non-Collapsed (see, Table 3.16) and Collapsed (see, Table 3.17) data partitions were pair-matched to corresponding subjects and assessed via a permutation test (see subsection 3.4.3.2) to determine the relative influence of the cross-trial sequence aggregation method (see subsection 3.4.3). Here the subject pair classification accuracies for the Overall and Target groupings were highly non-significant, this is attributed to the small, 1.3% and -0.1%, differences between the means across the respective groupings. For the Non-Target paired assessments, the result was relatively marginal ($p=0.09$), with the performance across the subjects dropping around 5% for the Collapsed (65.3%) *vs.* Non-Collapsed (70%) data partitions. It is asserted that the relative increase in data quality per sample introduced via the artificial sequence increase was offset by the huge reduction in the number of samples available to the respective classifiers.

3.6.4 Inversion Method Results: Non-Collapsed: Pipeline 2

The results reported in this subsection refer to analyses undertaken on the: Inversion Non-Collapsed data partition (refer to, Table 3.5). These data were collected using stimuli following the Inversion augmentation method involving the inversion of all black emoji coloured elements to white (for more information see, Figure 3.2). All data organisation, pre-processing and analysis were undertaken using the Pipeline 2 approach (see subsections 3.3.5.3 and 3.4.3.1). As stated above, all samples herein were computed via the averaging of all 5 sequences per trial.

	Overall		Target		Non-Target	
Subjects	Acc Mean	Std Dev	Acc Mean	Std Dev	Acc Mean	Std Dev
3	0.78*	0.04	0.86*	0.04	0.71*	0.08
5	0.79*	0.03	0.85*	0.05	0.73*	0.04
8	0.79*	0.03	0.85*	0.06	0.73*	0.04

Table 3.18: Here is presented a table showing the performance metrics associated with Subjects 3, 5 & 8 for the Inversion Non-Collapsed data partition (see, Table 3.14). All results were computed following the stages laid out in the Pipeline 2 data organisation, pre-processing and analysis methodology. Here all individual samples are composed of averages computed across all 5 augmentation sequences within each respective trial (see subsection 3.3.5.3 Data Pre-Processing: Pipeline 2). For additional information on table field headings and interpretation please refer to Table 3.16.

As seen in the table above (see, Table 3.18), all single-subject classifiers were shown to be significantly higher than chance (50%), based on the results of the individual 10-fold cross-validation procedures. When computing the percentage variation via the standard deviation and mean accuracy for each subject none dropped under the 70% functional usage threshold based on the Overall classification performance metrics. The same trend demonstrated for the Flash augmentation Pipeline 2 results is reported (see, Tables 3.16 & 3.17), here all single-subject trained LDA models assessed showed a bias towards the Target samples. When comparing the means across each accuracy grouping against chance via the permutation test, only the Target means reported a significant p-value (0.0481), with the Overall accuracies returning a highly marginal p-value (0.051).

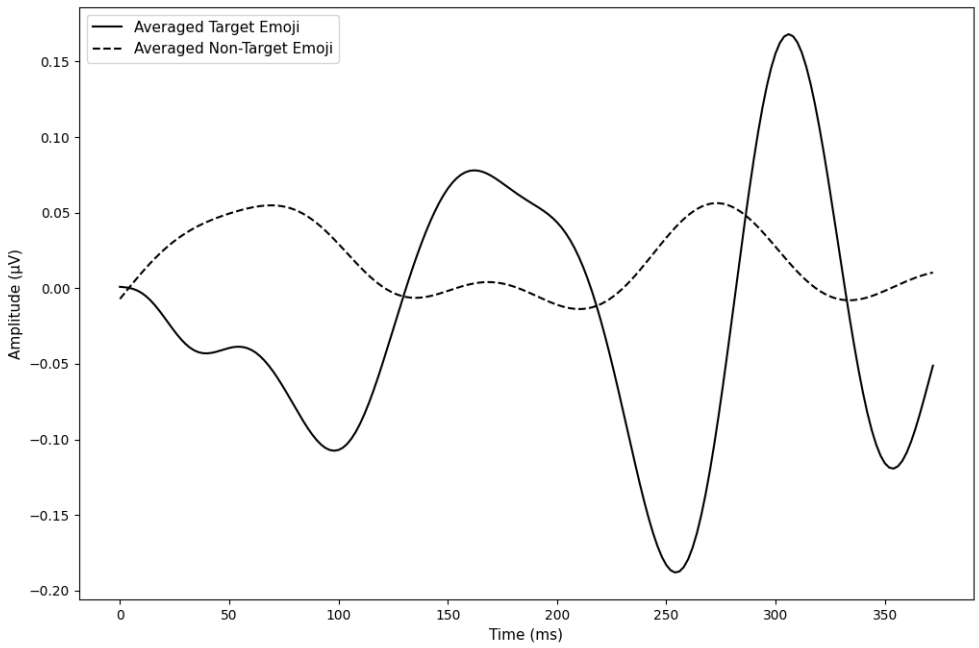

Figure 3.28: This is an average plot constructed exclusively from the Cz electrode for all P300, Averaged Target Emoji (solid line) and Non-P300 Averaged Non-Target Emoji (dashed line) samples collected across Subjects 3, 5 & 8 for the Inversion Non-Collapse data partition (see, Table 3.5). As can be seen, the time dimension is positioned on the x-axis and the micro-voltage range is oriented to the y-axis. Here, all samples were processed using the Pipeline 2 data pre-processing methodology (see subsection 3.3.5.3 Data Pre-Processing: Pipeline 2). For additional information regarding the interpretation of this plot please refer to Figure 3.27.

Here are shown the Cz grand average plots for samples collected during the Inversion method trials (see, Figure 3.28). The N200 component is delayed around 50ms and the crest of the P300 component peak is narrow, however crucially these key features have been extracted from the averaged target emoji P300 data segments to enable the effective separation of the samples as evidenced in the results table above (see, Table 3.18). Again, the averaged Non-Target emoji presents with a relatively flat profile with a small oscillatory component around 12Hz seen throughout and is observable given the low-pass cutoff threshold of 15Hz.

3.6.5 Inversion Method Results: Collapsed: Pipeline 2

The results reported in this subsection refer to analyses undertaken on the: Inversion Collapsed data partition (refer to, Table 3.5). All data organisation, pre-processing and analysis were undertaken using the Pipeline 2 approach (see subsections 3.3.5.3 and 3.4.3.1). Note, that all samples herein were computed via the averaging of 10 sequences across two neighbouring trials to artificially increase the number of sequences per trial and gauge the relative influence

of the method on the resulting classifier performance metrics.

	Overall		Target		Non-Target	
Subjects	Acc Mean	Std Dev	Acc Mean	Std Dev	Acc Mean	Std Dev
3	0.76*	0.06	0.87*	0.06	0.66*	0.07
5	0.76*	0.05	0.84*	0.07	0.67*	0.05
8	0.73*	0.06	0.81*	0.11	0.65*	0.07

Table 3.19: Here is presented a table showing the performance metrics associated with Subjects 3, 5 & 8 for the Inversion Collapsed data partition (see, Table 3.14). All results were computed following the stages laid out in the Pipeline 2 data organisation, pre-processing and analysis methodology. Here all individual samples are composed of averages computed across 10 augmentation sequences consisting of samples spanning two neighbouring trials (see subsection 3.3.5.3 Data Pre-Processing: Pipeline 2). For additional information on table field headings and interpretation please refer to Table 3.16.

The table positioned above reveals that all single-subject trained LDA models produced Overall classification accuracies significantly above the chance 50% level (see, Table 3.19). As was demonstrated for the Non-Collapsed Inversion data partition (see, Table 3.18), the Target samples were accurately classified to a higher degree than the Non-Target samples, with an 11% difference recorded for Subject 3. Despite this, all subjects performed well within the range (mean=75%), however in all instances given the relatively high standard deviations all dropped below the 70% functional performance threshold for a portion of the 10-fold cross-validation instances. The mean accuracies across subjects for the Target samples produced the only significant group-level result against the chance (50%), as computed via the permutation test. Further, the paired permutation test assessment comparing subject means collected between the Non-Collapsed and Collapsed results for each of the accuracy metrics listed did not produce any significant results. Despite a general trend of reduced performance for the Collapsed results relating to the Non-Target means, these findings suggest there is no significant difference in the data preparation methods in terms of end-point classification performance.

3.6.6 Flash *vs*. Inversion: Pipeline 2

In line with the methodology used to compare performance metrics between Non-Collapsed and Collapsed data partitions (refer to subsections 3.6.3 & 3.6.5), we now evaluate the differences between the Flash and Inversion augmentation methods. This analysis employs a similar permutation test to compare the overall accuracy means between subject pairs, focusing on the same data preparation method—whether Non-Collapsed or Collapsed. Specifically, the Flash method uses a white stimulus square overlay to elicit the visual-P300 response, whereas the

Inversion method involves flipping the black-coloured elements of a given emoji stimulus to white. By performing this permutation test, the author aims to discern if the observed differences in accuracy are statistically significant, thus providing insights into the relative efficacy of these augmentation techniques in enhancing BCI performance. This approach attempts to account for the inherent variability in subject responses and the specific impact of each augmentation method on accuracy, offering a robust comparison that aligns with the analytical framework established in previous sections.

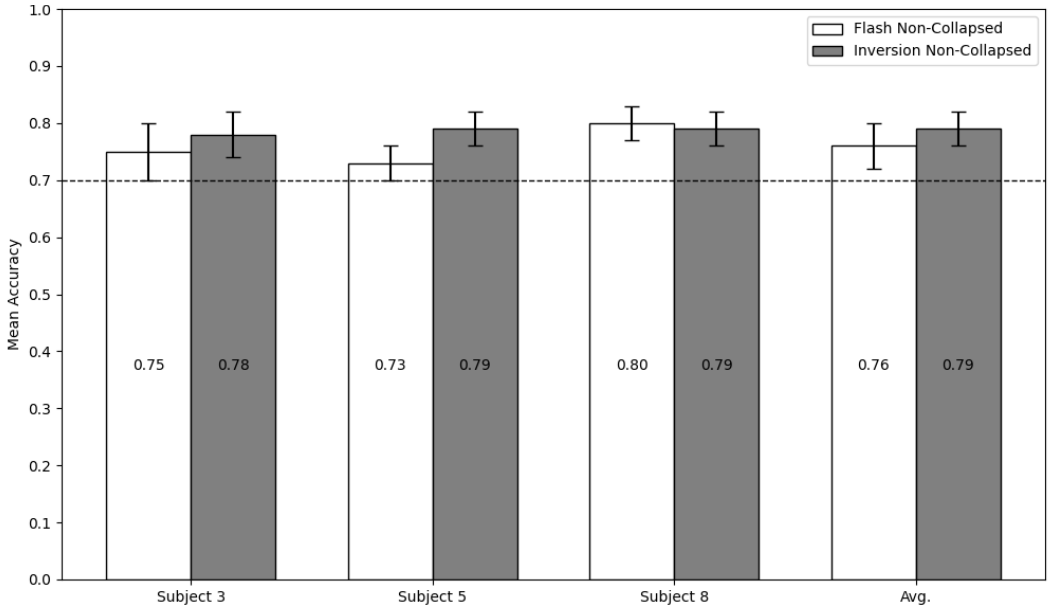

Figure 3.29: The plot displays a paired bar chart of the mean Overall accuracies and standard deviations for the Flash and Inversion methods using the Non-Collapsed data preparation technique in which each sample consisted of an average computed over 5 augmentation sequences (see subsection 3.3.5.3 for further information). The Flash augmentation method uses a white square overlay, while the Inversion method inverts black-coloured emoji elements to white (see subsection .3.3). These mean values are computed from a 10-fold cross-validation for each of the three subjects (3, 5 & 8) along with the pooled-subject average (Avg.) (see subsection 3.4.3.1). The figure also includes standard deviation bars to show variability in the results of the cross-validation. Each bar is also annotated with its corresponding mean accuracy value. A horizontal dashed line at 70% is included to help assess the performance of each method against this functional performance benchmark.

As can be seen in the plot positioned above (see, Figure 3.29), the difference between the Flash and Inversion augmentation methods for the Non-Collapsed data partition is marginal to non-existent for most subjects assessed with considerable overlapping of the standard deviations of mean Overall accuracies evident throughout. The visual similarity noted here is confirmed by the associated permutation significance test which reports a p-value of 0.4. This

is likely owing to the small difference in observed means (2.7%) between the Flash (76%) and Inversion (78.7%) average Overall classification accuracies. Notably, the accuracies relating to the single-subject LDA models for the Inversion methods never fell below the functional usage threshold of 70%. Further, the spread of related standard deviation bars appears marginally smaller for these instances.

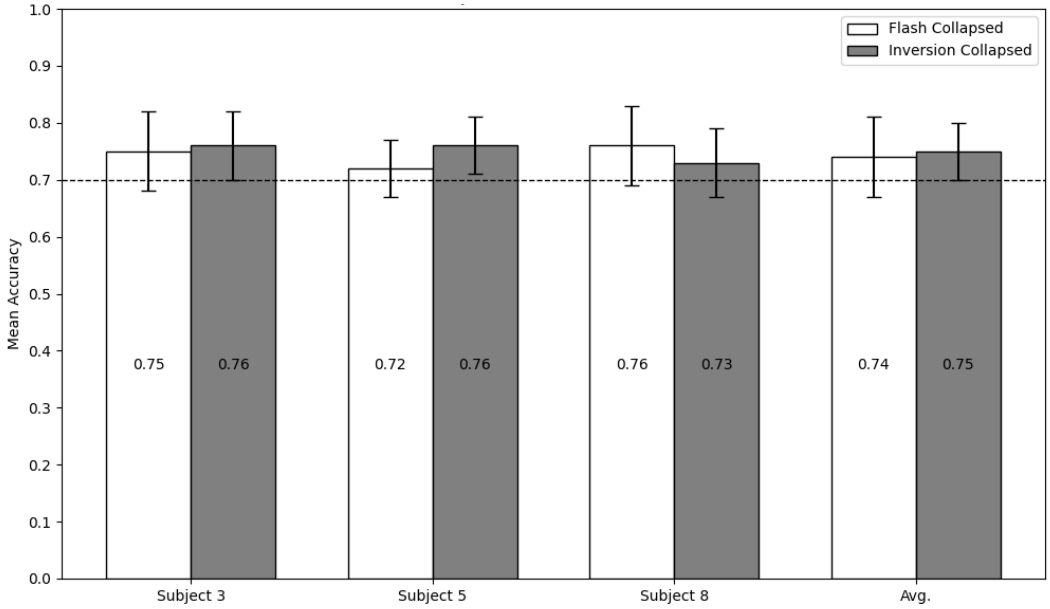

Figure 3.30: The plot displays a paired bar chart of the mean Overall accuracies and standard deviations for the Flash and Inversion methods using the Collapsed data preparation technique which each sample consisted of an average computed over 10 augmentation sequences (see subsection 3.3.5.3 for further information). The Flash augmentation method uses a white square overlay, while the Inversion method inverts black-coloured emoji elements to white (see subsection 3.3.3). These mean values are computed from a 10-fold cross-validation for each of the three subjects (3, 5 & 8) along with the pooled-subject average (Avg.) (see subsection 3.4.3.1). The figure also includes standard deviation bars to show variability in the results of the cross-validation. Each bar is also annotated with its corresponding mean accuracy value. A horizontal dashed line at 70% is included to help assess the performance of each method against this functional performance benchmark.

In the plot above (Figure 3.30), the paired-subject overall mean classification results for the Collapsed data partition, comparing the Flash method and the Inversion method, are presented. All instances, except for the Inversion method for Subject 5, fall below the 70% functional performance threshold. This is supported by the permutation test, which reports a p-value of 0.71. No significant difference is observed across subjects for either method, with overall mean accuracies of 74% for the Flash method and 75% for the Inversion method.

Both Non-Collapsed and Collapsed methods performed significantly above the 50% chance threshold at the single-subject level (see Tables 3.16-3.19). However, the small sample sizes, considerable overlap between paired subject results and inconsistencies present in group-level statistics, make it challenging to determine the relative performance of the two methods.

3.7 Conclusion: Pipeline 1

All findings and conclusions discussed herein relate to the Experiment 1 variant using the Pipeline 1 data organisation, pre-processing and analysis approach (see subsections 3.3.5.1 & 3.4). The corresponding tables and figures can be found in the text positioned above. Again, all analyses were conducted offline and correspond to a 7-emoji emotional communication BCI experimental paradigm outlined in the subsection Stimulus Presentation, 3.3.3.

3.7.1 Flash Method: No Class-Balancing Interpretation: Pipeline 1

These interpretations relate to the analysis of the Flash No Balance dataset (refer to, Table 3.5). This involved the implementation of the Flash augmentation method (see, Figure 3.2, upper panel) to generate the visual P300 waveform via a white square stimulus overlay. As noted in the dataset moniker, no class balancing was performed in the data preparation of these samples.

3.7.1.1 Pooled-Subject

As shown in Table 3.6, the pooled-subject dataset achieved above random performance (83.6%). This value is the result of a combination of the 100% classification accuracy for Non-P300 events in tandem with the 0% mean accuracy for the P300 events. These results suggest the classifier overfit during training. This can occur when there is a significant class imbalance in a training dataset. As the class balance for this training set data partition is 1:6, it is likely the LDA model overfit to the Non-P300 class. Along these very same lines, the confusion matrix (as seen in Figure 3.7) further demonstrates a significant bias in prediction selection for the Non-P300 event class.

As seen in Figure 3.8, there is a relatively uniform distribution of μV amplitude ranges across the data partition events spanning $+/-25\mu$V. The typical μV range difference present in a P300 waveform is characterised by around a 20μV deflection from the N100 negative component to the P300 positive component [120]. When considering this, it could be argued that the upper and lower μV amplitude bounds used in the channel rejection protocol of $+/-35$

μV are too large. Additionally, the presence of abnormally large deflections (as in Subject 5, see, Figure 3.8) raises the possibility that the inclusion of these events degraded the capacity of LDA classifiers to separate P300 signal features from redundant data or EEG noise. Importantly, this line of reasoning is not reflective of the pattern of results that remained. Such an explanation would present relatively low mean accuracies across both classes and not a biasing of prediction towards one class. Moreover, any attempts to enforce a stricter channel rejection protocol would ultimately have reduced the sample size of the total dataset so drastically as to prevent any meaningful performance from the LDA classifiers in question.

The distinct absence of typical P300 and Non-P300 waveform features in the pooled-subject average plots indicates there were significant issues in the experimental protocol (see, Figure 3.9). These could be related to subject attentiveness and instruction, augmentation durations and inter-stimulus interval parameters. The author must note that the most pressing issue relating to the data in this experimental variant (Experiment 1) involves data buffering. In a significant minority of trial sequences, the EEG buffer stream terminated the data capture link before the pre-coded experiment duration by around 150ms. This prevented the use of 500ms data windows for each respective augmentation event. To ensure a consistent size of data windows, necessary for averaging and analysis, the author had to reduce the size of this window down to 375ms.

This is not ideal as there exist numerous individual differences in the expression of the P300 waveform, with some components appearing in the time series after the 375 ms marker. This means the data acquisition potentially missed many crucial P300 peaking events, especially in the later stages of the experimental blocks. In these circumstances, subject P300 propagation is weaker and less responsive to augmentation stimuli. The absence of typical waveform features prevents the author from drawing firm conclusions relating to the performance of the LDA models evaluated for P300 classification.

3.7.1.2 Within-Subjects

As seen in Table 3.7, variance across subjects in terms of mean accuracy (86.58% $+/-1.39\%$), P300 accuracy (0%, $+/-0\%$) and Non-P300 accuracy are minimal (99.67% $+/-0.84\%$). These results suggest a similar pattern of overfitting as previously discussed above, with the LDA models demonstrating prediction bias for the more numerous Non-P300 events. For Subjects 6 and 7, the degree of overfitting may have been slightly lower, as shown by these subjects presenting sub-100% accuracy for the Non-P300 class. The Least Squares (lsqr) solver method was computed as the most effective for every instance. Note, that the uniformity in the

selection of the solver method (lsqr) is due to the parameter requirements of alternative solver methods. Other techniques such as the Eigen method require at least 2 variables to account for a significant amount of variance. As this threshold was not reached this restricted the selection to the lsqr solver method exclusively.

The vast majority of subjects presented with a relatively uniform distribution of μV amplitude ranges, typically between $+/-20\mu V$ (see, Figure 3.8). Subject 5 is an outlier in this instance, displaying by far the highest number of $>20\mu V$ amplitudes in terms of either positive or negative deflections (see, Figure 3.11). The presence of these higher ranges in Subject 5 did not seem to offer any negative or positive outcomes in terms of classification performance. This may be due to the same classifier overfitting issues as described earlier. In other words, the potentially confounding presence of classifier overfit prevents authors from making firm conclusions regarding the presence of values exceeding the $+/-20\mu V$ range in Subject 5 events.

In contrast, the total number of trials removed from Subject 5 is only 2 fewer than the maximum number of retainable test samples (67 *vs.* 69). The suggestion that the channel rejection was too stringent, does not hold up as this would assume that the inclusion of two more events could have dramatically influenced the performance of the LDA classifier. The possibility remains that the corresponding LDA model performance for these data would have been higher if the incidence of model overfitting had not negatively influenced the classifier training stage. The only way to validate the computational viability of these data is to artificially balance the ratio of P300 and Non-P300 waveforms.

3.7.1.3 Summary

It appears that the significant imbalance in the ratio of P300:Non-P300 events (1:6) severely impacted the LDA classifier during training and manifested in substantial overfitting. The highly reduced prevalence of incorrect predictions in the opposite direction, for example, the misclassification of a Non-P300 waveform as a P300 waveform suggests the overfitting process has been thoroughly completed and this significantly limits the extent of any interpretation based on these results at both the pooled-subject and within-subject levels. As noted in subsection 3.4.1, these effects were repeated for both the Inversion and Combined augmentation data variants. To avoid the repetition of overfit model discussion these results are positioned in Appendix A.1 and A.2. The following section aims to probe the influence of class balancing on model classification performance by enforcing a strict 1:1 ratio of P300:Non-P300 events.

3.7.2 Flash Method: Class-Balanced Interpretation: Pipeline 1

The following interpretations relate to the Flash Balanced data partition (see, Table 3.5). These data were gathered using the Flash augmentation method for P300 waveform elicitation (see, Figure 3.2, upper panel). All data here were class-balanced in an attempt to observe the performance of the LDA models for these data in the absence of overfitting effects. The data comprise all P300 samples collected as well as a subset of Non-P300 target waveforms. These were chosen according to a ranking method based on the spatial and temporal distances between target and non-target emoji stimuli. For further information see subsection 3.4.1.

3.7.2.1 Pooled-Subject

As shown in Table 3.8, the pooled-subject dataset achieved sub-random performance (43.8% *vs.* 50%). Notably, the class balancing protocols reduced the incidence of overfitting, with a slight bias in classification performance for the Non-P300 events. The low P300 classification performance could be due to the significantly reduced number of training examples available to the LDA classifier following the implementation of the class-balancing protocol. The confusion matrix (see, Figure 3.12) shows that the majority of predictions computed were incorrect, with the highest incidence of confusion relating to instances of misclassifying P300 events as Non-P300 events. This suggests that the LDA classifier was not able to identify the data features required to discriminate between the classes. Further, the visual appraisal of the pooled-subject amplitude range plot (see, Figure 3.14) reveals a highly uniform array of samples, as only a handful of events lie outside the $+/-15\mu$V range. This reduces the validity of any artefact-based reasoning for the low classification performances achieved.

The low LDA classifier performance is revealed in the pooled-subject data average plots (see, Figure 3.13). These demonstrate that the pooled-subject averages for the P300 class do not contain the fundamental features of the P300 waveform, namely a minor negative deflection at 200ms and a large positive deflection at 300ms. If it is assumed there exist no confounding outliers in the generation of these plots it is reasonable to assert that the events tagged as belonging to the P300 class do not possess enough P300 data features to be bourne out from the averaging process. In contrast, it is more likely that the absence of strong P300 waveform features is related to the individual differences in P300 propagation across subjects. The differences in P300 deflection onset and amplitude could have introduced a flattening effect across the entire signal profile, ultimately leading to the presence of some atypical features being amplified.

3.7.2.2 Within-Subject

The within-subject performance averages peaked above the random performance threshold (56% $+/-$16.45%) (see, Table 3.9). Classification performance for the P300 target class (50.13%) is however non-significantly different to the random performance threshold of 50%. The grid-optimized solver method tuned for each subject individually was universally selected as the lsqr technique. In contrast, the shrinkage values differ significantly across subjects, diverging from the trend observed in the Flash No Balance data partition results (see, Table 3.7). Specifically, the variance across subjects (average=0.63, $+/-$0.5) spans the entire 0-1 shrinkage value scale.

Crucially, it must be noted that Subjects 3, 6 and 9 all performed well above the random classification threshold, with Subject 3 achieving a respectable mean accuracy of 75%. Further, in all of these instances, there is considerably more balance in classification performance across P300 and Non-P300 target classes. The confusion matrix for Subject 3 (see, Figure 3.15) demonstrates this same pattern, suggesting that the corresponding LDA function is both accurate and robust, with no overfitting present. The sub-random performance demonstrated for Subject 8 (see, Table 3.8) alongside the high degree of misclassification observed in the respective confusion matrix (see, Figure 3.16) suggests the discriminant function was incapable of generating cohesive and accurate representations for the P300 and Non-P300 classes.

The author asserts that these results indicate a potential difference in Subject 8 task attentiveness. Further to this point, the corresponding amplitude range plot (see, Figure 3.18) negates any suggestions that the variance in subject performance can be attributed to significant differences in data quality. This is due to the plot broadly resembling the same distribution of maximum and minimum sample μV amplitudes as Subject 3 (see, Figure 3.17). Despite Subject 8 returning the lowest number of samples post-channel rejection (190), the class-balancing selection protocol removed all of the most volatile incidences. After removing these samples LDA classifier performance remained sub-optimal, even if the presence of overfitting had been largely mitigated.

3.7.2.3 Summary

In sum, the application of class balancing dramatically influenced the performance of all models at both the single and pooled-subject levels (see, Tables 3.7 & 3.9). Specifically, these methods removed the incidence of LDA model overfitting in nearly all evaluation instances. Note, that these data preparation techniques did not contribute to a universal increase in clas-

sification performance as numerous instances of random and sub-random classification accuracies are reported herein.

Note that the measures taken to balance the dataset have substantially reduced the ecological validity of these assessments. Ultimately, the data quality, experimental design and analysis configuration were not adequate to probe the original 7 Emoji-array design. The investigations herein were necessary principally to troubleshoot many of the obstacles faced during dry-EEG ERP acquisition. Ultimately, these analyses proved highly useful in the development of the systems outlined in the later stages of this series of experiments. In the subsequent experimental investigations (Experiments 2 & 3), methods to address the overfitting issues are explored.

3.7.3 Inversion Method: Class-Balanced Interpretation: Pipeline 1

The following interpretations relate to the Inversion Balanced data partition analyses (see, Table 3.5). All target P300 samples herein were generated via the use of the Inversion augmentation methods (see, Figure 3.2, lower panel). This involved the inversion of all elements in the emoji stimuli utilized augmenting from a black colouration to a white colouration. Note, as in the above subsection, all data were class-balanced according to a compound ranking method (see subsection 3.4.1).

3.7.3.1 Pooled-Subject

The sub-random mean accuracy performance (48.83%) of pooled-subject data for the Inversion Balanced data partition (see, 3.5) suggests that the inclusion of data across multiple subjects did not improve classification performance. This is likely related to the individual differences across the sample of subjects in the expression of P300 waveform features. The confusion matrix relating to the pooled-subject data (see, Figure 3.19) shows minimal evidence of overfitting. As stated in previous interpretation sections, is unlikely these low classification accuracies are related to movement artefacts. This is evidenced by the relatively normal distribution of pooled-subject amplitude range values seen in Figure 3.21.

Relating to the average plots generated from the Inversion Balanced data partition (see, Figure 3.20), both signals demonstrate some of the attributes typically associated with P300 and Non-P300 waveforms. Specifically, the increased positive deflection around 300ms for the P300 samples and the overall reduced volatility profile of the Non-P300. Despite this, the absence of an effective baselining method makes the comparison of these signal nonviable.

This process can be problematic when conducting iterative experimental data acquisition in which the inter-stimulus interval period is shorter than the typical refractory period of the P300 waveform.

3.7.3.2 Within-Subject

When inspecting the classification table for the Inversion Balanced data partition (see, Table 3.11) it is evident that only 1 subject (Subject 3) performed at an above-random classification performance (>50%) for both the P300 and Non-P300 samples. Further, these results suggest the Inversion method of augmentation is inferior to the Flash augmentation method (see, Table 3.9). It could be argued that the Flash method stimuli are more visually salient due to the increased 'area of augmentation'. This is owing to the white Flash method overlay square occupying 5-10% more of the presentation monitor as compared to the Inversion method of augmentation that only occupies the areas of each respective emoji stimuli containing black colourations (see, Figure 3.2).

Crucially, it is unlikely that the lower incidence of above-average classification performance for the Inversion method is due to a fatigue-based order effect as a strict counterbalanced method was applied when determining the presentation scheme of the two experimental augmentation variants. After inspecting these grand average signals (see Figure 3.20) is not possible to assert that the LDA classifier is distinguishing the samples based on the presence or absence of typical P300 waveform components. In contrast, the analysis has been guided by some non-specified and redundant data features that are present in this subsample of events. Crucially, this highlights the need for multiple analytical tools during the study of EEG-based ERP data.

3.7.3.3 Summary

On the whole, the given the high degree of overfitting and poor quality of the associated Cz grand average plots the author can not determine the relative quality of these stimuli for generating P300 waveforms. Crucially, the performance of related LDA models for the class-balanced data partitions was significantly lower as compared to the Flash-based augmentation results (see, Table 3.7). This could be due to the increased saliency of the overlay square implemented in the Flash augmentation method which engages a greater share of the presentation monitor. It is important to note that performance was not universally higher across the entire dataset for the Flash augmentation method, with some subjects performing better in the Inversion method data partition. The purpose of these experiments is not to determine indi-

vidual, tailored, methods of P300 elicitation, the aim is to develop an emoji-based emotion-communication BCI that has a high degree of usability across a large proportion of potential end-point users.

3.7.4 Combined Results: Class-Balanced: Pipeline 1

The following interpretations relate to the Combined Balanced data partition analyses (see, Table 3.5 and subsection 3.5.6). These data are comprised of samples collected across both the Flash and Inversion experimental augmentation variants (see, Figure 3.2). Note that these data were also processed via the compound ranking method to artificially enforce a class-balanced ratio of P300 and Non-P300 samples via downsampling.

3.7.4.1 Pooled-Subject

When inspecting the classification table (Table 3.12) for the Combined Balanced data partition, it can be observed that the combination of both class-balanced datasets did appear to remove the confounding effects of model overfitting observed for the non-class balanced variant (see Table A.2). Despite this, the increased size of the aggregated Combined dataset did not produce classification accuracies significantly above the random performance threshold (50%). The subjects undertook either the Flash or Inversion method of the experiment with short interleaved breaks, meaning some trials were conducted after a lengthy and attentionally taxing cognitive task. This would ultimately expose the LDA classifier to many more event instances in which the P300 waveform is significantly depressed due to fatigue order effects. A combination of these factors could have reduced the quality of the overall data that the LDA classifier was exposed to, ultimately reducing the capacity of the analyses to distinguish between the P300 and Non-P300 classes.

As seen in Figure 3.24 the model does not appear to have been trained effectively. The confusion matrix indicates there is both an absence of overfitting and learning, as seen by the near-random performance of the model. This suggests that class balance is not the only factor influencing the performance of the LDA classifier. As seen in the corresponding grand average plots (see, Figure 3.25) there exists a high degree of similarity between the P300 (solid line) and Non-P300 (dashed line) signals in terms of range and 300ms peaking event profile. This suggests that despite the efforts to implement a subsampling method to maximise the spatial and temporal difference between the target classes a significant degree of confounding bleed-over effects are present in the class-balanced Non-P300 dataset.

3.7.4.2 Within-Subject

The classification performance observed at the single-subject level (see, Table 3.12) for these data is arguably of lower quality than either of the alternative class-balanced data partitions discussed herein (see, Tables 3.7 & 3.10). There appear to be numerous instances in which the single-subject means and class-wise accuracies drop significantly below the threshold for purely random performance. This phenomenon is present for Subjects: 1, 2, 5 and 7. It is reasonable to assume that the increased prevalence of these events for the aggregated class-balanced data partition is owing to the combination of P300 and Non-P300 events across both visual augmentation variants of the experiment employed. This raises the possibility that the nature of P300 and Non-P300 events across the two variants differ substantively in the expression of typical features expected from the classes sampled.

The increased prevalence of this phenomenon for the aggregate data and its impact on the coherency of results across subjects tested demonstrate that the process of aggregation in this instance may not have been analytically sound.Ultimately, these results illustrate the substantial complexity of training ERP classifiers across multiple datasets. As can be seen in Figure 3.26, the confusion matrix for Subject 5 reveals some distinct patterns of misclassification. The P300 event class is predicted with a high degree of inaccuracy. In contrast, the Non-P300 events are misclassified as P300 events to a significantly higher degree. The imbalance in these plots suggests an absence of effective training for the LDA classifier utilized.

3.7.4.3 Summary

Overall, these findings indicate that the aggregation of data across the experimental variants did not lead to an increase in classification accuracies at either the pooled-subject or single-subject level. These explorations were undertaken principally to assess the impact of increasing the number of training samples for the LDA models evaluated following the reduction in available target events as per the class-balancing protocol. The results suggest that the reduction of within-class homogeneity for the P300 samples introduced following the aggregation of these datasets impeded the effective training of models even at the single-subject level. The potentially less salient Inversion stimulus variant could present P300 waveforms with a reduced positive deflection or a higher incidence of atypical features. This could lead to confusion during classifier training, as increased variance in the P300 target events would hinder the ability of the corresponding LDA models to effectively distinguish these samples from Non-P300 targets.

3.7.5 Conclusion: Pipeline 2

Here, all conclusions relating to the Flash and Inversion augmentation method samples processed using the Pipeline 2 approach are discussed. Please refer back to the relevant subsections for additional information regarding the Flash method white overlay square and Inversion black colouration inversion presentation method (see, Figure 3.3), the Pipeline 2 pre-processing method (see subsection 3.3.5.3), the associated Pipeline 2 cross-validation procedure (see subsection 3.4.3.1) and the associated tests of significance.

3.7.5.1 Flash Method Results: Pipeline 2

The results shown for samples collected with the Flash augmentation method via the Pipeline 2 re-analysis approach (see subsections 3.6.2 & 3.6.3) show a dramatic improvement in performance from the results collected for the Pipeline 1 approach used in the processing of Flash Non-Balanced (see Table 3.7) and Flash Balanced (see, Table 3.9) data partitions relating to both Subjects 5 and 8. As can be seen for Subject 3, despite the numerous adaptations implemented to improve the data organisation and pre-processing the performance of this subject in the Flash Non-Balanced instance (see, Table 3.7) has decreased slightly when comparing Overall mean accuracy, dropping from 75% to 74%. Note, however, that the Flash Non-Balanced result reported here was not evaluated using a 10-fold cross-validation procedure and therefore it remains tenuous to assert that this is a true representation of the actual classification accuracy for these subject samples.

For the remaining subjects, when comparing the Pipeline 2 oversampled results against the Pipeline 1 Flash Non-Balanced downsampled data partition assessments increased in mean (Overall) accuracy by 16% and 38% for Subjects 5 and 8 respectively. As can be seen from the corresponding grand average figures, the Flash Pipeline 2 method (see, Figure 3.27) is far more representative of a standard P300 waveform (see, Figure 3.1), as compared to the Pipeline 1 Flash Non-Balanced (Figure 3.9) and Flash Balanced (see, Figure 3.13). This is a clear visual indication that the pre-processing adjustments made in the Pipeline 2 approach allowed for the P300 waveform features to emerge from the background noise more effectively. Primarily, this is attributed to the lower high-pass filter cutoff of 0.1Hz (Pipeline 1= 1Hz), the SSVEP-notch filter at 8Hz, the transition to finite-impulse response filter designs and the more effective baselining methods.

As is discussed in subsection 3.6.2, the efficacy of aggregating emoji-specific time-locked data chunks across 2 trials in order to simulate an increase in the number of augmentation se-

quences from the original 5 (Non-Collapsed) to 10 (Collapsed) was assessed (see subsection 3.3.5.3). No significant difference ($p>0.05$) is reported from the paired permutation test conducted to compare the mean accuracy metrics between subjects for the Non-Collapsed (see, Tables 3.16 & 3.18) and Collapsed (see, Table 3.17 & 3.19) data preparation methods. Note, given the small sample size any broad conclusions regarding the significance of group-level effects are highly limited, however, is it evident from the range of values acquired across the Overall, Target and Non-Target accuracies that no substantive difference is observable between these stimulus methods.

Here, the author asserts that the simulated increase in the number of sequences per target should have dramatically improved the signal-to-noise ratio of associated samples, however, this adaptation was limited by the dramatic reduction in the number of samples remaining. As neighbouring trial data was aggregated and averaged into a single test or training sample this effectively halved the total number of samples available to the associated LDA model for the discrimination of Target and Non-Target classes.

Here, both the Non-Collapsed and Collapsed Flash method results demonstrate a similar pattern of performance characterized by a preference for all associated single-subject level LDA models for the Target class. This could be owing to the relatively higher degree of uniformity in the Target samples given that the corresponding data are parsed from the time-locked onset of the cued emoji augmentation event. In contrast, the Non-Target class should feasibly contain far less distinct and predictable waveform patterns, making the classification of these samples more complex. It is crucial however to note that following the SMOTE over-sampling methodology (see subsection 3.4.3.3) used to synthetically increase the number of Target samples up to the same number of Non-Target samples for each respective subject LDA classifier (see, Tables 3.14 & 3.15). Note, that there remains the possibility that these results are illustrative of overfitting.

Given the high degree of Target: Non-Target class imbalance (1:6), following all associated pre-processing and data preparation stages each subject-specific training dataset was composed of roughly 525 Target-P300 samples in the Non-Collapsed variant and 260 Target-P300 samples in the Collapsed variant. Each of these Target sample sets was composed of around 83.33% synthetically generated samples. It is possible that oversampling to this degree led to a dramatic imbalance in the relative amount of variance within the Target class, making these samples easier to parse from the Non-Target samples for the LDA model discriminative function. Despite this, all results were evaluated using a 10-fold cross-validation procedure.

At no point were any synthetic samples included in any of the respective test sets. By continually testing on real data while never exposing the LDA models to synthetic test samples, cross-validation serves as a safeguard against overfitting, making the model's results more trustworthy despite the heavy reliance on synthetic augmentation during training.

3.7.5.2 Inversion Method Results: Pipeline 2

The following subsection comprises all conclusions relating to the Inversion method samples processed using the Pipeline 2 approach. A substantial improvement (see, Tables 3.18 & 3.19) in all subjects assessed (3, 5 & 8) via the Pipeline 2 method is shown as compared to the corresponding Inversion Non-Class Balanced (see, Table A.1) and Balanced (see, Table 3.10) results. Specifically, the substantial incidence of overfitting observed for the Non-Balanced and the chance level results demonstrated for the Balanced Inversion data partition have both been addressed. In comparing the downsampled (Balanced) Pipeline 1 (see, Table 3.13) and oversampled Pipeline 2 Non-Collapsed (see, Table 3.18) results, subject overall classification accuracies increased by roughly 30% to a pooled-subject average of 78.6%. Further, when accommodating for the associated standard deviations, all subjects performed above the functional performance threshold of 70% over the course of the 10-fold cross-validation procedure.

In a similar vein to results reported above for the Flash augmentation method, the associated Cz grand average (see, Figure 3.28) presents with a far more typical profile for the Target and Non-Target waveforms expected when conducting a visual oddball paradigm as compared to the Inversion method results collected using the Pipeline 1 approach (see, Figures A.3 & 3.20). Notably, all single-subject level accuracy metric results comparing the spread of values collected during the respective cross-validated assessments produced accuracies significantly above chance (see, Tables 3.16 & 3.17). As stated above (see subsection 3.7.5.1), this is likely owing to the improvements in data pre-processing methods employed via the Pipeline 2 method (see subsection 3.3.5.3).

The paired-subject permutation tests conducted to discern the relative difference in mean classification accuracies (Overall, Target and Non-Target) across subjects for the Non-Collapsed (see, Table 3.16) and Collapsed (see, Table 3.17) demonstrated no signifcant differences. This is despite a mean drop in accuracy for the Non-Target class in the Collapsed data preparation variant of 6.3%. It is possible that the inclusion of additional subjects could have improved the resampling procedure associated with the permutation test or may have allowed the author to perform more robust parametric assessments to validate this downward trend more accurately. Again, the results demonstrate that the respective LDA models trained using single-subject

data showed a bias towards the classification of the Target class. The author acknowledges this could be owing to the large inclusion of synthetic samples in the Target class training subset. Note, that this addressed via a rigourous 10-fold cross-validation procedure in which all sub-models associated were never exposed to synthetic test samples to maximally nullify the possibility of overfitting.

3.7.5.3 Flash *vs*. Inversion Results: Pipeline 2

As discussed in subsection 3.4.3.2: Statistical Test of Significance, a series of paired permutation tests were performed to investigate the relative difference between the Flash and Inversion augmentation methods in terms of the mean classification accuracy metrics. None of these assessments revealed any significant difference between the conditions. Note, that this extends to both the Non-Collapsed (see, Figure 3.29) and Collapsed (see, Figure 3.30) results. From the associated figures it is clear that the Overall mean accuracies demonstrate significant overlap as evidenced by the subject-specific metric standard deviations. Here, the only implementation that produced mean classification results for all subjects well over the 70% functional threshold was the Inversion Non-Collapsed data partition. This trend of marginally higher mean classification performance is also evidenced in the Collapsed results to a lesser degree. As previously discussed, broad conclusions on the group-level results conducted via the paired permutation tests are highly limited given the small sample size tested.

3.7.5.4 Summary

A general improvement in the associated subject-level classification accuracies is reported here, with all subjects demonstrating Overall mean classification accuracies above chance level (see, Tables 3.16, 3.17, 3.18 & 3.19) as tested via one-tailed, One Sample t-test. These improvements in performance have been achieved without a dramatic increase in the incidence of overfitting. Notably, no significant difference in the Non-Collapsed *vs*. Collapsed data preparation method is evident from the results. Additionally, no significant difference is reported between the Overall mean classification accuracies observed for the Flash and Inversion augmentation methods. Crucially, some caution must be taken in the interpretation of these group-level results given the small sample sizes, the use of synthetic training samples for the Traget-P300 class and the associated non-parametric permutation test methods used in the tests of significance.

3.8 Reflections

This section outlines the key areas of import relating to Experiment 1. Any specific points raised related to the Pipeline 1 and 2 approaches are clearly denoted within the subsection titles. This covers interpretations regarding the key findings, a description of the experimental obstacles encountered, and a discussion surrounding the series of modifications the author implemented in the subsequent investigations (Experiment 2).

3.8.1 Data Capture Issues

As discussed in an earlier subsection (see 3.7.1), there were inconsistencies in the total sequence length of data capture across a substantial minority of samples. At times, the data capture duration was terminated up to 125ms before the intended termination point. This is owing to a data buffering issue that involved the failure of a pre-coded trigger for the buffer not to close after a default duration. Initially, researchers intended to parse the sequence-level data into 500ms event-locked chunks. For the data captures that experienced the aforementioned error this was not possible as the final event would only contain around 375ms data. This prevented the researchers from implementing consistent pre-processing, specifically relating to signal averaging. To ensure parity in pre-processing implementation across all samples the researchers decided to change the intended event-locked data chunk durations from 500 to 375ms.

This coding error and the resulting change to the data capture duration could have significantly influenced the classification performance of the LDA models utilized. Firstly, the reduction in data volume per event deprived the models of a more robust training dataset. Additionally, P300 waveform profiles exhibit large amounts of variance within subjects even for neighbouring experimental trials. It has been shown that P300 waveform features can change dramatically depending on age [242], time of day [243], fatigue [244] and many other factors. Often, these individual differences are expressed via delayed P300 waveform peaking, past the traditional 300ms timestamp. It could be argued that the early cut-off of the data capture to 350ms could have led to the large positive deflection characteristics being excluded from the analysis. Extending this cut-off to 500ms may have allowed for the averaging to capture more of the positive deflection expected from an oddball-based experimental design. Overall, this does not negate the absence of early less-stereotypical components such as the strong negative deflection at around 200ms (N200). Further, these considerations suggest that the principal issue of overfitting may not be the sole contributing factor diminishing the performance of the LDA classifiers described herein.

Crucially, it must be noted that the use of dry-EEG to perform the data capture of visual-P300s has been well-established in the past, despite some minor relative decreases in fidelity [245]. It is highly unlikely the specific hardware utilized is at fault as there exist numerous publications that demonstrate successfully collecting and averaging ERP waveforms, including the P300 [246]. A primary factor in the presence of these waveform features is subject attention not being solely fixated on the cued emoji location (spatial/temporal bleed-over effects) and movement artefacts not accommodated for by the amplitude-wise channel-rejection protocol. These possibilities are probed further in the following experimental adaptions.

3.8.2 Discussion on Performance for Pooled-Subject Data: Pipeline 1

The results of nearly all pooled-subject dataset aggregations reveal that this method was not conducive to the production of high-performing LDA models. The method, initially, implemented in this experiment was performed to try and replicate the findings in contemporaneous CNN-based BCI signal analyses [56, 58]. Despite these poor results, the issues relating to class balancing and data acquisition, alongside features of the LDA classifiers utilized could have placed significant bottlenecks on the viability of these techniques. Further, even in the most valid instances of data aggregation, cross-augmentation single-subject data combination, the influence on classification accuracies was marginal at best (see subject 9, Table 3.12) and detrimental on the whole, as seen when comparing the single-subject averages in Table 3.12 against the same metrics in Tables 3.8 and 3.11.

In contrast to the line of argumentation here stated, the only instance of pooled-subject data aggregation that increased both P300 and Non-P300 classification accuracies above the random performance threshold was found in the class-balanced aggregated data subset (see, Table 3.12). It could be asserted that the combination of class balancing and increased data volume contributed to these increases in classification performance. On reflection, the marginality of these results (P300 Accuracy=51.69% & Non-P300 Accuracy=51.58%) prevents solid conclusions from being drawn from the data at hand. In sum, these pooled-subject and cross-augmentation explorations will continue throughout the experimental series discussed herein.

Crucially, the combination of these data would be more accurately characterised as a the blending of one relatively higher and lower quality dataset, as opposed to the aggregation of two datasets containing completely different signals. On the whole, the combination of these data in principle is sound. Despite this, the relatively low quality of both datasets is likely the key factor in accounting for the poor performance observed in these analyses. Additionally, the

blending of data sources across multiple testing sessions would necessarily lead to the inclusion of data with higher relative subject attentiveness with data acquired from subjects experiencing fatigue. Further, the aggregation of data across subjects is also highly contentious for this specific style of analysis given the fact these methods are traditionally employed nearly exclusively with ultra-high volume datasets in conjunction with various highly-robust neural network methods.

3.8.3 Class-Balancing Considerations Across Data Partitions: Pipeline 1

The difference in the application of the class balancing protocols is clear in all datasets (Flash, Inversion and Combined). As previously discussed the classification performances for the class-balanced datasets display a substantial reduction in overfitting towards the Non-P300 class. Additional differences in shrinkage values across the subsets are also evident, with the class-balanced datasets demonstrating far higher single-subject variance (for reference see, Tables A.1 & 3.11 bottom-most rows). It is important to note that these differences can not be attributed exclusively to a drop in the number of samples held within the class balanced and non-class balanced subsets, as the same trends are observed across the Flash and Inversion subsets, as in the far larger Combined dataset. Alongside these considerations, the author asserts that the drop in data volume, owing to the class balancing protocol, did hamper the ability of LDA functions to separate the classes effectively. Following this argument, the Combined dataset should have produced higher levels of classification accuracy when exposed to the class balancing protocol as it contains far more samples. Despite this, the combining of data across augmentation types, as discussed above, introduced many issues for the LDA classifiers in the training stage and did not convert into better performance.

It must be noted that there is present substantial variance in classification performance across subjects. It appears that no one subject performed well across all class-balanced data partitions tested. Subjects 1, 2 and 7 scored very low overall. These individuals appear in the bottom four subjects for at least 2 of the class-balanced data variants assessed (Flash, Inversion or Combined datasets). It must be stated that these subjects did not produce significantly more rejected trials, as per the channel-rejection protocol. Interestingly, Subjects 2 and 7 are mentioned frequently in the preceding analysis as presenting with some of the higher max and min μV amplitude ranges noted in the interpretation of the signals (as per the amplitude plots). This could well explain some of the issues surrounding the low performance. Further, it must be stated that even though Subject 3 performed well in both augmentation variants this subject performed poorly in the Combined data partition assessments. This suggests, that even the aggregation of relatively higher-quality subject data across experimental variants can be

problematic.

3.8.4 Flash *vs.* Inversion Stimulus Augmentation Methods: Pipeline 1

The significant levels of overfitting present in nearly all instances of non-class-balanced analyses for the Pipeline 1 approach significantly hampered efforts to establish the unique performance characteristics of the two augmentation methods assessed. It is only when implementing the class-balanced controls that these differences are borne out. The introduction of these aforementioned methods for the Inversion augmentation format resulted in only one subject increasing both P300 and Non-P300 class accuracies above the respective random performance threshold (Subject 3, see, Tables A.1 & 3.11). In contrast, the same changes in data organisation and classifier training programme led to 3 times as many subjects breaching the same threshold for the Flash augmentation method (Subjects 3, 6 & 9, see, Tables 3.7 & 3.11). Crucially, the relative increases in classification accuracy for Subject 3 are shown to be significantly higher for the Flash method (P300 Accuracy=77.78% & Non-P300 Accuracy=72.73%), as compared to the Inversion method (P300 Accuracy=66.67% & Non-P300 Accuracy=63.34%).

After evaluating the cross-channel average signal amplitude range plots at both the pooled-subject and single-subject levels (refer to Figures 2.8 and A.2), it is evident that the observed variance in performance is not attributable to significant differences in data acquisition. The amplitude signal plots are generally represented within a range of ±25 μV, with occasional spikes exceeding ±30 μV in both datasets. Notably, these outlier events are even less frequent in the class-balanced datasets (see Figures 2.15 and 2.24). Consequently, the author determined that the impact of these outliers is minimal and unlikely to significantly confound the results. Further analysis of the samples retained after amplitude-based channel rejection (refer to Tables 3.7 and A.1, rightmost column) shows that 685 samples were retained for the Flash augmentation method, while 679 samples were retained for the Inversion method dataset. This represents a difference in retained data volume of less than 1%. Given these considerations, it is highly improbable that the variance in classification performance between the two augmentation methods stems from differences in data acquisition protocols.

Despite the challenges in interpreting the results, the author reasoned that the Flash overlay method might have introduced spatial bleed-over effects, potentially leading to adjacency errors due to the larger on-screen augmentation area (see subsection 3.8.6.1). In contrast, the Inversion augmentation method appeared to minimize the number of variables influencing the experimental data and results. This consideration was the primary factor in the decision to

adopt the Inversion method for all subsequent experimental implementations.

3.8.5 Flash *vs.* Inversion Stimulus Augmentation Methods: Pipeline 2

The primary objective of Experiment 1 was to evaluate the effectiveness of emoji stimuli within a visual oddball paradigm for potential applications in Brain-Computer Interface (BCI) systems. Due to various challenges associated with Pipeline 1's data organization, preprocessing, and analysis stages, the author was unable to validate claims effectively (see subsections 3.3.5.3 and 3.4.3). To address these limitations, Pipeline 2 was introduced, incorporating a new baselining method, 10-fold cross-validation, exclusive single-subject Linear Discriminant Analysis (LDA) model training and evaluation, SMOTE oversampling to correct class imbalance, and a lower (0.1 Hz) Finite Impulse Response (FIR) filter. These improvements aimed to enhance the differentiation between the various stimulus presentation methods.

The application of Pipeline 2 yielded results indicating that the emoji stimuli effectively induced the P300 waveform across all subjects, as evidenced by the Cz average plots (see Figures 3.27 and 3.28). Both Flash and Inversion augmentation methods were successful in generating P300 waveforms for these stimuli. The notable improvement in plot quality was attributed to the new baselining method, which removed DC drift components, and the adjustment of the high-pass filter cutoff to 0.1 Hz, which allowed P300-related signals to emerge more clearly post-averaging. However, some drift components remained visible, and the amplitude of the signals was still relatively low (±1 μV). These issues are likely connected to the non-continuous data collection method and the short inter-stimulus interval of 125 ms.

The non-continuous data collection method involved capturing data selectively during specific trials rather than continuously throughout the session. This approach eliminated the need for traditional markers but introduced several methodological challenges. Filtering stages were applied to individual segments rather than the entire session, increasing edge effects as the filter's impulse response could not fully develop at segment boundaries. Additionally, this method likely compromised the integrity of low-frequency content due to reduced frequency resolution. Although FIR filters generally avoid phase distortions, applying them to brief segments may still introduce phase inaccuracies.

Given the small data segments, traditional artifact rejection libraries, such as the Python MNE ICA method [247], were impractical, as they rely on continuous session data. Consequently, the author implemented a channel-based amplitude rejection method with a narrow ±35 microvolt window, diverging from the conventional ±150 microvolt range. This stringent criterion

likely led to the exclusion of many relevant P300 signals, potentially impacting the amplitude ranges observed in the Cz grand averages and affecting the accuracy of the results. Furthermore, the brief inter-augmentation interval of 125 ms between stimulus Flashes or Inversions likely resulted in significant temporal bleedover, which not only decreased P300 peak amplitudes but also prolonged the relative latency of the P300 signals [223]. This overlap may have obscured the precise timing and localization of the P300 peak, diminishing the clarity of the neural responses.

A thorough statistical analysis was conducted to assess the effectiveness of Flash and Inversion augmentation methods in improving classification accuracy. Despite the detailed analysis, no significant differences were found between the methods, whether the data was treated as separate (Non-Collapsed) or combined (Collapsed). The Inversion method, applied to Non-Collapsed data, consistently surpassed the 70% accuracy threshold for every subject, though this improvement was less pronounced with Collapsed data. These findings suggest some variability in performance, but the small sample size and reliance on synthetic data for the Target-P300 class necessitate caution in drawing broad conclusions. Importantly, the increased classification accuracy did not lead to dramatic overfitting as observed in the Pipeline 1 analysis (see subsections 3.5.1-3.5.6). Further, despite using a 10-fold cross-validation procedure, the author can not discount the potential confounding influence of the SMOTE linear interpolated samples used in class balancing given the high prevalence of these waveforms in the training set.

Regarding the secondary aim of discerning differences between Flash and Inversion augmentation methods, paired-subject permutation tests and comparative plots (see Figures 2.35 and 2.36) revealed no significant differences in overall mean classification accuracy. Despite Subject 8 achieving the highest overall mean accuracy with the Flash Non-Collapsed data partition (see Figure 2.35), there was no significant difference between the methods in terms of endpoint classification accuracy. This suggests the author's original decision to implement the Inversion method for all subsequent experimental stimulus variants on balance is the optimal decision, given the previously noted concerns relating to spatial bleedover for the larger Flash augmentation on-screen area.

3.8.6 Experimental Modifications

3.8.6.1 Discussion Relating to Augmentation Sizes

The differences in augmentation area across each technique are significant, with the Flash overlay square occupying 40mm^2 (see, Figure 3.2) and the Inversion method, on average, occupying 18mm^2. The reduced size of the inversion method augmentation stimuli could have led to a significant drop in visual salience. A reduction in visual saliency reduces the likelihood of maintaining high subject concentration and reduces the reliability of inducing the propagation of robust P300 waveforms [130, 248]. In other words, the Flash method may have led to a P300 waveform profile with a stronger peaking characteristic, owing to its higher visual salience. When inspecting the respective average wave plots (see, Figures 2.9, 2.16, A.3 & 2.25) the previous assertions do not support the results. The average plots resulting from the pooled-subject analysis for the Inversion method data present P300 average signals with more traditional waveform characteristics, in conjunction with lower volatility Non-P300 average signals. As mentioned above, the aggregation of pooled-subject data in these quantities is not typically implemented for these analyses owing to the significant individual differences present in P300 waveform expression.

When inspecting both the Flash and Inversion class-balanced instances (see, Figures 3.13 & 3.20), there are clear differences in the prevalence of P300 and Non-P300 characteristics between the Cz grand average signals. The signals suggest a large degree of bleed-over between the P300 and Non-P300 datasets as both plots share many of the same temporally offset waveform features. In other words, there appears to be a coherency between the data present in the P300 and Non-P300 subsets. This may indicate some instances of adjacency error, double flashing oversights in the stimulus programming augmentation order, or poor attentional focus from the subject. Given the similarity in signal quality, the data features utilized by the Linear Discriminant functions to separate and predict the respective classes remain unclear.

These differences in coherency between the P300 and Non-P300 waveform profiles could be related to the differences in the relative size of the augmented visual array across the emoji in the Inversion method. This is due to some emoji being quite simplistic requiring only minimal illustrative ornamentation, for example, the 'Neutral Face' central emoji stimulus. This is contrasted against more ornate emoji such as the far-right emoji 'Smiling Face with Heart Eyes' (see, Figure 3.1). This target features significantly more regions of black to illustrate the additional emotional content expressed within the emoji. This variance in total area augmented is not shared in the Flash method stimuli, as all receive a uniform 40mm^2 white overlay square.

The reduced coherency could be beneficial to classification performance, as the differentiation between Non-P300 and P300 waveforms is the key factor in this analysis pipeline. In the same way, having different colours assigned across emoji, the different spatial patterns of the emoji inversion augmentation could enhance emoji separability. However, it must be noted that the enhanced discriminability of the colour-differentiated emoji only works if the colour is vivid/vibrant enough to be consciously distinguished. Along these very same lines, the volume of inverted emoji colouring must first be big enough to induce a P300 waveform, once above this threshold the use of spatial patterns to enhance separability could be exploited to boost P300 peak amplitudes. To summarize, these results suggest, at least for some participants, that a reduced visual saliency for the Inversion method contributed to the absence of any subjects producing higher than random classification performance. The directionality of this visual salience and the relative influence on classification performance will be probed in the following experimental series.

3.8.6.2 Stimuli Colouration Justification

It must be noted that the author did not consider the coherency of emoji valence in the stimulus colour modification phase. In other words, no attempts were made to match emoji target colours based on perceived or real emotional connotations to the colours utilized. For example, the perception in Western culture to associate the colour red with negative emotional states did not inform the assignment of the colour red to any particular target emoji, rather the colour modification was done purely to enhance the ability of subjects to attend each respective target, relative to adjacent targets.

That is not to say the colouration modifications were performed entirely according to mathematical exactitude. The legibility of the emoji is important, irrespective of implicit positioning on a valence scale. Colours using darker hues would prove far more difficult to discern and may potentially confuse users. This can be observed within the stimuli set developed as seen in, Figure 3.1. For instance, when comparing the clarity of the most agreeable (yellow, 'Smiling Face with Heart Eyes', furthest right) and the second to most disagreeable (purple, 'Pensive Face', second from left) it is clear that the yellow base colour emoji is far easier to read than its darker-hued counterpart. Along these lines, brighter colours were used were possible to enhance emoji legibility.

The considerations outlined here do not preclude the future experimentation of other colour sets for those with colour blindness or even user preference-based modifications. Again, this

experimental paradigm was designed to serve as a universal baseline. On reflection, it may have been interesting to test if assigning colour based on perceived emotional connotation to the emoji targets can influence subject performance. Crucially, any increase in performance would likely be regionally specific, reducing the scope of any resulting EEG-based P300-speller. Further, the selection of emoji and assignment of these stimuli across the 1-dimensional emotional valence scale employed did not follow a strict methodology. Again, the process of constructing the emoji array was done to approximate a condensed scale of emotional reactions commonly utilized in daily life and not to systematically represent all the colours of human emotion available for expression in typical-healthy individuals.

3.8.6.3 Impedance-Based Channel Rejection

The use of amplitude-based channel rejection protocols is well established in the literature as an effective heuristical analogue for confounding acquisition events such as movement artefacts [249]. The author concluded that channels showing signal ranges exceeding $+/-35\mu V$ would negatively impact waveform averages and were removed to ensure data quality standards. After inspecting the numerous signal average plots the protocols implemented performed adequately given the restrictions imposed by the non-continuous sampling method, as no plots express outlier waveform features characterised by excessively high amplitude values. Despite this, the effectiveness of such an approach may not hold up well in real-world lab situations. Crucially, as noted above, such an approach may penalise those subjects with very strong P300 waveform deflections and thus reduce the overall dataset quality. In response to these considerations, an alternative method utilizing channel-wise impedance data for sample rejection is discussed in the following experimental variant (see subsection 4.3.5.4). It must be noted that impedance values are a more direct means of assessing movement artefacts and the quality of sensor seating against the skull.

3.8.6.4 Inter-Stimulus Interval Increase

The brief inter-augmentation interval of 125 ms between stimulus Flashes or Inversions likely caused significant temporal bleedover, where the effects of successive stimuli overlapped and interfered with the EEG signals. This overlap is expected to reduce the amplitude of the P300 peak and extend the relative latency of the P300 signals, thereby diminishing the clarity and accuracy of the neural responses. To mitigate these issues, the inter-augmentation interval was increased to 150 ms. However, this adjustment appears to be suboptimal, as evidenced by the findings of [223], where intervals of up to 250 ms between flashes were implemented. The increase implemented here of just 25ms was done in an attempt to keep the maximum achiev-

able information transfer rate low. Here, the author likely outweighed the relative importance of prospective ALS patient user opinions regarding operational speed [3, 4], as compared to the need for a system capable of functional performance, 70% classification accuracy, across a range of individuals [250]. Given this precedent, it would have been more effective to increase the intervals further to align more closely with those used in [223]. The implications and rationale for these decisions are detailed in the following chapter.

3.8.6.5 Subject Training Improvements

Throughout these conclusions, the author has questioned the attentiveness of subjects multiple times. The experimental design implemented herein features a low-skilled, highly repetitive task, for a relatively long duration. This style of experimentation can be hampered by a rapid decrease in data quality due to subject fatigue and disinterest. In all subsequent experimental designs, the researchers dramatically increased the amount of time spent pre-training and pre-testing subjects in concert with longer inter-block breaks to minimise these effects. Further, the introduction of a pre-testing localization task was implemented to assess subject P300-peaking profiles before data collection and to inform the application of a longer inter-block break.

Chapter 4

Experiment 2: Variable Array Density Assessments

4.1 Aims

The aim of Experiment 2 is to probe the influence of array density on classification performance. The subjects were presented with 3 stimulus variants differing in the number of emoji targets in the visual array. These involve 3 so-called 'levels' of array density each featuring 3, 5 and 7 Emoji targets respectively. The valance gradient of the emoji stimuli arrangement is kept consistent, with the emoji targets organized from disagreeable-to-neutral-to-agreeable in emotional content (left-to-right). Crucially, the analyses of Experiment 1 led to many new adaptations in experimental design and were implemented to address the shortfalls of the initial experimental implementation. Specifically, these relate to rectifying data buffering methods to ensure consistent 500ms sampling post-augmentation event, the application of an impedance-based channel rejection protocol and an increase in emoji-stimulus size. Further, as applied in Chapter 3, both the Pipeline 1 (3.3.5.1, 3.4) and Pipeline 2 (3.3.5.3 & 3.4.3) approaches to the data organisation, pre-processing and analysis are applied here, for further information see the noted subsections.

4.2 Stimulus Reduction Rationale

There are many counteracting influences in the design of oddball-style P300-based BCI speller experiments. As the number of visual targets increases, the likelihood of whether the next target augmented is the actual target being attended to decreases. This is discussed at length in the Literature Review subsection 2.5.1 with reference to [137, 138] and Figure 2.1 [125], relat-

ing directly to the inverse relationship between stimulus probability and P300 peak amplitude. As noted in [251], this should benefit systems implementing visual arrays with a high number of targets, as a greater number of targets decreases the chance of neighbouring targets being augmented successively, given the increase in potential spatial locations to be randomly selected in the presentation scheme. Along these very same lines, the incidence of temporal bleed-over effects should also theoretically decrease and ultimately improve the quality of the resulting EEG data captured.

Despite this, the handful of studies directly comparing speller systems utilizing different numbers of targets for P300-BCI applications often report that lower-density (fewer targets) arrays produce higher classification accuracies. This is demonstrated in [252], here the authors compared the performance between a reduced 3 *x* 3 character matrix and a complete 6 *x* 6 alphanumeric speller. The researchers also created additional stimulus variants by introducing two different inter-stimulus intervals for both layouts of 175ms and 350ms. The systems were tested during 5 sessions in 5 healthy subjects over the course of 3 weeks. The results revealed mean accuracies of 61.25% and 69.38% in the 3 *x* 3 matrices for the 175ms and 375ms variants respectively. These greatly outperformed the 6 *x* 6 175ms and 375ms variants, achieving just 53.75% and 48.13% respectively. Crucially, the 3 *x* 3 175ms (7.7bpm) inter-stimulus interval variant also outperformed the 6 *x* 6 175ms (5.83bpm) variant in terms of information transfer rate.

In line with these findings, [19] tested a variety of region-based (alphabetical and frequency groupings) and 6 *x* 6 matrix spellers (single character and row/column) across a range of performance metrics. Both region-based systems involved variants of presenting alphanumeric targets in 7 clusters on screen. The selection of one letter cluster is followed by the remapping of the cluster contents into the 7 onscreen locations and a second operation is required to select the target character. Here the authors demonstrate that the region-based methods (90.6% and 86.1%) outperformed the matrix-style layouts (72.2% and 85.0%) in terms of both accuracy and user acceptability when tested in 6 typical healthy subjects. Similar results are also reported in [20] comparing a 6 *x* 6 matrix speller with a region-based presentation method using a 6-cluster design for a larger 12-subject sample size. Here the authors report a significantly higher accuracy for the region-based (93.47%) speller, in contrast to a single-character 6 *x* 6 speller method (89.32%) for online trials. Further, the authors note that the P300s generated in the region-based method demonstrated higher amplitudes and lower latencies. This is attributed to the greater degree of separation between targets and the resulting reduction in spatial interference of said targets.

Notably, these findings suggest that an increase in the number of targets onscreen, if not compensated for by a relative decrease in the size of targets and a contingent increase in the distance between targets, can reduce the usability of visual-P300 speller systems. This is well illustrated in [141], here the authors compared a 6 x 6 alphanumeric P300 speller matrix of differing sizes between small, medium and large for symbol sizes (0.42cm, 0.79cm and 1.17cm), symbol separation (0.55cm, 1.04cm and 1.53cm) and overall matrix area (5.27cm, 9.98cm and 14.69cm). Both healthy and ALS patient subjects reported significant results in terms of the highest user satisfaction, and lowest loadings on the NASA-TLX [253] user preference scale items for effort, physical demand and temporal demand regarding the the medium-sized array. Interestingly, the most statistically significant result here was the separation between the medium and smallest-sized array in terms of user satisfaction, with subjects reporting discomfort with the latter layout. This suggests that there are limits to compensating for a higher array density by shrinking the symbol sizes to achieve greater target separation.

Along these very same lines, increasing the number of targets in a given visual array also inevitably increases the duration of the trial period and in turn, reduces the information-transfer rate of the system [142]. Researchers can compensate for the increased duration time per trial by reducing either the total time spent augmenting the target or the duration of the inter-stimulus intervals [128]. These modifications can also have negative side effects, as a reduction in the duration of augmentation time could potentially lead to issues in terms of P300 peaking amplitudes and decreased latencies [128]. Moreover, reducing the distance in time between the onset of augmentations could lead to an amplification of the previously mentioned temporal or spatial bleed-over effects, resulting in P300 waveform peaks shifting across yet more windowed data event chunks. These kinds of interactions highlight the numerous methodological design issues researchers can come across when attempting to develop a novel visual P300 speller experiment.

In sum, based on the findings of Experiment 1 and the literature noted above it was reasoned that simplifying the experiment would decrease the number of variables that are potentially contributing to the low-performance metrics observed. Further, this reduction could potentially improve the accuracy metrics of the prospective subjects via a predicted increase in P300 peak amplitudes and a drop in P300 latencies as per the findings of [252]. Moreover, the experiment herein utilizes 3 stimuli array variants, designed specifically to explore the influence of spatial bleed-over effects on classification performance by reducing the number of targets on-screen, in addition to increasing the distance between targets on-screen. As mentioned in a

previous subsection (see subsections 3.8.4 & 3.8.5), the Inversion augmentation method was utilized for this experiment as the quality of visual modification employed introduces a far lower risk of adjacency error owing to the smaller augmentation area.

4.3 Method

Here are outlined the methods employed in the investigations relating to Experiment 2. Broadly, this features the implementation of a staggered design to probe the influence of visual array density on visual P300 propagation using 3, 5 and 7 Emoji array designs (see, Figures 4.2-4.4). Further, the efficacy of a 1-emoji localizer task is evaluated in terms of subject training, pre-screening data quality improvement and as a tuning pre-stage for the LDA model used in the main experimental task.

4.3.1 Participants

A total of 5 neuro-typical subjects were recruited from the Durham University student population consisting of 3 males and 2 females (mean age of 28.8 years and age range of 24-35 years). The subjects were screened before the onset of the experiment to ensure all presented with normal or corrected to normal vision, had no history of clinical mental illness or epilepsy and were not currently experiencing a skin-based ailment of the scalp. No subject received payment to participate in the experiment. Ethical approval and oversight were granted by the Durham University Psychology Department Ethics Sub Committee.

4.3.2 Equipment

The acquisition of all EEG data collected for this study was acquired using the Cognionics Quick-20 headset (Cognionics, San Diego, USA). The μV amplitude and impedance (Ω) data streams (both 500 Hz) were handled via the LabStreamingLayer package. On-screen stimuli were rendered exclusively by a dedicated NVIDIA GTX 750ti GPU (2GB VRAM). Note, that before the onset of all trials the EEG headset sensors and rest points were thoroughly cleaned using anti-bacterial gel.

4.3.3 Stimulus Presentation

The stimuli were designed and controlled via the PsychoPy [221] Python library and displayed using a Samsung LED S27A35OH computer monitor (60 Hz refresh rate, 68.5cm diameter). These were displayed at a fixed 0.8m distance from each subject (seated). A total of 7 Emoji stimuli were used from the OpenMoji database [222]. All emoji stimuli utilized in these and proceeding experimental variants utilize the larger 27mm diameter sizing, this increase in scaling was easily achievable due to the use of SVG format image files.

4.3.4 Localizer Task

The localizer task employed in Experiment 2 is a simple 20-event pre-screening protocol (see, Figure 4.1). The purpose of this stage is to provide subjects with an additional phase of stimulus training, further, it enables the EEG headset electrodes to sit optimally against the scalp to ensure minimal signal impedance and lastly, it provides the experimenter with a pre-screening tool to assess P300 waveform quality before the onset of the main experiment. This task involved the presentation of a single emoji stimulus, onscreen for one second. Following this, the stimulus is either augmented or left in its original presentation format (duration 0.05 secs). As noted above, the Inversion stimulus augmentation method was utilized (see, Figure 4.1 central graphic). Briefly, this involved the inversion of all black colouration in each respective target emoji to white colouration. The onset of all augmentation events adheres to a randomised and partially stratified presentation schedule. This was employed to avoid excessive groupings of either class (augmentation *vs.* no augmentation events) throughout the course of the experimental period. After the augmentation or non-augmentation event, an inter-trial interval of 300ms is observed to allow for any post-propagation refractory processes to be concluded and prevent temporal bleed-over effects influencing subsequent trials (see, Figure 4.1 lower panel).

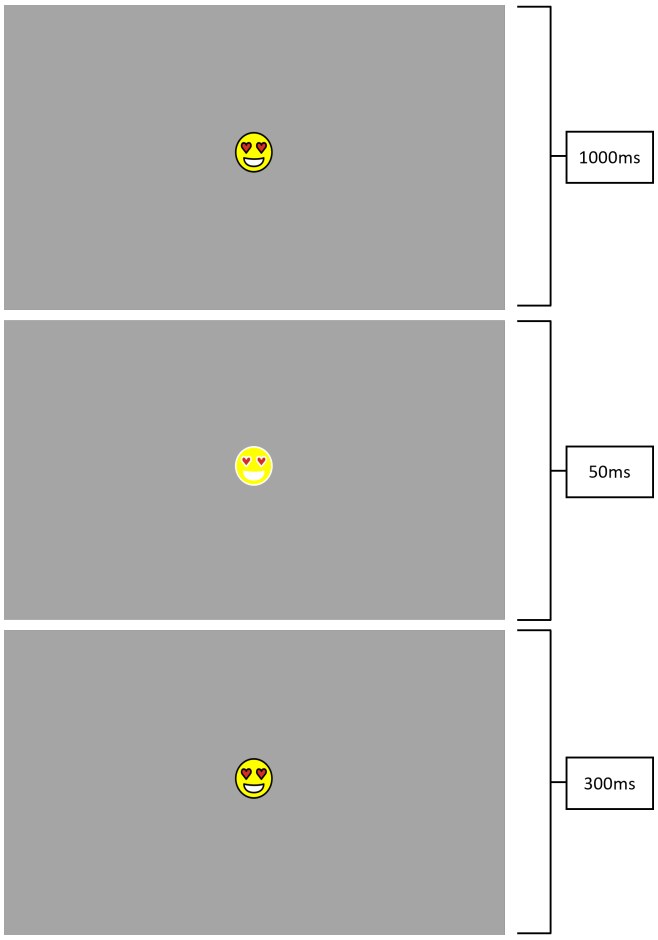

Figure 4.1: The upper plot of the figure displays the single default onscreen emoji (27mm diameter) utilized in the localizer task (1000ms). The central plot shows the visual augmentation phase of the emoji utilized to generate the P300 neural response (50ms). The final, lower plot shows the refractory pause period displaying the same standard, non-augmented emoji stimulus (300ms). Note, that the plots here illustrate a series of screenshots for the P300 trial augmentation variant. In all Non-P300 augmentation instances, no visual change to the default emoji stimuli is applied. Further, the order in which the P300 and Non-P300 trials occur is randomised by a structured control procedure to avoid large clusters of the same events occurring. This prevented repeatedly augmenting the stimulus at regular intervals and avoided the presentation of stimuli with long periods of inactivity. Please refer to the accompanying text in, subsection 4.3.4 Localizer Task, for additional information.

As stated above, this process is repeated for a total of 20 trials and occurs over a period of 27 seconds, as per (1000ms + 50ms + 300ms) × 20. The concurrent acquisition of EEG data was carried out simultaneously during the experimental sessions. This produced 20 trials of 500ms data windows from electrode locations: Fz, Cz, Pz, P4, P3, O1, O2, A1 and A2 using a 500 Hz sampling rate. All data gathered via this method were pre-processed using the same notch filter, bandpass, referencing and channel-rejection protocols employed in Experiment 1 (see subsection 3.3.5.2). Crucially, cross-trial signal averaging into grouped sequences was not performed, as per the main experiment.

In half of the trials (10 trials) the visual augmentation of the target stimulus is applied; the other half (10) involves no augmentation. These events are randomly arranged throughout the trial period. More traditional oddball methods require the distinction of a so-called 'deviant trial' which differs from the consistent baseline stimulus. The distinction is amplified by the reduced frequency of the deviant trial presentation. The author designed a 1:1 ratio for this localizer task to reduce any possible influences of overfit from class balancing issues as recorded in detail throughout Experiment 1.

It is crucial to clarify that the task described here does not simply involve the repeated augmentation of the singular emoji stimuli in discrete 1350ms blocks. If this were the case the task would more accurately be re-classified as an ultra-low frequency (0.741Hz) SSVEP task. The augmentation scheme is initialised with 10 P300 and 10 Non-P300 event codes. These are then shuffled randomly and the list is then checked with a conditional rule to assess if any clusters of the same event codes are neighbouring one another in groups of a length exceeding 4. If this is the case, the list is re-randomised and checked again until the conditional returns false. This combination of randomisation and cluster checking prevents the augmentation order from following a repetitive on-off procedure, as well as ensuring a relatively even distribution of P300 and Non-P300 events.

Following the completion of the localizer task, the data were grouped into classes and averaged across trials to create average signal plots. The experimenter utilized these plots to probe the viability of the subject before the onset of the main experimental variants and also further train the subjects before the acquisition of the main experimental data. The decision was made later in the analysis development process to try and utilize these localizer signals for the pre-training of the LDA classifiers as an initialization stage, this will be discussed further in subsequent subsections. This was hypothesized to potentially diminish some of the overfitting effects observed in the previous experimental implementation (Experiment 1).

Note, that this methodology deviates substantially from the traditional oddball paradigm, particularly concerning the ratio of P300 and Non-P300 events. Despite this, the Non-P300 event code involves no change to the default stimulus array (see, Figure 4.1, upper plot). Therefore, in any given experimental run the P300 augmentation event can never account for more than 500ms (50ms x 10 P300 event codes) out of the entire 27-second experimental period. Further, owing to the variance in the presentation sequence introduced by the randomisation protocol the author asserts that the augmentation of the stimulus via colour inversion constitutes a rare, unpredictable, yet expected change in the qualities of the attended stimulus. In sum, at best this method can be termed a variant in the oddball paradigm.

As discussed in the previous literature review (see subsection 2.5.1), the amplitude and latency of the P300 peak are inversely related to the probability of stimulus augmentation (see Figure 2.1). Given the 50% likelihood of stimulus augmentation per trial, the proposed localizer task is expected to produce lower-quality P300 waveforms. However, it is important to note that the perceived likelihood of augmentation may be lower, as the non-augmented state is displayed for a significantly longer duration during the experiment. This is because the non-augmented state here functions as the standard stimulus and remains unchanged on screen for the vast majority of the experimental duration. These compromises in P300 waveform quality were made to eliminate potential spatial and temporal bleed-over artefacts associated with the presentation of multiple stimuli in P300 oddball contexts. For further information see subsection 2.5.1 P300.

Given the pervasive issues related to class-balancing noted throughout the thesis thus far relating to the Pipeline 1 approach (see subsection 3.3.5.1), the author reasoned that the use of a binary system would avoid many of the associated complications. Arguably, the intention of the author to enforce class balance in the localizer task dataset impeded the effective design of the assessment. The utility of this technique in generating P300 waveforms, the efficacy of this system as a localizer pre-screening task and the adaptations necessary to modify this procedure into a valid P300 task are discussed further in subsections 4.6.7.4 Summary & 4.6.8.3 Localizer Data Pre-Training Considerations. Note, that all adaptations involving the Localizer task data are all undertaken using the Pipeline 1 approach and at no point feature in any of the associated Pipeline 2 evaluations.

4.3.5 Main Experiment

All three variants of the experiment and the pre-experimental localizer task defined herein subsample stimuli from the original 7 Emoji reference group (see, Figure 4.2). Further, all main experimental variants included the neutral ('Neutral Face'), most agreeable ('Smiling Face with Heart Eyes') and most disagreeable ('Persevering Face') emoji. This was done to ensure that the range of emotional valence was kept constant despite a loss in resolution of the emotional expressivity in the less dense 3 Emoji (see, Figure 4.4) and 5 Emoji (see, Figure 4.3) variants.

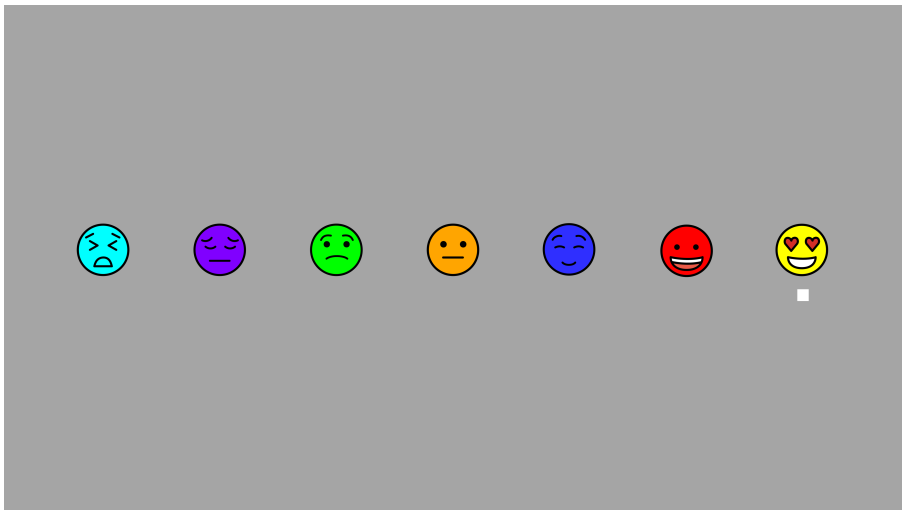

Figure 4.2: The image above is a screenshot of the 7 Emoji experimental variant. Emoji stimuli, diameter 27mm, are arranged in an approximated continuum of emotional valence from disagreeable to neutral to agreeable (left-to-right) in evenly spaced intervals of 85mm. The white square positioned below the rightmost emoji stimulus is a cue to indicate the target a subject must attend during the course of the following trial.

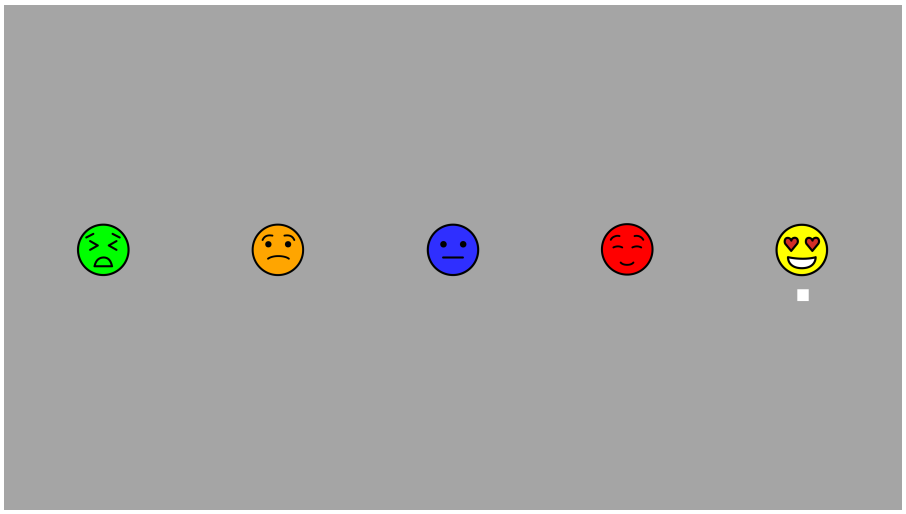

Figure 4.3: This figure displays a screenshot of the 5 Emoji experimental variant, with the emoji stimuli positioned at evenly spaced intervals of 110mm.

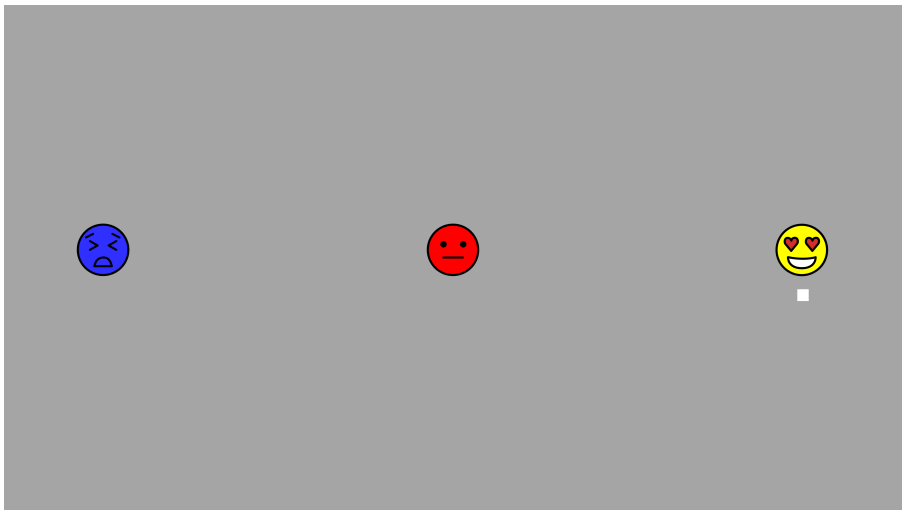

Figure 4.4: The above figure shows a screenshot of the 3 Emoji experimental variant, with the emoji stimuli positioned at evenly spaced intervals of 230mm.

The same colour modification protocol in Experiment 1 was extended in the design of this task (see subsection 3.3.3). In summary, the visual distinguishability of proximal neighbouring emoji was enhanced via the inclusion of a base colour. As can be seen in Figures 4.2, 4.3 and 4.4 each of the experimental variants uses the same colour modification protocol to maintain the distinguishability of target emojis relative to immediate neighbours.

4.3.5.1 Presentation Specifications

At the start of each trial, subjects are presented with a white cueing square (1000ms), positioned underneath the target emoji (see, Figure 4.2). Each stimulus is then augmented (duration=0.05s) and an inter-stimulus interval of 100ms is observed before the execution of subsequent augmentation events. Once all emoji have been augmented according to the randomisation protocol an inter-sequence interval of 500ms is undertaken. This process describes one sequence and is repeated 5 times in total to complete a full trial run. Once all 5 sequences have been completed, an inter-trial interval of 1 second is introduced. Each experimental variant consists of 30 trials, with each experimental variant: 3 Emoji, 5 Emoji and 7 Emoji requiring 3.3, 4.1 and 4.8 minutes to complete respectively. One trial for the 7 Emoji variant is computed as per: 1s + (((0.05s + 0.1s) $\times$ 7) + 0.5s)$\times$ 5) + 1s.Note, that given the duration between the onset of each augmentation event is equal is 0.05s and the duration of the inter-stimulus interval is 0.1s, it is likely that an SSVEP corresponding to $1/0.15s = 6.67Hz$ will emerge from any subsequent average signal plots. For further information and details of the efforts to address this in the pre-processing pipeline via notch filtering refer to Table 3.1 and subsection 3.3.5.3 Data Pre-Processing Pipeline 2. Subjects performed the 3 experimental variants according to a counter-balanced protocol, ensuring that any fatigue effects would be distributed evenly across subjects for each respective experimental type.

4.3.5.2 Randomisation Protocol Differences

The smaller number of visual targets per array dramatically decreased the total amount of non-consecutive randomised augmentation schedules available. This inevitably introduced some instances in the 5 Emoji variant that featured the augmentation of direct target neighbours. Further, for the 3 Emoji experimental variant, this condition of non-consecutive augmentations was not achievable as 2/3 of the targets neighbour 1/3 of all targets onscreen. For this experimental variant, a simplified process of pure randomisation was adopted. It must be noted that the phenomenon of double flashing (repeated augmentation of the same stimuli) was still consistently avoided.

4.3.5.3 Parameter and Data Window Modifications

In response to the quality of signals observed in Experiment 1, the specific inter-trial delays were modified. These were increased for the inter-stimulus interval (0.125 to 0.15 secs) and inter-sequence intervals (0.375 to 0.5 seconds) to address issues of temporal bleed-over across separate emoji events. The inter-trial intervals specifically were increased to widen the data window per event to ensure individuals presenting with delayed P300 peaking could have their

signals captured and analysed.

The inter-stimulus intervals were increased to enable subjects to follow the augmentation 'counting' strategy outlined in previous studies to increase subject attention during the task, as well as to boost P300 peak amplitudes and reduce P300 latencies. It must be noted that any increase in interval delay periods was introduced to the detriment of final ITR values. These were employed to compensate for P300 refractory and temporal bleed-over effects. In addition to probing the effects of array density, this experiment aimed to try and establish definitive parameters for the collection of high-quality P300 data, as opposed to developing a high-speed emotion-based P300-communication paradigm.

4.3.5.4 Data Pre-Processing: Pipeline 1

All data-organisation and pre-processing methods noted herein relate to the Pipeline 1 method, for further information please refer back to section 2.3.5.1, Data Pre-Processing: Pipeline 1. Each trial of the experiment consisted of 5 data sequences. These included 1 P300 event and either 2, 4 or 6 corresponding Non-P300 events for the 3 Emoji, 5 Emoji and 7 Emoji variants respectively. Every data event was parsed into 500ms time-series data chunks. A similar pre-processing pipeline described in Experiment 1 was employed for these data. All pre-processing steps described herein were conducted using the Python NumPy [254], Scipy [226] and Scikit-Learn [227] libraries. Firstly, this involved data baselining to zero and referencing all electrodes sampled to the A2 electrode. The second stage consisted of applying filters, namely a 50 Hz notch filter for powerline removal and a bandpass filter (1-15 Hz).

Subsequently, each electrode was passed through an impedance-based channel rejection protocol. Initially, the variance for each electrode sampled was computed (excluding the A2 reference electrode). Following this, a median variance value was calculated encompassing all electrodes sampled. The individual electrode variances were then compared to this median (cross-channel) variance value. If the electrode in question expressed a variance in ohms (Ω) value three orders of magnitude greater than that of the median then the channel was deemed as containing a high incidence of confounding data features and was removed from the analyses. In the event that a single channel was retained, the trial would be deemed unviable and discounted from the analyses.

The results of these analyses largely mirror those collected in Experiment 1, with a low incidence of retention for Pz and crucially a high incidence of retention for the Cz electrode.As noted in the previous chapter (see, subsection 3.3.5.2 Channel-Amplitude Rejection) and the

caption for Figure 3.6, the poor cap design of the Pz sensor mounting arm led to poor seating of the sensor to the skull and increased the prevalence of abnormally high sample amplitudes and associated impedance values for this location. These limitations ultimately prevented the vast majority of these Pz sensor samples from being included in the final analysis for most of the subjects tested.

Finally, the remaining channels were averaged to amplify embedded waveform features and improve the signal-to-noise ratios. Importantly, multiple electrode sites were included in the evaluation of the P300 waveform as the positioning of the headset differed across subjects. Experimenters adhered stringently to the methodical positioning of the kit, specifically over the Cz location. In some instances, this was not possible due to variations in head size and shape. It was reasoned that averaging signals across the central and parietal regions would provide a more coherent mapping pattern of the P300 waveform propagation across the scalp.

4.3.5.5 Data Pre-Processing: Pipeline 2

In a similar vein to that taken in Experiment 1, all data relating to Experiment 2 were re-organized and pre-processed using the alternative Pipeline 2 method (see subsection 3.3.5.3 Data Pre-Processing: Pipeline 2). This involved the application of numerous additional stages and adaptations, including a revised baselining method, a lower high-pass filter cutoff of 0.1Hz, and the implementation of a SMOTE-based oversampling method to artificially boost the number of Target training data samples. For additional information on the differences between the two pre-processing approaches, please refer to Table 3.1. Furthermore, as the stimulus presentation scheme for Experiment 2 involved an increase in the stimulus augmentation onset interval from 0.125 to 0.15s (see subsection 4.3.5.1 Presentation Specifications), a modified SSVEP-targeted notch filter at 6.67Hz was introduced to accommodate any signals induced by the updated stimulus presentation scheme. Note that the Pipeline 2 approach is applied exclusively to the main experimental data.

Note, that at no point was any Localizer task data used in relation to the Pipeline 2 approach. This comprises data orgnaisation, pre-processing and analysis. The integration of the Localizer samples was implemented as a strategy exclusivey in Pipeline 1 for addressing issues related to class balancing and the low quantities of samples for both the Target-P300 and Non-Target Non-P300 datasets. Here, the Pipeline 2 approach addresses these same issues via the implementation of the SMOTE linear interpolation-based oversampling method.

4.3.6 Analysis: Pipeline 1

All data herein were evaluated using the LDA and parameter grid search methods outlined in Experiment 1 (see subsection 3.4). The classifiers at all levels (pooled-subject and single-subject) and for all variants (Localizer, 3 Emoji, 5 Emoji and 7 Emoji) were trained using 90% of the respective dataset and evaluated using the remaining 10%. For the pooled-subject assessments, the test data was comprised of a blend of all 5 subjects sampled. Crucially, all data in the test set was novel to the classifier. Note, that despite differences in the number of targets on screen for the three stimulus variant described herein all analyses performed is offline and conducted at the sample-level leading to a random performance threshold of 50% for all models evaluated. For further information see subsection 3.4.2.

4.3.6.1 Localizer Data and Initialization

Principally the localizer task was employed to familiarise the subjects with the main experiment, allow experimenters to monitor changes in impedances between the tasks and provide a visual plot to assess the quality of P300 propagation before the onset of each experiment. Note, that the aforementioned preprocessing and LDA analyses pipeline were also applied to these data. The training and test datasets were comprised of 18 and 2 events respectively (9:1 train-test data partition). It was unlikely that such low data volumes would allow for highly accurate classification of waveforms, these tests were purely exploratory as researchers aimed to investigate whether any significant correlations could be discovered between performance at the localizer and main experiment levels.

Further, an alternative analysis within the Pipeline 1 approach involved utilizing these signals with a method more closely approximating those undertaken in real-time speller conditions. This involved using the data gathered during the localizer task in a pre-training stage before exposing the classifier to training data from the main experiment. It was hypothesized that this could help build out the software mechanics required for real-time speller paradigms as well as provide the classifiers with more subject-specific data.

4.3.6.2 Class-Balancing Considerations

The class balancing protocol adopted in Experiment 1 was not reimplemented here. It was previously concluded that the process of class balancing in the previous experiment led to the only significant results attained at the single-subject level. This was done primarily to remove the confounding influence of overfitting observed through the non-class-balanced results. Despite this, a far smaller number of trial events were collected per variant for Experiment 2. The

application of class balancing in this instance would have dramatically reduced the volume of events to evaluate the experiment. As mentioned in the section above, class balancing was addressed using an alternative classifier training initialization step that utilized pre-experimental localizer data. This is far more in line with current methods for P300-based EEG-speller applications.

Moreover, the process of class balancing dramatically reduces the ecological real-world feasibility of any resulting system tested within these limits. Ultimately, the aforementioned methods do not position this experimental series firmly for the planned real-time predictions explored in Experiment 3. In summary, each trial undertaken produced either 3, 5, or 7 averaged signals corresponding to one emoji in any given array. One in each group of averaged signals was labelled as the P300 event, with the rest labelled as Non-P300 events. All such events were fed into the classifier individually. At no point was any method of prediction ranking employed. In other words, it was assumed that only one signal would be classified as a P300 waveform, and no contingency was put in place to accommodate for the possibility that more than one emoji would be predicted as the target class. This is because the classifier, solver combination of LDA and Least Squares Method does not output a metric which can be ranked in terms of probability or similarity to a reference example.

4.3.7 Analysis: Pipeline 2

Here, in precisely the same manner employed for Experiment 1 (see subsection 3.4.3 Analysis: Pipeline 2), the main experimental data for three promising subjects (Subjects 1, 3 & 5) were re-analyzed using the Pipeline 2 approach. This involved exclusively training and assessing all related LDA models via a 10-fold cross-validation procedure using data only from individual subjects (see subsection 3.4.3.1). To address the class imbalance between Target and Non-Target classes for the 3-Emoji (1:2), 5-Emoji (1:4), and 7-Emoji (1:6) experimental variants discussed in this chapter, the same SMOTE oversampling technique was implemented (see subsection 3.4.3.3).

This involved initially partitioning the non-augmented subject trial data into a training set and a test set using a 9:1 split shuffled and stratified to ensure an even representation of Target and Non-Target samples throughout both sets. Following this, the Target-P300 samples within the training set underwent a linear interpolation process to generate enough new samples to match the relative amount of Non-Target samples in the training set. This process was repeated for all 10 folds in the cross-validation procedure, and at no point were any synthetic samples included in the test set. These analyses, conducted at the single-subject level, produced a

series of mean classification accuracy metrics (Overall, Target, and Non-Target) which were compared for significance against chance-level performance via one-tailed one-sample t-tests (see subsection 3.4.3.2).

The same Non-Collapsed and Collapsed data partitions are also applied to these data. This involves aggregating neighbouring trial samples to artificially simulate an increase in the number of sequences per trial from the default 5 (see subsection 3.3.5.1) to 10. This manipulation pertains to the Collapsed data partition and is performed before dividing the trial samples into the respective training and test splits for all 10 cross-validation folds. The relative efficacy of this method for enhancing the mean classification accuracies of the associated LDA models is investigated through a non-parametric permutation test on paired-subject matched mean classification accuracies. Additionally, a similar non-parametric permutation test is performed on paired-subject performance metrics across each of the 3 experimental variants. The mean classification accuracies are tested in iterative couplets, i.e., 3 *vs*. 5, 5 *vs*. 7, and 3 *vs*. 7. This approach is used to gauge the relative performance characteristics of the subjects across the different experimental variants.

4.4 Results: Pipeline 1

Here all results relate to data organised, pre-processed and analyzed using the Pipeline 1 approach, see subsection 3.3.5.1 and 4.3.5.4 for further information. A total of 2250 events were sampled over the course of all three experimental variants tested. Each emoji in all variants discussed produced a total of 150 events. Along these very same lines, the number of P300 target events captured per subject was fixed at 150 events, with the remaining data comprised exclusively of Non-P300 events. The same train-test (9:1) data partition employed in Experiment 1 was reimplemented here. Crucially, all data used for evaluation purposes was never included in the training data for each respective analysis variant. All trials sampled were included for analyses following the impedance-based channel rejection protocol. This is due to significant improvements in subject instruction and impedance monitoring as provided by the aforementioned localization tasks.

Experimental Variants	3 Emoji	5 Emoji	7 Emoji
Total Numer of Events	450	750	1050
Total Number of Test Events	45	75	105
Events per Subject	90	150	210
Test Events per Subject	9	15	21
Events per Subject per Emoji	30	30	30
Test Events per Subject per Emoji	3	3	3

Table 4.1: A table showing the differences in experimental variants in terms of the number of events. The Total Number of Events describes all data chunks sampled for each respective experimental variant across all subjects. For the 3 Emoji variant, this would be computed as $30\,Events\,per\,emoji\,x\,3Emoji\,Onscreen\,x\,5\,Subjects = 450\,Events$. The Total number of Test Events constitutes 10 % of the Total Number of Events for evaluation purposes. Additionally, information is provided on the number of Events per Subject, this denotes all emoji data chunks captured per subject and the Test Events per Subject constitutes 10% of the former, isolated as a test data subset for each individual that partook in the experiment. Further, the Events per Subject per Emoji and the Test Events per Subject per Emoji are also presented to clarify that irrespective of the experiment design the number of events per emoji is fixed. The only difference between the 3 experimental variants is the amount of emojis included. Again, any given emoji in any of the variants was augmented the same number of times and produced the same number of test events, it is only the number of emojis in each variant that differentiates the conditions.

4.4.1 Post-Processing Data Info: Pipeline 1

Each respective experimental variant features differing amounts of data due to the specific number of targets onscreen (either 3, 5, or 7 Emoji). The systems for reducing impedance

values across the entire scalp were improved with further consultations from the manufacturers. These included methods on seating sensors more effectively to the scalp in addition to instructing the subjects to count out the number of times each respective target emoji was augmented as a means of focusing subject concentration. Following the recommendations, all events sampled consisted of at least 2 electrodes following the channel-rejection protocols outlined above (see subsection 4.3.5.4).

4.4.2 Main: No Localizer Pre-Training: Pipeline 1

This subsection covers all the results generated via the LDA-based classification of data at the pooled-subject and single-subject level for the staggered array density variants outlined. These data are populated exclusively from the main experimental trials and do not utilize any of the events captured during the localizer task. Unless explicitly stated, the analysis subsections follow a standardized structure, this involves initially discussing the results generated using the pooled-subject aggregated data in terms of classifier performance metrics, class-wise accuracies and finally an appraisal of the grand average signal plots. This same format is then replicated for the results produced using single-subject data.

4.4.2.1 3 Emoji Variant: Pipeline 1

These results relate to the evaluations undertaken for the 3 Emoji stimulus variant described in Experiment 2 (see subsection 4.3.5). The investigations here involve the use of a reduced 3-target speller array (see, Figure 4.4) to probe the influence of array density on the resulting P300 waveform characteristics and associated LDA classifiers at the cross and single-subject levels. As noted above no localizer data was included in these evaluations.

Pooled-Subject

These results correspond to the aggregated, pooled-subject data partitions for the 3 Emoji experimental variant (see Table 4.2). A total of 405 and 45 training and test events were used in the evaluation of these data, respectively (see, Table 4.1).

	Mean Acc (%)	P300 Acc (%)	Non-P300 Acc (%)	Solver	Shrinkage
Pooled Subjects	73.33	46.67	86.67	lsqr	0.69

Table 4.2: This classification table displays the metrics relevant to the pooled-subject data for 3 Emoji variant computed without the inclusion of a localizer pre-training stage. The Pooled Subjects label refers to a data partition consisting of data aggregated across all subjects. Here the training data comprises the first 90% of all subject trials, with the final 10% of ecah subject dataset contributing an aggregated test set. In regards to the metrics reported, the 'Mean Acc (%)' column contains the overall accuracy obtained in the analysis of each respective data partition recorded across both classes. The 'P300 Acc (%)' and 'Non-P300 Acc (%)' columns record the classification performance for each respective data partition at the class level. Both the 'Solver' and 'Shrinkage' metrics relate to the grid-search optimized parameters computed to maximise classification performance. Note, that regions in which '-' is present are labelled as such to indicate that these values could or should not be computed. This could be due to the redundancy of said calculations, or due to the type of data required for specific computational operations.

At the pooled-subject level, the grid search method computed the optimal solver as the lsqr method and shrinkage factor at 0.69. The data evaluated demonstrated a mean accuracy (73.33%), substantially above the 50% random performance threshold. Further, this pattern was also shown at the class level for Non-P300 classification accuracy (86.67%). It must be noted that there is present a substantial bias and sub-random performance for the classification of P300 events (46.67%).

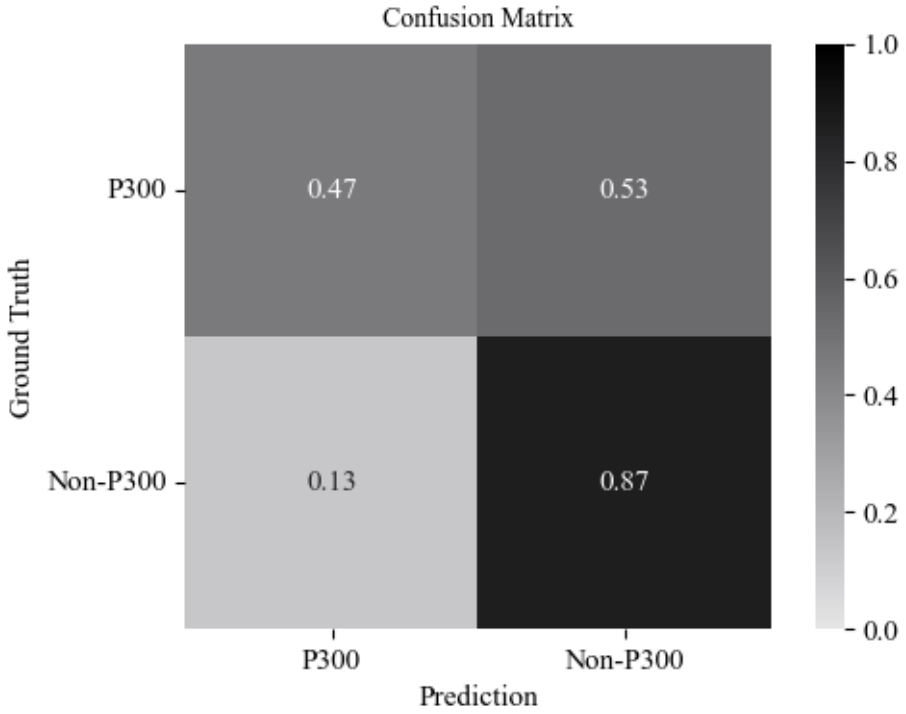

Figure 4.5: Here is displayed a normalized confusion matrix showing the classification performance of a trained LDA model for both P300 and Non-P300 classes relating to the pooled-subject 3 Emoji dataset (refer to, Table 4.1). Note, that these evaluations were computed using an aggregate dataset comprising all subjects sampled.

As can be seen in the above figure (see, Figure 4.5), the classification of Non-P300 events in the test set was performed at a relatively high level (86.67%) and the model demonstrated minimal confusion in the prediction of this class. P300 events were not classified with the same level of accuracy, with over half of all P300 events being misclassified as Non-P300 events.

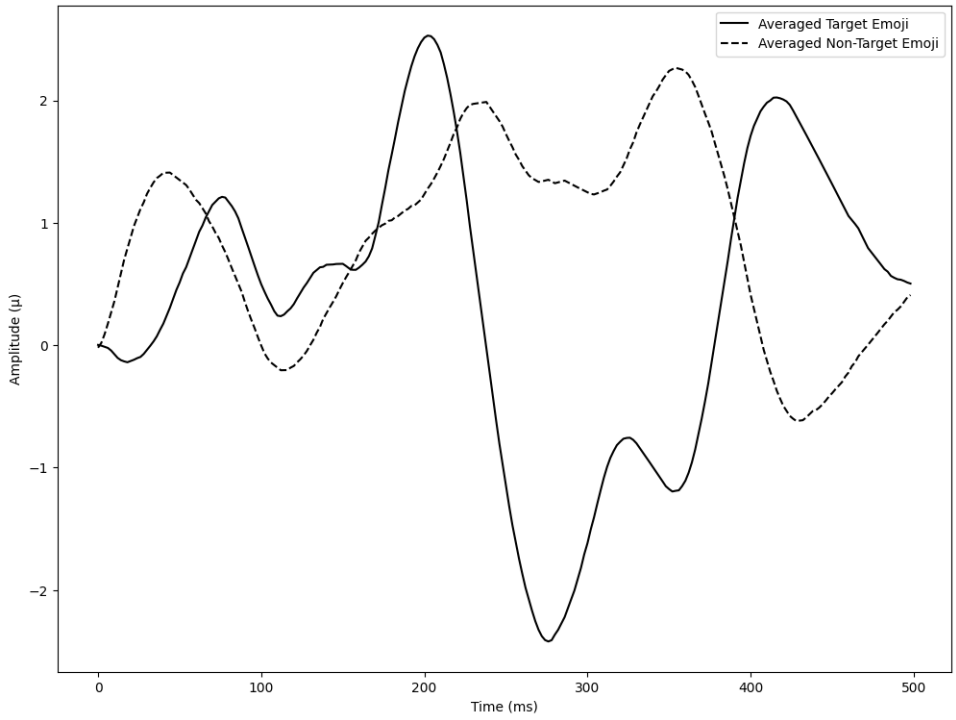

Figure 4.6: This figure presents a Cz grand average plot showing cross-trial P300 (solid line) and Non-P300 (dashed line) event signals for the 3 Emoji data partition (refer to Table 4.1). The x-axis represents time in milliseconds for each 500ms event data chunk, while the y-axis shows amplitude in μV of the EEG signal. The averages across these classes highlight underlying EEG waveform patterns embedded in the signals. It is important to note that the Cz channel was exclusively used for constructing these plots. Additionally, all signalswere baselined by averaging the first 50ms of collected samples. This baselining was done solely for presentation purposes and was not applied during the Pipeline 1 data pre-processing as outlined in subsection 3.3.5.3 (see Table 3.1).

The plot above (see Figure 4.6) reveals a marked difference in the averaged waveforms for the P300 (solid line) and Non-P300 (dashed line) classes. The P300 waveform displays greater variance in μV amplitude, characterized by a pronounced negative deflection around 200-300ms followed by a strong positive deflection at 350-400ms. In contrast, the Non-P300 waveform exhibits a narrower range of μV values, featuring a mild negative component around 100ms and minor positive deflections at 200ms and 350ms. Consistent with findings from Experiment 1 (see Figures 3.9, 3.13, & 3.20), the P300 samples show significant negative drift across the waveform, complicating comparisons with the Non-P300 signal averages. Additionally, the observed pattern, where the cued Target signal presents a lower P300 peak than the Non-Target signal, suggests that these waveform components were not effectively used in class separation, likely contributing to the observed poor results.

Within-Subject These results were generated using single-subject data for the 3 Emoji experimental variant. A total of 9 test events were used in the evaluation of these data. As can be seen from the above classification table (see, Table 4.3). All subjects excluding Subjects 1 and 5 demonstrate chance level results. The only subject that presented with a sub-RPT P300 class accuracy was Subject 2. The highest-performing subject sampled (Subject 5) reported a mean classification performance of 88.89% (P300 Acc= 66.67%, Non-P300 Acc=100%). During the grid search method, the lsqr solver method was shown to be the most optimal in maximizing classification performance.

	Mean Acc (%)	P300 Acc (%)	Non-P300 Acc (%)	Solver	Shrinkage
Subject 1	77.78	66.67	83.33	lsqr	0.52
Subject 2	55.56	33.33	66.67	lsqr	0.15
Subject 3	66.67	66.67	66.67	lsqr	0.06
Subject 4	55.56	66.67	50.00	lsqr	0.93
Subject 5	88.89	66.67	100.00	lsqr	0.85
Sub Avg	68.89	60.00	73.33	n/a	0.50
Sub Var	16.67	16.67	25.00	n/a	0.44

Table 4.3: This classification table holds all metrics relevant to the 3 Emoji dataset computed without the inclusion of a localizer pre-training stage. The individual subject monikers denote the performance relating to a single subject. The 'Sub Var' moniker denotes the range of single-subject metrics used to compute the 'Sub Avg' results. Note, that these do not represent the results of LDA models trained on aggregated data, these are the averages of the single-subject analyses performed here. For further information on the metric and parameter field headings please refer to Table 4.2.

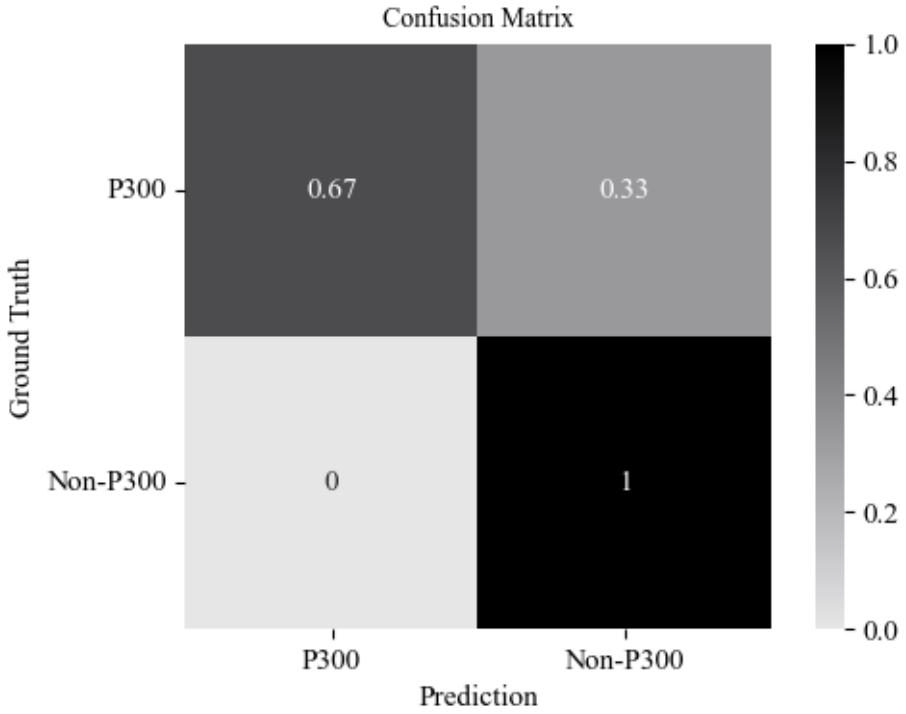

Figure 4.7: Here is displayed a normalized confusion matrix reporting the classification performance of a trained LDA model for both P300 and Non-P300 classes relating to Subject 5 in the 3 Emoji No Localizer dataset (refer to, Table 4.1).

The confusion matrix shown above (see, Figure 4.7) illustrates maximal levels of accuracy in the classification of Non-P300 events for Subject 5. Concerning P300 event classification, accuracies were substantially lower (66.67%). Note, that despite some evidence that the LDA classifier demonstrated a Non-P300 preferential bias, P300 accuracies were maintained significantly above the random performance threshold.

4.4.2.2 5 Emoji Variant: Pipeline 1

The investigations reported herein involve the use of a reduced 5-target speller array (see, Figure 4.3) to probe the influence of array density on the resulting P300 waveform characteristics and associated LDA classifiers at the cross and single-subject levels. As stated earlier, no localizer data was included in these evaluations, also, all pre-processing and analysis were conducted using the Pipeline 1 approach, see subsection 3.3.5.1.

Pooled-Subject The results described here correspond to the evaluations undertaken for the 5 Emoji variant described in Experiment 2 (see subsection 4.3.5). The investigations here

involve the use of a reduced 5-target speller array (see, Figure 4.3) to probe the influence of array density on the resulting P300 waveform characteristics and associated LDA classifiers at the cross and single-subject levels. As stated earlier, no localizer data was included in these evaluations. These results correspond to the aggregated, pooled-subject data partitions for the 5 Emoji experimental variant. Further they were pre-processed and analyzed using the Pipeline 1 method (see subsection 3.3.5.1.. A total of 75 test events were used in the evaluation of these data.

	Mean Acc (%)	P300 Acc (%)	Non-P300 Acc (%)	Solver	Shrinkage
Pooled Subjects	78.67	0.00	98.33	lsqr	0.12

Table 4.4: This classification table displays all metrics relevant to the 5 Emoji dataset computed using the pooled-subject data without the inclusion of localizer data (refer to, Table 4.1). For further information on field headings refer to, Tables 4.2 & 4.3.

The pooled-subject classification performance (as seen in, Table 4.4) for the 5 Emoji dataset is shown to achieve a mean accuracy of 78.67%, significantly above the random performance threshold of 50%. The performance at the class level reveals substantial signs of overfitting, with P300 class accuracies of 0% and Non-P300 accuracies of 98.33%.

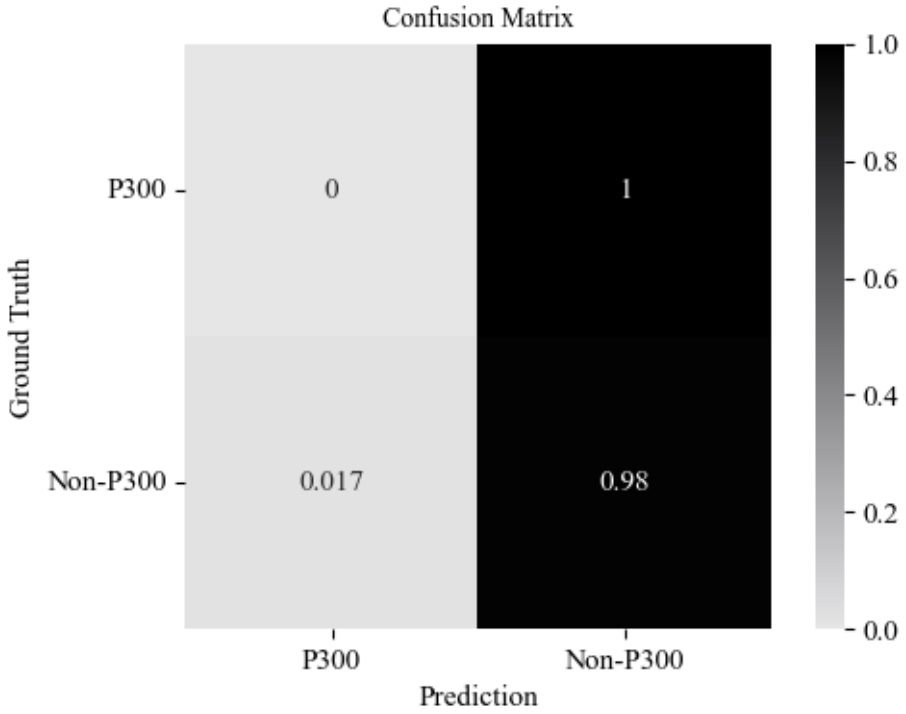

Figure 4.8: Here is displayed a normalized confusion matrix reporting the classification performance of a trained LDA model for both P300 and Non-P300 classes relating to the pooled-subject 5 Emoji No Localizer dataset (refer to, Table 4.1).

In the above confusion matrix (see, Figure 4.8) it can be observed that the Non-P300 event class was selected nearly exclusively by the trained LDA model. Substantial misclassification of the P300 waveform is evident, demonstrating no accurate predictions for this class. The only P300 predictions made were done erroneously, with some Non-P300 waveforms being misclassified as belonging to the P300 target class.

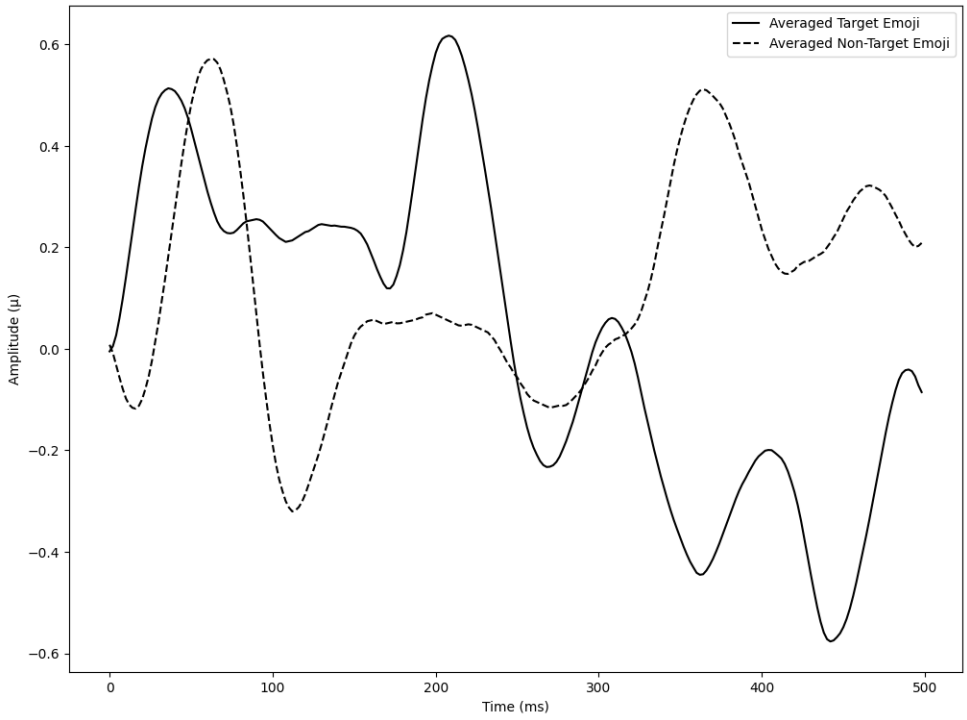

Figure 4.9: This figure displays a Cz grand average plot for pooled-subject data relating to the 5 Emoji No Localizer data partition (refer to, Table 4.1). The solid line shows an average signal for the P300 events and the dashed line is an average for the Non-P300 events. For further information on interpreting this figure please refer to Figure 4.6.

In reference to Figure 4.9, the average P300 signal (solid line) deviates markedly from the expected P300 waveform, primarily due to a pronounced negative drift observed throughout the 500 ms period. Conversely, the Non-P300 signal (dashed line) displays features more aligned with a typical oddball visual ERP, including a micro-voltage dip around 100 ms and a peak between 350-400 ms. The absence of appropriate baselining has impaired the accurate comparison between these two signal classes. Additionally, the appearance of a 6-8Hz signal within this plot is likely owing to the periodicity of the stimulus onset interval of 150ms (6.67Hz). For further information on how this is addressed in the Pipeline 2 method please refer to subsection 3.3.5.3 and 4.3.7.

Within-Subject These results were generated using single-subject data for the 5 Emoji experimental variant. A total of 15 test events were used in the evaluation of these data. As seen in Table 4.5, mean classification performance at the signal-subject level is well below the random-performance thresholds for all subjects sampled. The highest performing subjects (Subjects 1 & 3) demonstrate impressive classification accuracy for the Non-P300 class (91.67%) alongside sub-random performance at the P300 class level (<50%). Further, the

signs of overfit persist to a greater extent when examining the results from the remaining subjects sampled. Evidence of complete LDA model overfit is shown for Subjects 4 and 5, with 0% accuracies recorded for the P300 class. Note, that in all instances, the grid search optimization method assisted in identifying the lsqr solver method as the only viable technique for evaluation purposes.

	Mean Acc (%)	P300 Acc (%)	Non-P300 Acc (%)	Solver	Shrinkage
Subject 1	80.00	33.33	91.67	lsqr	0.18
Subject 2	73.33	33.33	83.33	lsqr	0.17
Subject 3	80.00	33.33	91.67	lsqr	0.30
Subject 4	60.00	0.00	75.00	lsqr	0.19
Subject 5	66.67	0.00	83.33	lsqr	0.06
Sub Avg	72.00	20.00	85.00	n/a	0.18
Sub Var	10.00	16.67	8.34	n/a	0.12

Table 4.5: This classification table displays all metrics relevant to the 5 Emoji dataset computed using single-subject data without the inclusion of localizer data (refer to, Table 4.1). For further information on field headings refer to, Tables 4.2 & 4.3.

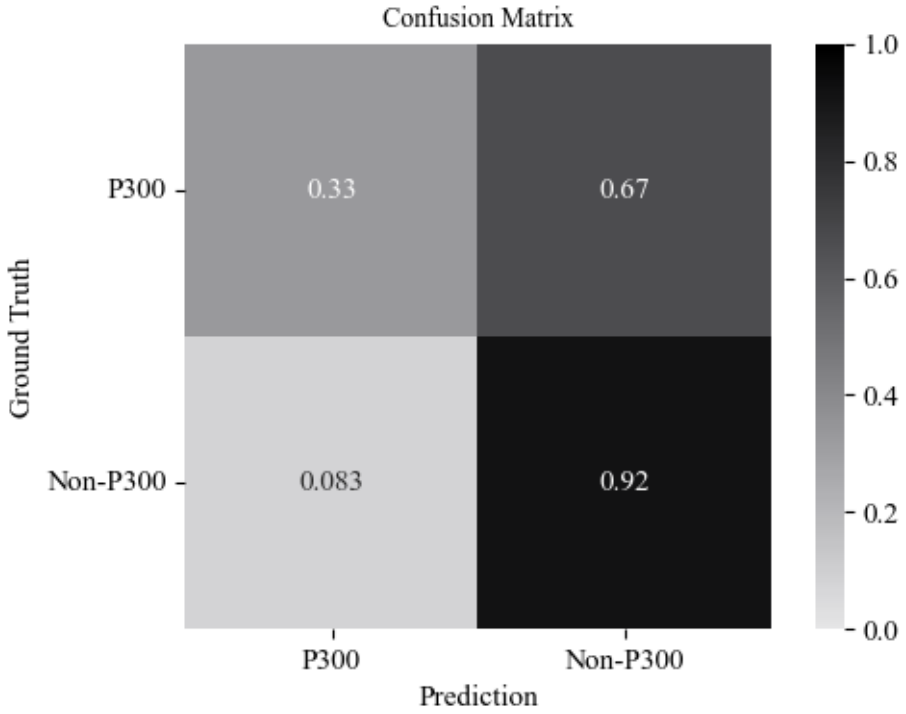

Figure 4.10: Here is displayed a normalized confusion matrix reporting the classification performance of a trained LDA model for both P300 and Non-P300 classes relating to Subject 1 in the 5 Emoji No Localizer dataset (refer to, Table 4.1).

In the above confusion matrix (see, Figure 4.10) it can be observed that significant confusion is demonstrated by the LDA model to accurately distinguish between the target P300 and Non-P300 classes. The direction of misclassification is primarily in relation to the Non-P300 samples being incorrectly predicted as P300 events, with very few instances being observed in the opposite direction.

4.4.2.3 7 Emoji Variant: Pipeline 1

The results detailed here relate to the evaluations undertaken for the 7 Emoji variant described in Experiment 2 computed via the Pipeline 1 method (see subsections 4.3.5.4 & 4.3.6). This involved the use of a 7-target speller array (see, Figure 4.2) to probe the influence of array density on the resulting P300 waveform characteristics and associated LDA classifiers at the cross and single-subject levels. As previously established, no localizer data was included in these evaluations

Pooled-Subject These results correspond to the aggregated, pooled-subject data partitions for the 7 Emoji No Localizer experimental variant (see, Table 4.6). A total of 105 test events were used in the evaluation of these data.

	Mean Acc (%)	P300 Acc (%)	Non-P300 Acc (%)	Solver	Shrinkage
Pooled Subjects	85.71	0.00	100.00	lsqr	0.09

Table 4.6: This classification table shows all relevant metrics for the 7 Emoji dataset computed via the Pipeline 1 method using the pooled-subjects data without the inclusion of localizer data (refer to, Tables 4.1 for data partition info). For further details on field headings refer to, Tables 4.3 & 4.3.

At the pooled-subject level, the LDA classifier trained using aggregate data from all 5 subjects sampled demonstrates significant signs of overfitting. The selective bias of the Non-P300 was complete, achieving 100% accuracy and a 0% classification accuracy for corresponding P300 event samples.

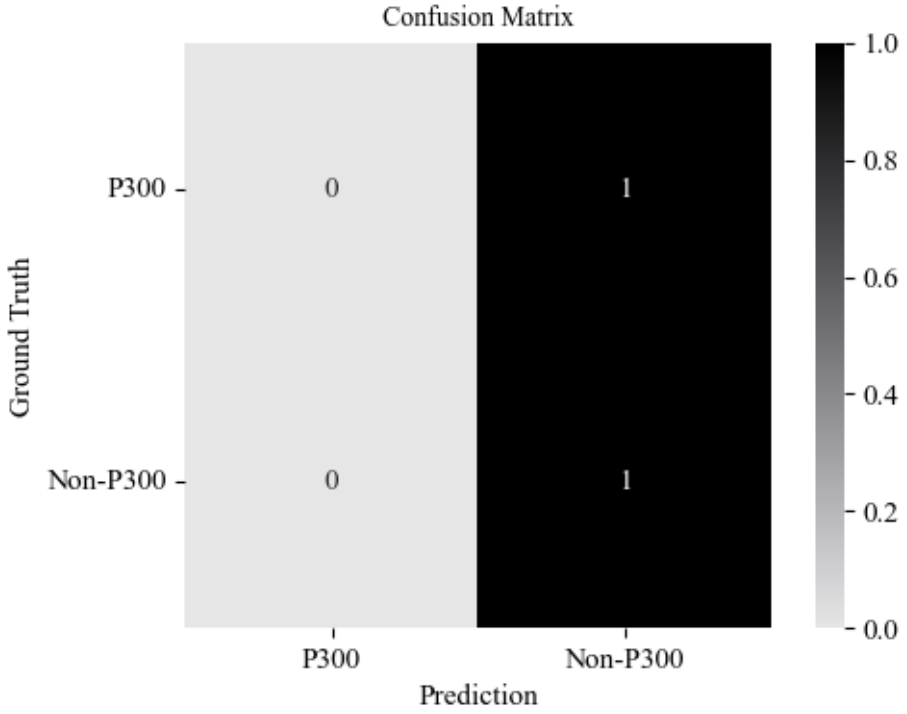

Figure 4.11: Displayed here is a normalized confusion matrix detailing the classification performance of a trained LDA model for both P300 and Non-P300 classes relating to the pooled-subject 7 Emoji No Localizer dataset (refer to, Table 4.6).

In the above confusion matrix (see, Figure 4.11) further evidence of significant overfitting can be observed. The models demonstrate a comprehensive selective bias towards the Non-

P300 class. At no point in the data evaluation was a single event classified as belonging to the P300 class, this includes 'misses' involving the erroneous classification of a Non-P300 event as a P300 event.

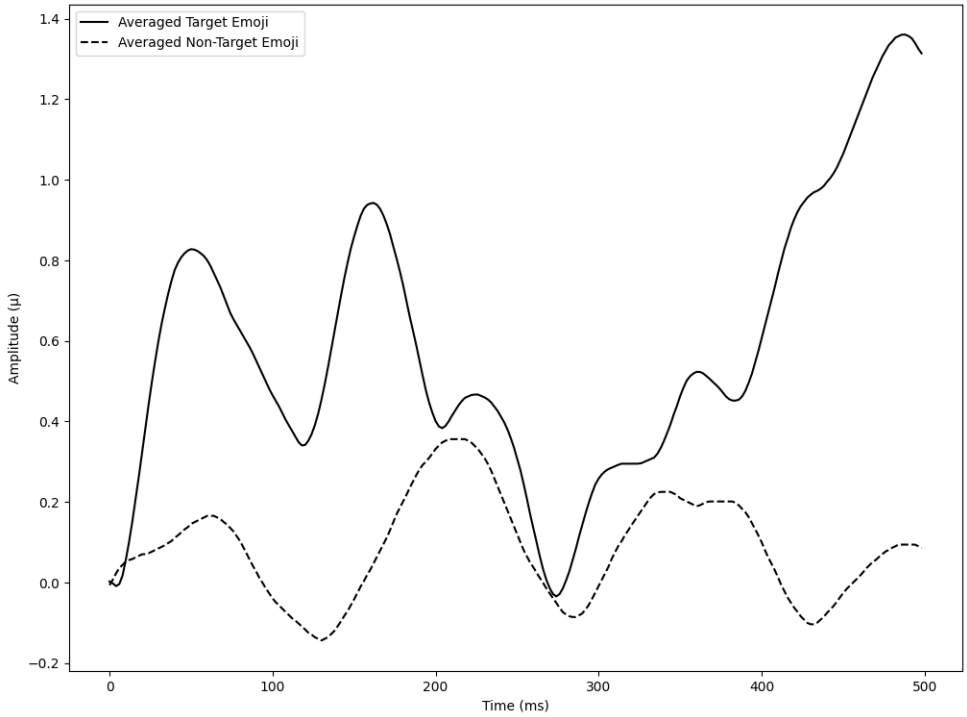

Figure 4.12: This figure presents a Cz grand average plot for pooled-subject data from the 7 Emoji No Localizer data partition (see Table 4.1). The solid line represents the average signal for P300 events, while the dashed line illustrates the average for Non-P300 events. For additional details, please refer to Figure 4.6.

As illustrated in the above plot (see Figure 4.12), the P300 waveform exhibits some of the characteristic features expected from an oddball visual ERP in this experimental design. Notably, both the N200 and P300 components appear significantly delayed, emerging around 250ms and 450ms, respectively. However, the signal also shows considerable positive drift, likely due to the lack of a robust baselining procedure.

Within-Subject These results were generated using single-subject data for the 7 Emoji No Localizer experimental variant. A total of 21 test events were used in the evaluation of these data. In all subjects evaluated the phenomenon of overfitting is present and broadly reflects the same pattern of classification behaviour identified from the previous pooled-subject analyses (see, Tables 4.2-4.6). In other words, every subject tested produced a P300 classification accuracy of 0%. As can be seen, there is some variance in the degree of accurate Non-P300 event classification, ranging from 100% in Subject 5 and 88.89% in Subject 4. It appears

that despite significant efforts to optimize model parameters via grid search techniques little variation in performance is observed, even with the substantial differences in shrinkage values.

	Mean Acc (%)	P300 Acc (%)	Non-P300 Acc (%)	Solver	Shrinkage
Subject 1	80.95	0.00	94.44	lsqr	0.20
Subject 2	80.95	0.00	94.44	lsqr	0.14
Subject 3	80.95	0.00	94.44	lsqr	0.18
Subject 4	76.19	0.00	88.89	lsqr	0.14
Subject 5	85.71	0.00	100.00	lsqr	0.32
Sub Avg	80.95	0.00	94.44	n/a	0.20
Sub Var	4.76	0.00	5.56	n/a	0.09

Table 4.7: This classification table shows all relevant metrics for the 7 Emoji dataset computed via the Pipeline 1 method using the single-subjects data without the inclusion of localizer data (refer to, Tables 4.1 for data partition info). For further details on field headings refer to, Tables 4.3 & 4.3..

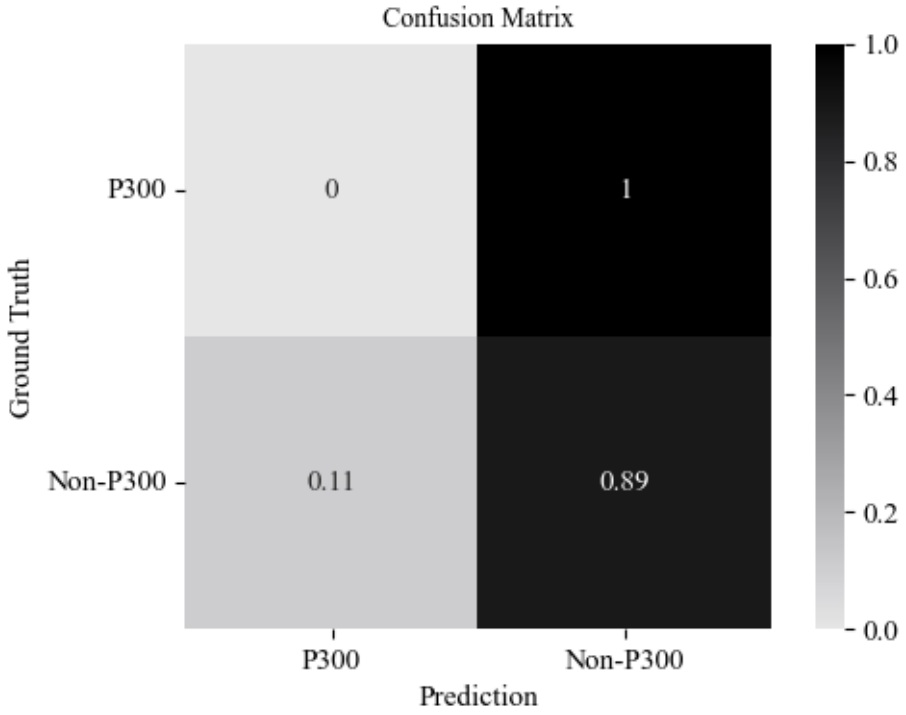

Figure 4.13: Displayed here is a normalized confusion matrix showing the classification performance of a trained LDA model for both P300 and Non-P300 classes relating to Subject 4 in the 7 Emoji No Localizer dataset (refer to, Table 4.1).

The confusion matrix positioned above (see, Figure 4.13) provides further insight into the

prediction behaviour of the trained LDA model evaluated for Subject 4 in the 7 Emoji No Localizer dataset. The plot indicates substantial selective bias via the LDA model for the Non-P300 class. There are no instances in which a P300 event was classified accurately. The only instance of P300 prediction was the misclassification of a Non-P300 event.

4.4.3 Main + Localizer Pre-Training Experiment: Pipeline 1

It must be noted that the LDA models discussed herein are initially trained using the corresponding (20 trial), class-balanced localizer dataset, used principally for visual signal appraisal before the onset of the main experiment (see subsection 4.3.4). The class-wise signal average plots discussed in the previous section will not be included in the analyses as these are fundamentally the same as those generated in the prior results sections. Note, that the training procedure, Pipeline 1, of the previous section is replicated. This involves partitioning the main experiment data into a 9:1 train and test dataset split for training and evaluation via a grid-search optimized LDA model.

As previously discussed (see subsection 4.3.4), the high probability of augmentation for the single emoji on screen (50%) raises concerns about whether the localizer task can reliably produce a distinguishable P300 waveform. To examine this, a Cz grand average plot was generated across all subjects, comparing the Target and Non-Target data segments collected during the localizer task. This task, consisting of 20 trials, was conducted before the onset of the three emoji design variants. Each subject contributed 60 trials, yielding 300 Target (P300) and Non-Target (Non-P300) samples per subject for the averaging process. The integration of these signals into a pre-training stage was done purely for exploratory purposes given the low number of samples available to the LDA models and the high degree of class imbalance between Target oddball and Non-Target standard trials in the main experimental data.

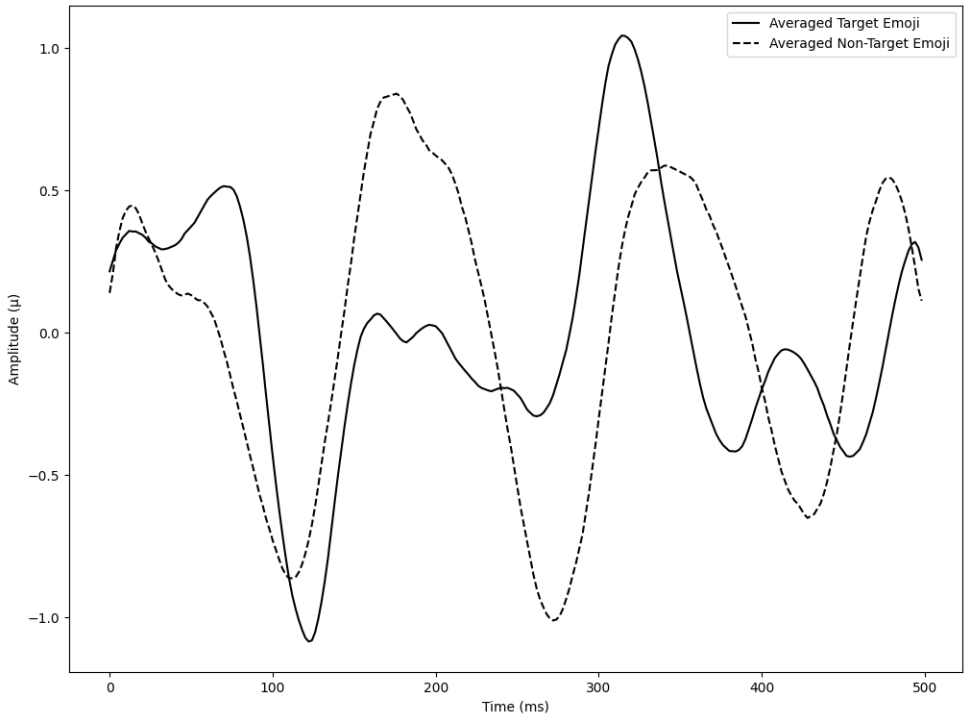

Figure 4.14: This figure presents a Cz grand average plot showing cross-trial P300 (solid line) and Non-P300 (dashed line) event signals for the Localizer data used (refer to Table 4.1). Note, that this includes samples collected prior to the onset of all stimulus variants covering the 3, 5 and 7-Emoji presenation methods.The x-axis represents time in milliseconds for each 500ms event data chunk, while the y-axis shows amplitude in μV of the EEG signal. The averages across these classes highlight underlying EEG waveform patterns embedded in the signals. It is important to note that the Cz channel was exclusively used for constructing these plots. Additionally, all signalswere baselined by averaging the first 50ms of collected samples. This baselining was done solely for presentation purposes and was not applied during the Pipeline 1 data pre-processing as outlined in subsection 3.3.5.3 (see Table 3.1).

As illustrated in the plot below (see Figure 4.14), there is minimal variance between the Target and Non-Target Cz grand-average signals. This lack of distinction suggests that using these data as pre-training samples for the corresponding LDA classifiers may not be justified. The author's hypothesis that the perceived probability of the localizer task might be lower than the actual probability of emoji augmentation is not supported by these results. Moreover, the pervasive and unexplained presence of a strong 5-6 Hz frequency in both samples further diminishes the discriminability between the two classes.

4.4.3.1 3 Emoji Variant: Pipeline 1

The results described here correspond to the evaluations undertaken for the 3 Emoji variant described in Experiment 2 (see subsection 4.3.5). The stimulus comprises a 3-target emoji array ranging in valance from left to right (see, Figure 4.4). All respective LDA models utilized the

associated localizer pre-screening task data as a pre-training dataset to tune the corresponding classifiers. Note, all data organisation, pre-processing and analysis were conducted using the Pipeline 1 approach (see subsection 3.3.5.1 & Table 4.1).

Pooled-Subject These results correspond to the aggregated, pooled-subject data partition for the 3 Emoji Main + Localizer Pre-Training experimental variant.

	Mean Acc (%)	P300 Acc (%)	Non-P300 Acc (%)	Solver	Shrinkage
Pooled Subjects	51.11	53.33	50.00	lsqr	0.00

Table 4.8: Here is shown a classification table reporting relevant evaluation metrics for the localizer-data-initialized LDA model copmuted using the pooled-subject data for the 3 Emoji dataset (refer to, Table 4.1 for data partition info). For further details on field headings refer to Tables 4.2 & 4.3.

In the classification table (see, Table 4.8) the pooled-subject data was evaluated to report a mean accuracy of 51.11%, orienting the performance marginally above the random threshold of 50%. This is largely mirrored in the results of the class-wise analyses.

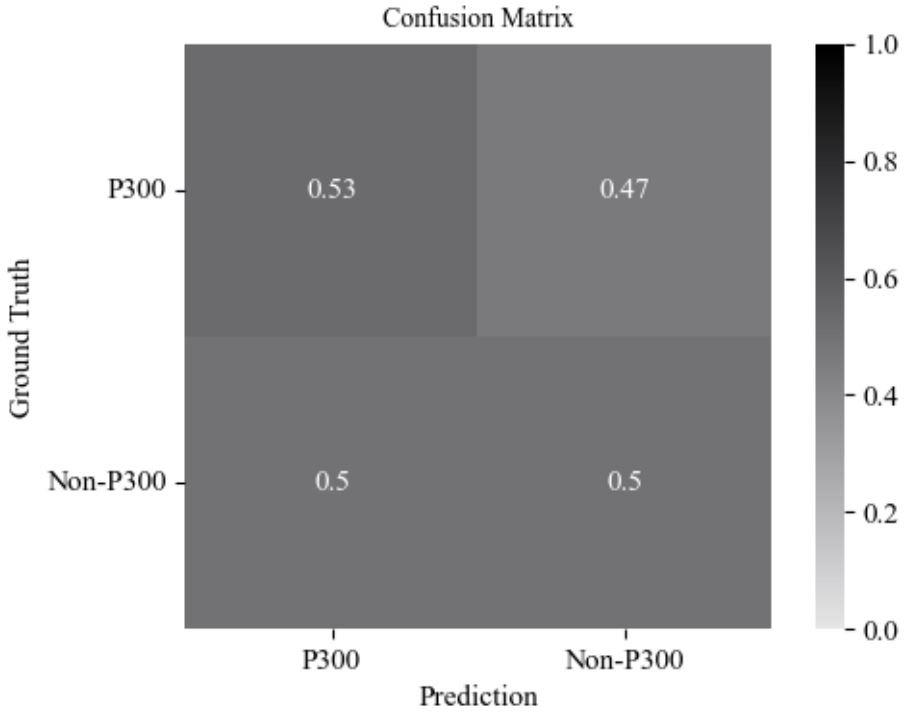

Figure 4.15: Here is shown a normalized confusion matrix showing the classification performance of a trained LDA model for both P300 and Non-P300 classes relating to the aggregated pooled-subject data for the 3 Emoji With Localizer dataset (see subsection 4.3.6.1).

As can be seen in the confusion matrix positioned above, substantial confusion is demonstrated for all aspects of the LDA model prediction behaviour. A near-perfect distribution of predictions is present for all combinations of classification and misclassification, with only a slight proclivity present for the accurate prediction of P300 events.

Within-Subject These results were generated using single-subject data for the 3 Emoji with Localizer experimental variant (see, Table 4.9). Broadly, at the single-subject level, the average mean accuracy falls marginally below the 50% random performance threshold (44.44%). The overall performance at the single-subject level is generally far lower than that achieved using the standard method of implementation outlined in the previous section (see, Table 4.3). Overfitting was observed in one subject (Subject 3) with marginal and poor performance throughout the remaining subjects (Subjects 2, 3 & 5).

	Mean Acc (%)	P300 Acc (%)	Non-P300 Acc (%)	Solver	Shrinkage
Subject 1	66.67	100.00	50.00	lsqr	0.94
Subject 2	22.22	33.33	16.67	lsqr	0.63
Subject 3	22.22	66.67	0.00	lsqr	0.00
Subject 4	88.89	100.00	83.33	lsqr	0.00
Subject 5	22.22	33.33	16.67	lsqr	0.01
Sub Avg	44.44	66.67	33.33	n/a	0.32
Sub Var	33.34	33.34	41.67	n/a	0.47

Table 4.9: Here is shown a classification table reporting all relevant evaluation metrics across subjects for the localizer-data-initialized LDA models trained on single-subject data from the 3 Emoji dataset (refer to, Table 4.1 for data partition info). For further details on field headings refer to, Table 4.2 & 4.3.

Of note, for the first time in the analyses thus far, the average single-subject classification performance for the P300 event class (66.67%) exceeds that observed in the Non-P300 event class (33.33%). Moreover, the classification performance of the P300 class is also above the random performance threshold. The highest-performing subject reported herein (Subject 4) achieved a mean accuracy of 88.89%. Crucially, 100% classification accuracy was reported for the P300 event class and a corresponding classification accuracy of 83.33% for Non-P300 event classes. Note, that the lsqr solver was identified as the optimal solver method for LDA model evaluations in all subject instances. It must be noted that given these results are not cross-validated (see subsection 3.3.5.1) and given the poor quality of associated P300 signals in the localization task (see Figure 4.14) this result is likely anomalous and not evidence of merit for the adoption of the localizer data pre-training approach.

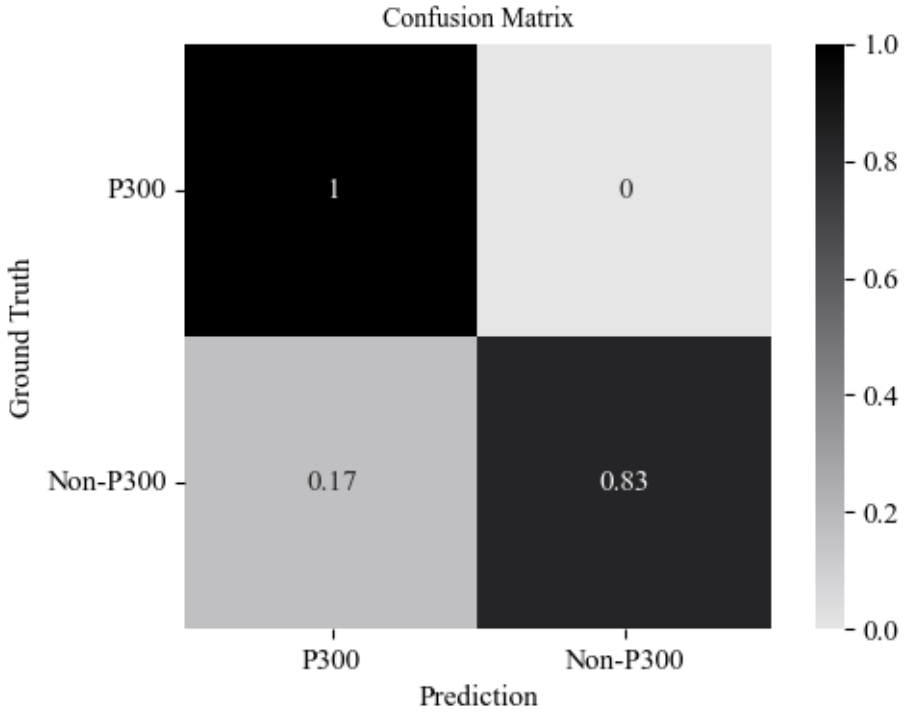

Figure 4.16: The figure displays a normalized confusion matrix showing the classification performance of a localizer-initialized and trained LDA model for both P300 and Non-P300 classes relating to Subject 4 in the 3 Emoji with Localizer dataset (see subsection 4.3.6.1).

The confusion matrix above (see, Figure 4.16) details the class-level prediction behaviour of the LDA model. Minimal misclassifications are reported and the pattern of incorrect predictions exclusively involved the erroneous selection of Non-P300 events as P300 events, noting a distinct change from previously discussed analyses.

4.4.3.2 5 Emoji Variant: Pipeline 1

The results detailed herein relate to the evaluations undertaken for the 5 Emoji with Localizer variant described in Experiment 2 (see subsection 4.3.5.4). The stimulus consists of a 5-target emoji array ranging in valance from left to right (see, Figure 4.3). For each respective LDA model, the associated localizer pre-screening task data was utilized as a pre-training dataset to tune the corresponding classifiers.

Pooled-Subject These results correspond to the aggregated, pooled-subject data partition for the 5 Emoji Main with Localizer experimental variant. The classification table in question (see, Table 4.10), shows a substantial improvement in performance at the pooled-subject

level as compared with the largely overfit results reported for the standard method previously implemented (see, Table 4.5). In greater detail, the mean accuracy reported (42.67%) is significantly lower than in the standard implementation (78.67%). The difference in quality is exemplified by the dramatic increase in performance at the P300-class level increasing from 0% in the standard method to 60% for the current localizer-initialization method. Despite this, no meaningful difference in the classification performance of the system has been made given the sub-random mean accuracy noted. This is likely due to the relative absence of key P300 waveform components from the localizer samples, as illustrated in Figure 4.14.

	Mean Acc (%	P300 Acc (%)	Non-P300 Acc (%)	Solver	Shrinkage
Pooled Subjects	42.67	60.00	38.33	lsqr	0.68

Table 4.10: Above is displayed a classification table denoting all associated evaluation metrics for the localizer-data-initialized LDA model conducted using the pooled-subjects samples from the 5 Emoji with Localizer dataset (see subsection 4.3.4). Note, all data organisation, pre-processing and analysis were conducted using the Pipeline 1 approach, see subsection 3.3.5.1. For more information concerning field headings refer to, Table 4.1.

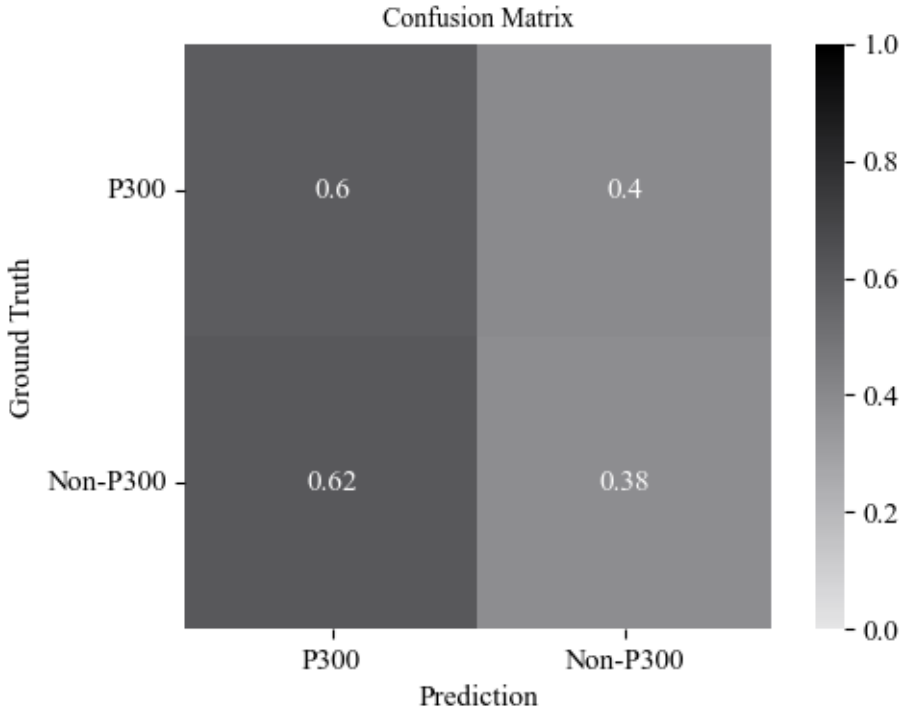

Figure 4.17: The figure displays a normalized confusion matrix showing the classification performance of a localizer-initialized and trained LDA model for both P300 and Non-P300 classes relating to aggregated pooled-subject 5 Emoji with Localizer dataset.

The above confusion matrix demonstrates that the majority of misclassifications recorded via the trained LDA model involve erroneously selecting Non-P300 events as P300 events. It must be noted that a substantial proportion of misclassifications also operate in the opposite direction, with many P300 samples being incorrectly identified as belonging to the Non-P300 class. In other words, the confusion present in the LDA classifier is bi-directional, with a marginal selective bias for the P300 event class. This differs dramatically from the confusion matrix computed for the standard analyses method as seen in Figure 4.8.

Within-Subject These results were generated using single-subject data for the 5 Emoji with Localizer experimental variant (see, Table 4.11). One subject (Subject 2) displays a negligible difference in performance, with signs of overfitting to the Non-P300 class (see, Table 4.5) reversing for the P300 class (see, Table 4.11) in both Subjects 1 and 2. Generally, across the other subjects sampled overfitting is still highly prevalent and is also accompanied by a substantial reduction in classification accuracy for the previously biased class. This is seen in Subjects 3, 4 and 5, where the previous application of the standard analysis method led to Non-P300 accuracies of $>$75% (see, Table 4.5) and dropped to below 66.67% (see, Table 4.11) for the localizer data-initialized method.

	Mean Acc (%	P300 Acc (%)	Non-P300 Acc (%)	Solver	Shrinkage
Subject 1	13.33	66.67	0.00	lsqr	0.95
Subject 2	46.67	100.00	33.33	lsqr	0.56
Subject 3	53.33	0.00	66.67	lsqr	0.00
Subject 4	40.00	0.00	50.00	lsqr	0.86
Subject 5	40.00	0.00	50.00	lsqr	0.01
Sub Avg	38.67	33.33	40.00	n/a	0.48
Sub Var	20.00	50.00	33.34	n/a	0.48

Table 4.11: Above is displayed a classification table denoting all associated evaluation metrics for the localizer-data-initialized LDA model conducted using the single-subject samples from the 5 Emoji with Localizer dataset (see subsection 4.3.4). Note, all data organisation, pre-processing and analysis were conducted using the Pipeline 1 apporach, see subsection 3.3.5.1. For more information concerning field headings refer to, Table 4.1.

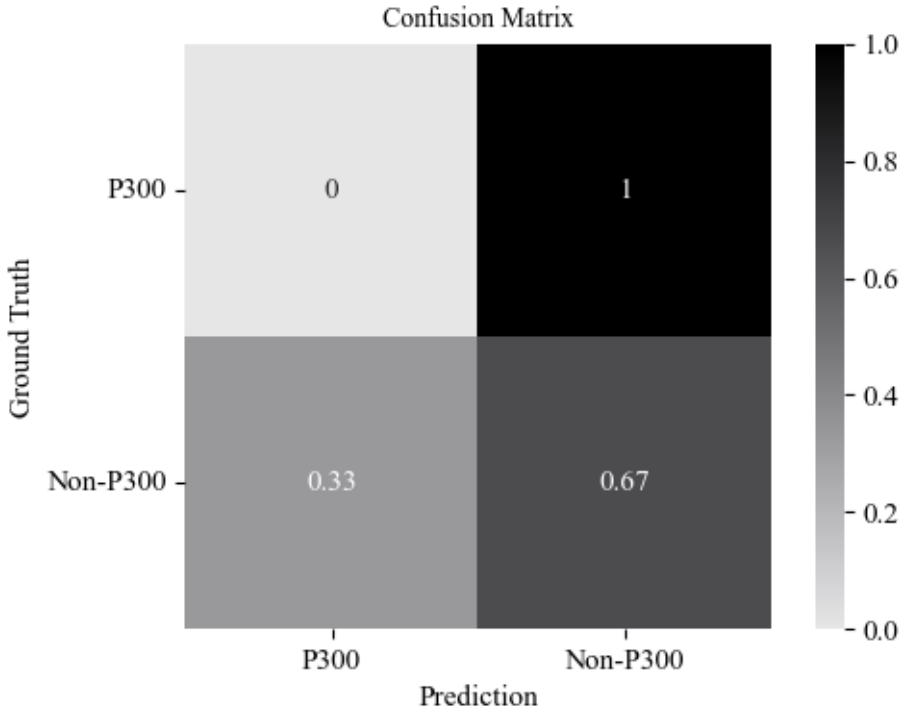

Figure 4.18: The figure herein shows a normalized confusion matrix displaying the classification performance of a localizer-initialized and trained LDA model for both P300 and Non-P300 classes relating to Subject 3 in the 5 Emoji with Localizer dataset.

As seen above (see, Figure 4.18), the confusion matrix provides a greater level of clarity in the prediction behaviours of the localizer-initialized LDA model for Subject 3. Non-P300 event accuracies are shown to exceed the random-performance threshold (>50%). Conversely, the P300 event class is not accurately selected for any instance evaluated in the test set. The primary direction of confusion expressed by the LDA model is characterized by a substantial bias for predicting P300 class events as Non-P300 events. Of note, some misclassifications of Non-P300 waveforms as P300 waveforms are also present. In sum, the poor classification results are likely owing to the low quality of associated P300 signals within the Localizer task data given the associated high oddball stimulus probability.

4.4.3.3 7 Emoji Variant: Pipeline 1

The results detailed herein relate to the evaluations undertaken for the 7 Emoji variant described in Experiment 2 (see subsection 4.3.5). The stimulus consists of a 7-target emoji array ranging in valance from left to right (see, Figure 4.2). For each respective LDA model, the associated localizer pre-screening task data was utilized as a pre-training dataset to tune the

corresponding classifiers.

Pooled-Subject The findings herein relate exclusively to the data acquired via the 7 Emoji Main with Localizer experimental variant. All LDA models discussed were initialized using the corresponding pre-experimental localizer data. As seen in the results table below (see, Table 4.12), pooled-subject performance marginally improved with the implementation of the LDA localizer-initialization method as compared to the standard method. The evaluation of Non-P300 waveforms saw a significant reduction in classification accuracy dropping from 100.00% (see, Table 4.6) to 52.22%, settling just above the random-performance threshold (50%). Conversely, P300 event classification accuracy increased from 0.00% (see, Table 4.6) to 73.33% (see, Table 4.12).

	Mean Acc (%)	P300 Acc (%)	Non-P300 Acc (%)	Solver	Shrinkage
Pooled Subjects	55.24	73.33	52.22	lsqr	0.01

Table 4.12: Here is displayed a classification table denoting all associated evaluation metrics for the localizer-data-initialized LDA models for the pooled-subject samples in the 7 Emoji with Localizer dataset (refer to, Table 4.1 for data partition info). For more information on field headings refer to, Tables 4.2 & 4.3.

The confusion matrix below (see, Figure 4.19), demonstrates improvements in classification performance for the 7 Emoji dataset at the pooled-subject level. When comparing the evaluations conducted here to those observed for the standard analysis method (see, Figure 4.13) a significant reduction in the prevalence of overfitting is seen. The principal direction of confusion expressed by the LDA model is the erroneous selection of Non-P300 waveforms as P300 waveforms, with a substantially reduced incidence of misclassification in the opposing direction. Given the poor quality of the associated localizer task Target and Non-Target Cz grand averages, it is difficult to assert that these accuracies are the result of any meaningful separation of the data based on waveform features associated with the visual oddball paradigm.

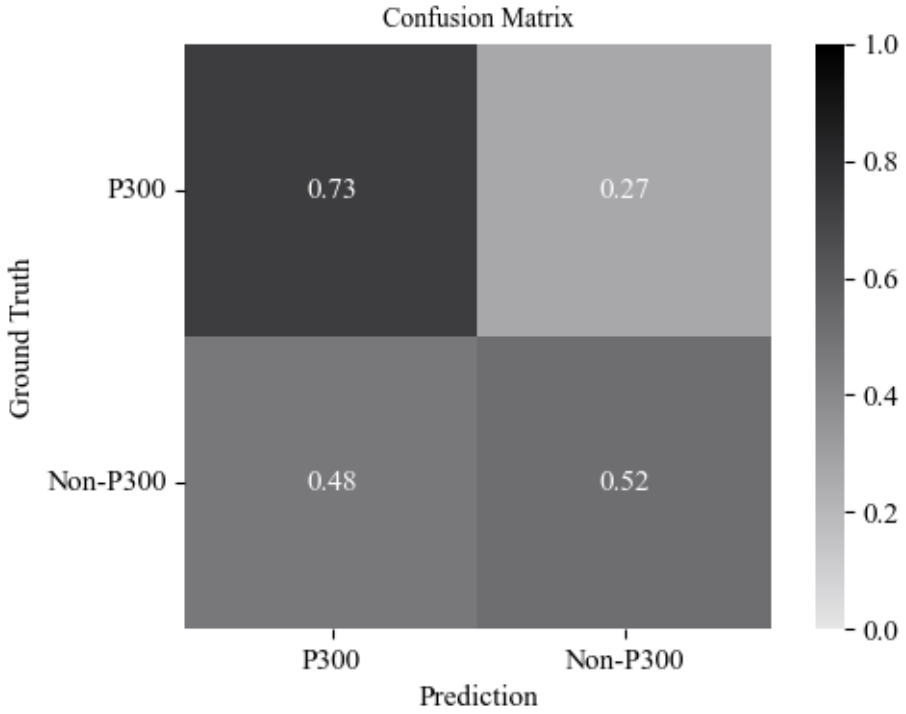

Figure 4.19: The figure herein shows a normalized confusion matrix displaying the classification performance of a localizer-initialized and trained LDA model for both P300 and Non-P300 classes relating to the aggregated pooled-subject 7 Emoji with Localizer dataset.

Within-Subject These results were generated using single-subject data for the 7 Emoji with Localizer experimental variant. At the single-subject level, only one subject (Subject 1) evaluated using the localizer-initialization method showed any signs of performance enhancement (see, Table 4.13). This is characterized by a reduction in the prevalence of selective bias towards the Non-P300 class, a significant increase in P300 class accuracy and the incidence of above random-performance for both class types sampled. Note, that unlike the results computed via the Pipeline 2 method (see subsection 3.3.5.3), these accuracies were not generated via a 10-fold cross-validation procedure, therefore the significance of this marginal result is highly questionable. For the other subjects sampled, the differences in performance metrics ranged from poor to chance level. Generally, this pattern involved a reduction in class preference by the LDA models.

	Mean Acc (%)	P300 Acc (%)	Non-P300 Acc (%)	Solver	Shrinkage
Subject 1	57.14	66.67	55.56	lsqr	0.01
Subject 2	38.10	66.67	33.33	lsqr	0.78
Subject 3	66.67	33.33	72.22	lsqr	0.00
Subject 4	85.71	0.00	100.00	lsqr	0.91
Subject 5	33.33	33.33	33.33	lsqr	0.00
Sub Avg	56.19	40.00	58.89	n/a	0.34
Sub Var	26.19	33.34	33.34	n/a	0.46

Table 4.13: Here is displayed a classification table denoting all associated evaluation metrics for the localizer-data-initialized LDA models for the 7 Emoji with Localizre dataset computed (refer to, Table 3.1 for data partition info). For more information on field headings refer to, Table 3.2.

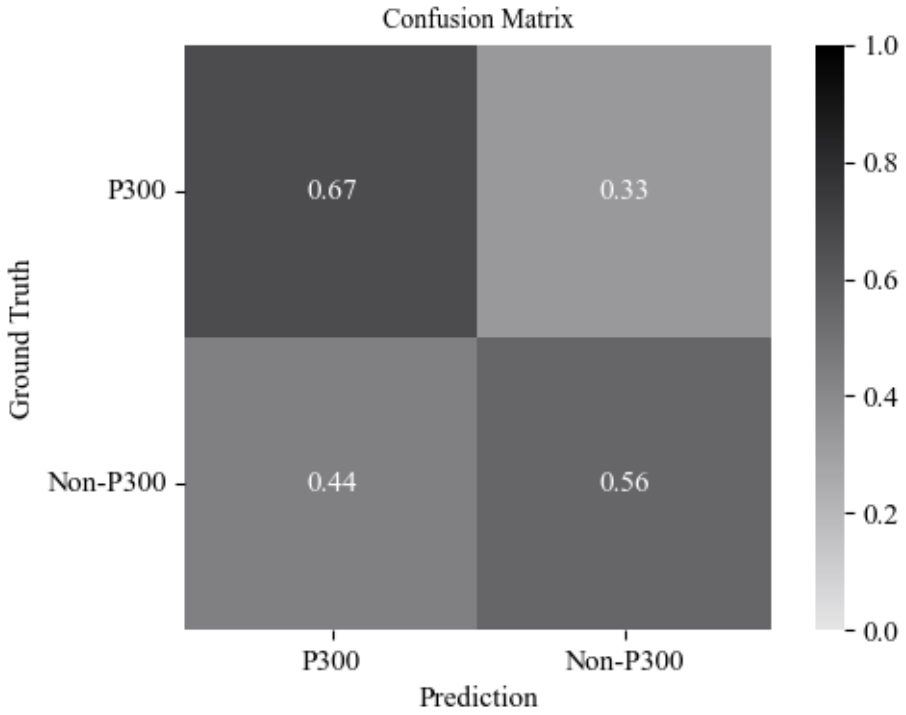

Figure 4.20: The figure displays a normalized confusion matrix showing the classification performance of a localizer-initialized and trained LDA model for both P300 and Non-P300 classes relating to Subject 1 in the 7 Emoji with Localizer dataset (refer to, Table 3.1).

In the confusion matrix positioned above (see, Figure 4.20), the prediction behaviour for the LDA model trained on data from Subject 1 is displayed. The confusion presented by the model is largely shared across both classes, with the Non-P300 events being misclassified as P300 events to a slightly greater degree. Overall, the prevalence of a selective bias is marginal

and the evaluation demonstrates that the LDA classifier did learn to distinguish some features of the class types tested. The extent of the learning achieved to accurately distinguish between the class types is shown to be far from complete and would not serve as a robust means of prediction.

4.5 Results: Pipeline 2

Here, are presented all results from Experiment 2 generated using the Pipeline 2 approach (for further information, see subsection 3.3.5.3). Broadly, these results describe an assessment of the staggered (3, 5, & 7) emoji variants. All results for the single-subject analysis are computed using parametric one-sample t-tests to gauge the performance of the associated models against the 50% chance threshold (see subsection 3.4.3.2). Following this, the relative difference in performance across the Non-Collapsed and Collapsed data partitions is evaluated using a non-parametric permutation test (see subsection 3.4.3.2). Finally, differences in mean overall classification accuracy across the three experimental variants are assessed using a similar paired-subject permutation test. For these data, class balancing was addressed through the application of a SMOTE oversampling method (see subsection 3.4.3.3).

4.5.1 Data Partitions

In this subsection, the data partitions associated with the analysis for corresponding stimulus variants (3, 5, & 7-Emoji) as discussed. Positioned below are a series of 3 tables relating to each of the stimuli assessed in the staggered emoji-array investigation (see, Tables 4.14, 4.15 & 4.16). These are presented to assist in the understanding of how each respective LDA model was trained and tested throughout the 10-fold cross-validation procedure. The quantities of Target-P300 and Non-Target Non-P300 samples for the subject are shown at each stage of the data preparation process. This information is critical, given the variance in the proportion of synthetic to real data for the Target-P300 samples included in the test set. As can be seen in the Train column listing the quantities of real and synthetic samples across all three tables the ratio of real to synthetic samples mirrors the same ratio of original Target and Non-Target samples. Here, around 66%, 80% and 85%, of all Target training samples are synthetic for the 3, 5 and 7-Emoji variants respectively.

To avoid excessive repetition of highly similar information, the corresponding Collapsed data distribution tables are positioned in the Appendix (see Tables A.5-A.7). The ratios between all samples in each respective grouping are highly comparable to those observed here. This

is because they are constructed from the same set of data, however the data preparation, as noted above, involves the aggregation of neighbouring trial sequences. Due to this process, the Collapsed data partitions feature roughly half as many samples per data partition grouping as presented here for each experimental variant analyzed. In real terms, the number of trials per subject for each emoji on-screen dropped from around 30 to around 15, hence this dramatically decreased the number of samples available for training and testing. In all Collapsed instances, as the number of P300 events stays the same irrespective of the emoji stimulus roughly 11-14 samples are retained following the channel-rejection procedure, 10% of this sample is reserved for testing and as the value must be rounded down it means all Collapsed models are evaluated in each k-fold with a single P300 test sample. For this reason, significant caution must be taken when interpreting the related results.

	3-Emoji Non-Collapsed						
	Total Post-Rejection		Test		Train		
Subjects	P300	Non-P300	P300	Non-P300	P300 (Real)	P300 (Synthetic)	Non-P300
Subject 1	29	58	3	6	26	26	52
Subject 3	27	55	3	6	24	25	49
Subject 5	27	56	3	6	24	26	50

Table 4.14: Here is presented a table detailing the distribution of sample quantities for the subject-specific datasets associated with the 3-Emoji (see Figure 4.4), Non-Collapsed Pipeline 2 approach (see subsection 3.3.5.3). All samples here are composed of signals collected over all 5 sequences of each trial (see subsection 4.3.7). For further information on field headings and interpretation please refer to Table 3.14. Note, that the ratios between Target and Non-Target samples for all datasets listed, including the proportion of Real vs. Synthetic P300 instances mirror those in the Collapsed data preparation variant. To avoid excessive repetition, the data distribution table related to the 3-Emoji Collapsed augmentation method is positioned in the Appendix Table A.5.

	5-Emoji Non-Collapsed						
	Total Post-Rejection		Test		Train		
Subjects	P300	Non-P300	P300	Non-P300	P300 (Real)	P300 (Synthetic)	Non-P300
Subject 1	28	116	3	12	25	79	104
Subject 3	26	113	3	11	23	78	101
Subject 5	26	114	3	11	23	79	102

Table 4.15: Here is presented a table detailing the distribution of sample quantities for the subject-specific datasets associated with the 5-Emoji (see Figure 4.3), Non-Collapsed Pipeline 2 approach (see subsection 3.3.5.3). All samples here are composed of signals collected over all 5 sequences of each trial (see subsection 4.3.7). For further information on field headings and interpretation please refer to the table Table 3.14. Note, that the ratios between Target and Non-Target samples for all datasets listed, including the proportion of Real vs. Synthetic P300 instances mirror those in the Collapsed data preparation variant. To avoid excessive repetition, the data distribution table related to the 5-Emoji Collapsed augmentation method is positioned in the Appendix Table A.6.

	7-Emoji Non-Collapsed						
	Total Post-Rejection		Test		Train		
Subjects	P300	Non-P300	P300	Non-P300	P300 (Real)	P300 (Synthetic)	Non-P300
Subject 1	28	172	2	17	26	128	154
Subject 3	29	174	3	17	26	130	156
Subject 5	29	169	3	17	26	126	152

Table 4.16: Here is presented a table detailing the distribution of sample quantities for the subject-specific datasets associated with the 7-Emoji (see Figure 4.2), Non-Collapsed Pipeline 2 approach (see subsection 3.3.5.3). All samples here are composed of signals collected over all 5 sequences of each trial (see subsection 4.3.7). For further information on field headings and interpretation please refer to Table 3.14. Note, that the ratios between Target and Non-Target samples for all datasets listed, including the proportion of Real vs. Synthetic P300 instances mirror those in the Collapsed data preparation variant. To avoid excessive repetition, the data distribution table related to the 7-Emoji Collapsed augmentation method is positioned in the Appendix Table A.7.

4.5.2 3-Emoji Variant: Non-Collapsed: Pipeline 2

The results in this subsection relate to the Non-Collapsed 3-Emoji experimental variant processed using the Pipeline 2 approach. As stated above, the Non-Collapsed samples were constructed using all 5 sequences within each experimental trial. This involved the presentation of 3 emojis on screen in an offline BCI communication speller investigation, using the Inversion augmentation method to produce time-locked P300 waveforms for a cued target emoji stimulus. Here the single-subject Overall, Target and Non-Target mean classification accuracies are discussed in terms of significance with reference to a one-sample t-test (threshold, $p<0.05$). Further, the group-level results are discussed in relation to a paired-subjects permutation test (see subsection 3.4.3.2).

	Overall		Target		Non-Target	
Subjects	Acc Mean	Std Dev	Acc Mean	Std Dev	Acc Mean	Std Dev
1	0.71*	0.11	0.84*	0.17	0.57	0.19
3	0.69*	0.15	0.85*	0.16	0.55	0.22
5	0.72*	0.12	0.81*	0.18	0.64	0.23

Table 4.17: Here is presented a table showing the performance metrics associated with Subjects 1, 3 & 5 for the Inversion Non-Collapsed data partition (see, Table 4.14). All results were computed following the stages laid out in the Pipeline 2 data organisation, pre-processing and analysis methodology. Here all individual samples are composed of averages computed across all 5 augmentation sequences within each respective trial (see subsection 3.3.5.3 Data Pre-Processing: Pipeline 2). Note, that all cell values denoted with a * indicate a significantly higher mean classification accuracy than the 50% chance level for the binary (Target *vs*. Non-Target) classification task. For additional information on table field headings and interpretation please refer to Table 3.16.

As seen from the table above (see, Table 4.17), all single subject-level Overall and Target mean classification accuracies were significantly above the 50% chance level ($p<0.05$) despite high standard deviations for the associated metrics. Further, the single-subject Non-Target mean classification accuracies computed for the results of the 10-fold cross-validation procedure all reported non-significant results following similar one-sample t-tests. Notably, here, the standard deviation is marked higher than the Overall and Target results suggesting that for a substantial majority of the cross-validation *k*-folds the subject-level accuracies dipped below the random performance threshold of 50%. Further, the group-level assessments involving the comparison of all subject mean classification accuracies for each accuracy metric variant computed via permutation test revealed a similar pattern. Here, only the accuracies relating to the Target samples were found to be significantly different to the chance ($p=0.049$). This result is highly marginal given the given associated mean classification accuracy standard deviations. Based on these results it can not be concluded that the implementation of the 3-Emoji variant in this configuration would serve as a capable BCI emoji-communication platform.

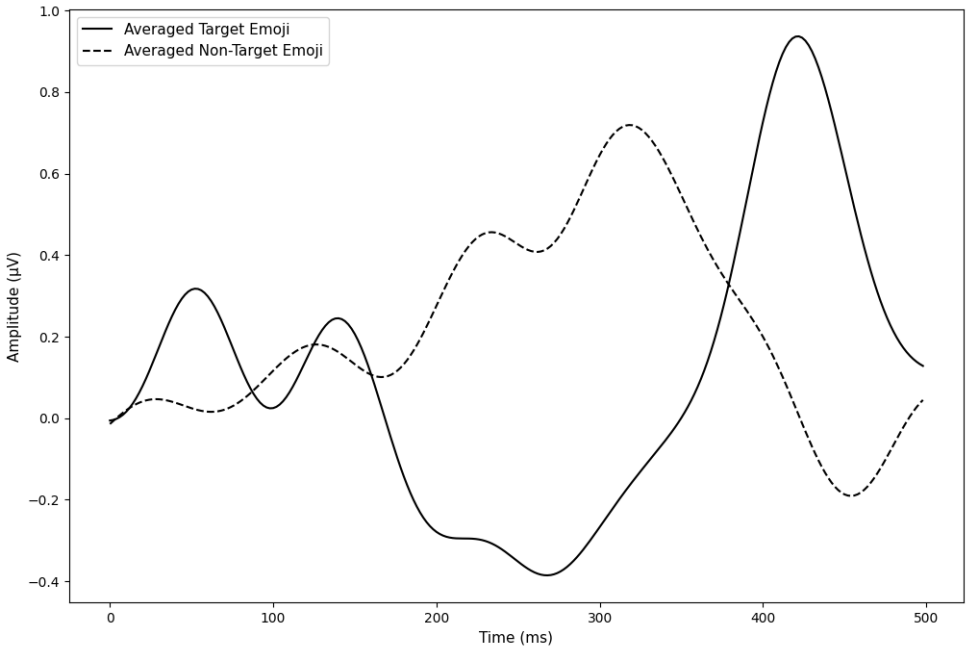

Figure 4.21: This is an average plot constructed exclusively from the Cz electrode for all P300, Averaged Target Emoji (solid line) and Non-P300 Averaged Non-Target Emoji (dashed line) samples collected across Subjects 1, 3 & 5 for the 3-Emoji Non-Collapse data partition (see, Table 4.14). As can be seen, the time dimension is positioned on the x-axis (0-500ms) and the micro-voltage range is oriented to the y-axis. Here, all samples were processed using the Pipeline 2 data pre-processing methodology (see subsections 3.3.5.3 & 3.3.5.5). Every Target P300 and Non-Target P300 averaged sample was aggregated into separate arrays and a grand pooled-subject mean signal was computed to generate the plot seen here. A total of 83 P300 samples and 169 Non-P300 samples were utilized respectively. At no point, were any synthetic P300 samples included in the construction of these average signals. Note, that the primary difference between the Non-Collapsed and Collapsed data variants relates to the number of sequences assigned to each sample average. As the plot here features all samples it represents a linear combination of all signals within the given augmentation variant meaning the corresponding average plot generated using the Collapsed data is effectively identical to the figure shown here.

The grand average plot positioned above (Figure 4.21) demonstrates a strong negative component at 250ms and a large positive deflection around 400ms. Both components are significantly delayed as compared to the plots seen in Experiment 1 (see, Figures 3.29 & 3.30) using the same Pipeline 2 data organisation pre-processing methodology. Notably, the Non-Target grand average also presents a large positive deflection, here at around 300ms. This could suggest that for this experimental variant, subjects failed more frequently in their attempts to maintain fixation on the cued emoji stimulus, this is surprising given the significantly greater distance between the targets on screen for this iteration (see, Figures 3.1, 4.2, 4.3 & 4.4).

4.5.3 3-Emoji Variant: Collapsed: Pipeline 2

The results reported here correspond to the Collapsed 3-Emoji data partition (see, Table in Appendix A.5) processed using the Pipeline 2 approach. For these analyses, all samples were constructed from 10 sequences collapsed across two experimental trials. For more information refer to subsections 3.3.5.3 and 4.3.5.5. As is shown in the table below (see, Table 4.18), all subjects returned Overall and Target mean classification accuracies above the chance 50% level as computed via one sample t-test. Notably, only Subject 3 returned mean accuracy significantly above chance for the Non-Target instances. Further, Subject 5 is the sole subject to reach the functional performance threshold for the Overall mean accuracy. This is driven primarily by a nearly maximal mean classification accuracy for the Target samples.

The 38% difference between the corresponding Target and Non-Target metrics could indicate bias from the classifier for the former. Here, the standard deviation reported for the Non-Target samples is extremely high (0.28), suggesting that for several of the 10 cross-validation *k*-fold associated LDA model accuracies dropped below the 50% threshold. Further, Subject 1 reported a mean classification accuracy of just 35% for the Non-Target instances, this relative drop in accuracy here and high relative accuracy for the Target samples is arguably a stronger indication of overfitting. In this instance, well over half of all cross-validation *k*-folds produced sub-50% classification accuracies. The author notes that there is a clear trend of potential overfitting exhibited by the LDA models for the Collapsed data preparation variant, as compared to the Non-Collapsed variant likely driven by the substantially lower number of samples available to associated models for training.

	Overall		Target		Non-Target	
Subjects	Acc Mean	Std Dev	Acc Mean	Std Dev	Acc Mean	Std Dev
1	0.62*	0.15	0.88*	0.16	0.35	0.25
3	0.63*	0.05	0.60*	0.02	0.67*	0.01
5	0.82*	0.12	0.98*	0.05	0.61	0.28

Table 4.18: Here is presented a table showing the performance metrics associated with Subjects 1, 3 & 5 for the 3-Emoji Collapsed data partition (see Table in Appendix A.5). All results were computed following the stages laid out in the Pipeline 2 data organisation, pre-processing and analysis methodology. Here all individual samples are composed of averages computed across 2 trials consisting of 10 sequences each (see subsection 3.3.5.3 Data Pre-Processing: Pipeline 2). Note, that all cell values denoted with a * indicate a significantly higher mean classification accuracy than the 50% chance level for the binary (Target *vs*. Non-Target) classification task. For additional information on table field headings and interpretation please refer to Table 3.16.

The group-level stats comparing the mean classification accuracies reported across subjects against a 50% random performance threshold revealed that Non-Target mean accuracies were highly non-significantly different(p=0.5), with the Overall mean accuracy providing a marginal result (0.051), likely owing to the single high accuracy observed for Subject 5. Further, the Target samples mean accuracies breached the significance threshold (p=0.044). Note, that the permutation tests performed here do not feature the consideration of the associated standard deviations. Hence, the results observed here for these highly volatile classification accuracies must be interpreted with caution. Finally, the paired-subject performance across the Non-Collapsed and Collapsed data preparation variants via permutation test here reported no significant difference for the Overall, Target or Non-Target samples mean classification results. Regarding the Non-Target samples, this is likely owing to the small observed mean difference (4.33%) between the Non-Collapsed (58.67%) and Collapsed (54.33%) pooled-subject averages.

4.5.4 5-Emoji Variant: Non-Collapsed: Pipeline 2

The results in this subsection relate to the Non-Collapsed 5-Emoji (see, Figure 4.3) experimental variant processed using the Pipeline 2 approach (see subsections 3.3.5.3 & 3.4.3). As stated above, the Non-Collapsed samples were constructed using all 5 sequences within each experimental trial (see subsection 3.3.5.1). As seen below, (see, Table 4.19), the one-samples t-test conducted on the classification accuracies for all 10 cross-validation folds against the 50% random performance threshold for each subject accuracy metric (Overall, Target & Non-Target) were found to be significantly higher than chance. Notably, both Subjects 1 and 5 report Overall mean classification accuracies above the 70% functional use threshold. Further, the group-level results computed across subjects via permutation test reveal both the Overall and Target samples accuracy means were significantly above chance levels.

	Overall		Target		Non-Target	
Subjects	Acc Mean	Std Dev	Acc Mean	Std Dev	Acc Mean	Std Dev
1	0.80*	0.10	0.94*	0.05	0.65*	0.19
3	0.74*	0.07	0.88*	0.06	0.60*	0.12
5	0.80*	0.08	0.90*	0.09	0.70*	0.14

Table 4.19: Here is presented a table showing the performance metrics associated with Subjects 1, 3 & 5 for the 5-Emoji Non-Collapsed data partition (see, Table 4.15). All results were computed following the stages laid out in the Pipeline 2 data organisation, pre-processing and analysis methodology. Here all individual samples are composed of averages computed across all 5 augmentation sequences within each respective trial (see subsection 3.3.5.3 Data Pre-Processing: Pipeline 2). Note, that all cell values denoted with a * indicate a significantly higher mean classification accuracy than the 50% chance level for the binary (Target *vs*. Non-Target) classification task. For additional information on table field headings and interpretation please refer to Table 3.16.

In the figure below (see, Figure 4.22), the quality of the Cz grand average means for the Target-P300 trials shows three distinct peaking events with separation of around 160-170ms, suggesting the presence of a 5.75-6.25Hz signal artefact. This is likely related to the inter-stimulus interval of 150ms (see subsection 3.3.5.1) leading to the propagation of a related SSVEP waveform. This is despite the explicit application of a 6.67Hz SSVEP-targetted notch filter (see subsection 3.3.5.5) applied to all collected samples. Here, the quality factor was set relatively high to avoid attenuating signals in the frequency range near the expected ERP waveforms, however, these plots suggest either reducing the quality factor or applying multiple notch filters around a mean of 6.67Hz would be a more effective pre-processing approach.

The plot suggests that this configuration of experimental stimuli potentially increased the difficulty for these subjects to attend exclusively to the cued emoji stimulus. Here, both the Target and Non-Target grand averages present with a fairly strong drifting component likely owing to the suboptimal baselining procedure implemented. Here, owing to the erroneous decision to select a non-continuous data acquisition style, the only samples available for baselining were the initial 50ms of the time-locked data segments. Further, given the relatively small number of samples collected per class as compared to Experiment 1 (30 *vs*. 98), the resulting averages demonstrate a lower signal-to-noise ratio. The presence of the strong SSVEP component here complicates the author's ability to effectively characterize the Event-Related Potentials embedded in the Target-P300 waveform. It is clear that the signal possesses more variance as compared to the Non-Target, Non-P300 signals and the absence of this oscillation in the Non-Target signals suggests that the LDA model classifiers may have learned to separate the class based on the absence or presence of a hybrid SSVEP-P300 signal.

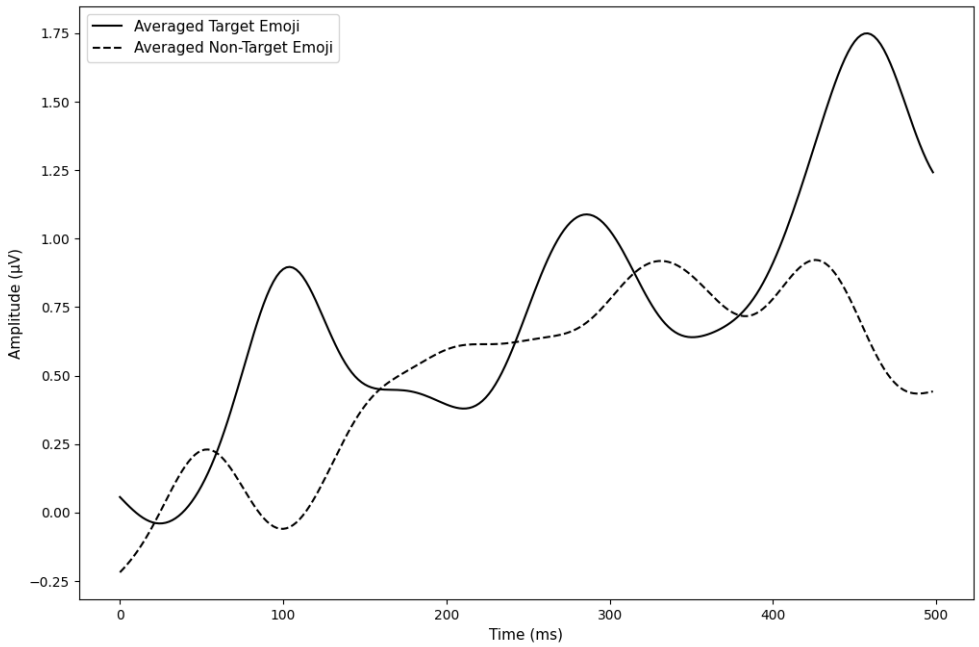

Figure 4.22: This is an average plot constructed exclusively from the Cz electrode for all P300, Averaged Target Emoji (solid line) and Non-P300 Averaged Non-Target Emoji (dashed line) samples collected across Subjects 1, 3 & 5 for the 5-Emoji Non-Collapse data partition (see, Table 4.15). As can be seen, the time dimension is positioned on the x-axis (0-500ms) and the micro-voltage range is oriented to the y-axis. Here, all samples were processed using the Pipeline 2 data pre-processing methodology (see subsections 3.3.5.3 & 4.3.5.5). Finally, every Target P300 and Non-Target P300 averaged sample are aggregated into separate arrays and a grand pooled-subject mean signal is computed to generate the plot seen here. A total of 80 P300 samples and 343 Non-P300 samples were utilized respectively. At no point, were any synthetic P300 samples included in the construction of these average signals. Note, that the primary difference between the Non-Collapsed and Collapsed data variants relates to the number of sequences assigned to each sample average. As the plot here features all samples it represents a linear combination of all signals within the given augmentation variant meaning the corresponding average plot generated using the Collapsed data is effectively identical to the figure shown here.

4.5.5 5-Emoji Variant: Collapsed: Pipeline 2

The results in this subsection relate to the Collapsed 5-Emoji (see, Figure 4.3) experimental variant processed using the Pipeline 2 approach (see subsections 3.4.3 & 4.3.5.5. For the Collapsed data preparation method all samples were constructed by collapsing neighbouring trials together and averaging over 10 sequences (see subsection 3.3.5.3). As shown below (see, Table 4.20), the mean classification accuracy metrics computed for this data partition were all shown to be greater than the 50% chance level via a one-sample t-test. Notably, the pooled-subject performance shown here for the Oveall and Target samples is the highest recorded for any of the three experimental variants discussed in this chapter. This is primarily driven by

the near-maximal classification accuracies seen for the Target samples. Further, the standard deviations of the associated values are extremely low, suggesting near-maximal performance for all 10 cross-validation *k*-folds.

This is contrasted against the high standard deviations for the Non-Target class. The group-level stats computed across mean classification accuracy metrics against the chance 50% report that both the Overall and Target samples showed a significant difference, with the marginal result for the Non-Target class (p=0.053). Finally, the Non-Collapsed and Collapsed data partitions were investigated using the paired-subject data across the two data preparation methods via the permutation test. Despite increases following the Collapse data aggregation procedure of 6.3%, 8.3% and 5.3% for the Overall, Target and Non-Target accuracy means none of the paired-subject assessments report significant differences, with the Target samples producing a p-value of 0.096. Here it is likely with the addition of more subjects these metrics would have breached the 0.05 significance threshold given the strong trend observed.

	Overall		Target		Non-Target	
Subjects	Acc Mean	Std Dev	Acc Mean	Std Dev	Acc Mean	Std Dev
1	0.88*	0.10	1.00*	0.00	0.76*	0.15
3	0.82*	0.11	0.99*	0.03	0.68*	0.22
5	0.83*	0.09	0.99*	0.03	0.67*	0.21

Table 4.20: Here is presented a table showing the performance metrics associated with Subjects 1, 3 & 5 for the 5-Emoji Collapsed data partition (see Table in Appendix A.6). All results were computed following the stages laid out in the Pipeline 2 data organisation, pre-processing and analysis methodology. Here all individual samples are composed of averages computed across 2 trials consisting of 10 sequences each (see subsection 3.3.5.3 Data Pre-Processing: Pipeline 2). Note, that all cell values denoted with a * indicate a significantly higher mean classification accuracy than the 50% chance level for the binary (Target *vs*. Non-Target) classification task. For additional information on table field headings and interpretation please refer to Table 3.16.

4.5.6 7-Emoji Variant: Non-Collapsed: Pipeline 2

The results in this subsection relate to the Non-Collapsed 7-Emoji (see, Figure 4.2) experimental variant processed using the Pipeline 2 approach (see subsections 3.4.3 & 4.3.5.5). As stated above, the Non-Collapsed samples were constructed using all 5 sequences within each experimental trial (see subsection 4.3.5.1). As seen below, all subjects reported mean classification accuracies significantly above chance. Notably, all metric variants display lower associated standard deviations as compared to the previous assessments conducted. Further, the same trend is observed characterised by a bias towards the Target-P300 class. Additionally, Subject

1 reveals substantially lower mean classification accuracies for this emoji stimulus variant, as compared to Subjects 3 and 5. Notably, this drop is attributed to lower relative classification accuracy for both the Target and Non-Target samples. The group-level stats reported that the mean accuracies across subjects follow the same trend with the Target samples computed as significantly different and the Overall and Non-Target accracse reporting non-significant differences to the chance 50% level. As stated previously, the validity of these permutation tests is limited by the small sample size and relative statistical power of this non-parametric test.

	Overall		Target		Non-Target	
Subjects	Acc Mean	Std Dev	Acc Mean	Std Dev	Acc Mean	Std Dev
1	0.74*	0.07	0.85*	0.10	0.64*	0.10
3	0.85*	0.05	0.94*	0.06	0.75*	0.07
5	0.82*	0.05	0.92*	0.07	0.71*	0.11

Table 4.21: Here is presented a table showing the performance metrics associated with Subjects 1, 3 & 5 for the 7-Emoji Non-Collapsed data partition (see, Table 4.16). All results were computed following the stages laid out in the Pipeline 2 data organisation, pre-processing and analysis methodology. Here all individual samples are composed of averages computed across all 5 augmentation sequences within each respective trial (see subsection 3.3.5.3 Data Pre-Processing: Pipeline 2). Note, that all cell values denoted with a * indicate a significantly higher mean classification accuracy than the 50% chance level for the binary (Target *vs*. Non-Target) classification task. For additional information on table field headings and interpretation please refer to Table 3.16.

The figure below shows the Cz grand average computed using all pooled-subject Target (oddball) and Non-Target samples. Here the Target waveform presents with a negative deflection around 200ms and two distinct positive components around 300 and 450ms. Notably, there is also present a large initial positive deflection around 150ms. There remains the possibility that the peaking observed at 150, 300 and 450ms is indicative of a 6.67Hz SSVEP signal. Further, for the Non-Target waveform an oscillation is observed with a periodicity of around 75ms, this is likely a harmonic of the same SSVEP waveform. Further, both signals demonstrate significant drift over the course of the time window indicating that again, the compromised baselining procedure was not sufficient to orient the signal around a zero mean. Despite this, the target samples do present with a larger area under the curve and could serve as a reliable means of separation via the corresponding LDA models, however, it is difficult to assert the data groupings are solely based on ERP features.

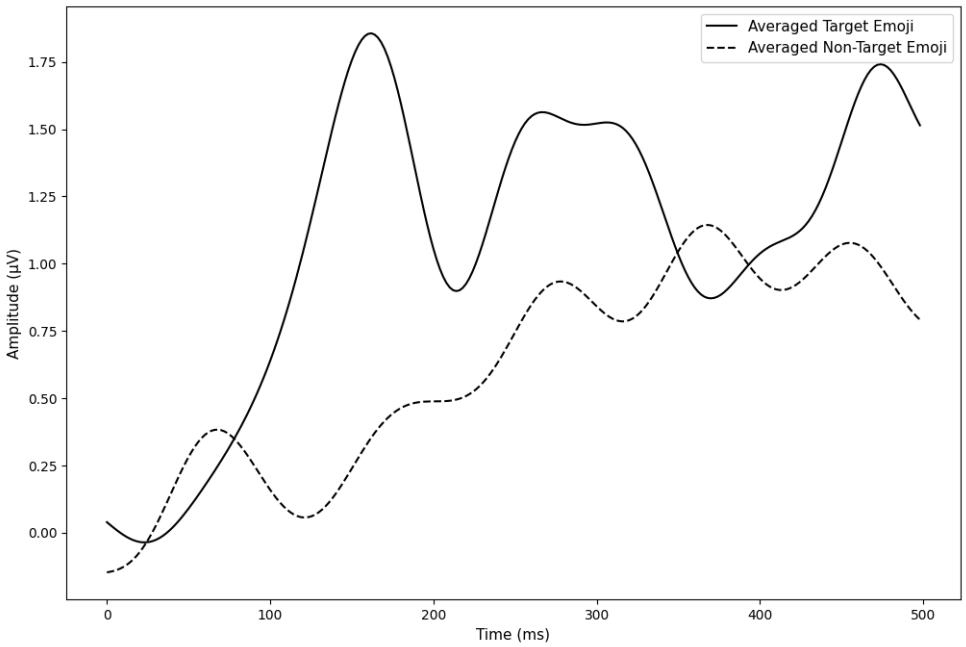

Figure 4.23: This is an average plot constructed exclusively from the Cz electrode for all P300, Averaged Target Emoji (solid line) and Non-P300 Averaged Non-Target Emoji (dashed line) samples collected across Subjects 1, 3 & 5 for the 7-Emoji Non-Collapse data partition (see, Table 4.16). As can be seen, the time dimension is positioned on the x-axis (0-500ms) and the micro-voltage range is oriented to the y-axis. Here, all samples were processed using the Pipeline 2 data pre-processing methodology (see subsections 3.3.5.3 & 4.3.5.5). Finally, every Target P300 and Non-Target P300 averaged sample are aggregated into separate arrays and a grand pooled-subject mean signal is computed to generate the plot seen here. A total of 86 P300 samples and 515 Non-P300 samples were utilized respectively. At no point, were any synthetic P300 samples included in the construction of these average signals. Note, that the primary difference between the Non-Collapsed and Collapsed data variants relates to the number of sequences assigned to each sample average. As the plot here features all samples it represents a linear combination of all signals within the given augmentation variant meaning the corresponding average plot generated using the Collapsed data is effectively identical to the figure shown here.

4.5.7 7-Emoji Variant: Collapsed: Pipeline 2

The results in this subsection relate to the Collapsed 7-Emoji (see, Figure 4.2) experimental variant processed using the Pipeline 2 approach (see subsections 3.4.3 & 4.3.5.5). For the 190 Experiment 2: Variable Array Density Assessments Collapsed data preparation method all samples were constructed by collapsing neighbouring trials together and averaging over 10 sequences (see subsection 4.3.5.1). The table below shows the mean classification accuracies for the 3 Subjects assessed. Here, all associated one-sample t-tests revealed the metrics were significantly higher than the 50% chance threshold. These results demonstrate a strong indication of overfitting towards the Target class. As seen in Subject 1, the mean accuracy

for Non-Target samples is 64% with a large standard deviation(0.19) suggesting that several of the cross-validation *k*-folds reported random performance-level accuracy metrics. This in tandem with the near maximal accuracy classification of the Target class suggests that up to 35-40% of all sample misclassifications involved the incorrect identification of Non-Target as Target samples. The group-level stats report that both the Overall and Target mean accuracy was significantly higher than chance, with the Non-Target mean accuracies reported as non-significantly different to chance. In the evaluation comparing the paired-subject performance between the Non-Collapsed and Collapsed data preparation methods, no significant difference was reported for the Overall or Non-Target mean accuracies, with a p-value of 0.097 reported for the Target Non-Collapsed (Avg.=90.33%) and Collapsed (Avg.=97.00%).

	Overall		Target		Non-Target	
Subjects	Acc Mean	Std Dev	Acc Mean	Std Dev	Acc Mean	Std Dev
1	0.82*	0.09	0.99*	0.04	0.64*	0.19
3	0.86*	0.07	0.96*	0.06	0.76*	0.12
5	0.81*	0.07	0.96*	0.09	0.67*	0.13

Table 4.22: Here is presented a table showing the performance metrics associated with Subjects 1, 3 & 5 for the 5-Emoji Collapsed data partition (see Table in Appendix A.6). All results were computed following the stages laid out in the Pipeline 2 data organisation, pre-processing and analysis methodology. Here all individual samples are composed of averages computed across 2 trials consisting of 10 sequences each (see subsection 3.3.5.3 Data Pre-Processing: Pipeline 2). Note, that all cell values denoted with a * indicate a significantly higher mean classification accuracy than the 50% chance level for the binary (Target *vs*. Non-Target) classification task. For additional information on table field headings and interpretation please refer to Table 3.16.

4.5.8 3 *vs.* 5 *vs.* 7 Emoji: Pipeline 2

Here is presented a cross-experimental comparison of the 3 emoji-stimulus variants in terms of Overall Accuracy assessed over the two data preparation methods applied, Non-Collapsed and Collapsed. The plot below (see, Figure 4.24) shows that performance for the 3-Emoji variant in Non-Collapsed instances was consistently the lowest-performing stimulus presentation method, dropping well below the functional performance threshold of 70% for all 3 Subjects evaluated. To probe the relative difference in performance between the conditions more thoroughly a series of paired-subjects permutation tests were conducted. Here, the paired-subject results for the 5-Emoji (Avg.=78.00%) and 7-Emoji (Avg.=80.33%) variants revealed no significant differences (p=0.59). Both comparisons made between the 5 and 7-Emoji against the 3-Emoji variants returned relatively marginal results, p=0.099 and p=0.096 respectively. None of the stimulus presentation methods produced results significantly above the random

performance threshold.

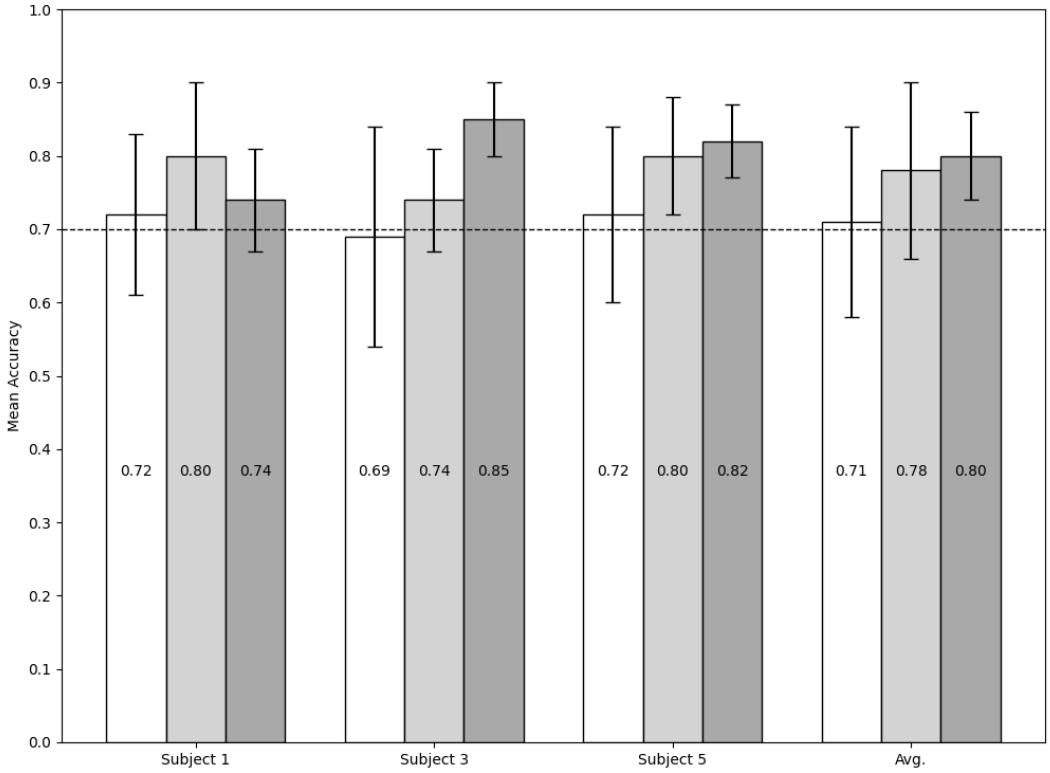

Figure 4.24: The plot displays a paired bar chart of the mean Overall accuracies and standard deviations for the 3 (white bars), 5 (light grey bars) and 7-Emoji (dark grey bars) stimulus presentation methods using the Non-Collapsed data preparation technique in which each sample consisted of an average computed over 5 augmentation sequences (see subsection 3.3.5.3 for further information). These mean values are computed from a 10-fold cross-validation for each of the three subjects (1, 3 & 5) along with the pooled-subject average (Avg.) (see subsection 3.4.3.1). The figure also includes standard deviation bars to show variability in the results of the cross-validation. Each bar is also annotated with its corresponding mean accuracy value. A horizontal dashed line at 70% is included to help assess the performance of each method against this functional performance benchmark.

The figure positioned below reveals similar patterns of results for the Collapsed data preparation variant. Here, it is clear that the Overall accuracies observed for Subjects 1 and 3 dropped dramatically as compared to the Non-Collapsed variant for the 3-Emoji stimulus presentation method. Notably, Subject 5 demonstrates highly stable and overlapping performance accuracies for all 3 stimulus variants. On the whole, for the 5 and 7-Emoji variants, all subjects demonstrated substantial increases in Overall mean classification accuracy, with a jump of 12% for Subject 3 relating to the 5-Emoji variant. The paired-subject permutation test was implemented here to observe the differences between stimulus variants as was discussed pre-

viously. As noted earlier, no significant difference is observed between the 5 (Avg.=84.33%) and 7-Emoji (Avg=83.00%) variants. Further, despite the pooled-subject average mean of the 3-Emoji variant of 69.00%, neither the 5 or 7-Emoji paired subject results revealed a significant difference. This is despite the absence of standard deviation overlap present in both Subjects 1 and 3. The author asserts that this is likely owing to the small sample size.

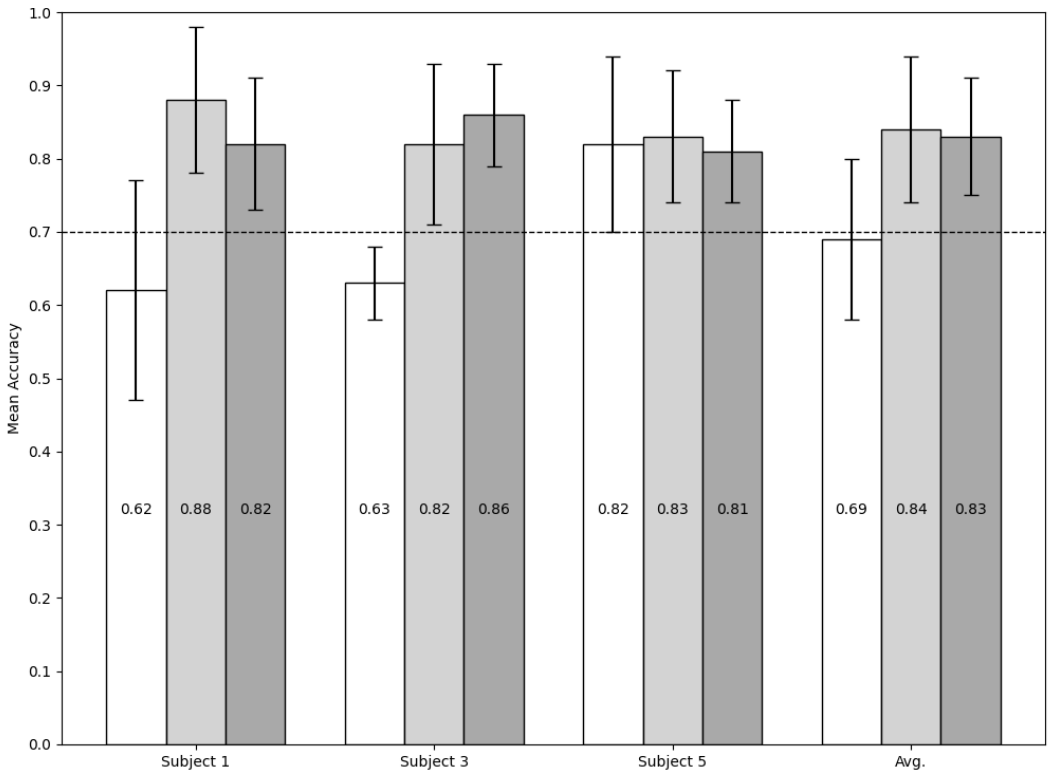

Figure 4.25: The plot displays a paired bar chart of the mean Overall accuracies and standard deviations for the 3 (white bars), 5 (light grey bars) and 7-Emoji (dark grey bars) stimulus presentation methods using the Collapsed data preparation technique in which each sample consisted of an average computed over 10 augmentation sequences (see subsection 3.3.5.3 for further information). These mean values are computed from a 10-fold cross-validation for each of the three subjects (1, 3 & 5) along with the pooled-subject average (Avg.) (see subsection 3.4.3.1). The figure also includes standard deviation bars to show variability in the results of the cross-validation. Each bar is also annotated with its corresponding mean accuracy value. A horizontal dashed line at 70% is included to help assess the performance of each method against this functional performance benchmark.

4.6 Conclusion: Pipeline 1

All interpretations and conclusions discussed herein relate to Experiment 2 computed via the Pipeline 1 approach (see subsections 3.3.5.1, 3.4, 4.3.5.4 & 4.3.6). All corresponding tables and figures can be found in the text positioned above. Again, all analyses were conducted offline and correspond to the three staggered experimental variants featuring 3, 5 and 7 Emoji targets. All experimental paradigm details are outlined in the corresponding section, 3.3.5.

4.6.1 3 Emoji Variant: No Localizer Pre-Training: Pipeline 1

Here are presented the conclusions relating to the 3 Emoji stimulus variant computed utilizing LDA models with no localizer data pre-training stage at either the single or pooled-subject levels.

4.6.1.1 Pooled-Subject

At the pooled-subject level (see, Table 4.2), mean accuracy (73.33%) is well above the random-performance threshold of 50%. Despite this, the imbalance in prediction selection for the Non-P300 class is still evident in the results, with nearly twice the within-class accuracy (86.67%) as compared to the P300 events (46.67%). The confusion matrix (see, Figure 4.5), demonstrates the lack of directionality for the P300 class with a near 1:1 ratio of hits (47%) and misses (53%). This suggests that the LDA models could broadly identify Non-P300 type waveforms despite the poor separation of the P300 class. As seen in the corresponding average plot (see, Figure 4.6) the expected drop in μV and subsequent peaking events are delayed by 50-100ms, further the Target signal (solid line) peak is lower than the Non-Target samples (dashed line) between 300-400ms. As mentioned in the above Method subsection, 4.3, the order of the task variants was counter-balanced. This could have led to more fatigue-related delaying of the P300 propagation influencing the position of the ERP crest post-cross trial averaging.

4.6.1.2 Within-Subject

The single-subject classification accuracies contain only one subject that produced a sub-RPT P300 classification accuracy (Subject 2, AoC $= 33\%$). These results demonstrate the same directional prediction preference for the Non-P300 class observed for both the class-balanced and non-class-balanced datasets in Experiment 1 (for reference see, Tables 3.7 & 3.9). Despite this, the degree of overfit observed is markedly reduced across the majority of subjects evaluated and classification accuracies extend significantly past the random performance threshold

in most instances. Further, the mean accuracy noted for Subject 5 of 88.89% is the highest class performance recorded.

4.6.1.3 Summary

Overall, these findings highlight the evident relationship between overfit incidence and visual array density. That is to say that the efficacy of class-balancing to mitigate overfitting reduces as the ratio of targets and non-targets approaches parity. The results presented here validate the original prediction, that a lower-density visual array would increase classification accuracies. This could be attributed to the reduction in computational complexity or a reduction in the degree of spatial bleed-over owing to the greater stimulus separability on screen. Despite the enhancement of performance metrics, the P300 class accuracies at the cross and single-subject levels never exceeded the 70% accuracy across both target classes needed for effective communication via a BCI speller as established in [250, 255]. Further, all assessments noted here were performed using the Pipeline 1 approach and do not feature cross-validation.

4.6.2 5 Emoji Variant: No Localizer Pre-Training: Pipeline 1

Here are listed the conclusions relating to the 5 Emoji stimulus variant computed utilizing LDA models with no localizer data pre-training stage at either the single or pooled-subject levels.

4.6.2.1 Pooled-Subject

The performance at the pooled-subject level for the 5 Emoji variant shows a marked departure from the accuracies achieved in the 3 Emoji variant (see, Tables 4.2 & 4.4, respectively). As shown, the P300 class accuracy is significantly lower at 0% and the Non-P300 class bias is near complete (98.33%). The confusion matrix (see, Figure 4.8), reveals that all P300 events were misclassified as Non-P300 events, repeating the pattern of overfitting observed for the non-class balanced analyses variants in Experiment 1. Interestingly, the pooled-subject average plots are also significantly different (see, Figure 4.9), displaying few of the characteristics typical to P300 (solid line) and Non-P300 waveforms (dashed line). Further, the magnitude of these signals is dramatically lower. The quality of these signals reduces the validity of any claims suggesting all the variance in classification performance can be wholly accounted for by the increase in array density. Equally, signal quality may be impeding the corresponding LDA models from effectively separating the classes and in turn reducing accuracy values.

4.6.2.2 Within-Subject

At the single-subject level, all subjects provided mean classification accuracies above the random performance threshold of 50%. Despite this, some returned P300 classification accuracies of 0% (Subjects 4 & 5). As seen when comparing the 3 and 5 Emoji variant classification tables (see, Tables 4.3 & 4.5), the increase in the number of non-target samples to the training array has re-introduced the issue of overfitting into the analyses. The highest-performing subject (see, Table 4.5) shows sub-random classification performance for the P300 class (33%) and as seen in the confusion matrix (see, Figure 4.10), the LDA model does not demonstrate any effective separation of the two classes based on the quality of the waveform features.

4.6.2.3 Summary

In sum, the increase in overfitting incidence between the 3 and 5 Emoji stimulus variants is clear at both the cross and single-subject levels. Further, the pattern of classification performance is not replicated here, with previously high-performing subjects now demonstrating at or below the random performance threshold (see subject 5, Tables 4.3 & 4.5). The author has determined these effects are owing either to the increase in the number of targets and resulting class imbalance in subsequent LDA model training datasets or a significant drop in data quality as evidenced by the grand average signal plots (see, Figure 4.9). It is possible that subjects may not be attending to the task as instructed and the expected P300 waveform propagation is not manifest in the plots. The author attempts to address this potential reduction in subject vigilance in the following experimental series (see subsection 5.4.9).

4.6.3 7 Emoji Variant: No Localizer Pre-Training: Pipeline 1

Here are presented the conclusions relating to the 7 Emoji stimulus variant computed utilizing LDA models with no localizer data pre-training stage at either the single or pooled-subject levels.

4.6.3.1 Pooled-Subject

As seen in comparisons between the 5 Emoji (see, Table 4.4) and 7 Emoji stimulus variants (see, Table 4.6), the degree of selective bias at the pooled-subject level for the Non-P300 class has increased in line with the addition of more targets to the visual array. It is possible the larger size of the dataset and the corresponding relative increase of Non-P300 training samples further exacerbated these overfitting effects. Additionally, the absence of regularization

measures to accommodate for this phenomenon such as the class-balancing protocols implemented in Experiment 1 (see subsection 3.4.1) ensures these confounding effects diminish the resulting classification performance.

4.6.3.2 Within-Subject

In all instances, an accuracy of 0% for the P300 class was reported at the single-subject level (see, Table 4.7). The variances in mean accuracy, Non-P300 accuracy and shrinkage values are all the lowest recorded in this experimental implementation thus far (Experiment 2). These results suggest all LDA models trained at the single-subject level underwent a similar process of overfitting to the more numerous Non-P300 class.

4.6.3.3 Summary

Overall, it is clear from the results (see, Tables 4.6 & 4.7) that these data depart from the previous experimental implementations in that the 3 Emoji variant displayed numerous high classification accuracies in conjunction with higher-quality average plots (see, Tables 4.2 & 4.3). In contrast, the 5 Emoji variant (see, Table 4.5) displayed a significantly lower quality of both classification accuracies and average plots. Concerning the 7 Emoji data, the classification accuracies are the lowest so far reported for this experimental implementation and also present with some higher quality signal averages (see, Figure 4.12). This demonstrates that to produce good classification results researchers must balance the processes of accommodating class event ratios at the analysis level and also ensure high-quality data collection during the acquisition phase. This can only be overcome with a regularization or priming stage for the LDA models trained in these more complex computational contexts.

4.6.4 3 Emoji Variant: With Localizer Pre-Training: Pipeline 1

Here are reported conclusions relating to the 3 Emoji stimulus variant computed utilizing LDA models initially tuned using a localizer data pre-training stage at both the single or pooled-subject levels.

4.6.4.1 Pooled-Subject

Concerning the pre-trained LDA model for the pooled-subject dataset evaluations, P300 class accuracies increased marginally from 46.67% to 53.33% (see, Tables 4.2 & 4.8). Despite this, a dramatic reduction in classification performance is noted for the Non-P300 class, dropping over 35%. For these data, the implementation of a localizer pre-training stage potentially

stunted the capability of the models to learn effective means of separating the waveforms tested. The parity in prediction selection across classes seen in the respective confusion matrix (see, Figure 4.15), shows that the reduction in overfit did not implicitly lead to a proportional increase in classification accuracies. This could be related to a distinct difference in data quality between the localizer 3 Emoji variant samples and 3 Emoji main experiment samples. Alternatively, the aggregation of samples across subjects and experimental formats (localizer and main experiment) could have further enhanced the prevalence of waveform features attributable to subject individual differences, effectively increasing the within-class diversity of samples and ultimately heightening the task complexity for LDA classifier models. Further, the high oddball probability of the localizer task likely resulted in weak P300 components (see Figure 4.14) and therefore these data were not suitable for the LDA models to learn data driven grouping for these classes.

4.6.4.2 Within-Subject

Broadly, at the single-subject level, the implementation of the LDA model localizer data pre-training reduced classification performance. Of note, Subjects 2 and 5 must be highlighted, as these evaluations demonstrate large decreases in classification performance for both the P300 and Non-P300 classes. The presence of low accuracies across both classes is unique to the analyses thus far and suggests for some subjects, the aggregation of data acquired over distinct experimental sessions (localizer and main experiment) is not viable. In contrast, Subject 4 returned the highest performance in cross-class accuracies for all subjects in any of the localizer pre-trained variants assessed.

4.6.4.3 Summary

On the whole, the implementation of the LDA localizer data pre-training stage resulted in a negligible increase in performance for the pooled-subject dataset. The reduction in overfitting previously observed was replaced by a convergence towards random classification performance. Further, the results observed at the single-subject level are arguably less favourable across the group with a notable exception for Subject 4. Note that this pattern of classification behaviour is expected given that the implementation of these regularization methods was targeted primarily towards the 5 and 7 Emoji variants to address the issue of overfitting that is dramatically less prevalent in these 3 Emoji variant data.

4.6.5 5 Emoji Variant: With Localizer Pre-Training: Pipeline 1

Here are noted conclusions relating to the 3 Emoji stimulus variant computed utilizing LDA models initially tuned using a localizer data pre-training stage at both the single or pooled-subject levels.

4.6.5.1 Pooled-Subject

The classification performance at the pooled-subject level after implementing the localizer data pre-training did demonstrate some improvements for the 5 Emoji variant dataset evaluations (see, Table 4.10). These are characterized by a significant decrease in the incidence of overfitting as compared to previous non-pre-trained results (see, Table 4.4). Further, for the P300 class specifically, the classification accuracies (60%) have been boosted beyond the RPT (50%). Despite, these improvements, a dramatic decrease in Non-P300 accuracies prevents the author from concluding the increased volume of training samples in concert with the pre-training stage produced a viable emoji-speller.

4.6.5.2 Within-Subject

As shown in the corresponding results table (see, Table 4.11) performance at the single-subject level after the implementation of the pre-training stage is substantially lower than in the non-pre-trained analyses variant (see, Table 4.5). The prevalence of overfitting is relatively similar, with large decreases in accuracy observed across both classes. The pattern of reduced classification performance is present in nearly all subjects, even in the instances that are characterized by an inversion of the selective bias towards the P300 class type. Based on these results, the localizer data pre-training stage cannot be said to have improved the classification performance for these subjects.

4.6.5.3 Summary

In sum, when looking at the 5 Emoji dataset, the performance for both the standardized and localizer-initialized methods was markedly lower than the 3 and 7 Emoji datasets. The lower performance cannot be attributed to fatigue effects, as the order of experimental variants was conducted across subjects using a pseudo-randomised protocol that evenly distributed the ordering of said variants across the subjects. It could be argued that the 3 Emoji variant is shorter and highly simplistic potentially meaning the task did not drain subject concentration over the experimental phase. In contrast, the 7 Emoji variant is more attentionally demanding meaning subjects could have been more engaged during the task. The 5 Emoji variant could be posi-

tioned in between these two opposing attentional states, ultimately reducing the engagement with the task and the resulting data quality.

4.6.6 7 Emoji Variant: With Localizer Pre-Training: Pipeline 1

Here are presented conclusions relating to the 7 Emoji stimulus variant computed utilizing LDA models initially tuned using a localizer data pre-training stage at both the single and pooled-subject levels.

4.6.6.1 Pooled-Subject

The performance of the pre-trained LDA models for the 7 Emoji variant (see, Table 4.12) outperformed similar classifiers for the corresponding non-pre-trained model (see, Table 4.6), the non-class balanced Experiment 1 Inversion pooled-subject model (see, Table A.1) and the class-balanced Experiment 1 Inversion pooled-subject model (see, Table 3.10). It could be argued that the difference in classification performance between the P300 (73.33%) and Non-P300 (52.22%) represents an instance of selective bias.

4.6.6.2 Within-Subject

The improvements in terms of overfitting reduction and P300 classification accuracy increases observed at the pooled-subject level are not replicated across all single-subject evaluations. The classification metrics of subjects 2, 3, 4 and 5 all demonstrate either model overfitting or highly atypical performance behaviour. As seen in the previous 3 Emoji variant analyses (see, Table 4.9) for Subject 4, large performance increases can be found for individual subjects following the localizer data pre-training stage, this is replicated here for Subject 1 to a far lesser extent.

4.6.6.3 Summary

Overall, any increases in performance following the model pre-training stage are restricted nearly exclusively to the pooled-subject dataset. The primary difference between these datasets is the overall size and diversity of waveform profiles present. It may be that the higher volume of samples in tandem with a greater array of P300 waveform profile expressions aided in the development of more robust class representations by the pooled-subject LDA model. Further, this aggregated localizer data was composed of significantly more samples, as compared to the single-subject datasets. This ensured that these pooled-subject LDA models were initially exposed to many more class-balanced examples than at the single-subject level due to

the localizer task being inherently class-balanced as it comprises 20 trials evenly split across augmented and non-augmented instances. When looking at the P300 class performance for the pooled-subject data the classification accuracy improves as the number of stimuli increases onscreen, 3 Emoji P300 accuracy = 53.33% (see, Table 4.8), 5 Emoji P300 accuracy = 60.00% (see, Table 4.10) and 7 Emoji P300 accuracy = 73.33% (see, Table 4.12). At each stage, as the size of the localizer-initialization dataset increases so does the classification performance at the pooled-subject level. These findings could suggest increasing the number of samples in the pre-training stage could improve performance in the main experiment. In contrast, the weak P300 components seen in Figure 4.14, the high associated oddball probability of the localizer task and the absence of cross-validation procedures for the results, as per the Pipeline 1 approach, prevent the formation of any broad conclusions regarding these findings.

4.6.7 Conclusion: Pipeline 2

Here, all conclusions relating to the 3-Emoji stimulus method samples processed using the Pipeline 2 approach are discussed. Please refer back to the relevant subsections for additional information regarding the specific presentation methods (see, Figure 4.4), the Pipeline 2 pre-processing method (see subsections 3.3.5.3 & 4.3.5.5), the associated Pipeline 2 cross-validation procedure (see subsections 3.4.3.1 & 4.3.7) and the associated tests of significance.

4.6.7.1 3-Emoji Results: Pipeline 2

As can be seen in Tables 4.17 and 4.18, the 3-Emoji variant produced the lowest mean classification accuracies of any of the stimulus presentation variants described in this chapter. This is surprising given the previous research outlined in [19, 20, 252]. Collectively, these studies suggest that a reduction in the number of targets on screen and greater spacing between targets typically lead to an increase in relative classification performance. In some instances, these increases boosted the information transfer rate of a 9 target array above that reported for a comparative 36 target array. Notably, for the experiments defined herein, the number of targets on screen were far smaller, ranging from just 3 to 7 emoji per design variant. Here it is possible that the relative increase in probability of augmentation per emoji for the 3-Emoji (33.33%), as compared to the 5 (20%) and 7 (14.3%) contributed to a decrease in P300 signal quality. This suggestion is however not likely after observing the differences between the variants for the Cz grand average plots (see, Figures 4.21, 4.22, 4.23). Here the 3-Emoji variants present with arguably the most prevalent waveform features characteristics of a visual oddball paradigm, with a strong negative component at 200ms and a large positive component at 300ms.

The primary difference between the 3-Emoji, as compared to the 5 and 7-Emoji variants, as seen in Tables 4.14, 4.15 and 4.16 relates to the quantity of Non-Target samples available to the respective LDA classifiers, as well as the ratio of synthetic to real data in the training samples, as afforded by the SMOTE over-sampling approach. Here roughly, 60% of all train samples were synthetic as compared to around 80% and 85% for the other respective methods. Based on this it could be asserted that the relative degree of variance within the oversampled Target-P300 training data for this method is higher, given the relatively lower incidence of linearly interpolated synthetic samples. Further, the average mean classification accuracy for the 3-Emoji method (59%) relating to the Non-Collapsed Non-Target samples is also lower than the 5 (65%) and 7-Emoji variants (70%). This value decreased following the Collapse data preparation method from 59% to just 54%, primarily driven by the results collected from Subject 1. Here a pattern emerges where the relative number of Non-Target samples available to the classifier also appears to influence the performance of the LDA models.

Notably, when comparing the results to the original Pipeline 1 implementation (see, Table 4.3), both Subject 1 and 5 dropped in relative overall classification accuracy, by 6.78% and 8.89% respectively when compared with the results of the Non-Collapased data partition (see Table 4.9). Further, when comparing the relative performance of these models against the metrics collected for the Localizer integrated variant of the Pipeline 1 method (see subsection 4.4.3.1) the Pipeline 2 approach discussed here resulted in a dramatic relative increase in performance for Subjects 1 (4.33%), 3 (46.78%) and 5 (49.75%). Crucially, it must be stated that these Pipeline 1 results are standalone stats and did not go through the same 10-fold cross-validation procedure.

Despite this, the data pre-processing methods outlined in the Pipeline 2 approach can not be said to have improved the classification performance in comparison to the results collected via the Pipeline 1 method. Concerning the Collapsed data partition (see, Table 4.18), Subject 5 demonstrated a substantial increase in Overall mean classification as compared to the Non-Collapsed condition. It is difficult to draw conclusions from a single anomalous result, however, it does appear, at least for some data, that the process of aggregating neighbouring trial samples to simulate the collection of 10 sequences per trial, as opposed to 5, can boost classification accuracies irrespective of the contingent drop in the number of training and test samples.

4.6.7.2 5-Emoji Results: Pipeline 2

The 5-Emoji subject-level results across both the Non-Collapsed and Collapsed data preparation methods all provided mean classification accuracies significantly above the 50% chance level (see, Tables 4.19 & 4.20). This is a dramatic improvement compared to the results collected via the Pipeline 1 approach for standard (see, Table 4.5) and Localizer-integrated method (see, Table 4.11). All included subjects demonstrated substantial (50%+) differences in Target *vs*. Non-Target sample class accuracies. Further, the group-level results revealed no significant differences between the Non-Collapsed and Collapsed data partitions via a paired-subject permutation assessment for any accuracy metrics. This is despite a substantial increase in the Overall (6.3%), Non-Target (5.3%) and Target (8.3%) metrics, with the latter leading to near maximal pooled-subject predictions for all three subjects (Avg.=99.33%). The author notes that this must be interpreted with caution, as seen in the corresponding data partition table (see Appendix Table A.6), the number of P300 target test samples available to the classifier for each of the 10 cross-validation folds is just 1. This is due to the aggregation of sequences across neighbouring trials in an effort to boost the data signal-to-noise ratio.

Notably, these samples presented with the highest incidence of stimulus augmentation-induced SSVEP artefacts for both the Target and Non-Target samples (see, Figure 4.22). Again, here the efforts undertaken to effectively pre-process the data via targeted notch filtering and baselining were not sufficient to nullify the presence of potentially confounding EEG artefacts. Given the higher relative quality of the plot generated for the 3-Emoji variant (see, Figure 4.21) data quantity is likely not a contributing factor here. Further, the significant drifting components suggest that the erroneous decision to perform a non-continuous data acquisition method has prevented the author from effectively removing DC drift components from the signal.

Despite the clear issues in data quality, the 5-Emoji variant produced some of the highest Target-P300 mean classification accuracies for this experimental series. The Target-P300 data windows may contain a strong SSVEP component with a higher relative coherence in phase, this signal is also likely to present in the Non-Target samples, however, given these are averaged across different time points for different emoji the phasic coherence is likely lower, leading to broad attenuation of the signal. In sum, the corresponding LDA model discriminative functions may have utilized the relative incidence of the unwanted 6.67Hz SSVEP induced via the 150ms stimulus onset interval in addition to the P300 waveform component when positioning the classes in the search space. Unfortunately, it is difficult from these plots come to any precise conclusions.

4.6.7.3 7-Emoji Results: Pipeline 2

The results relating to the 7-Emoji variant for the Non-Collapsed and Collapsed data preparation all returned mean classification accuracies above the chance level of 50%. As can be seen in Figures 4.24 and 4.25, the 7-Emoji variant demonstrates the highest overall accuracies for both Subjects 3 and 5 compared to both the 3-Emoji and 5-Emoji stimulus presentation variants. This could be due to the relative lower probability of augmentation per emoji on-screen as noted above (see subsection 4.6.7.1). Following the application of the Collapse data partition method all subjects Overall mean classification accuracies increased well above the 70% functional usage threshold (see, Tables 4.21 & 4.22). Here, the group-level stats did not reveal any significant differences for the paired subject results concerning any of the accuracy metrics assessed. Worryingly, these results demonstrate a strong indication of overfitting towards the Target class.

As seen in Subject 1, the mean accuracy for Non-Target samples is 64% with a large standard deviation (0.19) suggesting that several of the cross-validation *k*-folds reported random performance-level accuracy metrics. This in tandem with the near maximal mean classification accuracy of the Target class suggests that up to 35-40% of all sample misclassifications involved the incorrect identification of Non-Target as Target samples. Note, that the relative increase in the detection of Target samples did not lead to a drop in Non-Target mean classification accuracies.

Notably, the 7-Emoji variant is reported as the highest-performing experimental iteration detailed in this chapter, however, it is also the variant with the highest proportion of synthetic Target-P300 samples. As seen in Table 4.16, around 85% of all training samples were generated via the SMOTE oversampling method. This suggests that it is likely the associated P300 training dataset was populated with extremely similar signals despite efforts made to enforce the randomisation of interpolation power over 5 different nearest neighbours (see subsection 3.3.5.3). The high degree of homogeneity for this minority class could have afforded it a significant advantage during the training of the respective LDA models. The large imbalances in class predictions can ultimately reduce the usability of any BCI communication system due to the higher relative incidence of False positives, requiring users to correct these mistakes, adding additional time per communication received also decreasing user confidence in the system.

4.6.7.4 Summary: Pipeline 2

The intention of these investigations was to probe the relative influence of the the number of emoji on screen. This was done to discern whether a decrease in potential P300 peaking components could be offset by a reduction in the incidence of spatial bleedover effects. Along these very same lines, the distance between the number of stimuli increased as the number of stimuli decreased. Following the Pipeline 1 approach, the author was effectively unable to answer these questions owing to the confounding influence of model overfit. The author asserts that the low number of sequencs per trial lead to an extremely weak signal to noise to ratio. A far higher number of sequences, around 10 or 15 is necessary to effectively boost the trial-level P300 averages for accurate discrimination against Non-P300 data segements.

For the Pipeline 1 approach, data quality evaluation involved a series of fundamental assessments. Notably, no errors occurred during data loading, all metadata remained intact, and there were no instances of long stretches with NaNs or zeros, missing channels, or inconsistent sampling rates. This suggests that data packets were acquired regularly from the headset to the receiver. Visual inspection revealed no discontinuities, aside from issues related to the Pz sensor (see, Figure 3.6). As mentioned earlier, the author only had access to the medium-sized Cognionics Quick-20 system, leading to challenges in positioning the headset using cranial landmarks like the inion and nasion due to variations in head size among subjects.

Consequently, the analysis could not rely solely on the Cz electrode, as this would introduce substantial spatial positioning variance. Instead, a broader sampling approach was taken, using electrodes Fz, Cz, Pz, P3, P4, O1, O2, A1, and A2 (see, Figure 3.4). However, the Pz electrode frequently exhibited significant spiking events due to improper seating in several subjects. This was particularly problematic, as the Pz region is critical for capturing P300 waveforms. The removal of this channel in most cases due to amplitude-based rejection resulted in the loss of vital data for analysis (see subsection 3.3.5.2). To enhance the prevalence of P300 waveform features, the Pipeline 2 approach implemented corrections to the pre-processing methodology, specifically improving baselining and filtering techniques (see subsection 3.3.5.3).

The Localizer task was implemented with three primary objectives: to evaluate subject-specific P300 waveforms before the main experiment, to acclimate subjects to similar stimuli, and to utilize the data as pre-training samples for the associated LDA models, serving as a form of regularization before training with the main experimental data. However, each of these objectives was undermined by inherent flaws in the Localizer task design. First, the

decision to use a single target set to augment in 50% of the trials was not methodologically sound, given the inverse relationship between P300 quality and oddball probability. While this was intended to enforce an innate 1:1 class balance and reduce the impact of spatial bleed over artefacts, the 50% probability is too high to effectively generate a robust P300 signal, thereby failing to serve as a reliable pre-screening tool. Second, the rationale of using the Localizer to train subjects on similar data is flawed; a more effective approach would be to use a shorter version of the main experiment, allowing subjects to train on the exact task they would ultimately be assessed on.

Consequently, the premise of using the Localizer task data as class-balanced pre-training samples is fundamentally flawed, as the resulting P300 waveforms are not distinct enough from the Non-P300 data (see Figure 4.14). This is evidenced by the reduction in classification performance observed in Tables 4.9, 4.11, and 4.13, compared to the original implementation (see Tables 4.3, 4.5, and 4.7). The author acknowledges these methodological issues and admits that the Localizer task was erroneously implemented based on the mistaken assumption that the subjective probability of P300 events would be sufficient to induce a strong P300 waveform.

Overall, the Pipeline 2 system has addressed the principle issues of extreme overfit and low classification accuracies demonstrated in the same subject samples using Pipeline 1 for the standard and Localizer-integrated approaches (see, Tables 4.5, 4.7, 4.11 & 4.13). In the comparisons of all 3 stimulus variants for the Pipeline 2 approach, the series of paired-subject permutation assessments revealed no significant differences between the conditions (see subsection 4.5.8). This is likely owing to the significant degree of overlap in the Overall classification accuracies computed via respective 10-fold cross-validation results across the 3 subjects tested (see, Figures 4.24 & 4.25). The high degree of variance within subjects for each experimental variant ultimately suggests that an increase in the number of trials is key to boosting the relative signal-to-noise ratios of respective training and test datasets.

Further, as shown in the figures noted above a marginal increase in Overall classification accuracy is reported for the Collapsed data partition. This was undertaken principally to simulate the collection of 10 as opposed to the original 5 sequences per trial. Despite this, as noted both in the respective stimulus conclusion subsections, none of these effects were significant. It is possible that increasing the number of sequences per trial to 15 would have ultimately improved the resulting data quality, however, the limited samples available in these offline analyses prevented the author from assessing these in a similar simulated manner. Further,

the results relating to the Collapsed data preparation method must be interpreted cautiously given the extremely small number of test samples used in the evaluation of each model. Incorporating all these findings the author can only tentatively point towards a trend of increased performance as a function of reduced augmentaion probability per emoji. This is however qualified by acknowledging that the improvements observed also follow a similar orthogonal pattern where increased accuracy is related to a corresponding increase in the number of synthetically generated samples in the corresponding training datasets.

4.6.8 Reflections

Here, the author discusses the key findings of Experiment 2 relating to the Pipeline 1 and Pipeline 2 data organisation, pre-processing and analysis approaches. Further, a series of considerations regarding experimental issues and several paradigm improvements to address these obstacles are discussed.

4.6.8.1 Stimulus and Data Collection Adaptations

It is important to evaluate the influence of adaptations made to stimulus presentation and data acquisition hyper-parameters on resulting data quality. The introduction of increased inter-augmentation intervals and larger data windows appears to have had some influence on the quality of data collected. When inspecting the pooled-subject average plots for the 3 and 7 Emoji datasets, (see, Figures 4.6 and 4.12), large positive deflections are observed between 300 and 500ms after the onset of the time-locked augmentation events. The extension of the data window to capture delayed P300 peaking events potentially contributed to the increased separability of P300 and Non-P300 target classes for the respective variant analyses. It must be noted, in the 5 Emoji pooled-subject plot (see, Figure 4.9), a relative increase in μV amplitude for this target time window is not present. Interestingly, this aforementioned variant shows the lowest cross-variant performance for both pooled-subject and single-subject datasets in the non-pre-trained and pre-trained LDA model analyses sets (see, Tables, 4.4, 4.5, 4.10 & 4.11).

Concerning the changes to stimulus hyper-parameters, namely the increase in emoji diameter target size (18mm to 27mm diameter), these adaptations had little effect on the quality of signals collected. The prevalence of atypical oscillatory trends, poor baselining and specifically the volatile components positioned in the Non-P300 plots suggests that the measures taken to reduce the incidence of spatial and temporal bleed-over effects were not adequately minimised. Note, that many of the same non-characteristic waveform features are present in single-subject average signal plots not displayed herein. Ultimately, the presence of these

atypical waveform features is ever-present in EEG data collection owing to the non-stationary properties of the signals. The only feasible means of diminishing the presence in signal averages is to collect more samples to balance the noise across multiple trials. It is possible that the amount of sequences collected per trial was not sufficient to achieve to stated experimental goals.

4.6.8.2 Classification Performance and Array Density: Pipeline 1

As predicted at the onset of this experimental series, from the results of the Pipeline 1 approach a decrease in array density from 7 to 3 Emojis appears to have improved single-subject classification accuracies. The highest performing subset that achieved greater than random performance subject averages for both the P300 (60.00%) and Non-P300 (73.33%) classes was shown to be the non-pre-trained 3 Emoji variant (see, Table 4.3). Notably, as these results were collected via the Pipeline 1 approach, none of the findings are cross-validated and therefore caution must be taken in their interpretation.

The difference in performance across the variants assessed could be attributed to a reduction in task difficulty and the increase in balance between P300 and Non-P300 sample ratios. This is seen clearly by the positive relationship between overfitting incidence and array density across Tables 4.3, 4.5 and 4.7. Further, the task length differed significantly over the stimulus variants, meaning fatigue owing to experimental duration could have introduced more low-quality signals into the training and evaluation datasets. Despite this, the only viable means of assessing overall dataset quality are the pooled-subject average plots. These reveal the 7 Emoji data (see, Figure 4.12) present with significantly more P300 (solid line) and Non-P300 (dashed line) characteristic features than those shown in the respective 5 Emoji plots (see, Figure 4.9). It is clear that had a relationship between data quality and task length been present the increased quality of signals for the 7 Emoji variant would not be present. Further, as is noted at length in subsection 4.6.7.4, a complete reversal in this trend is observed for the Pipeline 2 approach results.

4.6.8.3 Localizer Data Pre-Training Considerations: Pipeline 1

The implementation of the localizer data pre-training stage was introduced primarily to address this issue of overfitting. As mentioned above, the increase in array density necessarily increases the degree of overfitting due to the ever-increasing imbalance in P300 and Non-P300 samples. In Experiment 1, the effect of overfitting was addressed using a class-balancing protocol. This process was implemented primarily as a means of studying the stimuli capacity

for P300 generation and LDA data separation without the confounding influences of overfitting on the resulting classification metrics. The author acknowledges the severe limitations of any interpretations based on these class-balanced results due to the marked decrease in ecological validity that must be employed. Additionally, these processes significantly diminish the volume of data used in LDA model training and the amount of Non-P300 samples used is restricted by the number of P300 events available. For these reasons, the author attempted to mitigate the overfitting due to class imbalances by pre-training the models with the data gathered during the localizer task conducted before each experimental variant. The implementation of these data in a pre-training stage was ultimately misguided given the high oddball stimulus probability of the associated Localizer task emoji augmentations., this was discussed earlier in greater detail (see subsection 4.6.7.4 Summary: Pipeline 2).

4.6.8.4 Increased Incidence of Model Overfitting in 3 and 5 Emoji Variants after Localizer Data Pre-Training: Pipeline 1

The influence of the localizer pre-training stage as compared to the initial assessments across all variants is highly varied, as mentioned in the corresponding conclusion subsections above (4.5.4 & 4.5.5). Crucially, it must be noted that on the whole, the incidence of overfitting increased for emoji variants 3 and 5 (see, Tables 4.8, 4.9, 4.10 & 4.11). The only marked decrease in this phenomenon is observed for the 7 Emoji variant (see, Tables 4.12 & 4.13). This could be due to the ratio of localizer trials to main experiment trials. The pre-training for all models at the single-subject level was conducted using 20 non-averaged and class-balanced localizer trials. For the 3, 5 and 7 Emoj variants the ratio of pre-training localizer trials to main experimental data samples was 1:4.5, 1:7.5 and 1:10.5 respectively. Note, that these ratios also apply to the pooled-subject datasets. The higher proportion of non-averaged localizer trials in the training scheme for the variants with lower visual density could have led to the resulting models being primed with data representations for P300 and Non-P300 waveforms that did not accurately cohere to the subsequent signals used for training from the main experimental dataset. Further the large difference in the oddball probability between the localizer task and the main experiment could have primed the LDA models for class-wise data representations that diverged from samples in the main experimental data in terms of P300 peak amplitudes and latencies.

Moreover, the reduction in overfitting incidence does not implicitly increase classification accuracies, as shown in many of the corresponding 7 Emoji variant LDA models trained that fail to effectively separate the classes evaluated (see, Table 4.13). Interestingly, the largest single-subject increase in classification performance after pre-training was observed for Sub-

ject 4 in the 3 Emoji variant. This subject showed a mean accuracy increase from 66.67% to 88.89% attributed principally to a significant increase in P300 classification performance. Ultimately, these analyses do not suggest the localizer data pre-training had a significant positive influence in boosting LDA model classification performance and in many instances, specifically concerning the 3 and 5 Emoji variants, the pre-training stage proved detrimental to the accurate prediction of P300 and Non-P300 target classes.

Note, that all other methods of pre-processing performed on the main experimental data were repeated for the localizer dataset excluding the intra-trial signal averaging in accordance with the original Pipeline 1 approach (see subsections 3.3.5.1 & 4.3.5.4). This was not conducted as the number of samples available for the pre-training stage was too low (20 trials). Had the same signal-averaging procedure comprising 5 separate data chunks been reimplemented here the author would have had just 4 trials for the class-balanced pre-training stage. This volume of trials would have a negligible effect on the resulting classifiers. It is likely the localizer data possessed substantially higher ranges and additional noisy non-stationary components as compared to the main data. It is reasonable to assume that this is the primary reason for the poor performance of models treated with the pre-training process.

In the subsequent localization task, the total number of trials was increased and the corresponding data was organised into class-based blocks of 5 waveforms according to their chronological order. This way a pseudo-sequenced localization task with just 1 emoji could be conducted. Arguably, an offline version of the main experiment as the localization task would have proven more effective. Importantly, the use of localizer task data was an apriori decision implemented by the author after the conclusion of the data collection period. The absence of signal averaging, the relatively small amount of pre-training trials used and primarily the extremely high oddball stimulus probability of the Localizer task contributed to the failure of these methods in this experimental iteration. These issues are addressed by adaptations implemented in Experiment 3 via the Pipeline 2 approach (see subsection 5.4.10).

4.6.8.5 Pipeline 2 Results Reflections

The results from the Pipeline 2 approach involving alternative pre-processing methods, SMOTE oversampling to address class balancing and a 10-fold cross-validation procedure to determine LDA model performance revealed a nearly opposite trend to those reported for the Pipeline 1 approach. Most notably, the initial trend observed for higher performance with lower density arrays (3 *vs*. 7-Emoji) was reversed here (see, Tables 4.2, 4.3, 4.6 & 4.7) for both the Non-Collapsed and Collapsed data partitions (see Tables 4.17, 4.18, 4.21 & 4.22). Importantly,

these are only trends, as none of the group-level stats comparing the different emoji stimulus variants revealed any significant differences. The author asserts that had more subjects been included in the analysis, this would likely have been demonstrated given the strong directionality of the results.

Despite this, even if such a result was reported the relative influence of synthetic data samples would not have been illuminated. It is likely that given that the 7-Emoji variant featured the highest proportion of synthetic data, the degree of inter-class cohesion for the Target-P300 samples would have made the separation of these samples easier to discern from the Non-Target data. It is important to state, that the evaluation of all LDA models was undertaken using an isolated subset of samples during the cross-validation procedure to mitigate the potential of overfitting. Based on this, it is difficult to conclude that the increase in classification accuracies observed for the higher-density stimulus arrays is wholly related to the higher degree of oversampled data in the respective training sets.

In contrast, it is possible that the lower relative probability of augmentation for the higher density arrays, 5-Emoji=20% and 7-Emoji=14.3%, could have contributed to the production of consistently higher quality P300 waveforms. These assertions however are not borne out from the corresponding Cz grand averages (see, Figures 4.21, 4.22 & 4.23). Here, the 3-Emoji variant samples present with arguably the most representative Target-P300 average signal of any of the stimulus variants tested. It is difficult to effectively evaluate the quality of the 5 and 7-Emoji variants due to the substantial presence of confounding SSVEP noise components induced via the presentation scheme augmentation intervals.

In future assessments, it is recommended that a series of notch filters are performed around a mean of the predicted SSVEP artefact frequency to more effectively nullify this component. Further, significant drifting artefacts are present indicating that the compromised baselining method involving the averaging of the 1st 50ms of each segment did not effectively centre the samples around a common reference point near zero. This dramatically decreased the ability of the author to evaluate the quality of the oddball-associated N200 and P300 components. Additionally, it must be stated that as the number of targets on screen increases and the spacing between targets decreases the incidence of adjacent error and subject distraction could have contributed to the relative drop in grand average signal quality, as compared to the 3-Emoji dataset.

Finally, the impact of the Collapsing procedure on the corresponding mean classification accu-

racies was inconclusive. The group-level comparisons revealed no significant differences due to the small 3-subject sample size, however, a clear trend for increased Target classification accuracies was revealed. These occurred without a corresponding increase in Non-Target sample accuracies. Effectively, the associated LDA model results demonstrated a higher degree of volatility within the Overall target accuracies following the implementation of the cross-trial sequence averaging to simulate more sequences per trial.

This is likely owing to the extremely small number of Target-P300 samples in the evaluation set (see Appendix Tables A.5-A.7). Had the Collapsed data preparation method not diminished the overall sample size for training and testing to such a degree the author asserts that this process would likely have proven effective in boosting classification accuracies by increasing the relative sample sign-to-noise ratios.

In sum, given the numerous qualifications for the interpretation of these data stated above it is difficult for the authors to definitively assert that a higher number of emoji on-screen ultimately leads to an increase in the quality of respective Target-P300 forms. Despite this, via the application of the Pipeline 2 approach, classification accuracies have been consistently boosted to levels significantly above chance for all subjects tested. Further incidences of confounding and extreme Non-Target class bias have been nullified and a strong, data-driven trend has been demonstrated to indicate that a higher number of targets is preferrable for this emoji-based BCI communication system. Ultimately, any future replication of these experiments must involve the implementation of a continuous sampling method, the collection of 10 sequences per trial and an increased number of trials per subject tested.

4.6.8.6 Summary

Overall, the implementation described using the Pipeline 1 approach initially led the author to believe that the results indicated agreement with previous literature suggesting a decrease in stimulus targets and an increase in stimulus separation. This was principally based on the performance of the 3-Emoji classifier with no localizer data pre-training (see Table 4.3). However, closer inspection of the Cz grand average plots revealed that the 7-Emoji data presented with the best P300 waveform features (see Figures 4.6 & 4.12). In the subsequent re-analysis undertaken via the Pipeline 2 method, the opposing trend emerged, whereby the highest density, 7-Emoji array, produced the best classification results (see Tables 4.14-4.16).

Notably, these data were handled with greater care in terms of pre-processing and also featured the application of a 10-fold cross-validation procedure (see subsection 3.3.5.3). Given

these alternate, more concrete findings the authors assert that these results indicate agreement with the basic principle of the oddball paradigm, whereby a higher number of targets following a randomised presentation scheme improves P300 signal generation due to a decrease in oddball stimulus probability. Along these very same lines, as noted in subsection 4.6.7.4, the intentions of implementing the localizer task as a training tool, pre-screening method and LDA model tuning to address class imbalance issues were ultimately flawed. This is owing to the high oddball probability of the task and the associated weak P300 signals (see Figure 4.14).

Chapter 5

Experiment 3: Real-Time Feedback Implementation

5.1 Aims

Experiment 3 aims to address the issues identified throughout the previous experiments conducted relating to overfitting, subject attentiveness and data quality. These areas are tackled by introducing an expanded localizer task for subject-specific model pre-training, real-time model prediction subject feedback and an active intra-experimental impedance monitoring method. Ultimately, it is intended that these adaptations increase model classification accuracies above functional percentage limits (70%) across all subjects assessed. As noted in the previous chapter the motivation to implement the localizer task as a means of pre-training LDA models questionnable given the high associated stimulus oddball probability. The merits and flaws of all decision undertaken in this chapter are discussed at length in the associated conclusions and reflections (see subsections 4.6 & 4.68)

5.2 7 Emoji Variant Selection Rationale

As discussed in subsection 4.6.8.2 the 3 Emoji variant produced the highest levels of classification accuracy in both the non-pre-trained and pre-trained LDA model implementations for the Pipeline 1 implementation. Note, that it is possible to utilize low-resolution (low number of targets) systems in specific clinical applications where patient end-point users present with increased sensitivity to intense stimuli and high susceptibility to fatigue. Further, the information transfer rate limitations placed on alphanumeric speller systems are somewhat mitigated during the use of pictorial-emoji style spellers, as the selection of a single target icon could

save the user significant time communicating the same intentions using traditional text characters.

Despite these considerations, after observing the dramatic changes in performance in the 7 Emoji variant following the localizer data pre-training scheme and the higher P300 signal quality (see Figures 4.6 & 4.12), the author decided to move forward with this stimulus design. Primarily this is owing to the increased functional capabilities of the higher-density system. Further, the use of a 7-emoji array design ensures more data is collected per trial and reduces the limitations placed on LDA models trained with low-volume datasets. Moreover, it is predicted that the inclusion of real-time feedback on LDA model predictions relayed to users on-screen during the task would heighten engagement and minimize the incidence of subject fatigue. As stated above, on reflection is is clear to the author that the use of a localizer task with a high oddball stimulus probability as LDA model pre-training stage is suboptimal owing to the weak P300 signals produced (see Figure 4.4). These considerations where only reached following the later implementation of the Pipeline 2 approach. Along these very same lines, all Pipeline 2 results noted herein utilize data exclusively from the main experiment.

5.3 COVID-19 Pandemic Comments

As referenced in the introduction of the thesis, this stage of the data collection process was significantly hampered by the COVID-19 pandemic restrictions observed across all university sites. The limitations imposed on in-person experimentation led to a necessary reduction in the projected sample size of 10 down to the 3 subjects presented herein. For further information on the impacts of the thesis from this point in time through to the following experimental series refer back to subsection 1.1.

5.4 Method

Here are outlined the methods employed in the investigations relating to Experiment 3. Broadly, this features the implementation of a 7-target emoji speller experiment featuring real-time classification and user prediction feedback (see, Figure 4.2). This also comprises an extended localizer pre-screening and data acquisition phase alongside the deployment of an online impedance monitoring system.

5.4.1 Participants

A total of 3 neuro-typical subjects were recruited from the Durham University student population consisting of 1 male and 2 females (mean age of 27.7 years and age range of 23-33 years). All subjects sampled were pre-screened to ensure all presented with normal or corrected to normal vision, had no history of clinical mental illness or epilepsy and were not currently experiencing a skin-based ailment of the scalp. No subject received payment to participate in the experiment. Ethical approval and oversight were granted by the Durham University Psychology Department Ethics Sub Committee.

5.4.2 Equipment

All EEG data acquired was collected using the Cognioincs Quick-20 headset (Cognionics, San Diego, USA). The concurrent sampling of amplitude (μV) and impedance (Ω) values were controlled using the LabStreamingLayer Python library (pylsl) [224]. The visual experimental stimuli were rendered using a dedicated NVIDIA GTX 750ti GPU (2GB VRAM). Before and after testing all EEG sensors and the headset housing was thoroughly cleaned with anti-bacterial gel. The stimulus array was presented to participants while seated at a desk (0.8m from central head position) via a Samsung LED computer monitor (Model: S27A35OH, 60 Hz refresh rate, 68.5cm diameter).

5.4.3 Stimulus Presentation

The stimuli used in the experiments detailed herein were developed and operated via the PsychoPy Python library [221]. All emoji stimulus targets were populated by the OpenMoji dataset, an open-source emoji repository developed from the outset for free use [222]. A total of 7 emoji were utilized from this repository alongside 7 visually modified versions of the very same emoji, forming the augmentation versions of said stimuli. A further 2 pictographic icons were also utilized from this repo, a thumbs up (emoji tag: 1F44D) and thumbs down (emoji tag: 1F44E). These were utilized to display visual feedback of the real-time LDA model classification performance to the subjects. Note, that both were colourized, thumbs up (green) and thumbs down (red), to enhance the readiness of information relayed to the subjects.

5.4.4 Localizer Task

The localizer task employed largely replicated the methodology implemented in Experiment 2 (see subsection 4.3.4). The task was substantially expanded to include more trials as per the recommendations cited in the previous summary subsection 4.6.8.6. All subjects were

instructed to attend and fixate on a single emoji presented centrally on the aforementioned presentation monitor (see, Figure 5.1). Initially, a 1-second delay period was employed, following this, the on-screen target emoji was either augmented (augmentation duration 0.05 seconds) via the inversion of all black emoji colouration to white, or no change in the appearance of the target emoji was initiated. The randomization protocol controlling the order of augmentations was computed before the onset of each localizer session according to a stratified procedure. This prevented large clusters of augmentations and non-augmentation trials throughout the experimental period. Following this stage of the experiment, an inter-trial interval of 1 second was observed to reduce the confounding influence of temporal bleed-over effects.

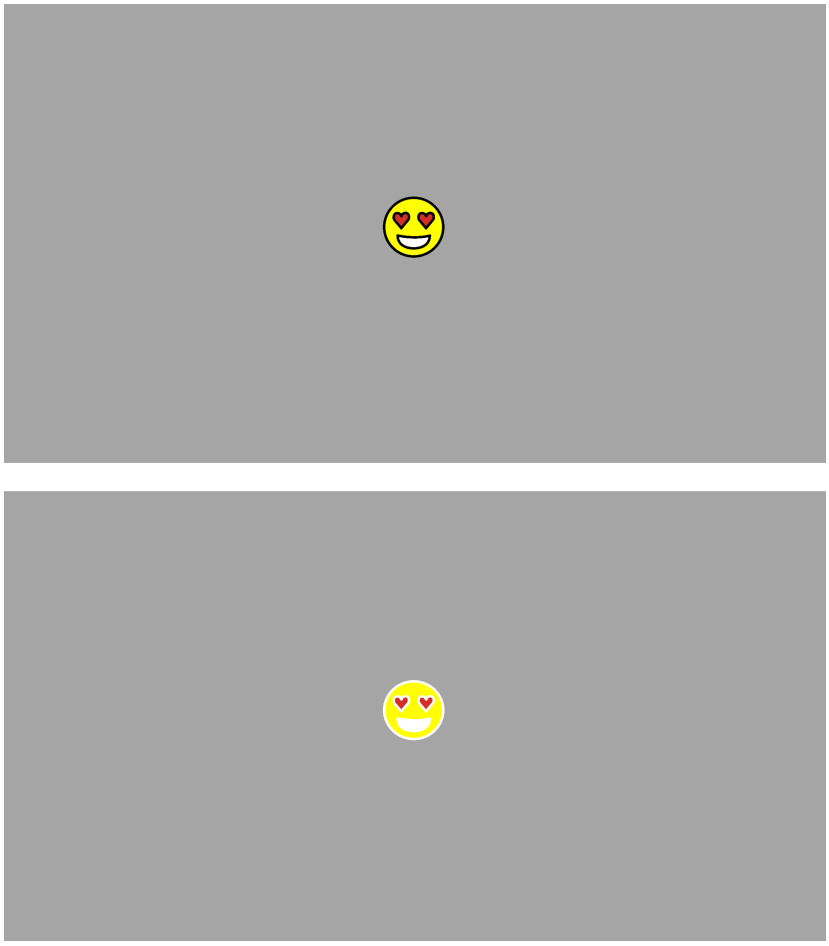

Figure 5.1: Here is displayed a dual figure showing the appearance of targets on-screen during the localizer task. The upper image displays the non-augmented localizer emoji state and the lower image shows the augmented localizer emoji state. This second image demonstrates the Inversion method of augmentation and involves modifying all black colourations of the emoji stimulus to white for 0.05ms. Note, that all stimulus sizing parameters remain consistent with the localizer task outlined in Experiment 2.

A total of 150 events, 75 Non-P300 (non-augmented) and 75 P300 events (augmented) were collected. One trial amounts to ((1s + 1s) × 75)+((1s + 0.05s + 1s) ×75) = 303.75 seconds, leading to an experimental duration of 5.06 minutes. Here the 1-second units relate to the initial trial rest period and post-trial delay stage. Further, the 0.05 second unit relates to the stimulus augmentation duration that occurs in all P300 trials. These data were treated with the same preprocessing pipeline, Pipeline 1, outlined in earlier experiments (see subsections 3.3.5.1 & 5.4.8). The event data from this variant were chronologically organized into groups of 5, all comprising either augmented (P300) or non-augmented (Non-P300) trials and averaged together, as per the standard oddball paradigm procedure. This resulted in a total of 30 trials from the original 150 events sampled.

These modifications were employed to enhance the similarity in characteristics between signals collected in the localizer and the main experiment. In the previous experiment, no signal averaging was done on the localizer data at the trial level as it was initially only intended for screening and visualisation purposes. In this instance, parity in the treatment of the localizer and main datasets was employed as the evaluation of the latter is, in some stages of the proceeding analysis, dependent on the former. The difference between averaged and non-averaged data is clear, with increased non-stationary components such as time-series drift, the heightened presence of movement artefacts and generally greater variance in the signal overall.

Further, these changes were made to address scaling discrepancies between the localizer and main experiment training data, as it was thought that having a large difference in scale between the localizer initialization training data and main experiment training data could mitigate effective LDA model training. Crucially, subjects were not presented with visual feedback at this point in the experimental procedure as presently no data had been collected to train a classifier capable of performing said predictions. Note, that subjects were familiarized with the feedback mechanism via multiple offline demonstrations of the main experiment.

As previously noted in subsection 4.3.4, originally the author believed that the subjective perception of stimulus augmentation probability would be lower than the actual 50% augmentation probability. This was due to the erroneous assumption that given the default non-augmented emoji stimulus was onscreen for the vast majority of the localizer task period and that the augmentation order was randomized and spread out over the trial period to avoid large clusters of non-augmentation and augmentation events, the perceived likelihood of stimulus change would be perceptually lower than the given 1:1 ratio. As will be noted in the corresponding results, conclusions and results sections this assumption was not borne out by

the analysis despite some indications of improved classification performance for this analysis variant (see all subsections relating to the LOCRT analysis variant).

This is principally owing to the inverse relationship between stimulus probability and P300 peaking and latency quality (for reference see Figure 2.1). The method of utilizing these Localizer data as LDA model training samples to ultimately predict the respective class of samples in the online main experiment was performed exclusively using the Pipeline 1 data organisation, pre-processing and analysis approach. All subsequent revisions to this method undertaken in the Pipeline 2 method involved the offline analysis comprising training and test samples collected using the main 7-Emoji experimental data described in the following subsection.

5.4.5 Main Experiment

In this final iteration of the main experiment, 7 emoji stimuli are presented across the horizontal centre of the presentation monitor (see, Figure 5.2 upper image). After a 1-second delay period, a white cueing square is presented indicating the target emoji the subject is to fixate on and attend to throughout the trial period (see, Figure 5.2 middle image). Following this, each of the 7 Emoji was augmented according to a non-consecutive randomisation protocol (augmentation duration = 0.05ms). This involves a modification of each emoji on-screen according to a specific time-locked program design to reduce the incidence of spatial bleed-over effects by preventing the consecutive augmentation of neighbouring stimulus targets. Once all emojis have been augmented a 500ms inter-sequence interval is observed. The process, as outlined, is repeated for a total of 5 sequences per trial.

Immediately following the conclusion of the final trial sequence, a 1-second inter-trial interval is employed. Finally, the real-time classification of the captured trial data is performed and a feedback icon is presented to the subject in the top-right corner of the presentation monitor (see, Figure 5.2 lower image). In the event the LDA classifier correctly identifies the target emoji a green thumbs-up is presented, conversely, a red thumbs-down icon is presented in the event any other target aside from the cued emoji is incorrectly identified as the target emoji (total computation time approx. 50ms). The stimulus protocol defined here closely mirrors the method used in the 7 Emoji variant employed in Experiment 2 (see subsection 4.3.5.1). The only differences between the methods in terms of timing relate to the inclusion of the computation duration and the user feedback stage (500ms).This leads to a per trial duration of, 9.8 seconds as per: $1\text{s} + (((0.05\text{s} + 0.1\text{s}) \times 7) + 0.5\text{s}) \times 5) + 1\text{s} + 0.05\text{s}$.

As can be seen in the respective grand average plots (see Figures 5.7, 5.10 & 5.14), a corresponding 6.67Hz SSVEP is observed in the data owing to the 150ms stimulus onset interval. As is discussed in the Pipeline 2 subsection (3.3.5.3), the author made efforts to remove this signal via a targetted notch filter. To clarify this was undertaken exclusively in the Pipeline 2 method following a re-evaluation of the pre-processing methods implemented for Pipeline 1. The author intended to ensure a maximal separability of the Target-P300 and Non-P300 target signals. The application of this filter was intended to remove a potential point of similarity between the classes and thus improve the accuracy of respective LDA classification models.

Further, the author implmented the same increase in stimulus onset interval, 125ms to 150ms, as was applied in Experiment 2. This was done to reduce the incidence of temporal overlap between the corresponding emoji-stimulus augmentation onsets. As per [223], it was reasoned the longer interval should improve the associated P300 peak amplitudes and reduce the latency. In the simplest terms, this minor adjustment was made in an attempt to improve the quality of the resulting EEG data. For additional information on the relationship between stimulus presentation parameters and P300 data quality in BCI context please refer back to subsection 3.3.3.

Initially, a 1-second rest period is implemented. Further, working out from the innermost brackets, 0.05 seconds denotes the augmentation duration and 0.1 seconds relates to the inter-stimulus interval, this is repeated for all 7 Emoji targets. Following this, an inter-sequence interval of 0.5 seconds is enforced and repeated for all 5 sequences making up a single trial. Finally, a 1-second inter-stimulus interval is observed, with 0.05 seconds allowed for real-time computation. A total of 30 trials were collected per subject leading to a summed experimental duration of 4.9 minutes, of which each trial is comprised of 5 underlying stimulus sequences and hence feature 5 signals for averaging per trial or 150 repetitions per emoji per subject. As has been discussed at length, this number of sequences was selected originally to maximise the potential information transfer rate of the final system. This oversight likely hampered the resulting signal-to-noise ratio of associated samples. To simulate the acquisition of 10 sequences the Pipeline 2 approach features the Collpased assessments, here samples across adjacent trials are aggregated to simulate the collection of more sequences per trial.

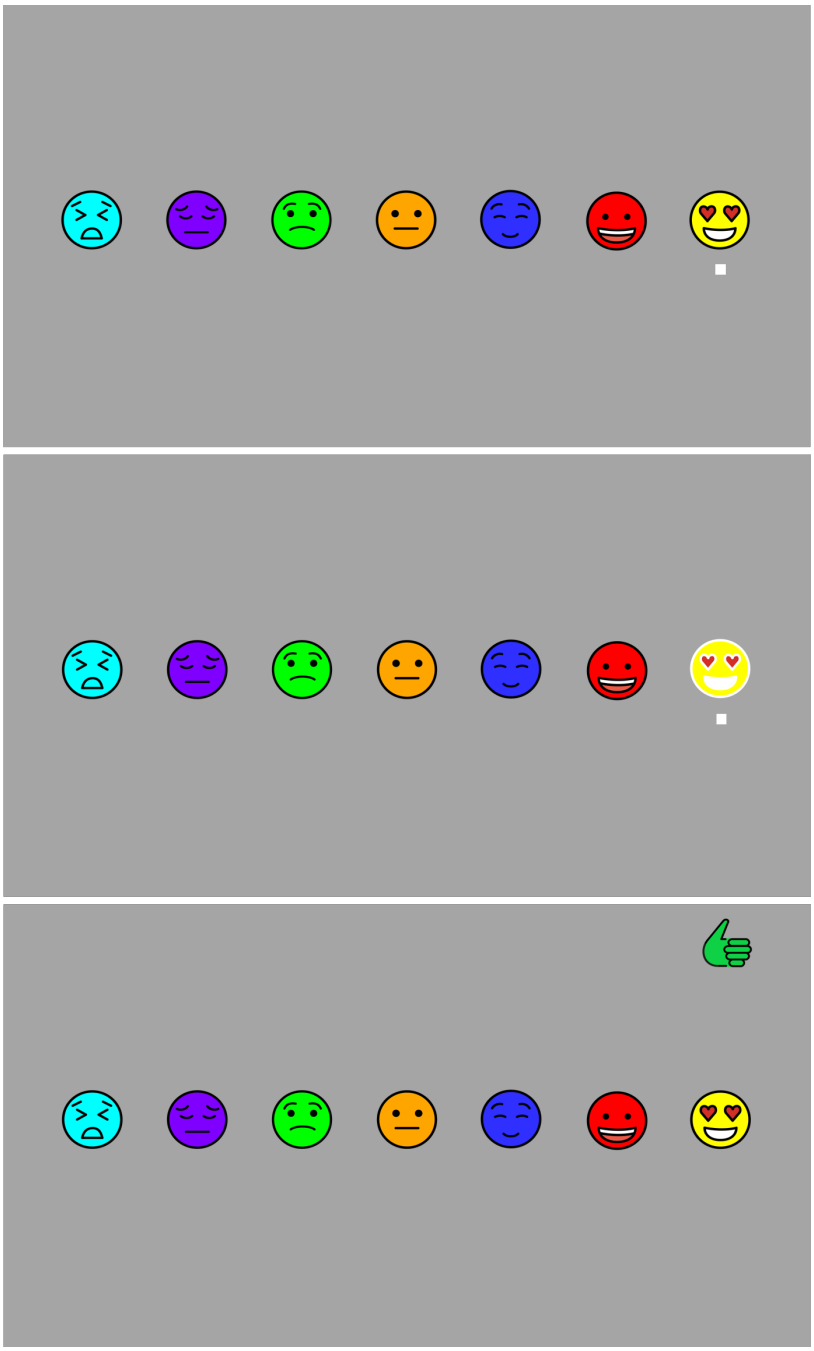

Figure 5.2: Here are presented three screenshots encompassing the primary stages of the real-time main experiment. The upper image shows the initial cue phase. This involves positioning a white cueing square directly below the target emoji subjects to fixate upon throughout the trial period. The second image in the sequence depicts the inversion method augmentation of the target emoji. This form of visual modification is applied to each emoji in the array according to a non-consecutive randomised procedure. The final (lower) image shows the real-time visual feedback presented (green 'Thumbs Up') to subjects after the subject-specific LDA model correctly identifies the target emoji. Note, that in the event of a misclassification, a corresponding red thumbs icon is presented. All of the stimulus parameter sizing is replicated from the 7 Emoji stimulus variant outlined in Experiment 2.

5.4.6 Data Acquisition

The same methods of data collection employed in Experiment 1 were reimplemented here (see subsection 3.3.4). Briefly, these comprise the use of the Cognionics Quick-20 Dry-EEG headset, sampling at 500 Hz across 9 channels (Fz, Cz, Pz, P4, P3, O1 O2, A1 and A2). As discussed above, a localizer task was performed before the main experiment. All subjects were provided with a 2-minute break before the onset of the main experiment to maximise attentiveness. This was not extended further unless explicitly requested by the subject, as due to the non-stationary nature of the EEG signal profile the degree of coherence in the reference signals gathered via the localizer task when compared against those sampled in the main experiment decreased as a function of time. Note, that two sets of data are stored in real-time during the data acquisition phase. Initially, the raw data is saved and stored for offline, pooled-subject assessments. A second set of data is pre-processed on-the-fly, stored and then processed via a real-time analyses pipeline.

5.4.7 Data Organisation

In the post-experimental period, all data (comprising 3 subjects) was processed into numerous distinct groups. These comprise an aggregated dataset including all signals sampled across all subjects (referred to as the Pooled-Subject data) and 3 separate subject-level subsets, built exclusively from individual data (referred to as the Single-Subject data in tandem with an identifying numeric for example, Subject 3). Additional class-balanced datasets were developed (see subsection 5.4.11.3) for the main experimental data. These protocols were not employed at the Localizer data level as the number of P300 and Non-P300 data samples already has a perfect 1:1 ratio.

5.4.8 Data Pre-Processing: Pipeline 1

All data including both the Main experiment and Localizer data were initially organised and pre-processed according to the stages laid in Pipeline 1 (see subsection 3.3.5.3). This same process was also implemented in the main experiment, at the single-subject level in real-time. Note, that all processing and analysis for the main experiment regarding aggregated datasets were performed offline.

5.4.9 Reactive Impedance Monitoring

During both the localizer and main experiment variants, each channel sampled was assessed after every trial to determine the variance in ohms across the given trial period. Initially, the

variance across all channels was calculated, following this the variance at each channel was then evaluated. If a given channel was found to have a variance three times greater than the median cross-channel variance the entire experiment was paused and the trial was repeated once the variance across all channels dropped below this threshold. To the subjects, a pause is indicated by the trial restarting and simply holding the screen at the initiation phase, with all non-augmented emoji onscreen and the standard white cueing square absent from the display. Arguably this could have been further reinforced with an additional onscreen cueing symbol (possibly a 'pause' symbol in the top left corner).

The experimenters overcome this limitation by explaining the mechanics of the system in detail to subjects before the onset of the main experiment, alongside demonstrations and a period for subjects to voice any questions regarding the procedure. Crucially, the same augmentation sequence is not repeated once the trial restarts, this was anticipated in advance by computing 4 times as many non-consecutive randomisation sequences as were necessary to complete the trial if completed flawlessly. This was employed to prevent the loss of data that occurs at the impedance-based channel rejection stage of the pre-processing pipeline. In effect, any data passing through this system should not require extensive additional post-processing for impedance spiking events as this would be completed on the fly.

Active impedance monitoring was also in place to reinforce the necessity to maintain a static head position throughout the trial. Any excessive movement that introduces impedance spiking would lead to the trial pausing and indicate to the subject that they must be seated with a neutral posture. Essentially, the protocol was implemented to ensure data interpretation could be simplified by reducing the possible confounding influence of movement artefacts. As the Cognionics headset is not fixed to the head with a chin strap as per alternative soft-cap-based EEG acquisition devices, this active monitoring also ensured that the experiment was conducted with all electrodes positioned over the correct cranial location.

Note that, during the real-time experimental monitoring of impedance values, the $3\times$ variance threshold was implemented, as stated above. The offline data pre-processing herein used a higher sensitivity $2\times$ median variance threshold. Initially, a $2\times$ threshold was employed during pilot testing and proved too sensitive. This led to subject confusion and enhanced the number of movement artefacts due to subjects communicating with the experimenters during the live main experiment. As discussed below, for the offline analyses undertaken, some trials were still excluded from the analysis in the post-processing using these higher sensitivity parameters. Further, owing to the previous assessments conducted the Pz electrode was excluded

from both the reactive impedance monitoring and experimental data collection. This is due to significant issues in electrode positioning and successful seating of the electrode against the scalp in nearly all subjects tested throughout the experimental series.

5.4.10 Data Pre-Processing: Pipeline 2

Following the initial data pre-processing method outlined in Pipeline 1, all Main experiment data a secondary approach following the stages defined in Pipeline 2 was implemented. These are a separate set of results exclusively computed to address the shortfalls of Pipeline 1. For further information on the differences in these approaches please see subsection 3.3.5.3 and the associated comparison table (see, Table 3.1). Here, given that the Pipeline 2 method was developed following the conclusion of Experiment 3, all analyses here are strictly offline. Notably, at no point are any Localizer samples or LDA models using pre-trained Localizer samples used in the computation of any Pipeline 2 results. The Localizer-integrated pre-training approach was employed originally as a means of addressing the issue of class imbalance for this 7-emoji real-time experimental variant. Here, Pipeline 2 utilizes a SMOTE-based oversampling method to address this issue.

5.4.11 Analysis Variants

The process of analysing this data involved the training of numerous LDA classifiers for each respective pooled-subject and single-subject dataset. These variants are organized into 4 main sections, with a series of results computed in real-time using an LDA classifier trained exclusively with the localizer data (LOCRT). A second set of results for the hybrid LDA classifiers (HYALL) pre-trained on both localizer data and a subset of the data from the main experiment in non-class-balanced and class-balanced formats. Note that, originally an additional data analysis variant was tested exclusively using non-class balanced data from the main experiment. These results are not discussed here as the results effectively mirror the same overfit performance metrics as observed for non-class balanced and non-localizer tuned data seen in Experiment 1 and 2 respectively, (see Appendix A.5). All of the 3 aforementioned analysis variants involved data organisation, pre-processing and analysis methods outlined in Pipeline 1 approach (see subsection 3.3.5.1). A final set of results is computed for the main experimental results exclusively using the Pipeline 2 approach, see below for further details.

5.4.11.1 LOCRT: Pipeline 1

Initially, the data from the localizer task is aggregated, pre-processed and averaged across 5 consecutive trial sequences. This was done in order to mirror the number of sequences utilized in the main experiment and ensure that the localizer data average signals were constructed with the same number of samples. As has been noted at length in the previous chapters, the author's decision to utilize 5 augmentation sequences is likely flawed given previous findings suggesting a minimum of 10 or 15 sequences are required for effective averaging. These data are then collated and used to train an LDA classifier specific to the subject in question. Following this, the main experiment initiates and data across the first 5 trial sequences are aggregated, pre-processed and averaged into unique emoji-specific signals. Each of the 7 emoji signals is then evaluated by the localizer pre-trained LDA classifier.

The results of this classification are then presented onscreen to the subject as a form of active visual feedback. This is repeated for the remaining 29 trials, against all subsequent emoji average signals. This means the classifier is evaluated using 210 test events (30 P300 & 180 Non-P300). Note, that this is the only analysis variant that can be classified as online or real-time, as all other methods required offline processes such as pooled-subject data aggregation and within-in-task (main experiment) data training. Note, that it was not possible to perform pooled-subject analysis in real-time as an even amount of subject data could not be guaranteed for each subject.

The use of the Localizer task as a pre-training method for the associated LDA models in the real-time lab setting was done principally to address the issue of overfitting and to avoid the potential confounding effects of spatial and temporal bleedover artefacts present in the main experiment data. At the time these considerations overly influenced the author and curical P300 oddball design features were poorly attended. Note, that the stimulus oddball probability of the Localizer task is 50%, contrasting against the 14.3% oddball probability in the 7-Emoji main experiment.

This suggests there could be a substantial mismatch in the quality of the associated P300s given the inverse relationship between oddball probability and P300 quality in terms of peak amplitudes and latencies. Along these very same lines, LDA models trained using Localizer data could therefore be poorly tuned for the identification of main experiment data. The merits and consequences of all decisions taken here are evaluated in the corresponding subsections 5.5.3, 5.6.1, and 5.7.

5.4.11.2 HYALL: No Class-Balancing: Pipeline 1

To probe the effects of a blended pre-training protocol the author also performed a final series of analyses including data from both the localizer and main experiment. This involves pre-processing both the localizer and main experimental data, followed by a separate data randomisation procedure applied to each subset individually. The two datasets are then aggregated, with the class-balanced localizer data positioned first. The default randomisation settings for the LDA model training protocol are deactivated to ensure the classifier is initially exposed to the localizer data, followed by the (non-class-balanced) main experimental training data. Once the LDA model is trained, the classifiers are then evaluated using a subset of the main experimental data (10% of the total main experimental data).

Note, that evaluations conducted for this analysis variant were undertaken using the Pipeline 1 approach. Hence, the training data was always comprised of the first 90% of all trials for any given single-subject or pooled-subject dataset. Along these very same lines, the remaining 10% of samples were utilized in the test data evaluations. The only application of a cross-validation procedure for the data collected in Experiment 3 relates to the Oversampled: Pipeline 2 analysis variant, see below for further details. This was undertaken to probe if the staged training protocol with a larger total training set could assist in enhancing classification performance. It was hypothesized that regularizing the LDA models with localizer data before training on the imbalanced main experiment training data could mitigate the overfitting effects previously discussed.

5.4.11.3 HYALL: Class-Balanced: Pipeline 1

A secondary HYALL analysis variant was assessed that involved class-balancing the main experimental data before aggregating the localizer and main experimental training datasets. This was performed using the same downsampling methodology implemented in Experiments 1 and 2 (see subsection 4.3.6.2). This involved isolating the Target samples from a similar number of Non-Target samples based on the relative temporal and spatial distance from the cued stimulus onset events.These additional assessments were performed to observe whether class-balancing across the entire training set would be more effective at increasing resultant LDA model performance, as opposed to only increasing the size of the training pool via the introduction of main experiment data as undertaken in the HYALL: No Class-Balancing variant. Note, that similarly to the models discussed in Experiments 1 and 2, both HYALL assessments were conducted offline.

5.4.11.4 Oversampled: Pipeline 2

In a similar approach to Experiments 1 & 2 (refer to subsections 3.4.3 & 4.3.7), the main experimental dataset for the three subjects assessed (Subjects 1, 2, and 3) in Experiment 3 was re-analyzed using Pipeline 2. This analysis involved training and validating all related LDA models through 10-fold cross-validation, applied independently to each subject's data (see subsection 3.4.3.1). To manage the imbalance between Target and Non-Target samples, the SMOTE oversampling technique was applied (see subsection 3.4.3.3). The data was organised into training and testing subsets using a stratified 9:1 ratio of Target and Non-Target samples to maintain a balanced distribution of classes across f-folds. For the training data, Target-P300 samples were generated using linear interpolation to match the number of Non-Target samples. This process was repeated for all 10 cross-validation folds.

Crucially, the no synthetic data was introduced into the test set. Mean accuracy scores (Overall, Target, and Non-Target) were calculated and compared against chance levels using one-tailed one-sample t-tests (see subsection 3.4.3.2). For the analysis, both Non-Collapsed and Collapsed data configurations were used. In the Collapsed setup, adjacent trial samples were merged to increase the number of sequences per trial from 5 (as outlined in subsection 3.3.5.3) to 10. This combination was done before the data was split into training and testing sets for cross-validation. The effectiveness of this collapsing method in improving classification accuracy was evaluated through non-parametric permutation tests on paired-subject data. Notably, all LDA models trained via the Pipeline 2 approach for either the Non-Collapsed or Collapsed data preparation methods were done with data exclusively from individual subjects.

5.5 Results: Pipeline 1

Across all subjects, for the localizer and main experiment, a total of 3600 samples were collected and later averaged down to 720 events total. The primary aim of this experiment, as mentioned above, is to probe the viability of a class-balanced localization task as a pre-training dataset for an LDA classifier implementing real-time P300 classification with onscreen visual feedback. Effectively, this aims to 'close the loop' by performing all operations live and relaying the results of the online analysis to subjects before the onset of the next trial. It is predicted that this combination of experimental adaptations will enhance the performance of the proposed emoji-based BCI system by improving subject vigilance and increasing the robustness of the LDA classifier.

5.5.1 Data Partitions: Pipeline 1

This final experiment uses just two datasets using samples from the localizer task and the main experiment (see, Table 5.1). As discussed above, the localizer task is significantly less extensive in duration, meaning there exists a large difference in the number of samples collected compared with the main experiment. This mirrors real-world applications which require relatively speedy initialization protocols to bring end-point users online as fast as possible. The majority of the differentiation in analyses comes by way of processing and combination, see below for more information.

Experimental Variants	Localizer	Main Experiment
Total Number of Events	90	630
Total Number of Test Events	9	63
Events per Subject	30	210
Test Events per Subject	3	21

Table 5.1: A table showing the differences between the localizer and main experiment in Experiment 3 in terms of the Total Number of Events (all data chunks sampled for each respective emoji across all subjects), Total number of Test Events (the 10% test data subset for evaluation purposes), Events per Subject (all emoji data chunks captured per subject) and Test Events per Subject (the 10% test data subset for each individual that partook in the experiment).

5.5.2 Analysis Partitions

As only one experimental variant (7 Emoji variant) and one Localizer task (1 Emoji variant) are utilized, the differences seen in datasets from the previous experiments are not present. The primary difference in this series of analyses is derived from the combination or isolation of respective data subsets. See the table below for further information (Table 5.2).

Analysis Variants	LOCRT	HYALL: No CLS-BAL	HYALL: CLS-BAL	Oversampled Non-Collapsed	Oversampled Collapsed
Total Number of LOC Train Events	81	81	81	0	0
Total Number of LOC TestEvents	9	9	9	0	0
Total Number of Main Exp Train Events	0	567	54	189	88
Total Number of Main Exp Test Events	630	63	6	21	10

Table 5.2: A table showing the differences between the five analysis iterations; LOCRT (Localizer Training + Real-Time Feedback), HYALL: No CLS-BAL (Hybrid Localizer + Main Data LDA training), HYALL: CLS-BAL (Hybrid Training Scheme with Class-Balanced Main Experiment Data) and the Oversampled Non-Collapsed and Collapsed variants. Note that the Oversampled methods utilized single-subject data exclusively and involved the implementation of 10-old cross-validation. Further, at no point used any Localizer task data and they were pre-processed and analyzed using the Pipeline 2 approach (see, 3.3.5.3). All other variants were processed via the Pipeline 1 method, for further information see subsections 3.3.5.1 and 5.5.2. The 'Total Number of LOC Training Events' refers to the number of samples used from the localization task to train corresponding LDA classifiers for the specific pooled-or-single-subject analyses in question. Similarly, the 'Total Number of LOC Testing Events' refers to the number of localizer samples used in evaluating a given LDA classifier. In the subsequent 'Main Exp' data volume and utilization metrics, the same principles apply concerning data gathered during the main experiment.

5.5.3 LOCRT: Pipeline 1

All results in this subsection relate to the LOCRT analysis variant (see, 5.4.11.1). This involved exclusively training single-subject LDA models using data from the respective localizer task to evaluate samples collected during the main experiment. This is the only real-time analysis discussed in this subsection, as all other variants involve post-experimental data reorganisation and training schemes. The result of these analyses was presented to all subjects during the main-experimental task using the visual feedback indicator, positioned in the upper right-hand portion of the stimulus monitor (see, Figure 5.2). Note, that the appropriate random performance thresholds here are 14.2% (see subsection 3.4.2) and ITR values are provided for these analyses exclusively owing to the real-time classification of the emoji samples collected. Further, the pooled-subject analysis of these data was not possible as data aggregation across all subjects assessed could not be completed uniformly until all individuals were tested.

5.5.3.1 Within-Subject

As seen in Table 5.3, greater than random performance was observed for all classification subcategories in each subject, excluding the within-class P300 accuracy (33.33%) for Subject 2. The highest-performing subject (Subject 1) in any 7 Emoji variant experiments conducted thus far (Subject 1) was revealed, achieving a mean classification accuracy of 80.95%. All subject LDA models were grid optimized for the solver method, demonstrating the highest performance with the lsqr method and a substantial variance in the degree of shrinkage employed (0.01-0.75). The subject average metrics also show the lowest variance in classification performance for any of the 7 Emoji variants assessed. Note, that a higher variance for the P300 waveform is still present. Note that none of these assessments were cross-validated due to the implementation of Pipeline 1 methodology. Consequently, interpretations should be made with caution, as the findings may not generalize beyond the scope of this analysis.

	Mean Acc (%)	P300 Acc (%)	Non-P300 Acc (%)	Solver	Shrinkage	ITR (bpm)
Subject 1	80.95	66.67	83.33	lsqr	0.01	9.39
Subject 2	55.00	33.33	58.82	lsqr	0.02	3.79
Subject 3	66.67	100.00	61.11	lsqr	0.75	5.99
Sub Avg	67.54	66.67	67.75	n/a	0.26	6.17
Sub Var	12.98	33.34	12.26	n/a	0.37	n/a

Table 5.3: The classification table contains the metrics relevant to the single-subject datasets tested for the LOCRT analysis variant. As stated above, the accuracy metrics listed here are the product of online classification. During the trial, EEG data was captured, pre-processed, and classified in real time. The predictions based on these analyses were then relayed back to the subject at the end of each trial via the presentation of a feedback indicator denoting the completion of either an accurate (green 'Thumbs Up') or inaccurate (red 'Thumbs Down') classification result. For more information see Figure 5.2, in subsection, 4.4.5: Main Experiment. Note, that these analyses do not include pooled-subject evaluations as all respective LDA models were generated using data exclusively from the localizer tasks employed to a specific subject. The 'Mean Acc (%)' column details the cross-class accuracy of all samples evaluated. This metric is broken down to the within-class level in the respective 'P300 Acc (%)' and 'Non-P300 Acc (%)' columns denoting LDA classifier accuracies for the separate target classes tested. The 'Solver' and 'Shrinkage' values relate to the LDA training grid-optimization scheme employed to determine the most effective solver method and shrinkage rate respectively. Finally, the ITR column provides the information transfer rate values of each LDA model trained in bit per minute computed according to the Wolpaw method [109] (see subsection 2.4). Note, that the 'Sub Avg' and 'Sub Var' are generated by calculating the mean value of each accuracy metric evaluated and the variance present within these values respectively.

A confusion matrix relating to the analysis conducted for Subject 1 is positioned below (see, Figure 5.3). This reveals high classification performance for the Non-P300 target class, alongside near-functional performance for the P300 waveform classifier. Importantly, the high classification accuracy in one class is not to the detriment of the other target class. The coupling of subject-specific localizer-based training data and a class-balanced data structure did appear to improve the classification performance of the evaluated LDA model. This suggests that the reduced noise present in the localizer data may be more suitable for the training of such classifiers, as the interference of temporal and spatial artefacts, present in the main experimental data are not confusing the direction of data separation computed by the LDA model. Despite this, the small sample size, the absence of statistical significance testing and the poor quality of the associated Localizer Cz grand-average plot restrict broad conclusions regarding P300-speller performance from being made based on these findings.

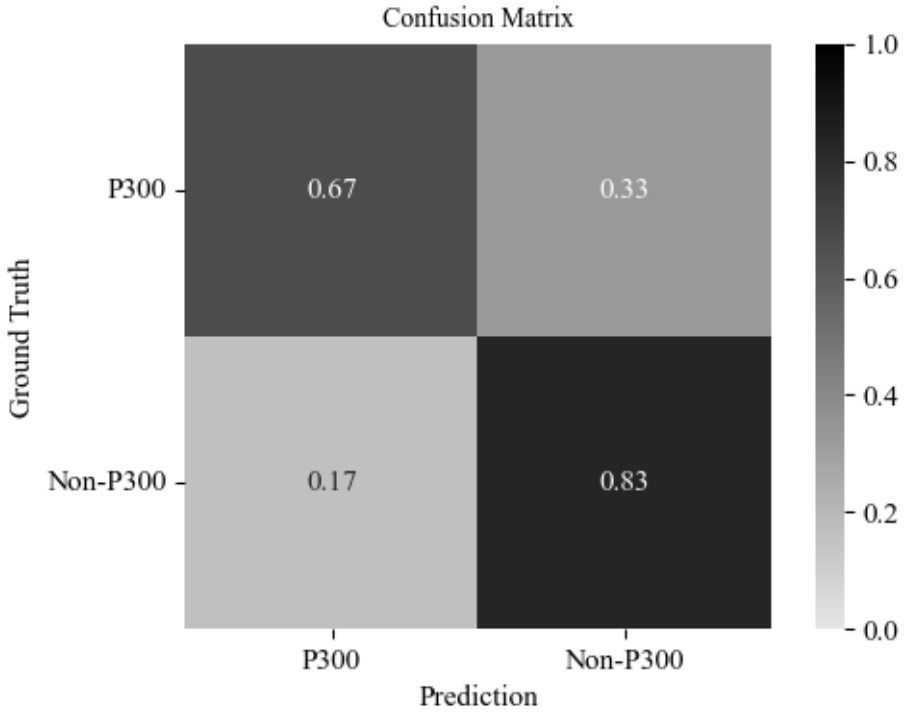

Figure 5.3: Here is displayed a normalized confusion matrix reporting the classification performance of the trained LDA model for both P300 and Non-P300 classes relating to Subject 1 in the LOCRT analyses block (refer to, Tables 5.2 & 5.3).

The figure positioned below (see, Figure 5.4) illustrates the below random-threshold performance for Subject 2 for the target P300 class. As seen in the upper right quadrant, the instances of confusion demonstrated by the classifier for the P300 class outweigh the instances of accurate classification. This may be due to the relatively small number of localizer trials captured, or potentially poor compliance with task instructions by the subject. Note, that at no point was any eye-tracking performed to monitor the subject gaze position during the task. These additional controls would help eliminate the potentially confounding effects of low-quality data inclusion alongside the online impedance assessments discussed in the method section (see subsection 5.4.9).

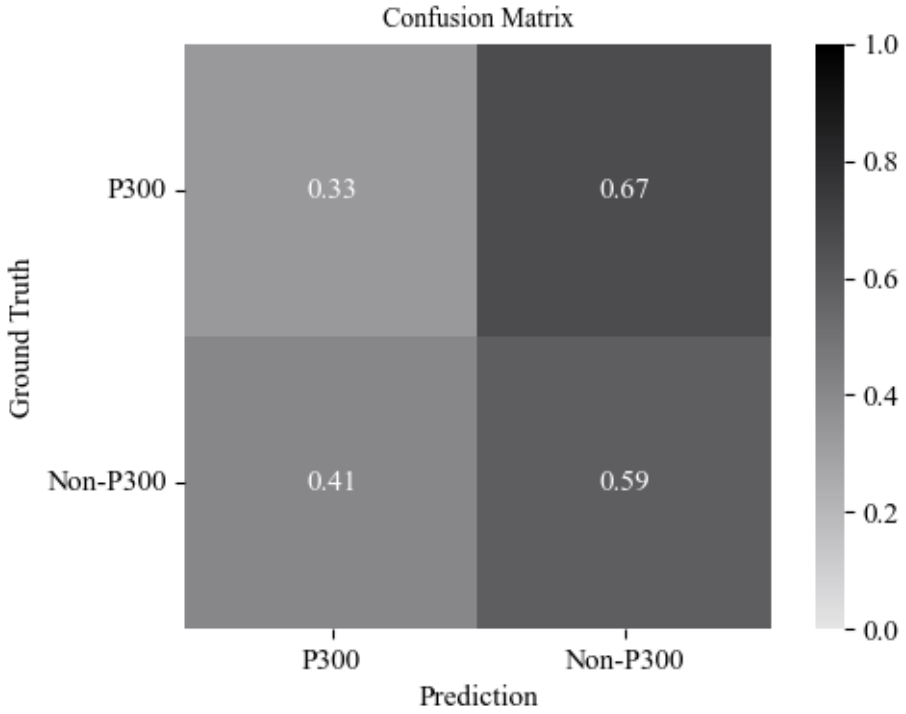

Figure 5.4: Here is displayed a normalized confusion matrix reporting the classification performance of the trained LDA model for both P300 and Non-P300 classes relating to Subject 2 in the LOCRT analyses block (refer to, Tables 5.2 & 5.3).

As illustrated in Figure 5.5, the Cz grand averages for the P300-Target trials reveal that the waveforms for single Localizer emoji stimuli diverge marginally from those collected during the Non-Target trials. This plot differs significantly from the Localizer average signals obtained in Experiment 2 (see Figure 4.14). Notably, a mild N200 component is observed, along with a distinct peaking event around 320 ms. This observation might suggest that the increased number of trials in the Localizer task variant of Experiment 3 allowed the predicted subjective probability of less than 50% to become more apparent. However, this is unlikely, given that the differences in amplitude are around 0.5 μV and individual samples are likely to exhibit considerable overlap. To clarify once again, the oddball probability of 0.50 for the augmentation (colour inversion) of the emoji stimulus is substantially higher than that employed in typical P300 experimental settings and it is predicted that implementing a variant with a lower probability would have resulted in substantially improved LOCRT classification results.

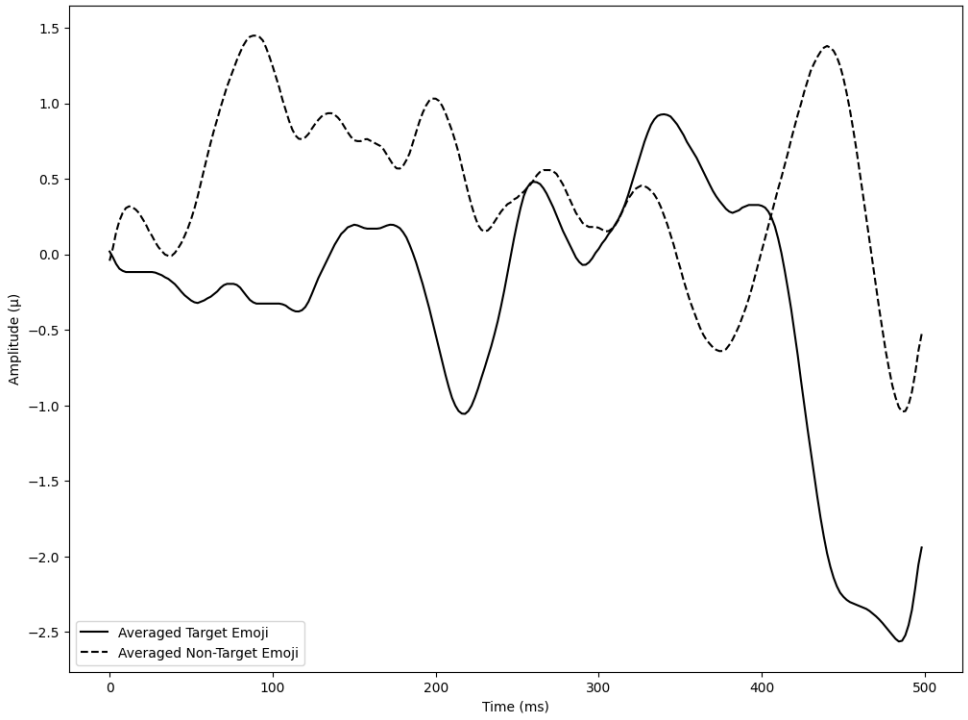

Figure 5.5: This figure presents a Cz grand average plot showing cross-trial P300 (solid line) and Non-P300 (dashed line) event signals for the Localizer data used (refer to subsection 5.4.11.1). The x-axis represents time in milliseconds for each 500ms event data chunk, while the y-axis shows amplitude in μV of the EEG signal. The averages across these classes highlight underlying EEG waveform patterns embedded in the signals. It is important to note that the Cz channel was exclusively used for constructing these plots. Additionally, all signalswere baselined by averaging the first 50ms of collected samples. This baselining was done solely for presentation purposes and was not applied during the Pipeline 1 data pre-processing as outlined in subsection 3.3.5.3 (see Table 3.1).

5.5.4 HYALL: No Class-Balancing: Pipeline 1

The results discussed in this subsection refer to the blended training data method outlined in the HYALL NO-BAL analysis variant (see, 5.4.11.2). At no point were the main experimental data subjected to any class balancing operations and retained the original 6:1 Non-P300:P300 sample ratio. Both cross and single-subject analyses are conducted herein to probe the influence of cross-trial and pooled-subject data aggregation in the development of classification method performance.

5.5.4.1 Pooled-Subject

All results reported in this subsection concern analyses conducted at the pooled-subject level and implemented using the Pipeline 1 approach (see subsection 3.3.5.1). The classification performance of the trained LDA model in the pooled-subject dataset for the P300 target class reports an accuracy of 0% (see, Table 5.4). Further, the Non-P300 target class was correctly identified in 96.55% of all instances. The grid optimization method revealed the combination of the lsqr solver method and shrinkage at 0.1 to be the nominal LDA training parameters for maximal classification performance. These findings indicate that overfitting is still a prevalent issue, irrespective of the initial LDA model training stage featuring class-balanced localizer data.

	Mean Acc (%)	P300 Acc (%)	Non-P300 Acc (%)	Solver	Shrinkage
Pooled Subject	78.87	0.00	96.55	lsqr	0.10

Table 5.4: The classification table herein presents the results generated from the HYALL: No Class-Balancing analysis variant for the pooled-subjects samples involving an LDA-model localizer pre-training and non-class balanced main experimental data training stage. For further information on field headings and interpretation refer to Tables 4.2 & 4.3.

Further, as seen in the confusion matrix below (see, Figure 5.6) the LDA model misclassified just 3.4% of all Non-P300 trials as belonging to the P300 class. In all samples evaluated, none of the P300 trial averages was correctly identified as belonging to this target class. Previously, the author asserted that greater data volumes in non-class-balanced pooled-subject datasets could reduce the incidence of overfitting. The increased data volumes in concert with a wide array of P300 and Non-P300 expression variants collected across multiple subjects could also have provided resultant classifiers with increased resilience to the characteristically noisy EEG data.

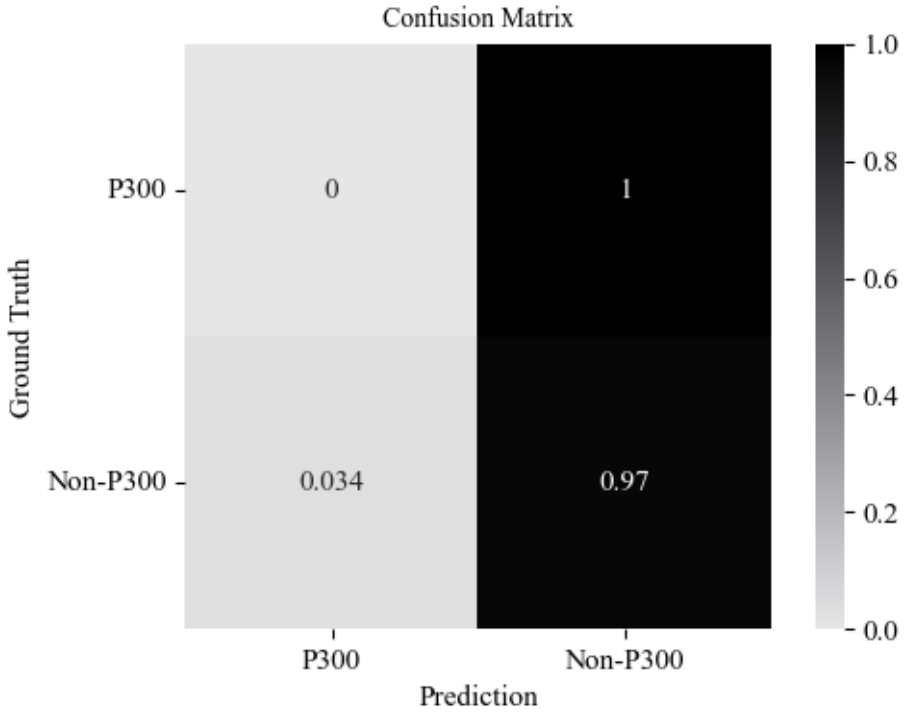

Figure 5.6: Here is presented a confusion matrix illustrating the classification pattern generated by the pooled-subject LDA model trained via the HYALL: No Class-Balancing analysis variant (see subsection 5.4.11.2).

In the grand average plot positioned below (see, Figure 5.7), the P300 average signal does not demonstrate the traditional visually generated-event-related waveform characteristics. This is primarily due to the absence of both a negative component around 100ms and a positive crest between 300-400ms. The Non-P300 average signal plot features both typical waveform characteristics, with the distinct absence of a final reduction in amplitude post-crest. Note, that it is possible subjects did not adhere strictly to experimenter instructions, varying gaze locations across the screen during trials. Due to the absence of concurrent eye-tracking, the evaluations of these possibilities lie beyond the scope of this analysis. Crucially, it must be noted that this does not indicate data mislabelling, as there are numerous instances of the opposing trend observed throughout the analysis for the same dataset.

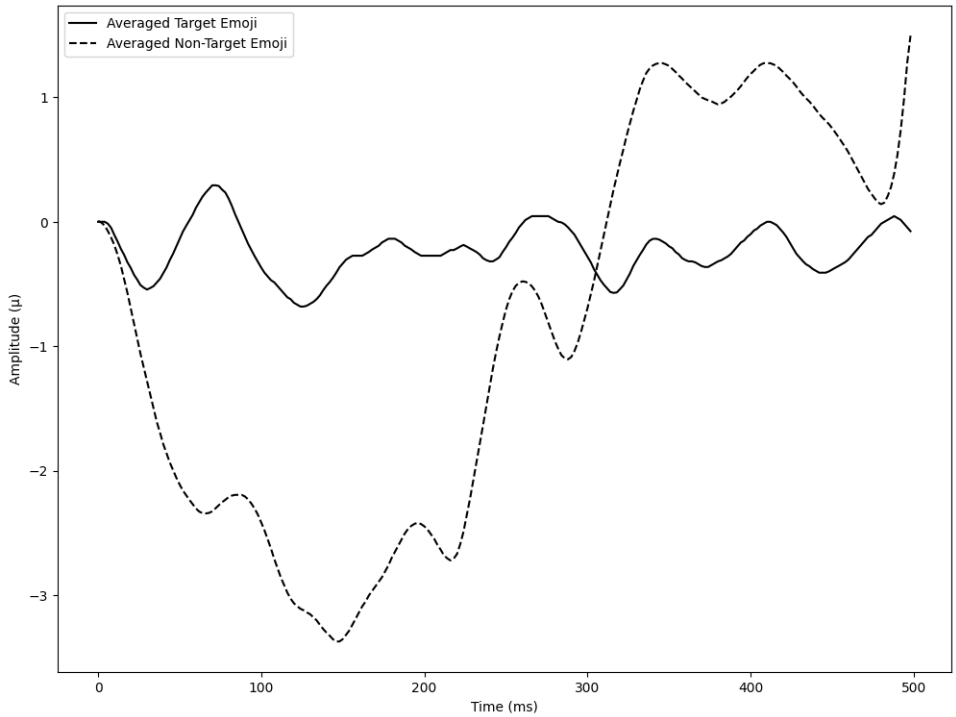

Figure 5.7: Here are presented the average-signal plots for the pooled-subject dataset relating to the HYALL: No Class-Balancing analysis variant. The plot contains the P300 (solid line) event average and the Non-P300 (dashed line) event average signal.

5.5.4.2 Within-Subject

This subsection focuses on the discussion of single-subject results for the HYALL: No Class-Balancing analysis variant (see subsection 5.4.11.2). All LDA models herein follow the same training methodology outlined earlier using data exclusively from individual subjects. The results of the single-subject analyses largely mirror the trends discussed at the pooled-subject level. As seen in Table 5.5, near-to-maximal classification accuracy is achieved for the Non-P300 target class for each subject evaluated. A sole subject (Subject 3) demonstrated marginally less susceptibility to the overfitting behaviour observed for the remaining subjects, with a P300 target class accuracy of 20%. All subject data revealed selection preferences for the lsqr grid search optimization solver method. It must be noted that additional variation for Subject 3 can also be seen in the dramatically higher shrinkage factor (0.86), as compared to Subjects 1 and 2.

	Mean Acc (%)	P300 Acc (%)	Non-P300 Acc (%)	Solver	Shrinkage
Subject 1	79.17	0.00	100.00	lsqr	0.29
Subject 2	82.61	0.00	100.00	lsqr	0.11
Subject 3	79.17	20.00	94.74	lsqr	0.33
Sub Avg	80.32	6.67	98.25	n/a	0.24
Sub Var	1.72	10.00	2.63	n/a	0.11

Table 5.5: The classification table herein presents the results generated from the HYALL: No Class-Balancing analysis variant, involving an LDA-model localizer pre-training and non-class balanced main experimental data training stage.

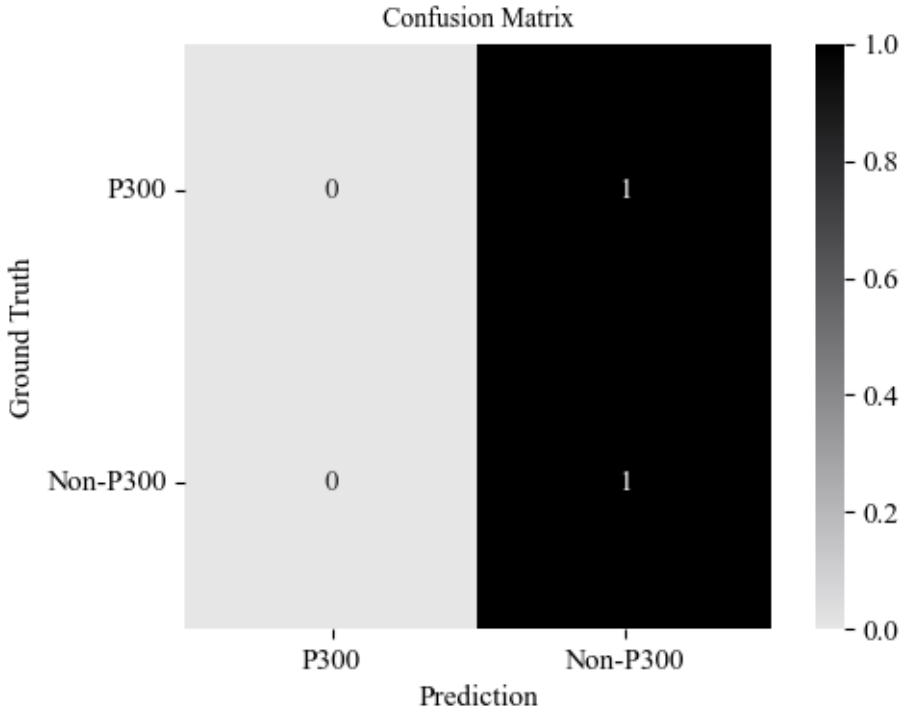

Figure 5.8: Confusion matrix illustrating the classification pattern generated by the LDA-model trained via the HYALL: No Class-Balancing analysis variant for Subject 2 (see subsection 5.4.11.2).

Above is presented a confusion matrix (see, Figure 5.8), illustrating the classification performance pattern of the LDA model trained via the HYALL: No Class-Balancing analysis variant. The model demonstrated a complete selective bias towards the Non-P300 target class for Subject 2. All instances of P300 prediction were confused and misclassified as Non-P300 events.

5.5.5 HYALL: Class-Balanced: Pipeline 1

All results reported in the following section relate to the HYALL: Class-Balanced analysis variant (see subsection 5.4.11.3). This involved training all respective LDA classifiers according to a scheduled training programme. Initially, models were trained using data captured from the localizer task and later trained using the data collected during the main experimental period. Note that, data randomisation was undertaken within these specific subsets to prevent the confounding influence of order effects. Crucially, the data utilized from the main experiment was processed via a class-balancing protocol to ensure the 1:1 ratio of P300 and Non-P300 event samples (see subsection 3.4.1).

5.5.5.1 Pooled-Subject

The evaluation of aggregated pooled-subject data was conducted in the same manner as in previous subsections. This involved collating samples across all subjects tested, with the first 90% of samples collected comprising the training data and the final 10% assigned to the test dataset.

	Mean Acc (%)	P300 Acc (%)	Non-P300 Acc (%)	Solver	Shrinkage
Pooled Subject	51.85	46.15	57.14	lsqr	0.46

Table 5.6: The classification table herein presents the results collected during the HYALL: Class-Balanced analysis variant for the pooled-subjects data samples involving an LDA-model localizer pre-training phase and a class-balanced main experimental data training stage.

The results presented in Table 5.6 demonstrate that the LDA model trained using the pooled-subject HYALL: Class-Balanced analysis variant achieved a mean classification accuracy of 51.85%. A similar result is reported for the Non-P300 target class (57.14%) with a sub-random accuracy shown for the P300 target class (46.15%). As seen in previous iterations of the analysis, the lsqr solver is selected by the hyper-parameter grid optimization method, with a shrinkage of 0.46. The consistent application of class balancing for the localizer and main experiment data undertaken for this analysis variant did not translate into higher classification performance. As the ratio between classes in the training data has been controlled for it suggests that the poor AoC observed are owing to individual differences in the expression of the P300 waveform across subjects. Alternatively, the differences in P300 expression across the two different tasks from which the data are aggregated could preclude the cohesive grouping of classes in feature space, reducing the capacity of the models to effectively separate the data.

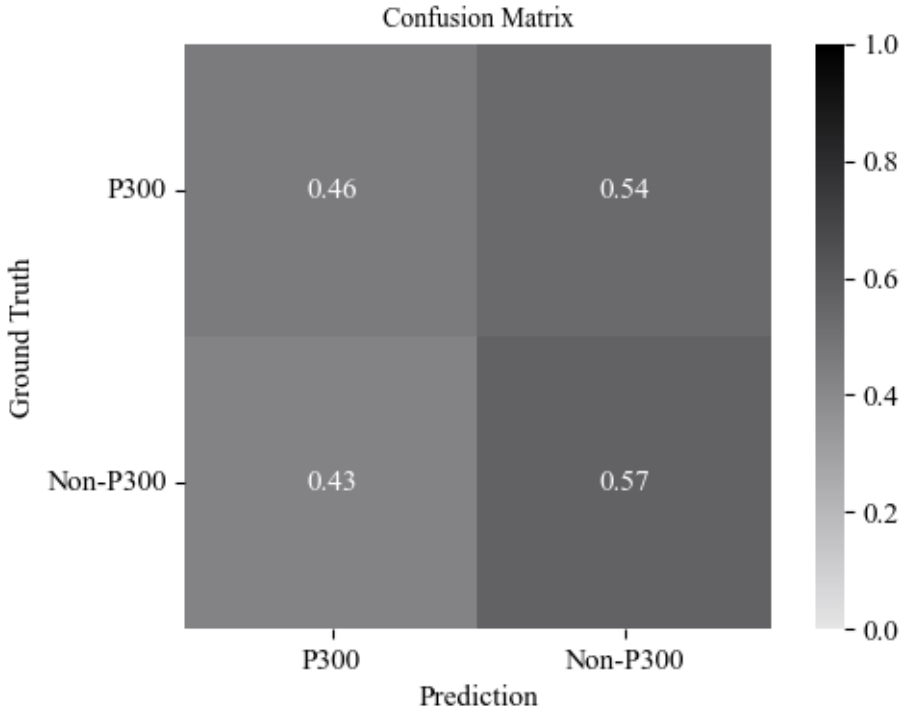

Figure 5.9: Here is presented a confusion matrix showing the classification performance generated by the pooled-subject LDA model trained via the HYALL: Class-Balanced analysis variant (see subsection 5.4.11.3).

As seen in the above confusion matrix plot (see, Figure 5.9), the respective LDA model demonstrates low classification performance for both the P300 and Non-P300 target classes evaluated. It could be argued that a slight selective bias in the target classification pattern can be observed for the Non-P300 target class. Despite this, the marginal differences in performance metrics exclude the ability to assert these claims with certainty.

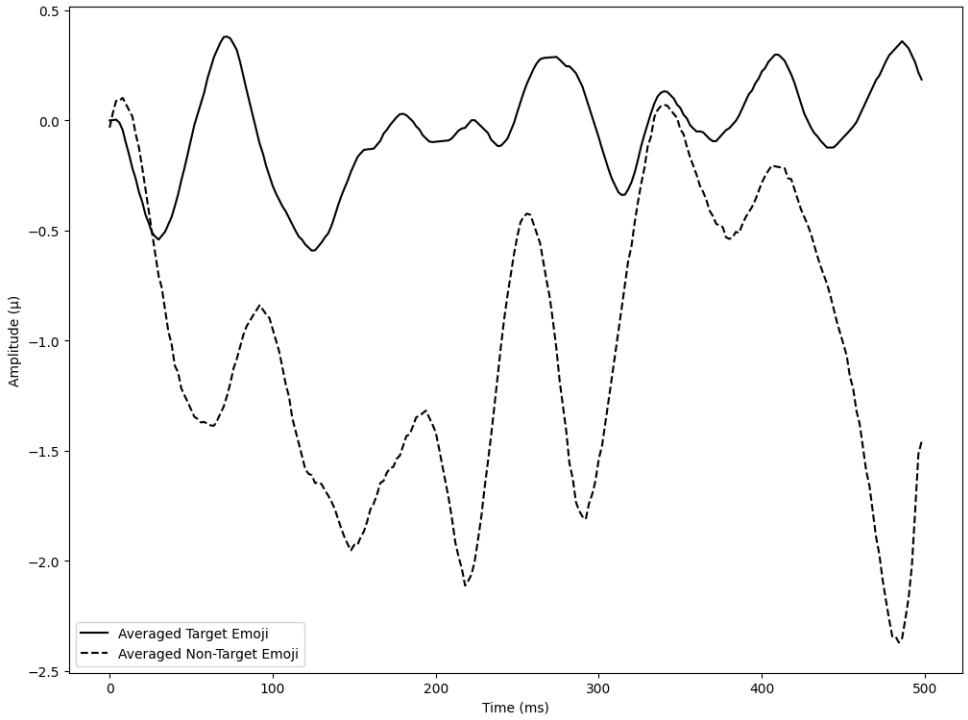

Figure 5.10: Here are presented the Cz grand average-signal plots for the pooled-subject dataset relating to the HYALL: Class-Balanced analysis variant. The solid line the P300 event average and the dashed line illustrates the Non-P300 event average signal. Note, that there is substantial similarity between these plots for the P300 waveform as these averages were computed in both instances using the same P300 event samples. The only difference here is the Non-Target samples used following the downsampling procedure used to enforce class-balancing (see Table 5.2 & subsection 5.4.11.3).

In the grand average positioned above (see, Figure 5.10), it appears the intended process of deleterious averaging to amplify inherent waveform features has reduced the prevalence of any strong and consistent differences in amplitude over time for the P300 waveform (solid line). In contrast, the Non-P300 waveform plot is a more typical example of the P300 waveform expected in precisely the opposing data type. The plot suggests that numerous samples containing P300 waveform features were subsampled from the main experiment data into the class-balanced Non-P300 group. This occurred despite efforts to select data from non-target emojis with the highest degree of spatial and temporal separation from the cued target emoji stimulus. To avoid this in the future, it may be preferable to generate multiple datasets subsampled from the Non-P300 events and perform a series of training and evaluation sessions, this is explored in the Pipeline 2 approach via k-fold cross-validation.

5.5.5.2 Within-Subject

The subsection herein relates to the single-subject results for the HYALL: Class-Balanced analysis variant. The classification table (see, Table 5.7) shows significant variance in LDA model performance across all subjects. Subject 1 demonstrates a selective preference for the P300 target class, the opposite trend is present for Subject 2 and Subject 3 illustrates a general lack of bias in concert with low classification accuracies for both classes tested. The highest mean accuracy attained is shared across Subjects 1 and 2 (66.67%), due to the mirrored classification performance for target classes. Additional similarities are observed in the rate of shrinkage (0.01 and 0.00) respectively. Subject 3 presents with a significantly higher grid-optimized shrinkage rate and all subjects share the selection of the lsqr as the solver method.

	Mean Acc (%)	P300 Acc (%)	Non-P300 Acc (%)	Solver	Shrinkage
Subject 1	66.67	80.00	50.00	lsqr	0.01
Subject 2	66.67	50.00	80.00	lsqr	0.00
Subject 3	44.44	40.00	50.00	lsqr	0.86
Sub Avg	59.26	56.67	60.00	n/a	0.29
Sub Var	11.12	20.00	15.00	n/a	0.43

Table 5.7: The classification table herein presents the results collected during the HYALL: Class-Balanced analysis variant for the single-subject samples involving an LDA-model localizer pre-training phase and a class-balanced main experimental data training stage.

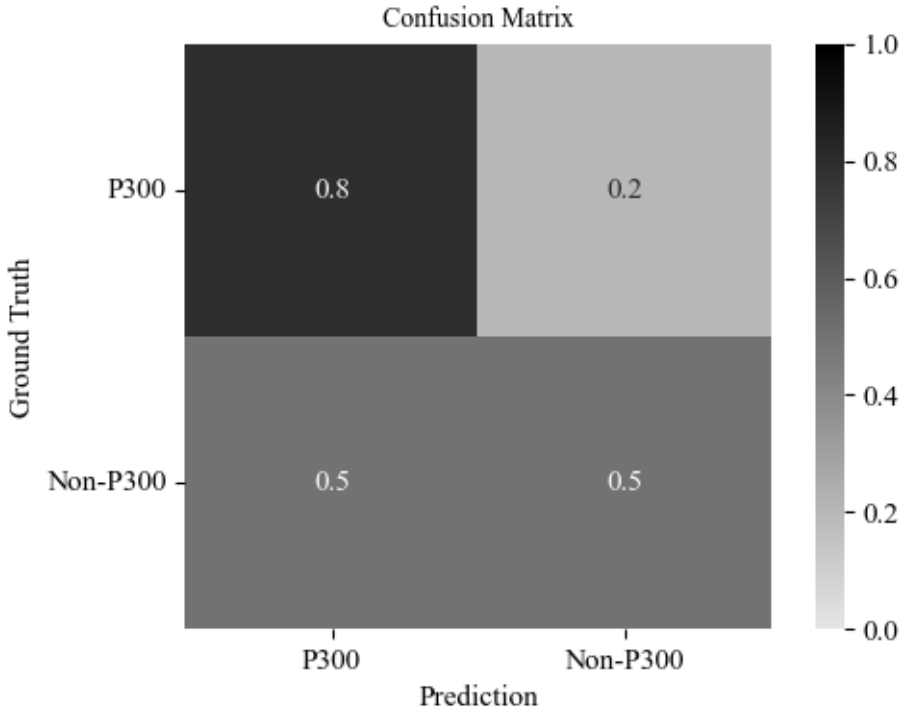

Figure 5.11: Here is presented a confusion matrix showing the classification performance gathered during the within-subject analysis of Subject 1 for the LDA model trained via the HYALL: Class-Balanced analysis variant (see subsection 5.4.11.3).

The confusion matrix positioned above (see, Figure 5.11) provides further insight into the classification performance of Subject 1 for the HYALL: Class-Balanced analysis variant. The upper two quadrants reveal the LDA model did demonstrate a high incidence of correctly classifying P300 waveforms. This trend is not replicated for the Non-P300 target class.

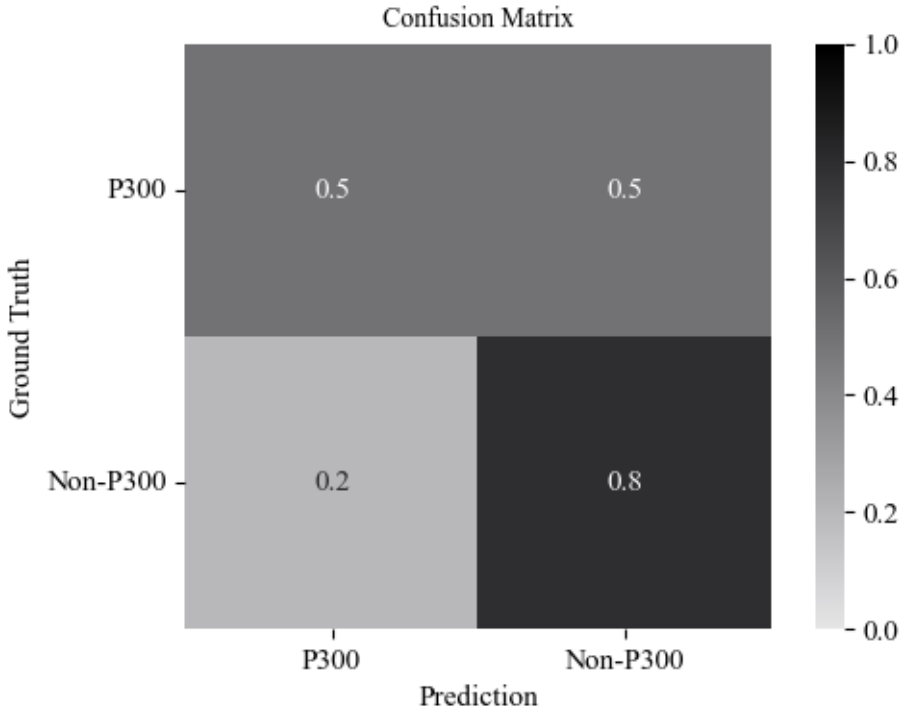

Figure 5.12: Here is shown a confusion matrix presenting the classification performance gathered during the within-subject analysis of Subject 2 for the LDA model trained via the HYALL: Class-Balanced analysis variant (see subsection 5.4.11.3).

The above figure (see, Figure 5.12) shows the inverse pattern of classification to the previous plot (see, Figure 5.11), with a high incidence of classification accuracy for the Non-P300 event samples and poor performance for the P300 events evaluated. Overall, these results suggest that the implementation of strict class-balancing protocols at the single-subject level can not remove all instances of class-wise selection bias.

5.5.6 Oversampled: Non-Collapsed: Pipeline 2

This subsection presents results for the Non-Collapsed Oversampled analysis variant implemented using Pipeline 2. As previously mentioned, Non-Collapsed samples were created using all 5 sequences per trial. Again, the class imbalance inherent to the data set was addressed here via the implementation of the SMOTE oversampling method involving cross-sample linear interpolation to generate synthetic data. The results focus on single-subject mean classification accuracies for Overall, Target, and Non-Target categories, assessed for significance using a one-sample t-test (threshold, $p<0.05$). Additionally, group-level results are examined through a paired-subjects permutation test (see subsections 3.4.3.2 and 5.4.11.4).

	Overall		Target		Non-Target	
Subjects	Acc Mean	Std Dev	Acc Mean	Std Dev	Acc Mean	Std Dev
1	0.83*	0.06	0.94*	0.10	0.74* (p=0.0002)	0.13
2	0.72*	0.05	0.82*	0.07	0.62* (p=0.0115)	0.12
3	0.82*	0.13	0.96*	0.07	0.69* (p=0.018)	0.21

Table 5.8: Here is presented a table showing the performance metrics associated with Subjects 1, 2 & 3 for the Oversampled Non-Collapsed data partition (see, Table 5.2). All results were computed following the stages laid out in the Pipeline 2 data organisation, pre-processing and analysis methodology (see subsection 3.3.5.3 Data Pre-Processing: Pipeline 2). Here all individual samples are composed of averages computed across all 5 augmentation sequences within each respective trial (see subsection 5.4.11.4). Note, that all cell values denoted with a * indicate a significantly higher mean classification accuracy than the 50% chance level for the binary (Target *vs*. Non-Target) classification task. For additional information on table field headings and interpretation please refer to Table 3.16.

As seen in the table above (see, Table 5.8), all single subject-level tests comparing the accuracy metrics (Overall, Target & Non-Target) collected during respective 10-fold cross-validation assessments revealed significant differences to the chance 50% level. Notably, Subject 2 provided a marginal result (p=0.015) owing to the low accuracy metric and relatively high associated standard deviation. Further, the group-level results comparing accuracy metrics across subjects against the chance level showed a significant difference from the threshold of 50%, with the Non-Target values providing another highly marginal result (p=0.048). Note, again that these permutation assessments have diminished statistical power given the small sample size (3 subjects).

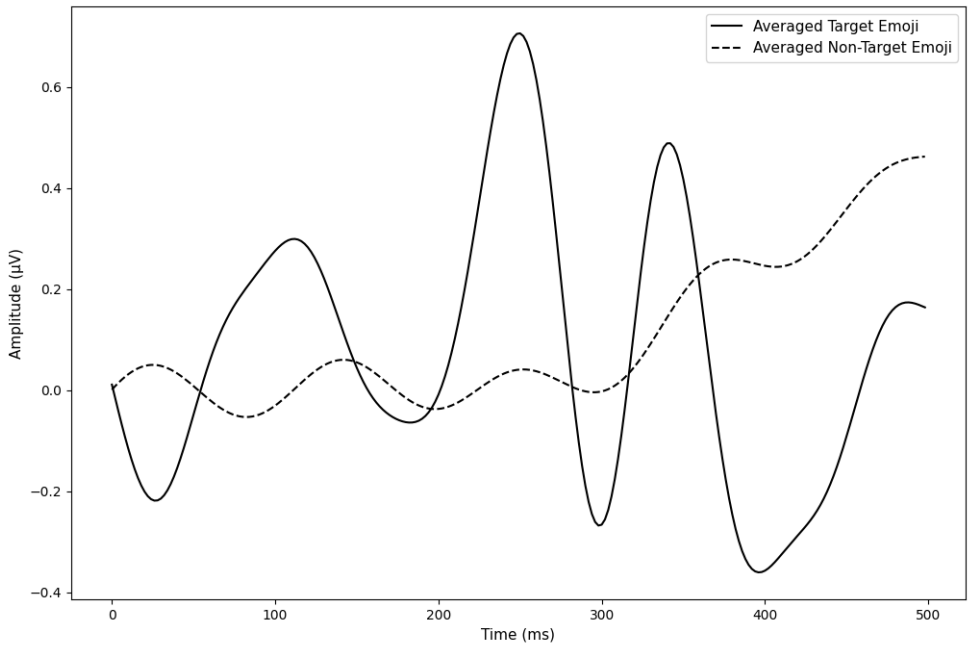

Figure 5.13: This is an average plot constructed exclusively from the Cz electrode for all P300, Averaged Target Emoji (solid line) and Non-P300 Averaged Non-Target Emoji (dashed line) samples collected across Subjects 1, 2 & 3 for the Non-Collapsed Oversampled data partition (see, Table 5.2). As can be seen, the time dimension is positioned on the x-axis (0-500ms) and the micro-voltage range is oriented to the y-axis. All samples were processed using the Pipeline 2 data pre-processing methods (see subsections 3.3.5.3 & 5.4.8). Target P300 and Non-Target P300 samples were averaged and aggregated into separate arrays to compute the grand pooled-subject mean signal shown in the plot. A total of 84 P300 and 547 Non-P300 samples were used, with no synthetic P300 samples included in these averages. The key difference between the Non-Collapsed and Collapsed data variants is the number of sequences per sample average. Since the plot here includes all samples, it represents a linear combination of all signals in the augmentation variant, making the Collapsed data average plot effectively identical to the one shown.

In the plot shown here, the Cz grand averages of the Target-P300 and Non-P300 reveal a similar pattern as noted in all previous Pipeline 2 EEG signal plots. Here, a mild negative component around 200ms is followed by a pair of large deflections around 300ms. The authors assert that these could relate to the aforementioned P300a and P300b waveform features. Notably, the Non-Target average displays the same lower degree of signal variance. Further, a signal with a periodicity of around 130ms is also observable, here this could be due to the ineffective application of a 6.67Hz filter to remove associated stimulus-induced SSVEPs. Overall, the quality of these signals is not impressive, however, these data should be adequate for the discrimination of the classes via the LDA, as evidenced in Table 5.8.

5.5.7 Oversampled: Collapsed: Pipeline 2

This subsection presents the results of the Collapsed Oversampled analysis variant, implemented using Pipeline 2. As previously described, Collapsed samples were created by averaging 10 sequences from consecutive experimental trials. In addition to one-sample t-tests and group-level results, the overall mean accuracies of the Non-Collapsed and Collapsed data are compared directly using a paired-subjects permutation test. As shown below (see, Table 5.9), all overall and Target mean classification accuracies were significantly above the 50% random performance threshold. However, after applying the Collapsed data preparation method, both Subject 1 (p=0.101) and Subject 3 (p=0.297) did not achieve accuracies significantly above random performance for the Non-Target samples. This outcome is attributed to a relative decrease in mean accuracy and an increase in associated standard deviations following the Collapsed data preparation method. At the group level, the only metric not significantly different from chance was the Non-Target accuracy metrics. This suggests that an increase in Target classification accuracies may have led to a relative decrease in Non-Target accuracies.

	Overall		Target		Non-Target	
Subjects	Acc Mean	Std Dev	Acc Mean	Std Dev	Acc Mean	Std Dev
1	0.77*	0.08	0.93*	0.10	0.61	0.16
2	0.83*	0.09	0.99*	0.03	0.66*	0.22
3	0.78*	0.11	1.00*	0.00	0.57	0.20

Table 5.9: Here is presented a table showing the performance metrics associated with Subjects 1, 2 & 3 for the Oversampled Collapsed data partition (see, Table 5.2). All results were computed following the stages laid out in the Pipeline 2 data organisation, pre-processing and analysis methodology (see subsection 3.3.5.3 Data Pre-Processing: Pipeline 2). Here all individual samples are composed of averages computed across 10 augmentation sequences from adjacent subject trials (see subsection 5.4.11.4). Note, that all cell values denoted with a * indicate a significantly higher mean classification accuracy than the 50% chance level for the binary (Target *vs*. Non-Target) classification task. For additional information on table field headings and interpretation please refer to Table 3.16.

Here, the differences in subject-matched Overall mean classification accuracies for the Non-Collapsed and Collaped data preparation methods are presented. The figure further reinforces the notion that, given the low number of samples remaining following the Collapsed method, the outcomes for individual subjects are fairly volatile, leading to an increase in relative standard deviations and most instances trends. Despite this, evidence of the contrary is present in the plot, with Subject 2 showing a dramatic increase in overall accuracy. This was manifested by a jump of 17% for the Target samples and a corresponding 4% increase for the Non-Target samples. This suggests, at least for some subjects, that the relative drop in

the number of trials available is not a hindrance to the respective classifier training. Notably, board conclusions can not be made given the small 3-subject dataset.

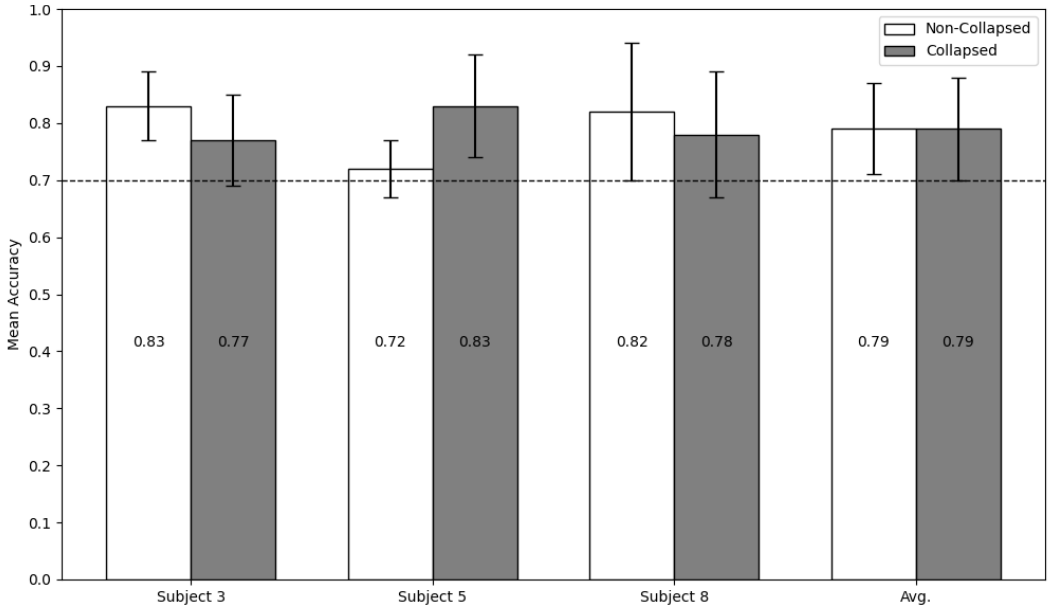

Figure 5.14: The plot presents a paired bar chart illustrating the mean Overall accuracies and standard deviations for the Non-Collapsed (white bars) and Collapsed (grey bars) data preparation methods across each subject, using the main experimental data. The Non-Collapsed bars represent datasets where each sample is an average of 5 augmentation sequences (see subsection 3.3.5.3 for details), while the Collapsed bars correspond to samples derived from 10 sequences by aggregating adjacent trials. These mean values are obtained from a 10-fold cross-validation for each of the three subjects (1, 3, & 5), as well as the pooled-subject average (Avg.) (see subsection 3.4.3.1). The plot also features standard deviation bars to indicate variability and annotations with the mean accuracy values. A horizontal dashed line at 70% serves as a performance benchmark for evaluating each method.

5.6 Conclusion

All findings and conclusions discussed herein relate to Experiment 3. The corresponding tables and figures can be found in the text positioned above. Again, these investigations featured real-time classification, user prediction feedback and an active impedance monitoring system for a 7-emoji target array. Notably, the use of either the Pipeline 1 or 2 approaches in the data organisation and preparation methods for each respective analysis variant is indicated in the respective subsection titles.

5.6.1 LOCRT: Pipeline 1

The analysis variant discussed here (LOCRT) was performed exclusively at the single-subject level and represents the only instance of a true real-time, closed-loop BCI speller system detailed in this thesis. The data used in the respective LDA model training sets was taken solely from the EEG time series acquired during the localizer pre-screening task. These very same models were then implemented in the online classification of samples collected during the main experimental period.

5.6.1.1 Within-Subject

The presence of classification accuracies above random performance (14.2%) for all subcategories, except the Subject 2 P300 Acc (%) (see, Table 5.3), is a substantial improvement when compared to the results of previous iterations discussed in earlier sections using the Pipeline 1 method. Crucially, the LOCRT analysis variant presents one of the highest performing single-subjects (Subject 1) in terms of ITR (9.39 bpm) and mean classification accuracy of 80.95% for the Pipeline 1 approach. This could be due to the decreased prevalence of spatial and temporal bleed-over effects, such as double flashing and adjacency error afforded by the 1-emoji target array of the localizer task training data. Despite this improvement in results, the variance in performance for P300 and Non-P300 target classification accuracies within subjects remains large.

Further, the P300 and Non-P300 grand average plot (see, Figure 5.5) generated across subjects demonstrate some atypical waveform characteristics as well as weak P300 peaking component and slightly delayed latency of around 350ms. This is likely owing to the high oddball stimulus probability of the Localizer task. (0.5). This reduces the ability of the author to confidently assert that the classification of these samples was performed successfully based on the separation of P300 and Non-P300 waveform features. Note, that these plots were generated using Localizer task data.

5.6.1.2 Variant Summary

The author initially designed the Localizer task with multiple objectives: to train subjects, prescreen the quality of P300 signals, and, in Experiment 3, provide pre-training data for LDA classifiers in online decoding. Additionally, the Localizer task employed a class-balanced presentation scheme and a single onscreen emoji to address the Target: Non-Target sample ratio issue and to mitigate potential spatial and temporal bleedover effects that could compromise the LDA model training. However, these objectives were significantly compromised due to

fundamental flaws in the task's design. Although the intention to minimize bleedover effects was sound, it could have been more effectively achieved by employing a 1-Emoji array with a lower oddball probability. The decision to use a high oddball probability of 50%, as depicted in the Cz grand average plot (Figure 5.5), substantially reduced the task's effectiveness in generating robust P300 signals.

Moreover, the discrepancy in oddball probability between the Localizer task and the main experiment led to the pre-training of LDA models on samples that likely exhibited different waveform characteristics. This inconsistency is evident when comparing Figure 5.5 with Figure 5.7, where the features of the waveforms differ significantly. Although the results from the Localizer task represent some of the best performance metrics recorded using the Pipeline 1 approach, it is important to note that this method did not include a 10-fold cross-validation procedure, as later adopted in the Pipeline 2 approach. Consequently, the validity of these results is inherently questionable. In hindsight, the author recognizes that a more effective strategy would have been to implement a shorter version of the main experimental task, increasing the number of sequences per trial in both instances. This approach would likely have improved averaging and enhanced the signal-to-noise ratio, leading to more reliable and valid results.

5.6.2 HYALL: No Class-Balancing: Pipeline 1

The analysis variant detailed herein (HYALL: No Class-Balancing), involved training the respective LDA models at the single and pooled-subject levels using all the data collected during both the localizer task and the main experiment. Note, that this produced significant class-imbalance in the corresponding main experiment training dataset. Again, all LDA models were first tuned with the localizer dataset and subsequently trained with samples from the main experiment. To clarify, owing to the use of main experiment data in the LDA training set, all analyses discussed here are categorized as offline.

5.6.2.1 Pooled-Subject

In consideration of the results collected for the pooled-subject HYALL: No Clas-Balancing analysis variant (see, Table 5.4) accuracies exceeding the random performance threshold for the P300 class are not observed. This suggests that the previous assertions surrounding the potential benefits of pooled-subject aggregation to increase training data volume and resultant model robustness do not apply in this context. Note, that these conclusions are given validity specifically by the pooled-subject HYALL: No Class-Balancing data evaluation, as this

analysis variant, contains the largest training dataset used in the LDA-model training for this experimental implementation (see, Table 5.2). Furthermore, the same pattern of atypical P300 and Non-P300 waveform features in the corresponding pooled-subject grand average plot (see, Figure 5.7) adds further weight to the suggestion that the aggregation of pooled-subject data for this analysis variant is unviable. Overall, the results suggest that the application of a pre-training stage utilizing class-balanced localizer data is not sufficient to prevent model overfitting during subsequent training with non-class-balanced data from the main experiment. This is likely owing to the large difference in oddball probability between the localizer and main experimental data. This likely introduce heterogeneity into the Target samples and reduced the effectiveness of the associated LDA models to accurately predict these data classes.

5.6.2.2 Within-Subject

It is clear that the HYALL: No Class-Balancing analysis variant performed similarly at both cross and single-subject levels (see, Table 5.5). The only deviation from this pattern of overfitting was found in Subject 3. The imbalanced ratio of P300 and Non-P300 events in the main experiment data precluded LDA-model class separation in nearly all single-subject instances. Both of the noted confusion matrices (see, Figures 5.6 & 5.8), to marginally varying degrees, broadly illustrate the same pattern of selective bias for the more numerous Non-P300 event class. Parity in the incidence of misclassification events across classes is a typical sign of effective classifier learning. The absence of this phenomenon here amplifies the likelihood of model overfitting.

5.6.2.3 Variant Summary

Overall, it is clear that the efforts undertaken to enhance LDA model robustness via the use of high-volume aggregated cross-experiment datasets were not successful. This is likely owing to the differences between data quality in the localizer task and the main experiment, in addition to the absence of class balancing in the main experiment data. As both single and pooled-subject data demonstrate poor performance it suggests that the issues of individual differences between subjects do not account for all the variance in determining classification accuracy. Had this been the case, the single-subject results would not demonstrate the same sub-random classification performance.

The possibility remains that two separate depressant effects are influencing each dataset individually. For example, individual differences may be reducing pooled-subject performance and low sample size could be driving the low classification accuracies at the single-subject

level. Despite these considerations, it is unlikely both sets of results would be expressed with such high levels of similarity in the absence of a shared confounding variable, principally the absence of a class balancing or fully regularized (weight) LDA training stage. Further, as seen in both plots (see Figures 5.5 and 5.7) neither the localizer or main experiment data contain the expected P300 waveform features. For the localizer task this is likely owing primarily to the high oddball stimulus probability of the augmentation events. Regarding the HYALL: No Class-Balancing plot, this is likely related to the suboptimal implementation of data organisation and pre-processing methods employed via the Pipeline 1 method (see subsection 3.3.5.1).

5.6.3 HYALL: Class-Balanced: Pipeline 1

The analysis variant detailed herein (HYALL: Class-Balanced), involved training the respective LDA models at the single and pooled-subject levels using data collected during the localizer task and an artificially class-balanced subsample of the main experiment data (see, Table 5.2). Again, all LDA models were first tuned with the localizer dataset and subsequently trained with the subsampled main experiment data. Note, that owing to the use of main experiment data, all analyses discussed here are categorized as offline.

5.6.3.1 Pooled-Subject

As shown in Table 5.6, the overfitting issues prevalent in many of the results sections presented herein do appear to have been mitigated after enforcing class balancing in the training dataset. Despite this, these contingencies were not adequate to advance the progress of the model past the random performance threshold (50%). When comparing these results to those gathered in the LOCRT analysis variant (see, Table 5.3), it could be suggested that the data collected during the main experiment is of inherently lower quality. Originally it was asserted that the integration of main experimental data into the training sample subset should improve model resilience to noise artefacts. On reflection, the inclusion of the localizer task data likely introduced significant variance within the Target P300 class samples and reduced the homogeneity of the associated grouping. This likely dropped the classification performance by complicating the task of data grouping by the associated LDA model.

5.6.3.2 Within-Subject

As seen in Table 5.7, there is a substantial reduction of overfit prevalence for all subjects tested as compared to the HYALL: No Class-Balancing variant (see, Table 5.5). Despite this, the presence of selective bias was not removed entirely, as shown by the 30% difference in

performance observed for Subjects 1 and 2 when comparing the P300 and Non-P300 classification accuracies. These results suggest that class-balancing is necessary, not sufficient, for classifier performance enhancement. Increased high-quality data volumes are likely needed to adequately train the LDA models to accurately identify these complex data features. As these models were trained using single-subject, class-balanced data the LDA model used can be evaluated in the absence of issues relating to individual differences and dataset class ratios. None of the models tested produced classification accuracies above the random performance threshold for both the Non-P300 and P300 classes. This suggests that the comingling of data across the localizer and main experiment prevented effective learning in the associated LDA models.

5.6.3.3 Variant Summary

In conclusion, the application of class balancing for both the cross and single-subject datasets for the aggregated localizer and main experiment data is not sufficient to produce LDA models with high classification accuracies. This is likely related to the mixing of samples across experimental variants. As noted above, the utilization of the localizer task data is likely flawed due to the difference in P300 waveform expression as a function of the higher relative oddball probability, as compared to the main task. This likely impacted the classification results negatively. Any reimplementation of a similar 1-Emoji localizer task must feature the same stimulus oddball presentation probability as the main task to ensure the generation of similar P300 amplitudes and latency, while also reducing the confounding noise components that are introduced via neighbouring targets.

5.6.4 Oversampled: Pipeline 2

The Non-Collapsed Oversampled analysis, involving the use of the SMOTE oversampling method to increase the relative number of minority class Target samples via linear interpolation (see subsection 5.4.11.4), produced the highest overall mean accuracies in all three subjects assessed for the data captured in Experiment 3. Notably, the Overall, Target and Non-Target offline accuracies were all reported as significantly higher than the 50% random performance threshold ($p<0.05$) as computed via one-sample t-tests. Here, as compared to the LOCRT method (see, Table 5.3), Subject 1 appears to have increased in mean classification performance by around just 2%. Further, a relative drop in performance is observed for Subject 1 for the Collapsed (cross-trial sequence averaging) data partition (see subsection 3.3.5.3). However, it must be stated that these initial assessments, LOCRT (5.5.3), HYALL: No Class-Balancing (5.5.4) and HYALL Class-Balanced (5.5.5), were undertaken via the Pipeline 1

Approach and were not performed according to a 10-fold cross-validation procedure (see subsection 3.3.5.3).

The influence of the Collapsed data preparation method appears broadly negative when compared to the Non-Collapsed method (see Figure 5.14). Specifically, two out of three subjects showed a significant drop in overall mean classification accuracy, along with a substantial increase in the associated standard deviations for the Non-Target class. Group-level statistics revealed a significant decline in classification performance, which was restricted exclusively to the Non-Target samples. At the single-subject level, mean Non-Target classification accuracies for Subjects 1 (61%) and 3 (57%) were not significantly different from chance. Moreover, the evidence of bias toward the Target-P300 class is substantially higher here compared to the Non-Collapsed data preparation (see Tables 5.8 & 5.9). This bias is reflected in the observed large increase in Target classification accuracies (14.7%) and a relative decrease in Non-Target classification accuracies (7.03%). In real-time classification contexts, this imbalance would likely lead to an increase in false positives, where Non-Target data segments are misclassified as belonging to the Target class, resulting in necessary corrections or miscommunication of key emoji-states. In summary, for this configuration of the 7-Emoji speller, the data do not support the assertion that the Collapsed data preparation method had a positive influence.

5.7 Reflections

This section outlines the main areas of import relating to Experiment 3. This includes key findings of the investigations from the Pipeline 1 and 2 evaluations, areas for improvement and future research, as well as some considerations on the P300 experiment series described across Chapters 3, 4 and 5.

5.7.1 Considerations on Pooled-Subject Data Aggregation

At numerous points throughout these analyses, concerns have been raised regarding the aggregation of data across subjects. It is likely that the LDA models utilized are not suitable for these kinds of pooled-subject training and classification. These conclusions are more adequately drawn from the results garnered in Experiments 1 and 2 (see subsections 3.5 & 4.4). This is primarily related to the sample size of Experiment 3. At just 3 subjects, the purported benefits of pooled-subject data aggregation related to increased data volumes and a greater diversity of P300 waveform profiles to enhance classifier robustness could not be satisfactorily meted out.

The author suggests that the most effective strategy for acquiring high-quality data would likely involve significantly increasing the number of sequences per trial to around 10 or 15, aligning with standard practices. This approach is expected to enhance the quality of the resulting averaged signals, potentially facilitating the use of more advanced analytical methods. Additionally, conducting extensive initial training sessions would have been preferable to the relatively brief and compromised Localizer task employed in subject training throughout this thesis. Furthermore, the adoption of a performance-gated screening procedure, where specific confidence thresholds must be met before a trial is deemed complete, would further refine data quality.

Expanding the sample size to include over 20 subjects would also allow for a more thorough investigation of key parameters such as augmentation onset intervals, optimal array density, and the sizing and spacing of stimuli. These measures would likely mitigate the confounding effects of subject fatigue and improve the performance of LDA models trained at the single-subject level. Moreover, they could enable the use of state-of-the-art convolutional neural networks, which have demonstrated the ability to leverage pooled-subject data aggregation benefits [58, 256]. It was the author's intention to collect data from this many subjects, however, due to the constraints imposed by COVID-19 pandemic restrictions on in-person testing, these data collection goals were not fully realized, as referenced in the Positioning Statement (see subsection 1.1).

5.7.2 Localizer Task Considerations: Pipeline 1

When evaluating the classification performance of the seven Emoji variants that included a localizer pre-training stage across Experiments 2 (see Table 4.13) and 3 (see Table 5.3) via the Pipeline 1 method, it appears that expanding this initialization task enhanced both mean and within-class (P300 and Non-P300) accuracies. However, these improvements are primarily limited to within-subject assessments, as the real-time nature of the LOCRT analysis variant constrains broader applicability. Specifically, the improvements observed at the average subject level include a greater than 11.4% increase in mean classification accuracy, a more than 26.67% rise in P300 accuracy, and an over 8.86% enhancement in Non-P300 accuracy.

While these gains may be partly attributed to the introduction of reactive impedance monitoring, it is unlikely that this factor had a significant impact beyond aiding subject-specific instructions during tasks, as protocols for excluding trials with large impedance variances were already established by Experiment 2. Furthermore, the presence of atypical features in both

P300 and Non-P300 signals across most grand average plots suggests that real-time feedback did not significantly improve the quality of data collected during the main experimental phase (see Figures 5.5, 5.7, 5.10, and 5.13). Further, closer inspection of the Localizer task grand average plot indicates that the high oddball probability of 50% depressed key P300 waveform features. These stimulus parameters likely reduced the potential difference between Target and Non-Target data dramatically increasing the difficult of class separation via the LDA models.

It can be argued that replicating the main experiment as the localization task could have addressed many of the challenges encountered in this experimental series. This approach would likely increase the similarity between target (P300) and non-target (Non-P300) data across the localizer and main experiments, as the class-balanced pre-training dataset would then encompass instances of double-flash, adjacency errors, and other temporal/spatial noise effects inherent in these oddball paradigm designs.

Additionally, maintaining consistent oddball probabilities across both datasets would ensure greater alignment in P300 peak amplitudes and latencies. However, the increased noise and complexity in these reference signals, particularly in Non-P300 samples, could potentially introduce confusion into the LDA classifier, thereby impairing model performance. This issue could be mitigated by increasing the number of sequences per trial and the number of trials per experimental session. Such methods would be most effective when applied to the development of a large benchmark dataset for pre-training convolutional neural network models focused on pooled-subject data, rather than relying on single-subject LDA classifiers.

5.7.3 P300 Waveform Quality: Pipeline 1

Upon examination of the Cz grand average plots presented for Experiment 3, it is evident that the Pipeline 1 approach employed does not ensure robust and consistent pre-processing of P300 waveform samples. The LOCRT method yielded notably weak P300 signals and exhibited a Non-Target average signal with a similar amplitude (see Figure 5.5). The HYALL: No Class-Balancing Cz grand averages, computed from all samples across subjects, display an unexpected inversion of results, with a pronounced P300 waveform observed in the Non-Target signal (see Figure 5.7). Similarly, the HYALL: Class-Balanced Cz grand average reveals a comparable pattern, characterized by minimal variance in the Target samples' average and an atypical drifting and peaking component in the Non-Target samples (see Figure 5.10). The ineffective application of an adequate high-pass cutoff value (1 Hz) likely resulted in the removal of significant P300 components during pre-processing.

Additionally, the use of an infinite impulse response filter on these short data segments introduced considerable drifting and edge effects. The suboptimal quality of these plots undermines the validity of any potentially promising results, as the Cz averages suggest that the classifiers may not have effectively distinguished samples based on P300 waveform characteristics. These findings are consistent across both the online Localizer results (see Table 5.3) and the evaluations of samples exclusively from the main experimental data. Consequently, the author is unable to provide a definitive conclusion regarding the primary objective of this experimental variant, specifically in determining the impact of visual feedback on real-time or offline classification performance metrics.

5.7.4 Pipeline 2: Relative Influence

The most appropriate means of comparison for the Pipeline 2 computed Experiment 3 results are arguably the findings of the Pipeline 2 method for the 7-Emoji variant from Experiment 2 (see Tables 4.6 & 4.7) and the Pipeline 2 implementation for Experiment 1 (see Tables 3.10 & 3.11). In these contexts, the LDA models for both Non-Collapsed and Collapsed data partitions across the subjects consistently produced a similar distribution of overall mean classification accuracies, ranging from 75-85%. Ultimately, given the low number of trials per subject the relative performance of the Pipeline 2 Collapsed data preparation method could not be effectively assessed, as the number of trials is so low. This reduced the number of samples available for training, diminishing the quality of the classifiers and reducing the samples for testing, limiting the scope of any findings that emerge from the associated analysis.

Notably, no significant difference was observed between the Non-Collapsed and Collapsed data preparation methods based on paired-subject permutation tests (see subsection 5.5.7). This finding contrasts sharply with the differences highlighted in the results and grand average plots for the Pipeline 1 and 2 approaches outlined in the previous three chapters. The principal factor contributing to the overall improvement in data quality is likely the transition to a finite-impulse response filter design, which avoids the severe reflections induced by applying infinite-impulse response filters to short data segments. Additionally, implementing a lower high-pass cutoff threshold likely increased the inclusion of signal frequencies associated with the P300 waveform. Furthermore, the application of a reasonable baselining method helped centre the data around a common reference, enhancing data consistency.

Despite these improvements in data quality and arguable contingent increases in overall mean accuracy, the substantial presence of Target-P300 classification bias is concerning. The bias observed towards the Target-P300 class consisting primarily of synthetic samples dramatically

reduces the veracity of any strong conclusion made on these data as the lower relative variance within this class could have offered a unique advantage, as compared to the Non-Target class, which would struggle to translate to a real-time classification context. It is not possible to rule out the possibility that the relative increase in performance is driven primarily by the large proportion of synthetic samples in the respective LDA model training datasets. This could have been addressed by exploring a combination of downsampling and upsampling techniques for the majority (Non-Target) and minority (Target) classes, or the implementation of weighted regularization methods to accommodate for the aforementioned class imbalance.

Irrespective of this previous point, the scope of such additional investigations is necessarily narrow given the small sample size. Owing to project time constraints it was not possible to continue these assessments, however, the author asserts that the Pipleline 2 results demonstrate a strong trend and indicate that further improvements to the experimental paradigm could lead to the production of a high-performing emoji-based BCI system. These improvements include implementing a more comprehensively balanced randomization scheme could have enabled more effective data labelling methods such as those discussed in subsections 3.4.3.4 & 3.4.3.5. Further, any future adaptations must feature both more trips and more sequences per trial, potentially extending to 50 and 15 respectively. This would allow researchers to train and test the respective models more robustly and enable the implementation of more powerful and reliable statistical assessments to evaluate the methods and draw stronger conclusions.

5.7.5 Cross-Experimental Grand Summary

This series of experiments details the efforts of the author to develop a simple emoji-based BCI speller system for potential applications in the development of a functional communication system, as well as provide a platform for the collection of high-quality P300 benchmark datasets using Dry-EEG methods. The numerous adaptations implemented comprising the mitigation of model overfitting via class-balancing, staggered stimulus parameter modifications, real-time feedback relaying, and reactive impedance monitoring to enhance data quality and subject performance did result in a trend towards results improvement over the course of the experiments conducted. Despite these efforts, significant issues remain in terms of data quality and quantity, low ITR metrics and sub-functional classification performances.

Ultimately, the aforementioned obstacles precluded the consistent generation of P300 waveforms employed in this series of experiments. The consistency in the generation, classification and deployment of any such system must be of the highest level to instil confidence in both users and technicians. Without the implicit level of trust in the communication system de-

ployed motivation levels will undoubtedly be lower than needed to maintain the attentional vigilance necessary for adequate P300 generation. Importantly, at no point can it be stated with certainty that an effective means of inducing and analysing P300 waveforms for emoji-based communication was achieved throughout this research.

Some of the issues present in the Pipeline 1 approach have been remedied, namely the application of a more appropriate baselining measure, the implementation of statistical significance testing, the use of 10-fold cross-validation and improved high-pass filter design. The author notes that any future implementation of these methods must feature a higher number of augmentation events per sequence. Notably, classification accuracies for the Collapsed data preparation variant did not differ significantly from the Non-Collapsed method. It is the author's belief that this is because of the dramatic drop in the number of available samples induced via the aggregating of adjacent trials. Further, reimplementation of the final iteration of this experimental series must be conducted using a shorter version of the main, 7-Emoji, experiment and not the original 50% oddball probability localizer task.

Finally, a more appropriate and sophisticated method of addressing the class imbalance issues and overfitting present throughout Pipeline 1 results in the form of the SMOTE oversampling method. However, the noted potential influence of overfit due to synthetic samples introduced via the SMOTE method and the fact that these re-assessments were done with a small sample size, 3 subjects each across all Chapters (3, 4 & 5), again limits the generalization of these results. This is compounded by fixed issues that could not be remedied in a posthoc re-analysis. Namely, the issues surrounding the non-continuous sampling method, the use of a high oddball stimulus probability in the Localizer assessments and the lack of subjects imposed by the COVID-19 pandemic. Despite this, the author asserts that the platform does advance the current research in the field of emoji-based BCI communication systems. Previous work concerning the use of emoji as P300-based oddball stimuli is extremely sparse and the work undertaken herein relating to the numerous modifications in stimulus size, augmentation durations and visual array density serve as guidelines for future work.

Previous research surrounding the development of simplistic icon-emoji integrated communication systems is highly limited [44, 257]. Further, the design of these systems is primarily aimed at affording the user a means of communication with care staff regarding environmental preferences, for example, 'turn off/on the radio' as depicted by a radio cartoon image [257], or personal states such as expressing that the patient is hungry, as depicted via an apple icon [44]. The paradigm herein outlines the first BCI speller system explicitly focused on emotional ex-

pression via emoji to date. Notably, the orientation of the speller layout, arranging emoji from agreeable to disagreeable in a Likert-inspired format is novel. As seen in the Conclusion chapter (see subsection 7.2.1), the author presents an adaptation of the experimental design. It is asserted that the increased range of emotional expression beyond the pleasure-centric valance scale will maximise patient life quality by clarifying the user's emotional state and adding a much-needed degree of humanization and personalization to the associated communications.

The most advanced system developed to date, seen in [258], for P300-based BCI speller communication using emoji stimuli, features a parallel system of input grids. The individual grids are populated with differing arrays of BCI device commands such as a standard keyboard and system navigation controls to interact with a range of PC applications. The study here details that subjects were capable of navigating a simplistic computer interface to open the WhatsApp messaging application and then select and type an emoji from the in-app keyboard, all using the system control grid. This article details an impressive hybrid methodology based on P300 waveforms for the target selection and limited muscular control for the grid selection. The platform is principally aimed at individuals with severe multiple sclerosis (MS). The availability of these systems is key to maintaining high patient life quality for MS patient users. Despite this, such methods are not viable for the near-complete locked-in target population of the BCI speller platform detailed herein given the previously stated mobility limitations (see subsections 1.3 & 1.4).

Relating to the emoji speller outlined in this thesis, the platform provides a system for individuals with the most severe forms of paralysis, as well as potentially providing a baseline experiment for the collection of P300 data. The author asserts that with the addition of more subject data collected via an adapted iteration of Experiment 3 would likely prove a viable means of P300-based BCI communication. Validation of these systems would allow for the widespread utilization of these methods in the development of high-volume P300 data repositories. The collection of additional data would increase the scope for analysis applications that require large sample sizes to function such as convolutional neural networks which have demonstrated promising results in P300 classification tasks [259, 260].

5.7.6 Future Research

Here the author outlines important areas of future research relating to the P300-based BCI speller field focused primarily on enhancing data quality and lab-based ecological validity considerations.

5.7.6.1 Pre-Screening and Online-Monitoring Development

The quality of EEG signals is controlled by numerous means, principally based on hardware resolution and scalp seating, subject instruction and engagement, as well as signal preprocessing. The phenomenon of so-called BCI illiteracy is a contentious issue in the surrounding literature. Certainly, there exist unique characteristics to individual subjects that could explicitly diminish the quality of signal harnessed, these include poor skin conductance and a previous history of mental illness. Even for subjects that do not present with the aforementioned characteristics, the EEG data acquired from these individuals could be low-quality owing to situational factors, for example, time of testing and levels of fatigue. The methods currently employed in the literature must be greatly expanded, especially concerning the development of larger P300 waveform benchmark datasets for the evaluation of novel preprocessing pipelines and classification methods.

Pre-screening assessments before the onset of any experiment to monitor impedance values, attention levels and signal propagation are valuable means for assessing the viability of resultant data collected for inclusion into a given dataset. It could be argued that the expansion of these methods, specifically relating to subject attention, was implemented in real-time during the experimental phase, as per the reactive impedance monitoring conducted in Experiment 3 (see subsection 5.4.9). Eye-tracking metrics concerning fixation position relative to the known position of targets on screen could provide an indirect measure of subject engagement. It is likely that dwell time and the distance between fixation positions for a given trial period could provide real-time metrics of subject task engagement during live trials. Further, whole-brain sampling during a baselining assessment before the onset of the main experiment could reveal subject-specific alpha-theta ratio metrics, alongside additional neural signatures, that could be compared to data captured online to inform experimenters of data quality in real-time.

This would ensure that subject performance during the localizer or baselining assessments could maintain parity with data captured in the main experimental phase. Further, these metrics could also provide experimenters with data to inform decisions regarding the implementation of break periods to ensure subjects provide quality signals. Additionally, this information and guidance could assist in the rapid training of test subjects and enhance performance during the main experimental trials.

Implementing these precautions and reporting these implicit measures of the subject attentional state alongside open-source P300 benchmark repositories could also assist any subsequent dataset users in the development of tools specifically aimed at classifying these wave-

forms under less-than-optimal conditions, as is the ultimate aim of these systems. In other words, datasets could be parsed in terms of attentional state metrics and classification methods compared in terms of efficacy for these lower-quality signals. Along these very same lines, the provision of these auxiliary data could assist in the transition of pre-processing and classification methods evaluated using typical healthy subjects to target clinical populations.

5.7.6.2 Exploration of Real-Time Active Emoji Selection

Currently, any conclusions regarding the clinical functionality of the speller system defined herein are restricted by the cueing feature of respective task designs. Any attempts to continue this specific line of research would benefit by exploring the implementation of an active subject selection method. This would involve the subjects using the system as intended by fixating on emoji targets in response to specific questions posed throughout the experimental period. Crucially, this would likely enhance both subject attention and motivation during the experimental phase.

The implementation of additional experimental measures would be necessary, for example, providing the ability to amend returned characters in the event of misclassification. There would be some obstacles related to order effects, as the strict non-consecutive randomised cueing protocol would no longer be in effect and could introduce ordering issues, for example, the subject could select the same emoji repeatedly. This would likely increase the amount of supervision by researchers and the degree of subject training before the task. Researchers could instead try to operate the task in a more ecological context, by first presenting a simple question to the subjects, for example, 'How does tiredness make you feel?', the answers to these responses would be obvious and the generic expected negative response to the question could be predicted and in turn inform the arrangement of subsequent questions following a relatively analogous randomised non-consecutive protocol as previously implemented. This solves the issue of the subject having no agency in the task, while also priming responses with a specific directionality to avoid unwanted order effects. Further, the task could simply continue indefinitely until the subject provides a balanced number of responses evenly selecting each target on the screen.

Finally, the integration of emoji or other iconographic targets into a pre-existing alpha-numeric array would dramatically enhance the functionality of any BCI-based speller system. The precision of text-based communications embellished with pictorial icons would enhance the speed of speller systems in addition to the richness of the messages generated.

Chapter 6

Subject-Specific Signal Pre-Processing Network Optimization

6.1 Chapter Outline

This thesis chapter aims to outline the efficacy of automated optimization procedures in signal pre-processing hyper-parameters for enhancing convolutional neural network performance in the classification of SSVEP bio-signals. Recent advances in this field have primarily been explored via the development of novel architectures for the specialized classification of these time-series data. There are now a myriad of different convolutional neural networks currently available to researchers, with well over 30+ design configurations detailed to date [214, 261]. This often involves increasing network depth or the application of previously unviable, computationally expensive regularization measures such as drop-out [199] and batch normalization [262] layers which ultimately increase the number of system parameters [214]. It must be noted that additional advances include the development of novel operational units or network modules, for example, the blending of data across the temporal and spatial dimensions via the so-called Depth-Wise operation [60] or the use of Inception-style modules for advanced intra-network layer connections [263]. The author recognises the value of these techniques in specializing networks for the classification of modality-specific inputs.

Substantial efforts have been made to integrate numerous model design improvements that enhance classifier performance in computer vision tasks. This includes adaptive learning rates (Adam [264], AdamW [265]), weight initialization schemes (Xavier [266], Kaiming [267]), transfer learning and network ensembles. This systematic approach to identifying the optimal network parameters has greatly improved cutting-edge models in the field of CNN-based bio-

signal classification. Along these very same lines, the author believes that extending the same diligence to the pre-processing of input data could also improve the performance of current networks available for study in addition to allowing for a greater level of clarity in comparing the performance across said models.

The key defining characteristic of EEG data is the presence of non-stationarity and a high degree of individual differences. For BCI systems that utilize single-subject calibration procedures the non-stationarity of the EEG signal often leads to a drop in performance over time as the baseline samples collected before the experimental period begin to differ in their quality of expression, resulting in lower classification accuracy [268]. Understandably, given the complexity of accounting for intra-subject variability, the task of optimizing a classifier to accommodate inter-subject variance in cross-subject decoding applications is compounded.

Despite this, the current EEG classification literature spanning BCI, affective decoding and epilepsy detection has seen the widespread adoption of cross-subject classification methods [269]. This is primarily driven by the aim to develop plug-in-and-play, hyper-generalizable methods that do not require extensive subject-specific calibration. Further, it has been asserted that the training of high-accuracy networks using cross-subject, aggregated datasets is possible as long as adequately large volumes of data are utilized in combination with the appropriate data balancing and test-set isolation protocols [269]. Moreover, it is argued that the use of cross-subject datasets can even afford the resultant models a greater degree of robustness, given the exposure to a larger incidence of variance in the training samples [270].

It must be noted that, despite the aforementioned advantages in cross-subject training schemes, it does not preclude the possibility that all preparations of these aggregate datasets for a specific subject should result in comparative levels of performance. The tuning of global signal pre-processing parameters employed across all subject data such as zeroing, channel averaging, smoothing and filtering could all greatly influence end-point performance for the end-point user. Further, given the high degree of individual differences in the expression of brain-based bio-signals, it may be profitable to prepare cross-subject training data for specific individuals, combining the strengths of both aggregated dataset volume training and bespoke model development. As shown in [53, 54, 57], this combined approach has resulted in the highest performing SSVEP-based BCI speller classification results thus far. For further information see subsection, 2.6.4.

In the following sections, the author will compare the performance of four CNN models on raw (notch filtered, 50 Hz), fixed-parameter (9-30 Hz bandpass filtered) and pre-processing parameter optimized data using three algorithmic search methods for high-pass filter (0-9 Hz) and low-pass filter (15-85 Hz) cutoff values. This final optimized parameter search will involve the comparison of the Optuna Python package [271] hyper-parameter tuning methods: Median, Percentile and Successive Halving pruners (SHP) [272]. The following investigations will reveal the most time-effective means for performing optimization tuning to boost classification accuracies. Further, these processes will enable a more thorough evaluation of the networks tested (EEGNet [59], EEGNetSSVEP [60], DeepConvNet [61] & ShallowConvNet [61]) as currently the selection of fixed filtering parameter thresholds is either arbitrary (industrial standards), theoretical or vague, given the absence of data on these topics in the current literature.

Note, that all the models assessed herein are relatively well-established and have arguably seen some of the most widespread reimplementation of any contemporary convolutional neural networks in BCI contexts. Despite this, at the time of writing these models were not made publically available alongside corresponding sets of pre-trained weights. Given these circumstances, all models were trained from the ground up using the open-source SSVEP data repository acquired from [180]. For more information on the origins, design and performance of these networks please refer to subsection 6.6.4, Convolutional Neural Network Summaries.

The application of hard-coded pre-processing parameters may be possible due to the highly flexible nature of CNN models. It could be argued these capabilities can overcome any small adjustments in parameters. The author would argue that any significant performance improvement is crucial to effectively compare models. For example, larger networks could potentially be optimized for higher low-pass filter cut-offs and be capable of harnessing additional latent information within the high frequency ranges such as second and third order harmonics as is considered in the design of Filter-Bank Canonical Correlation Analysis methods. Further, the restriction of low and high pass filters to the edges of the target frequency space might be more optimal for shallow networks, by removing higher complexity signals embedded in the low and high-frequency space, effectively focusing the training data.

The ultimate aim of these experiments is to determine the effects of signal pre-processing parameter optimization schemes on subject-level performance for a given SSVEP dataset [180]. Secondly, the author aims to gauge the differences in optimal parameter selections across shallow and deep convolutional neural network configurations. Finally, the author will pro-

vide guidance and improvements on how the method can be performed correctly and suggestions for multiple improvements in this methodology that lie outside the scope of the current research.

6.2 Bandpass Filtering Background

The optimal detectable range of SSVEP waveforms is located between the 8 to 15 Hz frequency range [161] (see subsection 2.6.1). The aim of bandpass filtering is to maximise the power of these signals relative to data acquisition artefacts and other associated noise components to facilitate higher performance in the respective waveform classifiers. A high-pass filter is employed primarily to reduce the incidence of eye-movement artefacts and sub-target frequency harmonic reflections in the ultra-low frequency range (0 to 4 Hz) [273]. In comparison, the low-pass filter is designed to remove non-essential so-called high-frequency noise [274]. Currently, researchers have deployed SSVEP bandpass signal pre-processing frequency filters between several boundaries ranging from 0.1 and 80 Hz.

Typically, FBCCA methods configured to harness high-frequency harmonic components utilize larger low-pass filter cutoffs, between 70 and 90 Hz, to ensure the availability of these signals to the corresponding classifier method [47, 56, 179, 180]. In contrast, CNN-based methods have traditionally utilized a narrower frequency band with low-pass cutoffs mainly positioned between 30 and 40 Hz [59, 60, 213, 275–277]. This is related to the reduced efficacy of these methods for the extraction of task-relevant harmonic components above these frequency thresholds. Despite these apparent restrictions, the frequency filter bandwidth employed has slowly increased over time due to improvements in the quality of EEG data acquisition, data pre-processing and the relative power of the models implemented in terms of architecture, trainable parameters and network layer count [50, 57].

Note, that the upper low-pass value threshold is dependent on the relationship between the stimulus frequency and the data acquisition sampling rate. For the sampled data to accurately reflect the signals propagated from the source the sampling rate must be double the frequency of the highest frequency stimulus flicker rate. This limit is termed the Nyquist frequency. In the simplest terms, an SSVEP signal oscillating at 50 Hz would require at minimum a sampling rate of 100 Hz to ensure the effective representation of the source signal in the captured data. Any decrease in the sampling rate would introduce aliasing noise, this describes the failure to accurately reconstruct a signal due to missing data.

Recent advancements in EEG hardware mean that sampling rates of up to 1000 Hz are achievable in clinical-research-grade devices. In turn, the upper limit of this low-pass filter can be increased dramatically. The well-established practice of signal downsampling to reduce the expression of redundant waveform features such as drift and movement artefacts precludes researchers from taking full advantage of these increased sampling rates. Despite this, the opportunity to capture latent information within these higher frequency ranges is possible yet remains relatively unexplored in the context of convolutional neural networks.

Several studies demonstrate the potential of higher low-pass filter thresholds to improve classification performance via the inclusion of additional target SSVEP frequency harmonics [52, 54, 180, 256]. As seen in the table below, a dramatic increase in available harmonics (60+) is achievable with the raising of the low-pass cutoff threshold from 30-80 Hz. Notably, there exists a crucial trade-off in terms of signal-to-noise ratio following the increase in cutoff value. Any increase in the low-pass filter threshold ultimately reduces the proportional expression of the target waveforms in the input data. This places additional demands on the classification systems employed to parse the redundant data and extract useful harmonic information.

Target Hz	H 1	H 2	H 3	H 4	H 5	H 6	H 7	H 8
9.25	18.5	27.75	**37**	**46.25**	**55.5**	**64.75**	**74**	*83.25*
9.75	19.5	29.25	**39**	**48.75**	**58.5**	**68.25**	**78**	*87.75*
10.25	20.5	**30.75**	**41**	**51.25**	**61.5**	**71.75**	**82**	*92.25*
10.75	21.5	**32.25**	**43**	**53.75**	**64.5**	**75.25**	*86*	*96.75*
11.25	22.5	**33.75**	**45**	**56.25**	**67.5**	**78.75**	*90*	*101.25*
11.75	23.5	**35.25**	**47**	**58.75**	**70.5**	**82.25**	*94*	*105.75*
12.25	24.5	**36.75**	**49**	**61.25**	**73.5**	*85.75*	*98*	*110.25*
12.75	25.5	**38.25**	**51**	**63.75**	**76.5**	*89.25*	*102*	*114.75*
13.25	26.5	**39.75**	**53**	**66.25**	**79.5**	*92.75*	*106*	*119.25*
13.75	27.5	**41.25**	**55**	**68.75**	**82.5**	*96.25*	*110*	*123.75*
14.25	28.5	**42.75**	**57**	**71.25**	85.5	*99.75*	*114*	~~128.25~~
14.75	29.5	**44.25**	**59**	**73.75**	88.5	*103.25*	*118*	~~132.75~~

Table 6.1: Here is shown a table of harmonics for the selection of target SSVEP frequencies (far-left column) used in [180] to the 8th order. Each respective column, denoted H X (X representing the order), is a multiple of the target frequency (Target Hz). The table cells displayed in plain (unbolded or italicized) type represent all the harmonics captured from an EEG signal sampled at 256 Hz and band-pass filtered between 9 and 30 Hz, as per the pre-processing methods implemented by the developers of the EEGNetSSVEP network [60]. The bold cells indicate the harmonics available for inclusion following an increase in the low-pass filter threshold from 30 to 85 Hz. Note, that access to these additional harmonic frequency components has been shown to dramatically increase classification performance for the FBCCA method. The italic cells show the harmonics that remain excluded despite the increase in the low-pass filter threshold, as these frequencies are below the 85 Hz low-pass filter cutoff. Finally, the crossed-through cells highlight the filtering limits imposed via the Nyquist frequency as these are greater than half of the data sampling rate (256 Hz) and in turn can not be accurately represented owing to the downsampling operation implemented.

6.3 Model Optimization *vs*. Architecture Development

As mentioned above, the current emphasis in the literature is primarily related to the development of novel statistical methods, particularly neural network architecture design as seen for the 1DSCU [212], EEGNet [59], EEGNetSSVEP [60], ShallowConvNet and DeepConvNet [61], IENET [263, 278], FBCNN [279] and TRCA-Net [54]. The author suggests that more attention is required to evaluate the performance of current networks by comparing the many variants under so-called ideal conditions. In other words, it is suggested that significant improvements in the current model architectures could be enhanced via bespoke data signal pre-processing at the single-subject level. This involves developing cross-subject aggregated

datasets tailored for the training of networks for the individual end-point user. To determine the ideal parameters for these data preparation stages a systematic optimization method is presented herein.

Optuna is a Python software wrapper used for hyper-parameter tuning typically in neural networks and similarly aligned classification methods [271]. Typically, the wrapper is positioned in a top-down orientation with a specific statistical protocol (see, Figure 6.2). For example, the validation loss of a neural network could be iteratively input into a corresponding pruner algorithm to log the performance of the model for a classification task. Once the model is trained a second run is performed using a new instance of the model (reset filter weights). Before the onset of the new run, a specific parameter of the network training scheme is altered. This could be any one of the innumerable metrics used in the configuration of the network, popular optimization training parameters include: learning rate, batch size and loss function.

Alternatively, the network architectural parameters could be optimized, such as drop out, momentum, stride and even filter size. Note, that these can require more complex reactive adjustments if embedded in an automated optimization approach due to changes in the dimensions of the data through the model. During this process, the pruner algorithm is programmed to compare the current model performance epoch-for-epoch with the previous model performance. In the instance that the given change in optimized parameter leads to a relative reduction in performance, the trial would be terminated, and a new metric value selected. In contrast, a relative increase in performance as displayed by a larger drop in validation loss over the same number of epochs would lead to the continuation of the optimization run. Through this process, the classification performance of the network is maximized.

The application of these automated parameter search methods requires significant computational and time resources [280]. Despite this, the methodology is far more effective than the common instance of researchers copying the parameters utilized in previous articles for data involving similar tasks or data modalities [281]. The utilization of prior knowledge from previous research for effective boundary pre-setting is of course crucial in saving time, as the implementation of the the widest possible search ranges would incur significant penalties in resource usage [282]. The author here suggests that the optimized search of signal pre-processing filter cutoffs within and marginally beyond the bounds previously explored in the associated literature would prove beneficial in clarifying several open-ended questions. Namely, this would assist in verifying whether the commonly implemented ranges are valid, further, it would aid in identifying if there exist significant range preferences between networks of differing com-

plexity and also establish the degree of individual differences in these optimized filter ranges. Moreover, outlining a standardized method for the use of a tuning wrapper blended into a network training and evaluation procedure could save researchers countless hours of valuable research time by automating the process.

It must be noted that the relationship between network parameters is sometimes adversarial. In other words, the assignment of one value can undo the efficacy or intent of another. For example, the merits of a low batch size have been discussed at length in the literature [283] in addition to the benefits of a low learning rate [284], as both can contribute to the development of models with a greater degree of generalizability and higher task performance. Despite this, setting the batch size too low can mitigate the benefits of a low learning rate by increasing the tendency of trained model weights to become stunted inside local as opposed to global minima. In other words, the interaction of these parameters can prevent models from attaining the lowest achievable loss values and in turn fail to attain the highest possible classification accuracies.

It is these intricacies of hyper-parameter value selection that slow down the already often arduous process of neural network architecture development. The ability to accurately assess the best parameters for a given task from the outset is highly dependent on the skills present in the researchers involved. The capacity to perform this feat adequately drops dramatically during the implementation of novel techniques in tandem with a lesser-studied data modality. It is ultimately in the interests of researchers to speed up and systemize the network development process as well as provide tools to new users of these models to enhance the adoption of neural networks across disciplines.

The author argues herein that the optimization of network training should be extended to all facets of the project that account for end-point classification performance, including data preparation. There exist several pre-processing stages available to researchers for SSVEP-based EEG preparation, the focus of these analyses is the selection of bandpass filter frequency cutoff values. Crucially, by applying these principles to CNN training schemes it will allow researchers to effectively evaluate the current state-of-the-art models with enhanced clarity.

Currently, there are cloud-based solutions for the evaluation of network architectures to compare model efficacy for a range of different bio-signal classes, for example, the Mother of all Benchmarks Python-based software platform [216] (see subsection 2.6.1). These are critical tools for the research community to assess the viability of any given model for a specific classi-

fication task. Despite this, little consideration has been given to the task-specific optimization of these models before deployment. This could be due to a project pre-requisite for models to have a wide range of applicability to any number of bio-signals utilized out-of-the-box. In other words, the aim of these tools could be to evaluate the plug-in-and-play functionality of the respective model, without laborious optimization procedures. It is asserted by the author that the optimization of networks for these same goals is still necessary, as it is feasible that a specific confluence of hyper-parameter values and conditions is necessary to optimize model performance for these forms of implementation. In other words, models likely still require optimization for broad out-of-the-box functionality contexts, and the absence of such assessments before conclusions on the efficacy of a given model could leave viable network architectures erroneously labelled as sub-optimal. Essentially, the hyper-parameter tuning of models for several bio-signal classification contexts is necessary to improve the validity of any claims regarding the efficacy of the evaluated networks.

6.4 Current State-of-the-Art Classification Techniques

In BCI research SSVEP-based EEG classification methods are designed to identify specific frequencies by isolating target waveforms from background noise. The current benchmark, Filter-Bank Canonical Correlation Analysis (FBCCA) and its variations employ bandpass filtered data matrices to correlate with reference signals, ultimately determining the most likely target class. Notably, these methods achieved information transfer rates over 100 bpm [285], and recent advancements have pushed these rates beyond 300 bpm [55] (see subsection 2.6.1). Recent studies indicate that convolutional neural networks (CNNs) have significant potential to surpass FBCCA methods, especially when deep architectures are used and trained on large datasets, often leveraging transfer learning [215]. Unlike FBCCA methods, which sometimes integrate reference signals with subject data [180, 181], CNNs develop internal representations of target classes through iterative training on aggregated data. This research focuses on evaluating different CNN architectures for SSVEP classification and optimizing hyper-parameters for both cross-subject and single-subject scenarios, aiming to enhance classifier performance and robustness.

In EEG signal processing, common methods include temporal segmentation, time correction, signal referencing, and active electrode selection. While these techniques are standardized, the choice of filter cutoffs can still be somewhat arbitrary. Although some studies have explored low-pass filter values and stimulus frequency harmonics [218–220], there has been limited research on how these filters impact CNN performance. It is hypothesized that setting high-

pass filters close to target frequencies helps reduce unwanted low-frequency noise, though the optimal low-pass cutoff remains less clear. A higher low-pass cutoff may benefit models depending on their complexity and depth [47, 56, 180].

The relationship between high and low-pass filter cutoffs has not been extensively explored, suggesting the need for further investigation. This study aims to clarify the impact of preprocessing on CNN performance and guide future model calibration. Given the variable nature of EEG signals, finding the 'perfect' filter settings is challenging, but automated parameter optimization could identify effective values for individual subjects. This research presents a methodology for simultaneously optimizing signal processing and neural network parameters to improve classification performance at the single-subject level.

6.5 Experimental Investigation

Throughout the research defined herein, the author investigated the performance of 4 well-established convolutional neural networks, EEGNet [59], EEGNetSSVEP [60], ShallowConvNet [61] and DeepConvNet [61] for the classification of a globally recognized SSVEP repository in 3 contexts: raw/minimal signal pre-processing, standard literature-derived pre-processing and automated signal processing using optimized hyper-parameters at the single-subject level. Further, 3 different optimization pruning algorithms are assessed for optimization speed and quality. The following subsections outline the data repository utilized, a brief description of the models employed and a summary of the pruners applied for network optimization.

6.6 Methods

Here is presented the methodology for implementing the signal pre-processing hyper-parameter optimization of the 4 convolutional neural network architectures selected. This includes a description of the SSVEP online data repository utilized, a breakdown of the classification models and a discussion surrounding the software features of the optimization system deployed.

6.6.1 Online Data Repository

The data used herein are derived from a well-established online SSVEP repository [180]. This consists of data collected from 10 subjects (1 female, 9 males, mean age: 28 years) for 12 SSVEP target waveforms ranging between 9.25-14.75 Hz in 0.5 Hz increments (see, Table 6.2). Note, groups of distally positioned target waveforms are modified in terms of phase to

minimise target correlations. These data were collected via visual presentation of flickering SSVEP stimuli on-screen in a virtual numpad using the Joint-Phase-Frequency Modulation (JPFM) method (see, Figure 2.2). The JPFM enhances the number of flicker rates by adjusting the phase angles of target flicker profiles and rapidly switching between different signal frequencies, such as 10 Hz and 15 Hz. This approach minimizes signal correlation and produces a 12.5 Hz SSVEP oscillation through an averaging effect. Subjects, cued by a randomized red square overlay, completed 180 trials each, resulting in 1,800 total trials across 10 participants. EEG data were recorded at 2048 Hz using the BioSemi ActiveTwo system across 8 channels (O1, Oz, O2, PO7, PO3, POz, PO4 and PO8) and were downsampled to 256 Hz from 2048Hz.

Target Class	Frequency (Hz)	Phase (π)
1	9.25	0
2	11.25	0
3	13.25	0
4	9.75	0.5
5	11.75	0.5
6	13.75	0.5
7	10.25	1.0
8	12.25	1.0
9	14.25	1.0
10	10.75	1.5
11	12.75	1.5
12	14.75	1.5

Table 6.2: The table displays the corresponding frequency and phase angle for each of the 12 target classes. The frequency value here denotes the rate at which targets flicker from all white to an inverted black per second. The phase angle values denote the point in the frequency cycle a given signal is initiated at. A phase 0 indicates that the signal is initiated at the start trough of the frequency cycle (0°). Further, the phase metrics of 0.5, 1 and 1.5 represent phase angles of 90°, 180° and 270° respectively.

Crucially, it must be noted that the training scheme used to develop the results reported here differs significantly from those detailed for the networks herein. The maximum number of training samples required to generate the subject-specific Combination CCA reference signals was just 12. Further, all signals were bandpass filtered between 6-80 Hz. This differs markedly from the fixed parameters used for the Fixed Parameter assessments detailed herein (see subsection 6.7.2).

6.6.2 Software and Equipment

All the results generated herein were initialized with the environment requirements outlined in the Lawhern Army Research Labs GitHub repository [286]. Specifically, this featured the use of Miniconda Python 3.8 alongside CUDA Toolkit 12.1. The Keras (version: 2.10.0) [287] Python module was used to deploy all models noted in this analysis. Optuna (version: 3.1.0) [271] was implemented to perform the pre-processing hyper-parameter optimization search methods and SciPy (version: 1.10.1) [226] was utilized for all signal filtering applications. Note, that all the networks defined herein were trained and evaluated using a NVIDIA 1080ti GPU, 11GB VRAM.

6.6.3 Optimization Datasets

Here is presented an outline of all datasets utilized in the optimization studies discussed. This includes a description of the raw and fixed parameter data alongside a breakdown of the pre-processing stages implemented.

6.6.3.1 Raw Data

To provide a baseline for the comparative evaluation of the fixed and optimized pre-processing parameter datasets the author developed a so-called 'raw' iteration of the data. These assessments were conducted to establish the robustness of CNN models specialized for the classification of inherently noisy EEG data. Previous research has demonstrated the capacity of CNN architectures to develop highly complex data representations. Along these very same lines, some model variants may not require significant pre-processing of input data. The author thought it prudent to assess this possibility from the offset.

Despite the indications of the moniker here assigned ('Raw Data'), the corresponding samples did undergo some fundamental pre-processing treatments. As in all subsequent datasets, a redundant section (0.15ms) was removed from the start of the trial. Following this, each 4-second trial was split into 1-second data chunks. This was done primarily to increase subsequent information transfer rates (ITR) and in turn, the total number of trials for model training and testing [61]. Further, the data as provided via the repo was already pre-treated with a 50 Hz notch filter to remove powerline noise (see, Figure 6.1, upper left quadrant). To clarify, all subsequent datasets underwent this set of pre-processing stages.

Owing to the data segmentation here noted, the total number of samples increased to 7200. Further, 720 samples are assigned to each subject, constituting 60 samples for each of the 12 target frequencies. This reflects the volume of samples utilized in all datasets throughout these analyses. Additionally, all analyses are evaluated according to a leave-one-out cross-validation method. All models are trained for one subject exclusively and all samples relating to this subject are isolated into a test set after all pre-processing stages. The remaining data are then randomised into 3 distinct k-fold sets for the purpose of metric averaging and evaluation (see subsection 6.6.6.2).

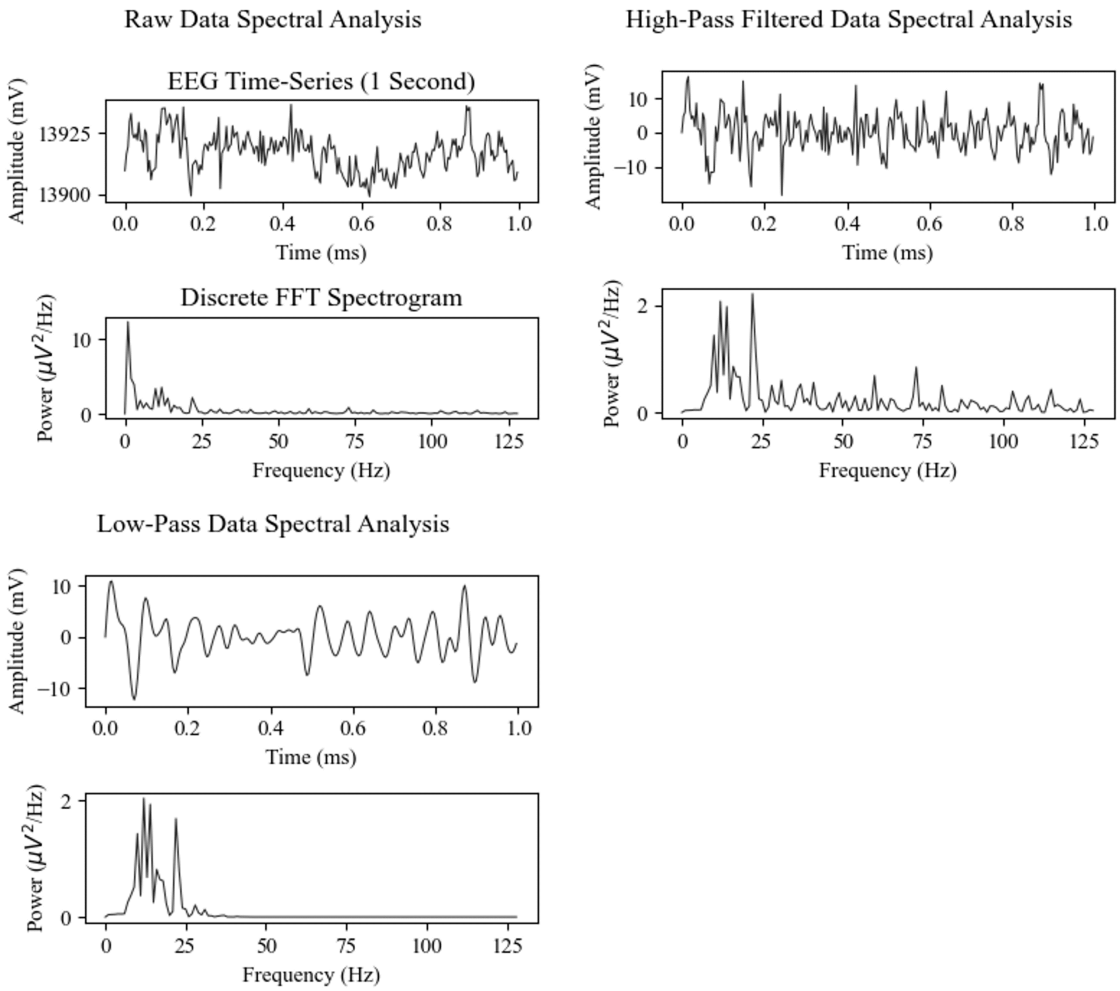

Figure 6.1: Here is presented a series of dual plots consisting of an input EEG signal (1-second) and corresponding discrete FFT spectrogram. The signal is derived from Subject 3 as a representative sample to illustrate the differences between a raw (top-left), 9 Hz high-pass filtered (top-right) and finally 9 Hz high-pass + 30 Hz low-pass filtered (bottom-left) EEG signal. The dual plot in the top left is an example of the data loaded immediately from the online repository. As seen in the lower of the two plots, the absence of any 50 Hz powerline noise spiking component reveals the data has already been notch-filtered. Further, the presence of some small peaks towards the upper end of the frequency spectrum and numerous high-magnitude peaks around 0-1 Hz suggest no other filtering steps have been implemented. Further, the influence of the filtering stages in terms of baselining and frequency representations is clear. The top-right upper plot demonstrates substantially reduced drifting and a mean of 0μV, as compared to the top-left upper plot. The reduction of sub-9 Hz components is also confirmed in the lower spectrogram plot. The lower-left dual plot displays a far cleaner EEG signal ready for classification, possessing frequencies restricted to the 9-30 Hz sub-band.

In the case of the so-called Raw Data variant, each data chunk was baselined to zero via mean subtraction. This was done as, in the absence of filtering, the data present significant differences in EEG μV amplitude ranges (see, Figure 6.1, top-left). This alone could preclude any effective training for the network models as the CNN filters can prove highly sensitive to input data scaling. Note, that these steps were not undertaken for either the fixed or optimizer-selected hyper-parameter datasets. This was avoided as the step is essentially redundant given the application of the high-pass and low-pass filters (see, Figure 6.1, top-right quadrant) corrects for these differences in μV amplitude.

6.6.3.2 Fixed-Parameter Data

As an additional point of comparison, a fixed parameter data variant was generated by pre-processing the samples using the same parameters implemented in the training of the EEGNet SSVEP network [60]. This was done as it is assumed these authors optimized both the network architecture and the pre-processing parameters specifically for the classification of the SSVEP bio-signal. Further, the architecture referenced is the highest performing in terms of classification accuracy and ITR of all networks implemented in this study and crucially, for the same SSVEP dataset repository [180]. It is intended that the use of these parameters as a fixed baseline assessment will allow for the comparison of the models tested under deployment conditions that maximise respective performance metrics. For this fixed parameter variant all data chunks were pre-processed bidirectionally between 9 (high-pass cut-off) and 30 Hz (low-pass cut-off) using a zero-phase forward 3rd-order Butterworth filter (see, Figure 6.2). These filters were employed via the SciPy filtfilt function [226]. Note, as stated above the data were not zeroed to baseline via mean subtraction.

6.6.3.3 Optimized-Parameter Data

The so-called Optimized Parameter data refers to samples pre-processed on the fly during the optimized frequency filter cut-off value search. These data undergo the same pre-processing stages as noted above (see subsection 6.6.3.2), differing only in the cut-off values utilized for the low-pass and high-pass filtering stages. Further information on the integration of this process with the wider parameter search optimization method is presented below (see subsection 6.6.5).

6.6.4 Convolutional Neural Network Summaries

In the following analyses, four convolutional neural network variants are trained and evaluated utilizing the three data variants outlined above. These four networks were selected due to ease of implementation, diversity of network sizes (number of trainable parameters and network layers) and degree of SSVEP-based architecture specialization. In the discussions below a brief description of each architecture composition and the respective development contexts are presented.

To clarify, given the lack of availability for model pre-trained weights at the time of writing all networks detailed herein were trained from the ground up using the open-source data repository provided by [180]. All models were trained using cross-subject data via a 3-fold cross-validation method (see subsection 6.6.6.2: k-Folding). Here the training data was exclusively comprised of samples from 9 out of all 10 subjects and then evaluated using the remaining subject. The duration of the training period varied substantially depending on the network and the optimization pruner algorithm implemented (see, Table 6.12). This is primarily attributed to an interaction between the number of trainable network parameters and the features of the optimization protocol methodology, for further information see subsection 6.8.4.3: Optimizer Study Durations.

Note, that as all pruner-network combinations were trained on datasets only differing in terms of the exclusion of one subject at a time there were no substantial intra-combination differences in the duration of the optimization studies between them. Here the subject-specific model training durations can reliably be computed as a fraction of the total given study duration. As shown in Table 6.12, for the EEGNet model using the Median Pruner this can be computed as $37.94\,hours/10\,Subjects = 3.8\,hours$.

6.6.4.1 ShallowConvNet

The ShallowConvNet variant is derived from research undertaken to determine the appropriate conditions for replicating so-called end-to-end CNN training methods, originally developed for computer vision tasks, in EEG-time series data. The original study aimed to demonstrate that minimal data pre-processing for deep convolutional neural networks is sufficient to exceed the classification performance of, at the time, a state-of-the-art filter-bank common spatial pattern (FBCSP) analysis method for motor-imagery-based bio-signals [61]. Traditionally, CSP is typically applied in the binary classification of synchronized and desynchronized states in motor-imagery-based data [288, 289].

Moreover, extensions to the method that enable the classification of more target signals (>2) have been developed including the Pair-Wise, One-Versus-Rest and Divide-and-Conquer methods [290]. Specifically, the FBCSP method initially involves band-pass filtering the data into numerous bins. Typically, these are limited to 4 Hz increments between 4-40 Hz. Following this, the separate bins are processed via the common spatial pattern algorithm to calculate data features unique to the filter frequency bounds used. A discriminative feature selection method is then implemented to identify components with the greatest variance, and finally, these results are input into a classifier.

The ShallowConvNet was developed as a baseline for the CNN model assessments and the design of the network was intended to replicate many of the processes found in the FBCSP technique. The first layer consists of a convolutional layer implemented as a temporal filter (13 $\times$ 1 kernel), followed by a spatial convolution (40 $\times$ 44), an average pooling operation and a logarithmic activation function. A final dense layer paired with a SoftMax operation generates the output network prediction (see, Appendix: Figure A.13). Note, that some adaptations to model kernel sizes are employed to ensure the functional operation of the network, namely the initial and second convolutional kernel dimensions are modified from 1 $\times$ 25 to 1 $\times$ 17 and 44 $\times$ 40 to 40 $\times$ 30. These were introduced via the authors of [59] and as stated in the relevant GitHub repository and associated open-source documentation [286], these modifications were not directly verified by the original research team. Despite this, owing to minor adjustments in kernel size, as opposed to significant architectural changes, these modifications are deemed by the authors of [286] and this thesis as minimal. In sum, the implementation and resultant performance metrics are broadly representative of the original network iteration.

The model was initially evaluated utilizing two exclusively motor-imagery-based datasets, BCI IV dataset 2a [291] and 2b [292] alongside an in-house high-gamma dataset (HGD) (0-125 Hz & 4-125 Hz bandpass filtered data). Each dataset consisted of 2, 4 and 6 motor imagery states respectively including, resting, or lifting right/left arms or feet. Further, the models were evaluated on an open-source blended dataset, Mixed Imagery Dataset (MID), that also included mental rotation and word generation. A brief outline of these datasets is necessary to clarify the development context of the ShallowConvNet and DeepConvNet (see, the following subsection 6.6.4.2). Crucially, these models do not represent SSVEP-specialized networks.

In typical applications, CSP analysis involves whole-brain electrode arrangements (18+ locations) to identify robust differences in time-series data over proximal and distal spatial locations. Despite this, both the ShallowConvNet and FBCSP are designed for the classification of oscillatory brain-based bio-signals. Further, as stated concerning the MID, these models were also evaluated in non-motor-imagery bio-signal contexts, providing credence to the assertion that the techniques are not domain-specific, even for non-oscillatory neural patterns. Finally, despite being inspired by FBCSP methods, the CNN models herein are not explicitly or purely performing FBCSP-based operations and as shown in the subsequent results sections, the models perform well outside of the originally intended context.

Interestingly, the data pre-processing undertaken across all available samples was, in nearly all instances, highly minimal. The time-series data was either input in full-bandwidth format, with no frequency-based filtering or only with a high-pass filter at 4 Hz to remove eye-movement artefacts. Relating to the HGD, the absence of low-pass filtering was undertaken primarily as a means of increasing the possibility of extracting high-frequency movement execution-related components. Similar principles were applied in the advancement of the FBCCA method (see subsection 6.6.1), this involved applying a far higher low-pass filter bound (80 Hz) to extract target SSVEP frequency harmonics and boost classification performance. Based on these design considerations the author asserts that the ShallowConvNet [61] and DeepConvNet [61] are likely to outperform the more specialized EEGNetSSVEP [60] model in the raw data variant evaluations.

6.6.4.2 DeepConvNet

As noted above, the CNN models developed in the reference article [61] significantly outperformed comparative FBCSP methods in the classification of oscillatory, multi-class, motor-imagery-based data. In the case of the DeepConvNet architecture, this was achieved via the implementation of leading contemporary CNN training and design features including, batch normalization [262] and novel optimizers (Adam [264] and Adagrad [293]) as well as operational architecture advances including drop-out [199] and exponential linear units (eLU) [294]. Crucially, as the available pool of training data for motor imagery tasks at the time prevented the utilization of the most powerful computer-vision-based networks, the authors designed numerous models of differing complexity and assessed these classifiers using alternative training schemes.

The DeepConvNet features four operational blocks arranged in a series of convolutions and max pooling layers (see, Appendix: Figure A.14). The first layer convolutional kernel is 1

dimensional or 'channel-wise' in design (10 $\times$ 1). As noted in the ShallowConvNet, this functions as a temporal filter and reduces the dimensionality of the input EEG data. Before max pooling, the input is convolved by a secondary unit (kernel 25 $\times$ 25) and subsequently fed to the ELU activation function. Following this, traditional 2-D kernels and max pooling layers are employed. These increase in size as per standard deep convolutional neural network architectural conventions (25 $\times$ 50, 50 $\times$ 100, and 100 $\times$ 200, respectively). Finally, the same dense layer is implemented in combination with 12 SoftMax units corresponding to each target SSVEP frequency class in the data repository. Note, as discussed above, the kernel sizes for each convolution have been modified to accommodate the new input data format. Despite this, the design of the network broadly mirrors the original.

Crucially, the authors state that the architecture of the network was purposefully developed using standardized DCNN architecture design principles. This was done to allow for the seamless integration of novel, cutting-edge advances in DCNN operations and to maximise the scope of network application across numerous bio-signal classification tasks. Since the publication of this article, these adaptations could include weight initialization methods for example the Xavier [266] or Kaiming [267] methods and the application of AdamW [265] optimizers.

6.6.4.3 EEGNet

The EEGNet architecture was initially developed to provide a highly flexible EEG-based bio-signal classification method across P300, sensory-motor rhythms, error-related negativity, and movement-related cortical potentials [59]. The approach was benchmarked against a comparable DCNN model as a baseline and equalled or surpassed the DCNN in both standard and restricted training data conditions. The EEGNet model is composed of two operational blocks, Firstly data are convolved across the temporal axis using a 1D kernel (1 $\times$ 32), to generate numerous band-pass filtered EEG signal feature maps. Following this a 'depth-wise' convolution is applied overall all channels (8 $\times$ 1) for a single timestamp to fit a spatial filter for the EEG data input. This block essentially mimics the computational stages of many filter bank approaches that have previously proven successful in the classification of BCI bio-signals. Further, by combining these convolutional operations the developers reduced the number of the model training parameters, effectively increasing the scope of network deployment options. Following this is a separable convolutional block, that applies a 1D (1 $\times$ 16) filter over the temporal axis to generate a summary for each feature map. These representations are later combined in an optimized arrangement to determine to most effective relationships within and between the feature maps present.

Note, that all convolutional units were followed by a batch normalization operation. Further, after both depth-wise and separable convolutional blocks, the inputs are processed via an exponential linear unit (activation function), average pooling and drop-out operations (0.5). For further information on the architecture specifications see, Appendix, Figure A.15. Crucially, this network demonstrated significant potential in the classification of oscillatory brain-based bio-signals and notably outperformed the current state-of-the-art FBCSP method outlined above. Further, comparisons against similar convolutional neural networks, namely, the ShallowConvNet and DeepConvNet [61], revealed fewer differences in performance. Despite this, significant advantages over the ShallowConvNet [61] model was observed for non-oscillatory bio-signal classification and a dramatically reduced computational load as compared to the DeepConvNet deployment [61].

6.6.4.4 EEGNetSSVEP

As stated above, the EEGNet architectural format demonstrated world-class performance across a range of oscillatory and non-oscillatory brain-based bio-signals. The same authors later extended the findings of this research by developing EEGNetSSVEP or CompactCNN [60]. This model was developed specifically to demonstrate the potential of neural networks for SSVEP-based classification. The network contains the same architectural layout discussed above (see, Appendix: Figure A.16). This comprises two convolutional blocks, the first is composed of a temporal and later depth-wise unit, followed by the separable convolutional block. The primary modifications implemented relate to the size of the initial 1D convolutional filter (increased from 1×32 to 1×256) and the number of filters applied at each stage also increased significantly (8 to 96). These adaptations effectively enhance the computational power of the network. Note, that the term 'power' concerning neural networks can be relatively nebulous and opaque, in this context, it relates to a large number of corresponding trainable network parameters. Alternatively, the same term could be applied to a network with a substantial number of layers. In sum, power denotes overall network complexity and hence representational capacity as a decision-making function.

The larger number of connections and parameters increases the resource cost of the model as a trade-off for enabling the extraction of more complex and diverse EEG components from the input data. As mentioned above, the FBCSP and FBCCA methods informed the design of the models herein. Specifically, the Combined FBCAA methods described earlier (see subsection 6.6.1) dramatically increased classification accuracy due to the use of subject-specific CCA reference signal templates and the expansion of the filter bank analysis to 2nd and 3rd

EEG target frequency harmonics. The authors applied a 9-30 Hz 3rd order Butterworth bandpass filter pre-processing stage to all samples assessed. Using this restricted bandpass range impeded the ability of the CompactCNN to extract latent high-frequency components, for example, 3rd-order harmonic information. In the article corresponding to the GitHub SSVEP repo used in this study [180] the authors implemented a much wider bandpass filtering range, between 6-80 Hz. Replicating this wider filter range in combination with the greater number of filters could enhance network performance beyond that observed in the original article. Here the author attempts to verify and extend this method to the models defined herein. In other words, the authors will explore the implementation of greater bandpass filter thresholds in the signal pre-processing stage on network performance, by allowing for the extraction of 2nd and 3rd harmonic components embedded in the EEG data (see, Table 6.1).

Further, the author aims to evaluate the possibility of diminishing returns concerning ever-increasing threshold values. To explore this systematically a method of optimized signal-pre-processing parameter search has been developed using the Optuna library [271]. It is predicted that the increase in low-pass filter cut-offs could enhance performance in higher complexity networks, with greater computational power. The increased cut-off values could likely hamper the smaller networks as potentially useful, yet noisy data containing harmonic information could ultimately compromise model training.

6.6.5 Optuna Optimization Process

Numerous methods have been employed to optimize neural networks in the past. These are typically borne from new theoretical considerations and evaluated within respective computational bounds. Often, these guidelines are shifted towards a brute force trial and error methodology until a requisite amount of research eventually determines rough, base values for the parameter in question. For example, the length of a convolutional filter for the classification of EEG time-series data is constrained by the limitations of the Nyquist frequency and the optimal sizes of these filters in architectures for these applications vary for the many bio-signals available for classification. This can manifest in real-world applications with larger filter sizes for slower EEG components and vice versa for higher-frequency signals.

The process of neural network parameter optimization can be extended to nearly every aspect of the model. This can include different metric ranges for drop-rate rates and learning rates, it can cover alternative operational methods for average or max pooling as well as optimizer methods covering Adam [264], AdamW [265], stochastic gradient descent (SGD) [295] and Adagrad [293]. Further, the optimizations can extend into the architectural features of the

model varying the number of filters employed, the number of layers used and the connectivity between respective layers. The research deviates from these methods by optimizing the data input into the network, by searching the parameter space for optimal signal pre-processing thresholds, specifically low-pass and high-pass filter cut-offs.

Many optimization tools are available to researchers ranging from open-source Python-based Scipy methods [226] to paid services as provided by Neptune [296]. The Python library Optuna [271] was utilized to run all optimization processes described herein. Essentially, a wrapper is positioned overtop the typical CNN deployment code to control the iterative evaluation of pre-processing parameters in terms of network performance metrics (see, Figure 6.2). Note, that a single optimized parameter search is termed a study and each run within this process is known as a trial. Each study in the analyses is unique to one specific subject for a certain model, consisting of 100 trials each. Firstly, the search parameters are selected, in this case, the high-pass and low-pass cut-off values. Following this, the value range is specified, here a 0-9 Hz and 15-85 Hz range are selected for the respective parameters. This high-pass parameter search range was chosen as it encompasses all signals available in the EEG time series up to the lowest frequency target waveform used, 9.25 Hz. The author asserts it is likely that the optimization process will select a cut-off value as close to this frequency as possible.

Further, the low-pass filter range implemented is significantly larger extending from 15-85 Hz. The original data repository article [180] employed an 80 Hz low-pass filter to allow for the capture of higher-order harmonic information, as noted above (see subsection 6.6.1). The author extended the range to explore the effects of retaining additional more high-frequency information on network classification performance. Note, that the selection of alternative cut-off values was capped in 0.25 Hz increments. This reduced the number of potential parameter values available to the Optuna pruner and was employed to expedite the algorithmic parameter selection through the search space. Further, the use of non-integer frequency filter values can produce unexpected complications in deployment. As the system is intended to run for prolonged periods uninterrupted the reduction of such unforeseeable issues was nullified via this decision.

It is crucial to note that the optimization process is composed of two key elements, the sampler and the pruner algorithm. Here the sampler refers to the method by which the value is selected from the search space, this is typically based on the processing of previous monitoring metrics such as loss or accuracy values. The pruner algorithm actively compares the performance of the monitor values to the metrics collected during each training epoch to determine the prof-

itability of continuing or halting training. In all instances, the models in this thesis used the same sampler method. This is based on the default parameter setting in the Optuna library [271], namely the Tree-structured Parzen Estimator (TPE) technique [297]. This is characterized by the application of probabilistic modelling to perform highly efficient parameter space searching.

The technique involves ranking the parameters, in our case the low-pass and high-pass filter cutoff values, using a network performance measure, in this study the loss value computed for each network against the evaluation dataset was used. The ranking procedure is handled by the implementation of two Gaussian Mixture Models (GMM). This type of model is computed to represent an array of parameter metrics as the sum of a series of Gaussian distributions [297]. Through this method, one model is computed to represent the values associated with good performance metrics, i.e. low-loss values, and another is computed to represent values associated with all other performance metrics computed thus far, i.e. higher-loss values.

Effectively, each model generates likelihood values corresponding to the entire search space for every combination of parameters relative to the associated parameter performance metric dimension. The ratio between each parameter value in the search for these two models is used to compute the next set of values. Here the TPE method aims to select parameter values by maximizing the ratio between the likelihood values from GMM computed on the high-performing parameters against the likelihood values computed from all other parameters tested. As seen in Figures 6.6, 6.7, 6.8 and 6.9, the initial parameter selections are highly volatile, this is because the GMMs have yet to acquire an adequate sample size of parameter-performance metric pairings. Over the course of the optimization session, the sampler typically converges towards a specific region of the high-pass or low-pass filter cutoff search space.

To clarify, following the selection of the new frequency cutoff values via the respective optimization sampler-pruner algorithm combination all subject samples are pre-processed using these filter parameters. This was done to exclusively probe the effect of EEG data filtering cut-off values on a range of neural network performance metrics. Any additional changes to the network parameters during the optimization stage would have dramatically increased the optimization search space and would have prevented any strong conclusions on the influence of the frequency cutoff optimization on end-point classification accuracies. During each optimization study, the weights for the model currently being optimized were reinitialized and the network was trained from the ground up using these newly pre-processed data. Note, that all other network parameters including the learning rate, drop-out rate, batch size, num-

ber of kernels and stride were kept constant. Understandably, these parameters only differ between the model configurations (e.g. EEGNet *vs.* EENetSSVEP), as opposed to the same network variant within the same optimization study (EEGNet Optimization Run 1 *vs.* EEGNet Optimization Run 2). For more information please refer to Figure 6.2 and subsection 6.6.6: Pruners.

As noted above in subsection 6.6.3.1: Raw Data and 5.6.3.3: Convolutional Neural Network Summaries, all models defined herein were trained on aggregated, cross-subject data using a leave-one-out 3-fold cross-validation procedure (see subsection 6.6.6.3: *k*-Folding). This continues the current trend established in cutting-edge CNN-based SSVEP classification methods [269, 270] that operate under the assumption that given the adequate data volume, the variance within a cross-subject dataset could present advantages over a single-subject dataset by introducing a greater array of example waveform expressions. The increased variance in the quality of waveform expression could lead to more robust models and crucially these systems have significant design benefits including substantially higher generalizability across subjects as well as the ability to classify signals without the need for a subject-specific calibration period [270].

Along these very same lines, in all instances, the training data herein is composed exclusively of samples collected from 9 of the 10 subjects comprising the open-source SSVEP repository [180]. The remaining subject data is used only for evaluation purposes. To clarify, the model is trained using data from 9 of the 10 subjects and the performance of the model, for the given configuration of frequency filter parameters, is monitored during the optimization process by intermittent testing of the model on the single remaining subject. For further information on how this intra-study test performance is handled by each of the three pruners assessed please refer to the subsection below, 6.6.6: Pruners.

Note, that each model-pruner combination required several hours to process for each of the optimization studies performed.Further, the networks differ considerably in terms of total computation time across the entire 10-subject dataset. As can be seen in Table 6.12, the EEGNetSSVEP required 66.41 hours of processing time (6.7 hours per subject) when optimized with the Median pruner algorithm, as compared to the EEGNet (3.8 hours per subject), DeepConvetNet (3.6 hours per subject) and ShallowConvNet (4.6 hours per subject). It is asserted that this is due in part to the fact that the EEGNetSSVEP model possesses the highest number of trainable parameters. For further information on these differences see subsection, 6.7.4.3: Computational Resources and for more interpretation of these findings see subsection, 6.8.4.3:

Optimizer Study Durations.

The purpose of undertaking these highly computationally expensive investigations is to determine optimal ranges for the implementation of filter frequency cutoffs across the individual subjects and the model-pruner combinations assessed.The high computational cost incurred in this study is done to assist researchers in performing convolutional neural network optimizations in future studies. Here I aim to illuminate the interaction between filter frequencies, model configuration and pruner selection in terms of computational duration and end-point performance. I intend for these results to serve as a guide to the selection of these parameters based on thorough, systematic testing to help save countless hours of crucial research time and resources.

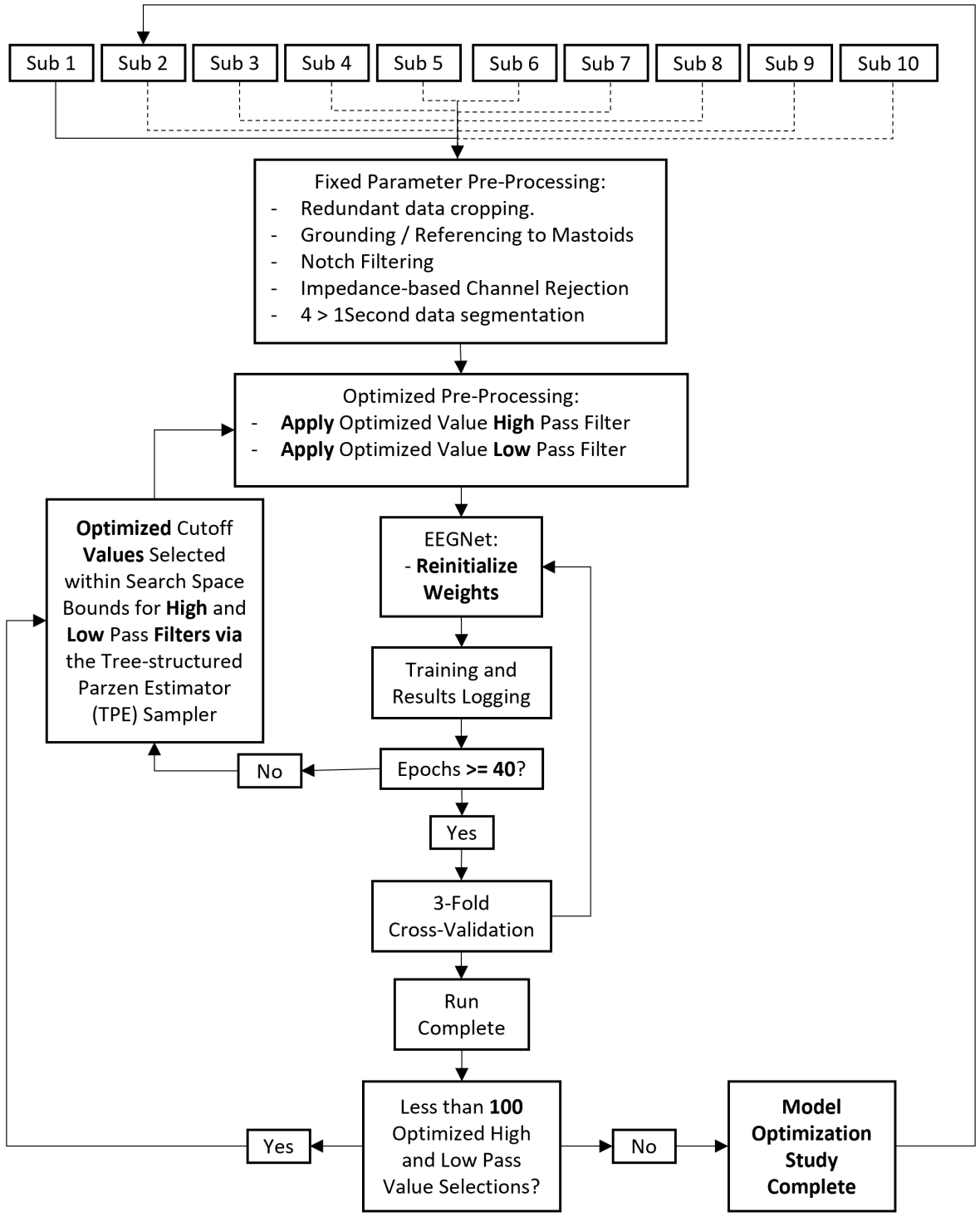
Sub 1
Sub 2
Sub 3
Sub 4
Sub 5
Sub 6
Sub 7
Sub 8
Sub 9
Sub 10
Fixed Parameter Pre-Processing:
- Redundant data cropping.
- Grounding / Referencing to Mastoids
- Notch Filtering
- Impedance-based Channel Rejection
- 4 > 1Second data segmentation
Optimized Pre-Processing:
- Apply Optimized Value High Pass Filter
- Apply Optimized Value Low Pass Filter
Optimized Cutoff Values Selected within Search Space Bounds for High and Low Pass Filters via the Tree-structured Parzen Estimator (TPE) Sampler
EEGNet:
- Reinitialize Weights
Training and Results Logging
Epochs >= 40?
No
Yes
3-Fold Cross-Validation
Run Complete
Less than 100 Optimized High and Low Pass Value Selections?
Yes
No
Model Optimization Study Complete

Figure 6.2: Here is presented a figure illustrating the automated workflow of the optimisation code. The caption is positioned on the following page to accommodate the size of the figure. In the example detailed, the data for Subject 1 is loaded and pre-processed via the fixed parameter stages. Following this, the high-pass and low-pass frequency filters are applied using cutoff values selected from the associated frequency ranges, 0-9 Hz and 15-85 Hz, respectively via the Tree-structured Parzen Estimation (TPE) method. During the first few runs, the values are highly varied as the associated Gaussian Mixture Models (GMM) have not yet been provided enough samples to guide the selection process effectively. At this point, the pre-processed data is then used to train the corresponding CNN model. In this instance, the EEGNet architecture is selected. Over the course of this process, the model is trained for a maximum of 500 epochs. During this training phase either the pruner or early stopping function could establish, via comparison to previous runs or rate of validation loss decrease, that the current study does not possess the optimal configuration of parameters to maximise network performance. This is based on the iterative assessment of the loss computed on the evaluation data at the end of each epoch. As can be seen in Appendix A.6: Optimizer Loss Profiles, Figure A.20, in most instances of the networks assessed herein, once a given model reaches above 40 epochs the loss metric becomes highly stabilised (see subsection, 6.6.6.2: k-Folding). This suggests that the models have successfully learned to extract and classify samples based on features inherent to the SSVEP data. For optimization runs in which this threshold was breached, the model was allowed to continue training to 500 epochs and went on to be re-tested via the 3-fold cross-validation method (see, subsection 6.6.6.2: k-Folding). This involves re-initializing a new iteration of the target network and training it using a randomized version of the training dataset. This is repeated twice to generate enough performance values for metric averaging and model validation. The heuristic of 40 was implemented as a means of balancing the need to thoroughly explore as many viable network-filter-cutoff parameter configurations as possible, without wasting time optimizing models with a low likelihood of producing good performance values. For instances where the model is pruned early, i.e. before the 40-epoch mark, the training is stopped, and a new series of high and low-pass filter cutoffs are selected. As noted above, the process of selecting new high-pass and low-pass cutoff value pairs is performed over 100 optimization studies via the Tree-structured Parzen Estimator (TPE) (for further information, see subsection 6.6.5: Optimization Process). Once the optimization study is complete, all results and associated plots are generated to determine the best trial consisting of the optimal high-pass and low-pass metric pair. Note, that the dotted lines displayed here represent the relative subsequent set of analyses for each subject in the dataset.

6.6.6 Pruners

The pruning method defines the algorithm used to select novel parameter values at the trial level over the course of each study. A suite of three pruners was utilized from the Optuna library [271], namely the Median, Percentile and Successive Halving Pruner. These were selected as a representative group of commonly employed algorithms to gauge differences in speed (time taken to complete one study) and quality (final performance of optimized networks). A brief outline of each method is presented below, for further information see, the related Optuna documentation [271]. Note, that all pruners effectively perform the same two basic tasks, namely parameter value selection for a specific metric directionality and the exe-

cution of an early stopping mechanism for resource-saving purposes (see, Figure 6.2).

In this instance, the parameter values selected are for the high-pass and low-pass cut-off values and the monitored performance metric is validation accuracy. The preferred direction is to increase this value. In many instances, the performance metric input to the pruner is validation loss. This is principally used for unbalanced datasets to ensure biased network performance for a specific data class does not impede effective network training. As the data utilized in this study is perfectly balanced the author was not obliged to follow this convention and the preliminary results gained in project troubleshooting revealed the end-point performance increase was greater for the original monitor and directionality configuration.

During the deployment of the Median pruner, trial rejection is controlled by repeatedly calculating and comparing the median metric value of intermediate epochs against those gathered during previous optimization trials. Firstly, the network is trained and optimized without pruning to initialize the optimization algorithm and provide a baseline of performance at each respective trial epoch. During the training period of the following network, the newly generated metric values are compared to those of previous runs. In this instance, if the median of metric values for the current number of epochs on the new run is lower than the median value obtained by previous runs for the same number of epochs, the trial is pruned. If the network is capable of consistently matching or exceeding the median metric values of previous runs, then the trial will be completed for all epochs programmed for training.

In contrast, for the Percentile pruner, trial rejection is determined by comparing the current performance metric specified to all other trials collated thus far. Like all pruners, a baseline must first be established. Following this, in subsequent trials, current network performance is compared against previous trials in terms of performance metrics for corresponding epochs. If the current trial epoch produces a performance metric that positions the parameter configuration in the e.g., bottom 25% of all previous trials, then the trial is pruned. During the pruning of network parameters, this pruner selects values that enhance the rate of convergence (e.g., low drop-out and high learning rates) and can lead to a large percentage of trials being pruned.

Finally, the Successive Halving pruner was evaluated. This method is primarily suited to the rapid calculation of hyper-parameter configurations distributed over a network of optimization protocols running in parallel. The process is a hierarchical search program in which all net configurations are first tested with a minimum of resources (low number of epochs), those networks achieving above-average performance are promoted to a second run in which

these networks are retested using twice as many resources (number of epochs). This iterative training and evaluation process is repeated until a single network configuration is decided. The parameter search algorithm is controlled by a multi-armed bandit-style function which aims to maximize performance metrics by directing parameter configurations towards optimality via recall to previous trial performances.

6.6.6.1 Early Stopping *vs*. Pruning

During the training phase of each network, the monitor metric (validation accuracy) is relayed back to the pruner to assess the trial performance against currently logged study-level metrics. In addition to this, an Early Stopping feature monitoring validation loss was implemented to prevent the excessive over-training of the networks, specifically during the early stages of the study. It was crucial to employ both of these methods simultaneously to reduce the total amount of computational time used per study. Further, despite the EEGNetSSVEP article [60] listing the number of epochs trained for at 500, the author herein found that the EEGNetSSVEP converged in most instances before the 100-epoch mark. The only architecture that seemed to substantively benefit from the full 500 epoch training scheme was EEGNet as will be discussed in the subsequent results and methods sections (see subsection 6.7.3.1).

It is well established that the continued training of models beyond the point of convergence can lead to over-training and an increased risk of overfitting the training data. These obstacles are averted in the current study due to a combination of the pruning method, early stopping (both based upon validation loss monitoring) and the use of so-called best-weights saving. This involves only overwriting the current network weights save state for the batches that produce the maximal classification performance, as opposed to simply saving each weight state after every epoch.

6.6.6.2 *k*-Folding

Following the completion of all 100 trials in a given study, all networks that accrued at least 40 epochs are separated from the other runs for further evaluation (see, Figure 6.2). This was selected as typically at this number of epochs all networks, irrespective of model type, are likely at or approaching convergence (see, Appendix: Figure A.20). The remaining models then undergo further evaluation via a *k*-fold method. Each training dataset is re-processed using the same corresponding high-pass and low-pass values and the respective samples are randomised alongside the corresponding class labels. A newly initialized network (zeroed weights) is then trained using this randomised data for the same number of epochs previously achieved by the

original model. This was repeated twice to produce three sets of performance metrics per retained model and allowed researchers a means of calculating standard error values for each network assessed. This process was not integrated directly into the code encased in the Optuna wrapper as the iterative logging of multiple monitor values (validation accuracy) within the same trial is not presently a compatible feature of the Optuna optimization library [271].

6.6.6.3 Acknowledgements on the Re-Implementation of CNN Models

The Lawhern GitHub repository was a crucial resource in the implementation of the analyses described herein [286]. Despite this, some network training parameters had to be estimated given a lack of detail in the repo and original corresponding articles. This is related to the specific learning rate and optimizer implemented. This is reflected in the fact that the exact performance metrics recorded by the original authors using the same models, dataset, and fixed pre-processing stages, could not be replicated here. The performance is marginally lower for nearly all methods tested at the single-subject and cross-subject levels. Due to the absence of these details, standardized values were implemented across all models assessed, the learning rate was set at 0.0001 and the optimizer utilized was the Adam method [264].

6.7 Results

Here are presented the results concerning all studies conducted using the raw, fixed-parameter and optimized datasets. For further information regarding the configuration of these datasets please refer back to subsection 6.6.3. Following these evaluations, a discussion relating to the comparison of the three different optimization pruner methods tested is undertaken. This covers the differences in endpoint classification accuracy, optimized frequency filter cutoffs and computational resource usage.

6.7.1 Raw Data: Assessments

The analysis of network performance on the Raw dataset (see subsection 6.5.6.3.1) was undertaken to provide baseline performance statistics and explore the computational limitations of the models tested for the inherently noisy EEG input data. The samples taken from all 10 subjects consist of 6-second data chunks relating to a specific SSVEP target frequency. As noted earlier, the data are spliced into 1-second packets and then normalized to a zero mean. This involves calculating the mean μV amplitude of each EEG channel signal and then subtracting this metric from all values in the respective waveform.

The intention of applying this normalization stage is to ensure all input signals are in a similar μV amplitude range. Previous literature has demonstrated the sensitivity of CNNs to dramatic differences in training data scaling and the absence of care in this instance can significantly reduce the rate of convergence and end-point performance metrics of the model evaluated. Note, the researchers also performed the same iterative subject-level training scheme for purely raw data signals and non-significantly different results were attained. Further, as stated previously (see, Figure 6.1) all data were pre-filtered at 50 Hz as per standard powerline noise removal. In light of these circumstances, the effects of forgoing any notch filtering before network training could not be evaluated.

All models herein were trained for 500 epochs with the same learning rate (0.001), batch size (64), optimizer (Adam) and loss function (Categorical Cross-Entropy). Note, given the number of target classes (12) the random performance threshold is 8.33%. Again, test set data comprised all samples for one individual subject and at no point was this data present in the training set. This was done to develop subject-specific models in a simulated real-world, offline analysis context.

As seen in the table below (see, Table 6.3), no subject-model combination achieved a classification performance significantly above the random performance threshold (8.33%). There are some instances of 9% AoC at the subject level, as shown by Subject 10 for the DeepConvNet architecture. Further inspection of these results reveals that this is principally owing to the model overfitting (see, Figure 6.3) for two target classes, in this instance class 2 (11.25 Hz signal, Phase: 0π) and class 9 (14.25 Hz signal, Phase: 1.0π). The instances of 8% AoC that are present in most subject-model combinations are slightly more diffuse (less overfit) and present with a lower hit rate for the biased class selections. Additionally, the standard error values computed for the averaged *k*-fold stats relating to each subject AoC metric are not reported here. This is owing to the fact the differences between maximum and minimum values for all stats recorded were $<= 1\%$.

	Sub 1	Sub 2	Sub 3	Sub 4	Sub 5	Sub 6	Sub 7	Sub 8	Sub 9	Sub 10	Mean
EEGNet	0.09	0.08	0.08	0.09	0.08	0.08	0.08	0.08	0.08	0.08	0.08
EEGNetSSVEP	0.08	0.08	0.08	0.09	0.08	0.08	0.08	0.08	0.08	0.08	0.08
DeepConvNet	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	**0.09**	0.08
ShallowConvNet	0.09	0.08	0.09	0.09	0.08	0.08	0.08	0.09	0.08	0.09	0.09

Table 6.3: Here is presented a table containing the classification accuracies of the four CNN models assessed for each of the 10 test subjects (Sub) in the Nakanishi SSVEP data repository [180]. The AoC is presented from 0-1.0, with 0 indicating 0% classification accuracy and 1 indicating 100% classification accuracy. The mean column positioned to the far right of the table shows the cross-subject average score relating to each model tested. Note, that the standard deviations are not displayed as all subject-model combinations possessed near identical values ($0.004 +/- 0.001$).

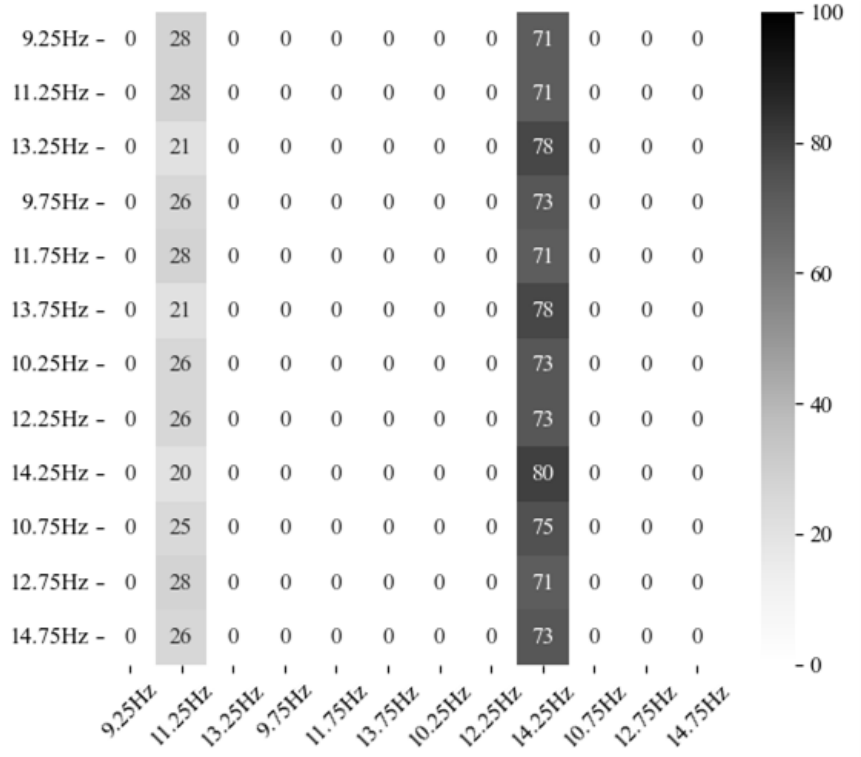

Figure 6.3: Here is presented a confusion matrix detailing the performance of the DeepConvNet model for the classification of Subject 10 data. The columns and rows listed refer to each of the 12 SSVEP signal classes utilized in the Nakanishi data repo article [180]. For further information on the phase angle content of the respective signals used refer to Table 6.2 positioned in the Methods section (6.6.1). To accommodate visual interpretation a higher incidence of classification accuracy for a given target class is indicated by the use of a darker corresponding matrix cell. As noted in the legend (positioned on the right), 100% classification accuracy is denoted via the use of black as the cell background and conversely, a white background is used to indicate 0% accuracy.

The different levels of network complexity and number of parameters seemingly offered no advantage to the classification of the signals in the absence of any frequency-based pre-filtering. These results suggest some filtering of these data is necessary before the training of the networks assessed herein. As can be seen in the original data repository article [180], the EEGNet [59] and EEGNetSSVEP [60] articles and the subsequent analyses noted herein indicate Subjects 4, 6 and 8 as the highest-performing individuals for this task (see, Table 2.1). The metrics observed for these raw analyses demonstrate no significant differences between these subjects or any of the other subjects assessed. These findings indicate that, despite some evidence for the adaptability of CNN models to noisy EEG bio-signals, a comparable pattern of performance was not replicated here. This is likely due to the complexity of the task, given the high number of targets and the narrow range that the target frequencies occupy (9-15 Hz).

6.7.2 Fixed-Parameter Data: Assessments

This subsection relates to the performance of the networks assessed using the fixed parameter dataset (see subsection 6.6.3.2). This involved applying the same high-pass (9 Hz) and low-pass (30 Hz) filtering methods implemented by the developers of the EEGNetSSVEP model [60] for the same online repository dataset [180]. As stated previously, the fixed parameter classification performance metrics serve as a baseline for comparison to gauge the efficacy of the automated hyper-parameter optimization techniques defined herein. The networks evaluated did not outperform the Combined CCA methods reported in the original article [180] that introduced the data repository utilized here (see, Table 2.1). Further, the models performed marginally below the levels presented by the developers of the EEGNetSSVEP architecture [60]. Despite this, the reported statistics are only marginally lower and crucially the same trend of classification performance is observed at the model and subject levels.

	Sub 1	Sub 2	Sub 3	Sub 4	Sub 5	Sub 6	Sub 7	Sub 8	Sub 9	Sub 10	Mean
EEGNet	0.34	0.18	0.48	0.68	0.63	0.72	0.60	0.82	0.61	0.68	0.57
EEGNetSSVEP	0.56	0.24	0.69	0.91	0.84	0.91	0.81	0.97	0.84	0.80	0.76
DeepConvNet	0.40	0.21	0.52	0.75	0.79	0.77	0.68	0.87	0.78	0.72	0.65
ShallowConvNet	0.25	0.16	0.30	0.36	0.34	0.46	0.35	0.55	0.41	0.45	0.36

Table 6.4: Here is presented a table consisting of all subject-level classification accuracies (Sub) for each of the 4 respective CNNs assessed. The corresponding AoC values are denoted between 0 and 1, with 1 representing a 100% hit rate. For further information regarding the interpretation of these results see, Table 6.3.

As seen in Table 6.4, the EEGNetSSVEP model variant produced the highest classification accuracy both at the single-subject (Subject 8) and cross-subject (Mean) levels. The ShallowConvNet produced the lowest mean accuracy of classification and the lowest individual AoC reported herein (Subject 2). Additionally, the DeepConvNet model outperformed the standard EEGNet model for all subjects evaluated. This network arguably presents with a higher level of architectural complexity as compared to the EEGNet and EEGNetSSVEP, due to the greater number of convolutional layers (see, Appendix: Figures A.14, A.15 & A.16). Despite this, the dramatically higher volume of convolutional filters alongside the inclusion of Separable and Depth-Wise convolutional operation blocks affords the EEGNetSSVEP model significantly greater computational power.

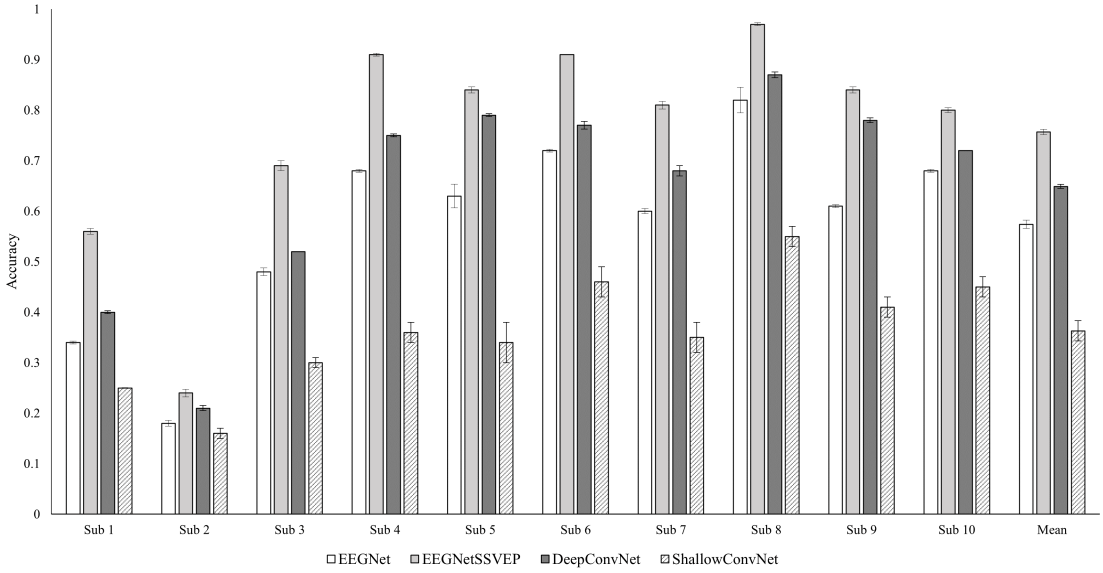

Figure 6.4: The plot above shows a bar chart depicting the accuracy of classification values for all subjects (Sub) and models assessed using the Fixed Parameter dataset (see subsection 6.6.3.2). To clarify this plot is generated using the same metrics that populate Table 6.4, any comparison between these results and the baseline raw assessments can be done by inspecting these tables. Each subject block reports the respective model performance alongside error bars computed via the standard deviation of all accuracy metrics generated from the 3 k-fold randomized evaluations. Further, a series of means computed from all the individual subject performance values is positioned to the right of the subject data.

When inspecting Figure 6.4, a general pattern of subject-related performance is present across all individuals assessed. Specifically, Subject 8 is the highest-performing individual for all models tested, likewise, Subject 2 is the lowest-performing subject for all networks evaluated. Further, the variance in subject performance within the 3 randomised k-folds is generally below 2% , as only a marginal number of instances can be seen that report variance $>$

4% . The coherence in model performance metrics across the subjects evaluated suggests that there exist specific data characteristics conducive to the CNN analysis method. This is realized as a high relative correlation between performance metrics gathered across subjects for all models. The results suggest that the rapid troubleshooting of novel network architectural features can be effectively tested using models with a low number of parameters. Moreover, the minimal amount of variance between networks within the *k*-fold reports suggests that the models are relatively consistent during deployment. This is key to the stable operation of any communication-based BCI.

6.7.3 Median Pruner Optimization: Assessments

The following section contains all analyses relating to networks optimized using the Median Pruner method (see subsection 6.6.6). A summary of the results can be seen in Figure 6.5. The same pattern of mean classification performance found for the fixed parameter assessments is replicated here. The EEGNetSSVEP performs at the highest level and the ShallowConvNet demonstrates the lowest classification accuracies.

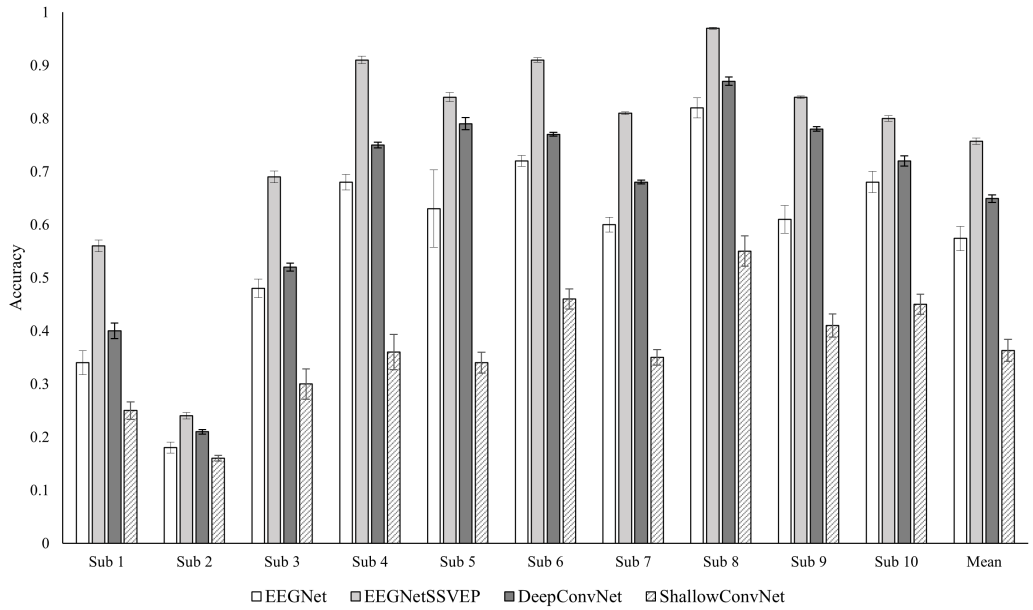

Figure 6.5: The plot above shows a bar chart depicting the accuracy of classification values for all subjects (Sub) and models assessed using the Median pruner optimization method (see subsection 6.6.3.3). The metrics are presented in terms of classification accuracy between 0-1.0, this involves 0 representing 0% AoC and 1 representing 100% AoC. Each subject block reports the respective model performance alongside standard deviation error bars. Further, a series of means computed from all the individual subject performance values is positioned to the right of the subject data.

6.7.3.1 EEGNet

The performance metrics for the EEGNet model post-filter cutoff optimization differ markedly from the original fixed parameter study. Paired t-tests revealed a significant ($p > 0.005$) drop in classification accuracies, as calculated from the mean subject-specific accuracies and are reflected in the decrease of mean cross-subject AoCs from 57% to 52.6% (see, Tables 6.5 & 6.5). Only one subject (Subject 2) presented with a marginal increase in AoC (0.8%) and many subjects noted large decreases as in the case of Subject 3 (-8.6%) and Subject 5 (-11.5%). This optimization required on average 227 minutes per subject and a total of 37.9 hours to complete.

	AoC	Min K	Max K	Standard Deviation	High-Pass Cutoff	Low-Pass Cutoff	Epochs
Sub 1	0.29	0.26	0.34	0.045	9	15.5	145
Sub 2	0.19	0.17	0.21	0.021	9	15.5	75
Sub 3	0.39	0.37	0.43	0.035	7.75	17.25	128
Sub 4	0.66	0.63	0.68	0.029	9	15	208
Sub 5	0.52	0.40	0.68	0.145	1.5	41.75	241
Sub 6	0.72	0.69	0.73	0.021	8.75	15	212
Sub 7	0.55	0.53	0.58	0.028	9	15.25	139
Sub 8	0.81	0.79	0.85	0.038	8.25	15	239
Sub 9	0.55	0.52	0.61	0.052	8.5	21.75	271
Sub 10	0.59	0.55	0.63	0.040	8.75	17.25	166
Mean	0.53	0.49	0.58	0.045	7.95	18.93	182.4

Table 6.5: Here is presented a table of results for the EEGNet model in the Median Pruner optimization study. The metrics reported related to all 10 subjects (Sub) assessed in addition to a mean computed from all respective individuals. These metrics correspond to the highest-performing model variant generated from the optimization study. The AoC header refers to the average classification accuracy achieved over the 3 k-fold randomisation runs, with Max K and Min K referring to the highest and lowest AoC values achieved in all 3 runs performed. The standard deviation computed from the 3 accuracy scores is also presented to provide insight into model performance variance. Further, the optimized high and low pass filters of the best-performing model are listed alongside the number of epochs the network was trained for before pruning. Note, that the number of epochs trained for was dependent on both the Early Stopping checkpoint function and the respective pruner algorithm selected. In >90% of all trials assessed the pruner algorithm was responsible for trial terminations. The early stopping protocol merely served as a time-efficiency backup.

As shown in Table 6.5, the high-pass cutoff values have coalesced around the lower limit of the SSVEP target frequencies assessed. A clear trend towards the 9 Hz boundary can be observed for Subject 8 in Figure 6.6 (left plot). A similar weaker pattern of algorithm parameter search progression in the orthogonal direction, towards the 15 Hz limit of the low-pass filter bound, can also be seen in Figure 6.6 (right-plot). Ultimately, the series of parameter search selections for the low-pass cutoff value demonstrate an increased incidence of bi-modal parameter search shifting. This is characterized by the rapid study-by-study fluctuations in cutoff values.

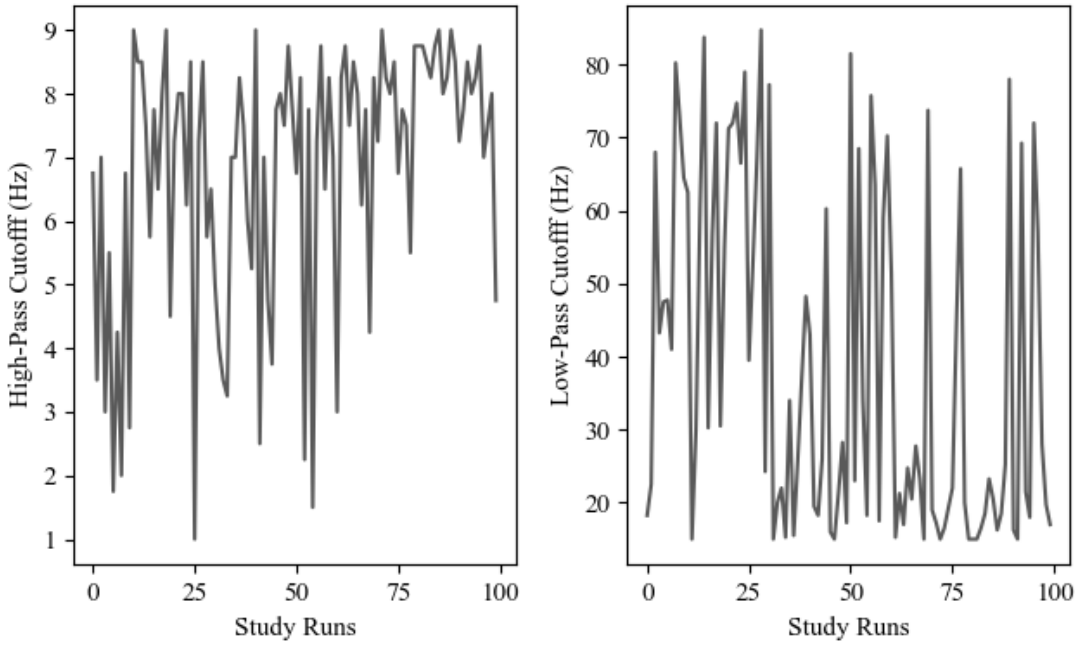

Figure 6.6: The figure presents all high-pass (left plot) and low-pass (right plot) Median Pruner algorithmic parameter search selections for Subject 8 in the respective EEGNet model optimization study (100 trials). The frequency (Hz) of the respective filter cutoff value is positioned on the y-axis alongside each study run (trials) plotted on the x-axis.

As seen in Figure 6.6 (right plot), the final 25 runs feature selections ranging between 20 and 78 Hz. This indicates that the optimization algorithm has yet to adequately converge and likely requires additional runs to find the optimal frequency range. Note, that this pattern is replicated independent of subject AoCs (see, Appendix: Figure A.17), with some outliers present in the data showing a reduced directional trend (see, Appendix: Figure A.18). Further, the number of epochs each subject-specific network was trained for demonstrates substantial variability, ranging from 75 (Subject 2) to 271 epochs (Subject 9).

6.7.3.2 EEGNetSSVEP

In relation to the EEGNetSSVEP model, the increase in cross-subject mean accuracy post-optimization, as compared to the fixed parameter results, is marginal (+ 1.69%) and is shown to be non-significant (p = 0.075). In most cases, the percentage increase in classification performance stood at around 1% . More substantial performance increases were noted for a handful of individuals, namely Subjects 2 (+ 7.4%), 3 (+ 4.8%) and 9 (+ 2.4%) (see, Table 6.6). Crucially, these performance increases were not exclusively restricted to low or high-performing subjects.

	AoC	Min K	Max K	Standard Deviation	High-Pass Cutoff	Low-Pass Cutoff	Epochs
Sub 1	0.54	0.53	0.57	0.023	7.75	52.5	52
Sub 2	0.31	0.30	0.33	0.012	5.75	16.75	69
Sub 3	0.74	0.72	0.76	0.023	8.5	75	93
Sub 4	0.910	0.90	0.93	0.014	5	45.5	79
Sub 5	0.85	0.83	0.87	0.017	8.75	58.25	53
Sub 6	0.92	0.91	0.93	0.009	8	58	107
Sub 7	0.82	0.81	0.82	0.005	8.25	16.75	59
Sub 8	0.97	0.97	0.98	0.003	8.75	69.75	66
Sub 9	0.86	0.86	0.87	0.005	5	70.5	151
Sub 10	0.81	0.80	0.82	0.010	6.25	80	42
Mean	0.77	0.76	0.79	0.012	7.2	54.3	77.1

Table 6.6: Here is presented a table of results for the EEGNetSSVEP model in the Median Pruner optimization study across all subjects (Sub) tested. The metrics reported herein relate to the highest-performing model variant generated from the optimization study. For further information on column headings and table interpretation refer to, Table 6.5.

A similar trend in high-pass filter selection directionality is present (see, Figure 6.7 left plot), with a lower average optimized mean of 7.2 Hz as compared to the EEGNet optimization study (see, Figure 6.7, left plot). This is due to the presence of some lower optimized metrics at around 5 Hz from Subjects 2, 4 and 9. The higher number of convolutional filters may have enabled the EEGNetSSVEP model to more effectively parse redundant or noisy waveform features from the target frequencies. This could have lowered the need for a strict high-pass cutoff value at the very lower limit of the stimuli frequencies utilized. Despite this, the author still asserts that the soundest method for any future optimization projects in this field would be to focus solely on the optimization of the low-pass filter cutoff and other relevant parameters. Fixing the high-pass value at 9 Hz would avoid wasting time due to exploring unviable high-pass rates and focus the research purely on the inclusion of additional harmonic

information.

As shown in Figure 6.7 (right plot), the relatively strong directional trend towards the 15 Hz low-pass cutoff limit observed in the EEGNet results has been replaced with an upward trend at the other end of the low-pass cutoff bound (85 Hz) for the EEGNetSSVEP architecture. It could be argued that the noise present in the signals above the target SSVEP range of 15 Hz is easily parsed by the convolutional filters and the selection of these larger values is the function of randomness. Conversely, the networks may be utilizing the SSVEP harmonic information afforded by the higher low pass filter cutoff values. The fixed parameter low-pass cutoff at 30 Hz only allowed for the inclusion of 1st-order harmonic information and some 2nd-order waveforms in the lower frequency range (see, Table 6.1). The increase to an 85 Hz threshold allows for the holistic inclusion of harmonics up to the 4th order, with some target frequencies extending to the 7th order.

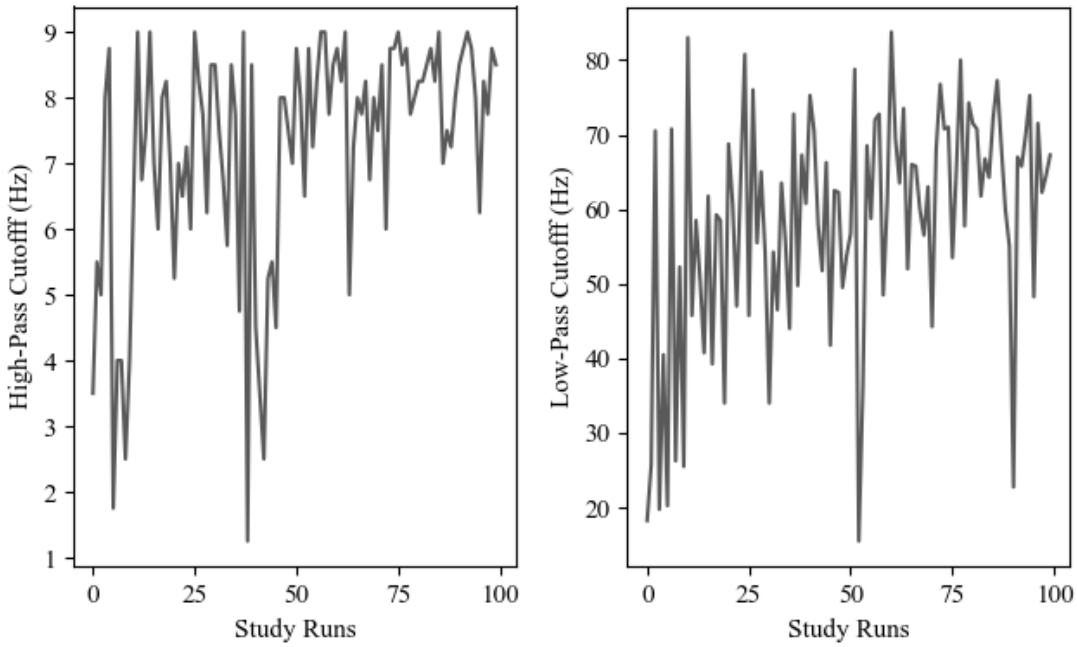

Figure 6.7: The figure presents all high-pass (left plot) and low-pass (right plot) Median Pruner algorithmic parameter search selections for Subject 9 in the respective EEGNetSSVEP model optimization study (100 trials). For further information refer to, Figure 6.6. Note this subject was selected for presentation purposes, similar results were found for nearly all subjects with optimized low-pass filter cutoffs >30 Hz.

The presence of this trend in multiple subjects suggests that some advantage is offered by this wider bandpass filter, otherwise, the pattern of parameter selection would be entirely random and no directional trend over the course of the study would be observed (see, Figure 6.7). Note, that this pattern is not present in all subjects assessed, as seen for Subjects 7 and 2 (see, Appendix: Figures A.19 & A.18). These both demonstrate a similar pattern of parameter selection towards 15 Hz, as observed in the EEGNet optimization study. This suggests that even for higher-powered CNNs the individual differences in EEG profiles influence optimal pre-processing filter cutoff parameters.

Additionally, the average time taken per subject to complete the 100-trial optimization study was 398 minutes and required a total of 66.41 hours to complete. Despite the larger number of trainable network parameters, the average number of epochs required for training is substantially lower (77.1) than that reported for the EEGNet optimization study (182.4).

6.7.3.3 DeepConvNet

As seen in Table 6.7, at the cross-subject level, mean classification accuracies for the DeepConvNet model increased by 2.3% as compared to the original fixed parameter results (see, Table 6.4). Crucially, the enhancement in performance was significant ($p < 0.05$). Note, that the cross-subject mean AoC of 67.2% nearly breaches the 70% usage threshold for BCI functionality. This is achieved using a model with significantly fewer trainable parameters and optimized in nearly half the time as the EEGNetSSVEP network, with the entire optimization study requiring 36.42 hours to complete, operating at an average of 218 minutes per subject.

This boost in performance is observed uniformly across 8 subjects excluding Subject 3 which demonstrated a marginal increase ($+$ 0.9%) and Subject 5 which showed a substantial decrease in accuracy ($-$ 4%). The effect is most pronounced in Subjects 2 ($+$ 3.5%), 7 ($+$ 3.4%) and 9 ($+$ 4.7%). Crucially, these subjects demonstrate significant variance in AoC performance, suggesting that the enhancement effects of the optimization process are not dependent on subject data quality.

	AoC	Min K	Max K	Standard Deviation	High-Pass Cutoff	Low-Pass Cutoff	Epochs
Sub 1	0.42	0.40	0.45	0.029	9	29.75	76
Sub 2	0.25	0.23	0.24	0.009	7.25	23.25	41
Sub 3	0.53	0.52	0.55	0.016	7.75	21.25	157
Sub 4	0.81	0.80	0.82	0.010	6.75	15	166
Sub 5	0.75	0.74	0.78	0.023	8.5	37.5	107
Sub 6	0.79	0.78	0.80	0.007	8.5	24	118
Sub 7	0.71	0.71	0.72	0.007	9	17.5	116
Sub 8	0.89	0.88	0.91	0.016	8.75	19.25	87
Sub 9	0.83	0.82	0.83	0.009	8.25	36	113
Sub 10	0.75	0.73	0.77	0.019	6.5	33	117
Mean	0.67	0.66	0.69	0.014	8.025	25.65	109.8

Table 6.7: Here is presented a table of results for the DeepConvNet model in the Median Pruner optimization study for all subjects (Sub) tested. The metrics reported herein relate to the highest-performing model variant generated from the optimization study. For further information on column headings and table interpretation refer to, Table 6.5.

Both the low-pass and high-pass optimization parameter search patterns reveal that the models successfully converged around unique and narrow frequency ranges. As seen in Figure 6.8 (left plot), a similar parameter value selection pattern emerges for the high-pass cutoff near the 9 Hz boundary. Further, the low-pass selections (right plot) demonstrate a weaker overall trend towards the 15-25 Hz filter range. Interestingly, one subject (Subject 1) reported near-identical parameter selections (high-pass $=$ 9 Hz, low-pass $=$ 29.75 Hz) to the original fixed parameter cutoff values. Notably, the increase in performance reported ($+$ 1.8%) is highly marginal.

Despite this, the number of epochs used to train each model differs greatly as the Median Pruner optimized network was terminated at just 76 epochs, compared to the 500 epochs used in the training of the original fixed parameter network. These findings suggest aggressive early stopping mechanisms have significant potential to enhance optimization efficiencies in this analysis context.

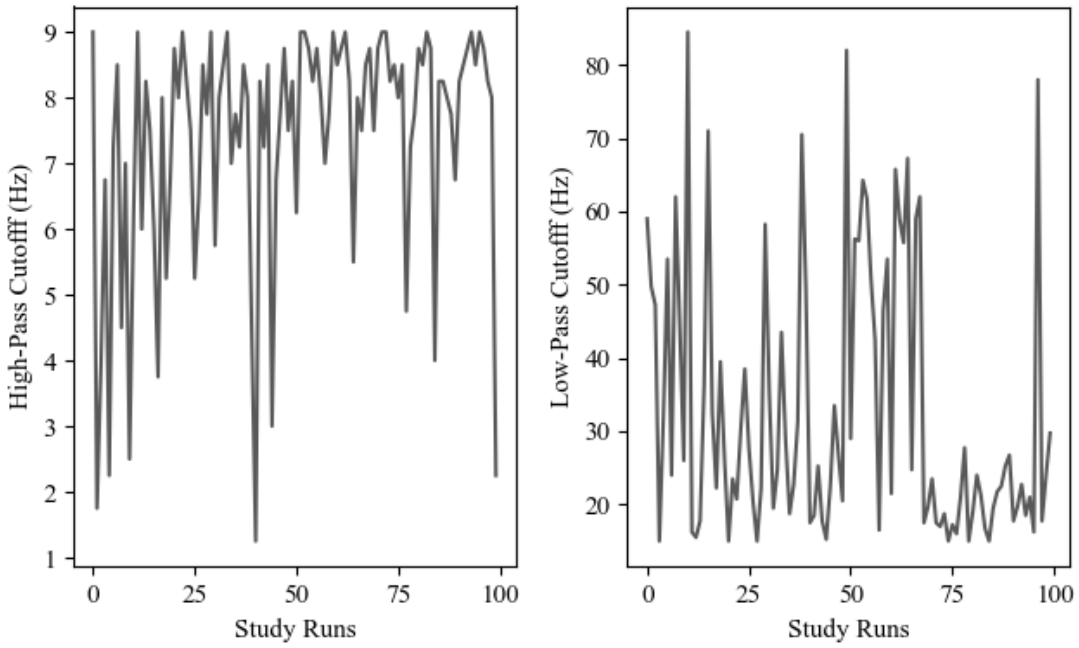

Figure 6.8: The figure presents all high-pass (left plot) and low-pass (right plot) Median Pruner algorithmic parameter search selections for Subject 6 in the respective DeepConvNet model optimization study (100 trials). For further information refer to, Figure 6.6.

6.7.3.4 ShallowConvNet

The mean cross-subject performance increase post-optimization for the ShallowConvNet model is the largest of all networks reported herein at 6.73% (see, Table 6.8). This is a significant improvement in classification accuracy as compared to the original fixed parameter results ($p < 0.005$). The largest subject-level enhancements in AoC are seen in Subjects 4 ($+$ 15.7%) and 6 ($+$ 12.7%) respectively, with Subject 8 reaching an AoC of 61%. Despite this, the results here clearly demonstrate that this model in its current configuration is underpowered for these applications. This is further evidenced by the dramatic difference in classification performance for a comparatively lightweight alternative, the Combined CCA method reported in [180] (see, Table 2.1).

	AoC	Min K	Max K	Standard Deviation	High-Pass Cutoff	Low-Pass Cutoff	Epochs
Sub 1	0.24	0.23	0.29	0.033	8.25	15.25	88
Sub 2	0.18	0.17	0.19	0.010	7.5	15	49
Sub 3	0.33	0.29	0.40	0.057	9	17.5	80
Sub 4	0.52	0.47	0.59	0.066	8.5	15	112
Sub 5	0.41	0.37	0.45	0.040	9	16	75
Sub 6	0.58	0.56	0.63	0.038	9	15	147
Sub 7	0.47	0.45	0.50	0.029	9	15	162
Sub 8	0.61	0.57	0.68	0.057	8.75	15	100
Sub 9	0.45	0.42	0.50	0.043	9	15.25	129
Sub 10	0.52	0.50	0.57	0.038	8.5	15	146
Mean	0.43	0.40	0.48	0.041	8.65	15.4	108.8

Table 6.8: Here is presented a table of results for the ShallowConvNet model in the Median Pruner optimization study for all 10 subjects (Sub) tested. The metrics reported herein relate to the highest-performing model variant generated from the optimization study. For further information on column headings and table interpretation refer to, Table 6.5.

The relationship between the number of trainable parameters and optimized frequency filter cutoff values is repeated here again. As shown in the minimal cross-subject variance for optimized high-pass (Mean = 8.65 Hz) and low-pass (Mean = 15.4 Hz) cutoffs, the ShallowConvNet was consistently optimized to narrow frequency ranges at the very limit of the target SSVEPs presented to the subjects. This suggests relatively low-power models benefit from a more heavily processed input EEG signal as the reduced number of convolutional filters prevents the effective parsing of additional harmonic information from redundant waveforms in the frequency space above 15 Hz.

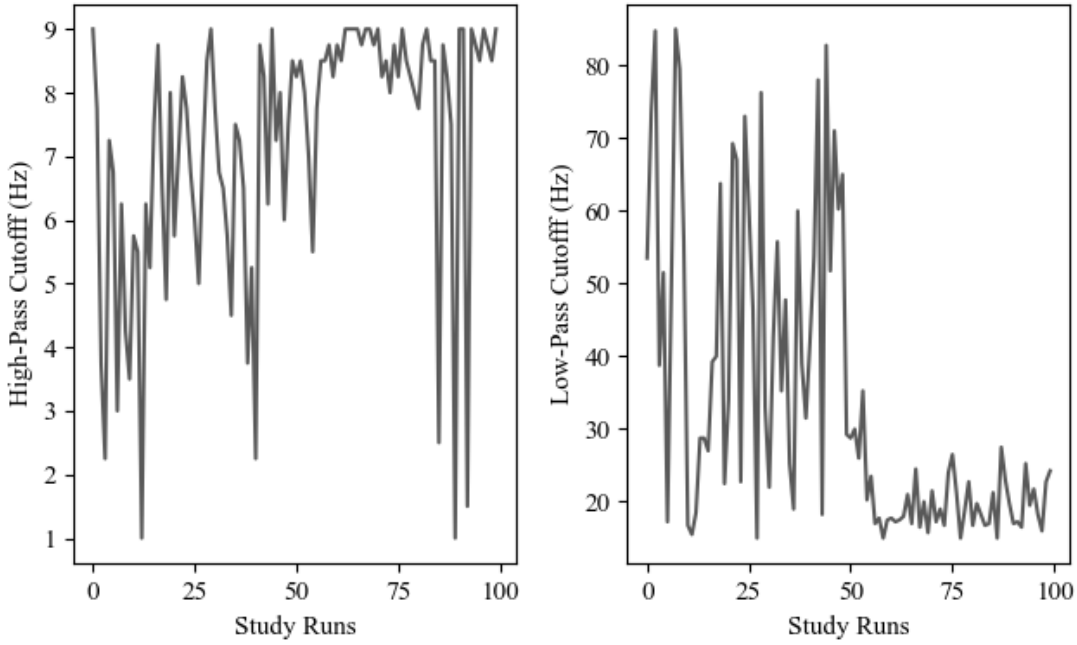

Figure 6.9: The figure presents all high-pass (left plot) and low-pass (right plot) Median Pruner algorithmic parameter search selections for Subject 3 in the respective ShallowConvNet model optimization study (100 trials). For further information refer to, Figure 6.6.

The time required for the optimization study totalled 46.35 hours, at an average of 278 minutes per subject. The higher optimization duration for the ShallowConvNet is notable when compared to the DeepConvNet and EEGNet. It would be reasonable to assume that a higher number of trainable parameters in the aforementioned models would necessitate more optimization, especially given the similar average epochs per subject. This could be accounted for by the unique log and square activation functions present in the ShallowConvNet.

On the contrary, it is the author's opinion that the ShallowConvNet has a slower and more consistent rate of loss reduction over the training period (see, Appendix: Figure A.19). As the optimization procedure required at least 40 epochs for the k-fold operation to be engaged, this slower more consistent rate ensured the ShallowConvNet had a greater chance of not being at the convergence point or in a local minimum once the 40-epoch point was breached (see, Figure 6.2). Essentially, this ensured that ShallowConvNet study trials across subjects were trained for more epochs on average. Given these findings, future efforts to optimize neural networks for bio-signal classification can not safely assume that the relative complexity of a model correlates with the total optimization study duration. Ultimately, considerations for efficiency saving must be applied uniformly irrespective of model size or depth.

6.7.4 Pruner Assessments

Following the initial optimization studies performed using the Median pruner parameter search algorithm additional studies were run utilizing the Percentile and Successive Halving pruners (for further information see subsection 6.6.6).

6.7.4.1 Classification Accuracy

As seen in Table 6.9, there exists minimal variance in the best trial AoCs reported for each network assessed. The differences computed across all individual subject AoCs for each respective pruner algorithm were shown to be highly non-significant ($p > 0.05$). A detailed breakdown of all subject-specific classification accuracies for the Percentile and Successive Halving pruners is presented in Appendix A.7.

	EEGNet	EEGNetSSVEP	DeepConvNet	ShallowConvNet	Mean
Non-Optimized	0.57	0.76	0.65	0.36	0.59
Median	0.53	0.77	0.67	0.43	0.60
Percentile	0.51	**0.78**	0.65	0.42	0.59
Successive Halving	0.51	0.77	0.66	0.41	0.59

Table 6.9: Here is presented a table of cross-subject mean classification accuracies for each respective model assessed (upper row) showing the non-optimized baseline study results (see Table 6.4) and all optimization pruner algorithms tested (left column). A compound metric (mean) is positioned to the right of these values to clarify pruner performance across all networks evaluated.

6.7.4.2 Optimized Filter Frequency Cutoffs

The table positioned below (see, Table 6.10) shows each optimal high-pass filter value computed following every respective subject, pruner and model study combination evaluated. Nearly all networks tested, irrespective of depth or number of trainable parameters, demonstrated a directional preference for frequency values close to the lower boundary of the target SSVEPs assessed (9.25 Hz). Despite this, the relative strength of this pattern does differ across the networks evaluated. It is clear from Table 6.10 that as network complexity increases, this upper boundary directional trend weakens. This can be seen by ranking the models in terms of relative power alongside the within-subject, cross-pruner optimized filter means, respectively EEGNetSSVEP (6.93 Hz), DeepConvNet (7.90 Hz), EEGNet (7.48 Hz) and ShallowConvNet (8.63 Hz) (see, **bold** values).

Note, that the variance across cutoff values within each model tested is minimal. The only comparison approaching significance is the difference between the Percentile (8.23 Hz) and SHP (6.28 Hz) high-pass cutoff values for the EEGNet model variant (see, **bold** values). All other differences in mean optimized filter values between pruners for the same models (see, BS Mean column) vary less than 1 Hz.

Of the numerous network, pruner and subject combinations assessed all presented with mean high-pass values exceeding 6.5 Hz. This validates the suggestion raised in [61] that the removal of these frequencies, previously shown to represent eye-movement artefacts, dramatically improves resultant data signal-to-noise ratios. Along these very same lines, the removal of redundant and potentially confounding signals in the training and validation datasets enables network learning to focus on the extraction of task-relevant neural patterns, as opposed to filtering unnecessary noise. Based on these observations the author recommends the hard-coded application of a threshold near the lower bound of the target SSVEP frequencies utilized (9.25 Hz) in the stimulus, irrespective of the pruner algorithm.

		Sub 1	Sub 2	Sub 3	Sub 4	Sub 5	Sub 6	Sub 7	Sub 8	Sub 9	Sub 10	WS Mean	BS Mean	BS Diff
EEGNet	Median	9.00	9.00	7.75	9.00	1.50	8.75	9.00	8.25	8.50	8.75	n/a	7.95	3.75
	Percentile	9.00	7.50	9.00	8.75	6.25	8.00	8.75	7.50	8.50	9.00	n/a	**8.23**	1.38
	SHP	8.75	8.75	3.50	5.75	1.75	8.75	8.50	6.75	1.25	9.00	n/a	**6.28**	3.88
	Mean	8.92	8.42	6.75	7.83	3.17	8.50	8.75	7.50	6.08	8.92	**7.48**	n/a	n/a
	Diff+/−	0.13	0.75	2.75	1.63	2.38	0.38	0.25	0.75	3.63	0.13	1.28	n/a	n/a
SSVEP	Median	7.75	5.75	8.50	5.00	8.75	8.00	8.25	8.75	5.00	6.25	n/a	7.20	1.88
	Percentile	8.50	4.75	8.50	3.25	8.75	5.00	9.00	8.75	8.50	6.00	n/a	7.10	2.88
	SHP	7.75	3.50	7.75	5.25	7.50	6.00	5.25	8.50	7.50	5.75	n/a	6.48	2.50
	Mean	8.00	4.67	8.25	4.50	8.33	6.33	7.50	8.67	7.00	6.00	**6.93**	n/a	n/a
	Diff +/−	0.38	1.13	0.38	1.00	0.63	1.50	1.88	0.13	1.75	0.25	0.90	n/a	n/a
DCN	Median	9.00	7.25	7.75	6.75	8.50	8.50	9.00	8.75	8.25	6.50	n/a	8.03	1.25
	Percentile	8.25	5.50	8.25	9.00	8.75	8.50	9.00	8.50	8.25	6.50	n/a	8.05	1.75
	SHP	9.00	7.50	8.50	8.25	3.50	7.75	8.25	8.50	7.75	7.25	n/a	7.63	2.75
	Mean	8.75	6.75	8.17	8.00	6.92	8.25	8.75	8.58	8.08	6.75	**7.90**	n/a	n/a
	Diff +/−	0.38	1.00	0.38	1.13	2.63	0.38	0.38	0.13	0.25	0.38	0.70	n/a	n/a
SCN	Median	8.25	7.50	9.00	8.50	9.00	9.00	9.00	8.75	9.00	8.50	n/a	8.65	0.75
	Percentile	7.75	8.50	8.00	9.00	8.25	8.75	9.00	8.75	9.00	9.00	n/a	8.60	0.63
	SHP	8.75	8.50	7.50	9.00	8.50	8.75	8.50	9.00	9.00	9.00	n/a	8.65	0.75
	Mean	8.25	8.17	8.17	8.83	8.58	8.83	8.83	8.83	9.00	8.83	**8.63**	n/a	n/a
	Diff +/−	0.50	0.50	0.75	0.25	0.38	0.13	0.25	0.13	0.00	0.25	0.31	n/a	n/a

Table 6.10: This table contains all the best trial-optimized high-pass filter cutoffs for each model and pruner pair tested. Here, the DCN and SCN rows relate to the DeepConvNet and ShallowConvNet. The mean positioned below the Successive Halving Pruner (SHP) metric row is calculated from all within-subject values across each pruner assessed. The 'Diff +/−' row header relates to the computation of the range of within-subject values across all pruners assessed divided by 2. The WS Mean column reports the mean of the averaged within-subject frequency cutoff values and the mean of within-subject differences. The BS Mean header reports the between-subject optimized high-pass frequency cutoff value means for each pruner tested. The BS Diff column shows the difference of these cross-subject values (range / 2), these are in **bold** to aid in comparing the variances of within-subject and between-subject optimized values.

Similarly, the table positioned below (see, Table 6.11) shows the optimized low-pass cutoff values for all subject, model and pruner combined assessments. In almost all instances a large range of values can be observed between the subjects evaluated, with some of the larger differences driven primarily by outliers, as seen in the Percentile and SHP pruner mean high-pass values for ShallowConvNet (see, **bold** values). The increased variance between subjects for the same network is expected for the low-pass values, as it follows that the individual differences in the expression of SSVEPs would create unique neural signatures containing alternative levels of harmonic expression. As shown for the DeepConvNet SHP pruner (see, **bold** values), a substantial range is present between optimized values for Subjects 7 (15.25 Hz) and 5 (80.50 Hz).

		Sub 1	Sub 2	Sub 3	Sub 4	Sub 5	Sub 6	Sub 7	Sub 8	Sub 9	Sub 10	WS Mean	BS Mean	BS Diff
EEGNet	Median	15.50	15.50	17.25	15.00	41.75	15.00	15.25	15.00	21.75	17.25	n/a	18.93	**13.38**
	Percentile	23.00	15.25	16.25	15.00	61.25	35.75	20.75	16.25	20.00	16.00	n/a	23.95	**23.13**
	SHP	20.50	15.00	16.00	80.25	22.25	15.00	15.00	22.25	48.75	15.50	n/a	27.05	**32.63**
	Mean	19.67	15.25	16.50	36.75	41.75	21.92	17.00	17.83	30.17	16.25	**23.31**	n/a	n/a
	Diff +/−	**3.75**	0.25	0.63	32.63	19.50	10.38	2.88	3.63	14.38	0.88	**8.89**	n/a	n/a
SSVEP	Median	52.50	16.75	75.00	45.50	58.25	58.00	16.75	69.75	70.50	80.00	n/a	54.30	**31.63**
	Percentile	73.25	17.00	50.00	43.25	83.25	54.50	50.75	68.25	62.00	67.75	n/a	57.00	**33.13**
	SHP	74.00	18.00	83.50	64.75	70.50	54.50	63.25	47.50	53.25	32.50	n/a	56.18	**32.75**
	Mean	66.58	17.25	69.50	51.17	70.67	55.67	43.58	61.83	61.92	60.08	**55.83**	n/a	n/a
	Diff +/−	10.75	0.63	16.75	10.75	12.50	1.75	23.25	11.13	8.63	23.75	**11.99**	n/a	n/a
DCN	Median	29.75	23.25	21.25	15.00	37.50	24.00	17.50	19.25	36.00	33.00	n/a	25.65	11.25
	Percentile	78.00	17.25	46.75	16.50	32.75	21.25	40.00	20.00	40.00	74.25	n/a	38.68	30.75
	SHP	29.50	16.75	20.00	19.00	**80.50**	17.25	**15.25**	64.75	46.25	29.50	n/a	33.88	32.63
	Mean	45.75	19.08	29.33	16.83	50.25	20.83	24.25	34.67	40.75	45.58	**32.73**	n/a	n/a
	Diff +/−	24.25	3.25	13.38	2.00	23.88	3.38	12.38	22.75	5.13	22.38	13.28	n/a	n/a
SCN	Median	15.25	15.00	17.50	15.00	16.00	15.00	15.00	15.00	15.25	15.00	n/a	15.40	1.25
	Percentile	15.25	**72.00**	15.75	15.00	19.25	15.00	15.00	15.00	18.50	16.00	n/a	21.68	28.50
	SHP	16.75	**75.50**	16.50	16.00	21.75	15.25	15.50	17.00	17.50	15.00	n/a	22.68	30.25
	Mean	15.75	54.17	16.58	15.33	19.00	15.08	15.17	15.67	17.08	15.33	**19.92**	n/a	n/a
	Diff +/−	0.75	30.25	0.88	0.50	2.88	0.13	0.25	1.00	1.63	0.50	3.88	n/a	n/a

Table 6.11: Here is shown a table containing all the best trial-optimized low-pass filter cutoffs for each subject, model and pruner combination assessed. For further information on table headers and interpretation see, Table 6.10.

In the majority of instances, the variance across the same subject for different pruner methods is substantially lower than the variance in optimized values across subjects. As shown, the EEGNet WS Mean difference is computed by averaging the range in optimized frequencies across all pruners tested. As shown for Subject 1, the value is reported as $+/-$ 3.75 Hz (see, **bold** value). All these within-subject differences are then averaged again to produce a mean within-subject difference of 8.89 Hz (see, **bold** value). This is substantially lower than the between-subject differences computed across all individuals tested via the Median (13.38 Hz), Percentile (23.13 Hz) and SHP (32.63 Hz) pruners. Again, this pattern is also repeated in the EEGNetSSVEP model. The mean difference within subjects across the 3 pruners tested is just $+/-11.99$ Hz (see, **bold** value), as compared to the mean difference calculated across all subjects computed at 31.63 Hz, 33.13 Hz and 32.75 Hz respectively for the Median, Percentile and Successive Halving Pruners (see, **bold** values).

This suggests that all pruning methods are coalescing around a similar range of filter cutoff values for each subject-model combination. Further, the same pattern of network complexity and optimized filter frequency range is observed for these data. Specifically, as relative network power increases the mean optimized low-pass filter values also increase, as seen for EEGNetSSVEP (55.83 Hz), DeepConvNet (32.73 Hz), EEGNet (23.31 Hz) and ShallowConvNet (19.92 Hz). This indicates that individual differences in subject data and the level of network complexity account for the majority of the variance seen in these low-pass filter optimization results, as opposed to the different pruner algorithms utilized.

6.7.4.3 Computational Resources

The table below (see, Table 6.12) reveals the amount of time (hours) required for the completion of all subject-specific optimization studies for each respective model and pruner assessed. The Median pruner is shown to require 30-70% more computational processing time as compared to the Percentile and SHP pruners. This culminates in around 50 + hours more to complete the same series of tasks. Despite this, as seen in Table 6.9, these differences in computational time do not translate into higher classification performance metrics.

	Median (hours)	Percentile (hours)	Successive Halving (hours)
EEGNet	37.94	30.09	28.37
EEGNetSSVEP	66.41	47.87	51.29
DeepConvNet	36.42	28.05	25.88
ShallowConvNet	46.35	30.29	26.56
Total	187.12	136.29	132.10

Table 6.12: Here is presented a table containing the total amount of computational processing time required for each respective model using all three algorithmic pruners assessed, the Median, Percentile and Successive Halving (for more information see subsection 6.6.6). All durations are presented in hours and the sum of all individual model pruner assessments is positioned on the bottom row.

The Median Pruner functions by retaining any trial providing the current loss metric is lower than the median of all other trials conducted for the given number of epochs. In other words, any network configuration must only perform better than half the trials reported for training to continue. This is a substantially less rigorous criterion than the Percentile pruner, as the rejection threshold was set at 25% , meaning to be retained, the current trial was required to perform at or above the top 25% results currently collected. These findings indicate that the less aggressive configuration of the Median pruner led to many more epochs per trial.

Notably, the SHP pruner method records the lowest total optimization duration (132.10 hours), requiring two fewer days of processing time than the Median pruner. Despite these impressive results, as seen in both Tables 6.10 and 6.11, the method arguably produced the most outliers in optimized filter cutoff values. Further, this method reported relatively higher computation resource demands for the highest-performing network tested, the EEGNetSSVEP model, as compared to the Percentile method. On balance, the author asserts that the Percentile pruner is arguably the most capable parameter selection method due to the combination of lower computational processing duration, consistent between-subject parameter variance and the highest, albeit non-significant, mean cross-subject classification accuracy produced in any of the studies performed at 78% (see, Table 6.9). Further exploration is needed as all pruners were implemented using the default Optuna library [271] settings.

Given the extensive computational resources required to optimize neural networks, replicating the exact methodology employed in this study would be impractical for future work. The primary aim of this study was to establish a systematic approach for optimizing signal pre-processing parameters in SSVEP-based BCI speller applications. Based on the findings presented, future researchers can circumvent similar computational challenges by leveraging the insights provided here. Specifically, the results indicate that optimizing filter values

within the frequency range of 0 to 9 Hz, or the lowest frequency present in the stimulus array, yields limited benefits. Additionally, extending the low-pass filter cutoff beyond 30 Hz is advisable only when employing a more robust neural network with additional layers and trainable parameters. It is estimated that these insights alone could reduce the search time for optimization by over 50%, by constraining the optimization space to align with stimulus parameters and network capabilities. Furthermore, it is plausible that an initial deployment of a non-optimized SSVEP system could be employed by new users. Subsequent data collection could then facilitate the application of the optimization process to enhance the classification performance of the model.

6.8 Conclusion

Here are presented the observational interpretations relating to all previously outlined research. These will be discussed chronologically and a final summary of the work will be presented alongside suggestions for improvements and future research.

6.8.1 Raw Data

In the raw data training and evaluations, no subject or model showed any significant increase in the random performance threshold of 8.33% (see, Table 6.3). This pattern was replicated independently of the relative network complexity or prior subject-specific classification accuracies observed in the original FBCCA [180], subsequent EEGNet [59] or EEGNetSSVEP analyses [60]. In summary, for the models tested here, it is clear that the application of some frequency-based signal filtering is necessary to generate effective classification methods for the SSVEP waveforms assessed.

Further, subsequent analysis reveals that the high-pass filter level is typically pegged at or around 9 Hz for all models evaluated (see, Table 6.10). In most instances, models demonstrating optimized low-pass filter cutoff values above 30 Hz in this research relate to the EEGNetSSVEP model. Given these findings, it is possible that models with a greater number of convolutional filters only require high-pass frequency filtering to remove eye movement and associated noise artefacts ($<=$ 9 Hz). The author asserts that future research would likely reveal that the increased computational power of these networks could allow for the useful extraction of harmonic information in higher frequency ranges and the application of low-pass filters set at or below 15 Hz could hinder this process.

6.8.2 Fixed-Parameter Data

Notably, the networks defined herein did not reach the same level of performance described in the original study [180].Here the authors implemented a Combined-CCA method to achieve a mean accuracy of 92.78% and a mean ITR of 91.68bpm. This involved integrating reference sinusoids with user-specific data for all target stimulation frequencies utilized, for further information see subsection 2.6.2 Cutting-Edge Classification Methods for SSVEPs: Filter-Bank Canonical Correlation Analysis and Table 2.1. These results, alongside other FBCCA methods, might suggest there is little need for the development of alternative SSVEP classifiers, however, given the dramatic advances in neural networks for similar signal-processing tasks it is feasible that given the requisite academic scrutiny, more robust and higher accuracy decoder systems could be achievable. Ultimately, improved classification methods translate into better system performance and a more satisfactory user outcome for potential future patient populations. For more information regarding these topics please refer to subsections 2.6.1, 2.6.2 and 2.6.4.

Along these very same lines, in the original paper introducing the EEGNetSSVEP [60], this model was compared against a variant of the Combined-CCA where no user-specific data was utilized and tested using the same data repository as in [180], the same open-source 12-target SSVEP repository used in this optimization study. This involved integrating reference sinusoids with cross-subject averaged data for the target frequencies tested, allowing researchers to bypass the need for a calibration stage. Ultimately, the alternate implementation of the Combined-CCA method performed substantially lower than the original, dropping as much as 60% in classification accuracy for some subjects. The EEGNetSSVEP model performed much better in this non-calibration data context achieving a mean of around 80% across all subjects.

Crucially, the EEGNet and EEGNetSSVEP models tested herein did not replicate the classification performance metrics attained in the original EEGNet [59] or EEGNetSSVEP [60] articles. This is likely owing to the absence of specific details relating to the training scheme and data preparation methods. Despite these results, there is a high degree of similarity between the reported stats and crucially the same pattern of relative subject-specific AoCs is present. This is evidenced by the analyses recording the same highest (Subject 8) and lowest-performing (Subject 2) subjects. The author notes that many subjects demonstrate performance values below the functional operation threshold of 70%. Despite this, the consistency in the operation of the models under fixed parameter signal pre-processing conditions suggests the networks can function within narrow performance metric windows. These analytical

features are crucial as they suggest further improvements in network design could lead to the development of a BCI speller methodology with high levels of repeatability.

Further, as expected, the mean model-level performance displays a strong positive relationship between network complexity and classification accuracies, with the EEGNetSSVEP model reporting the highest cross-subject mean AoC (76%) and the ShallowConvNet returning the lowest mean AoC (36%). Interestingly, the EEGNetSSVEP model significantly outperformed the DeepConvNet architecture (Mean AoC = 65%) despite this network possessing a greater number of model layers (see, Appendix: Figure A.14). This is likely owing to the significantly higher number of convolutional filters in addition to the initial Separable and Depth-wise convolutional operations featured in the EEGNetSSVEP model.

6.8.3 Optimized-Parameter Data

Here all results are reported concerning the optimization methods implemented using the Median pruner.

6.8.3.1 Median Pruner | EEGNet

As stated earlier, the EEGNet optimizations failed to improve the classification accuracy of any individual subject assessed (see, Table 6.5) and reported a significantly lower cross-subject mean AoC than those initially reported for the fixed parameter results (see, Table 6.4). Firstly, this suggests that the fixed parameters utilized for the study were potentially optimized by the original authors for this very model. Note, that this is presumed from the metrics reported here and these claims are not explicitly attested to in the reference article [60]. These findings suggest that a low learning rate and a higher number of training epochs are required for the maximization of some convolutional neural network configurations, specifically those with a lower relative number of convolutional filters, as in the EEGNet model.

The absence of a strong trend towards lower or higher cutoff values, as observed for the other models (see, Figures 6.7, 6.8 & 6.9, right plot), suggests the optimization method did not converge towards an optimized frequency range for the low-pass cutoff value (see, Figure 6.6, right plot). Given that, nearly all models expressed a preference for a high-pass cut-off value around 9 Hz, future optimizations could solely focus on the parameter search of the low-pass filter threshold. This would ensure that significantly fewer trials are wasted as in the numerous instances of extremely low high-pass filter selections. The freeing up of computational resources exclusively to the search of low-pass filter values could decrease the likelihood of

the study terminating before convergence around an optimal value range.

6.8.3.2 Median Pruner | EEGNetSSVEP

Concerning the EEGNetSSVEP model performance following the optimization process, some increase in AoCs reported is observed at the cross-subject level (see, Table 6.6) as compared to the fixed parameter results (see, Table 6.4). Despite this, the differences remain non-significant. This could be due to the models requiring slightly more epochs of training than the early stopping and pruner afforded.

The largest boots to performance are found at the single subject level, with Subject 2 reporting an increase of 7.4% . Interestingly, the few improvements noted were not restricted solely to low or high-performing subjects. This suggests there is scope for these methods to boost classification performance and in turn BCI speller functionality across a wide range of end-point users. Moreover, this suggests that the bespoke tailoring of pre-processing stages is viable for users presenting with classification accuracies furthest from functional usage. It must be noted that the optimization methods detailed did not lead to said functional usage thresholds. The author recognises that these results merely indicate that optimization of filter cutoffs in addition to the parameter search of for example optimal electrode arrangements could enable some users previously classified as BCI illiterate to communicate using speller devices powered via CNN classification methods.

The EEGNetSSVEP model provided lower optimized high-pass filter cutoff values as compared to any of the alternative networks assessed. The increased relative power of this model likely enabled the network to more effectively parse redundant information in this frequency range. Note, that the author maintains the removal of frequencies below the lowest stimulus target frequency is the most valid approach for future investigations. Further, the mean cross-subject optimized low-pass filter cutoff value of 54.3 Hz is unique to the EEGNetSSVEP model variant. This is likely due to the higher relative power of the model, as the network is comprised of substantially more convolutional filters and associated trainable parameters than the EEGNet, DeepConvNet or ShallowConvNet. The author asserts this increased network power resulted in the utilization of higher-order harmonic frequency components.

It must be noted that the EEGNetSSVEP model reported a significantly higher total optimization study duration than the other networks assessed here. At nearly double the amount of time required, as compared to the EEGNet, the process of optimizing this style of deep, high convolutional filter volume networks, presents significant challenges in terms of compu-

tational resources and time. The previous suggestion to increase the number of filters further is likely, not viable utilizing the hardware deployed herein. Instead, it is recommended for similar future investigations to utilize GPU hardware possessing over 14GB of VRAM at the highest attainable clock speed and at a minimum, 32GB of RAM.

6.8.3.3 Median Pruner | DeepConvNet

As seen in Table 6.7, at the cross-subject level, mean classification accuracies for the DeepConvNet model increased by 2.3% as compared to the original fixed parameter results (see, Table 6.4). Crucially, the enhancement in performance was significant ($p < 0.05$). Note, that the cross-subject mean AoC of 67.2% nearly breaches the 70% usage threshold for BCI functionality. This is achieved using a model with significantly fewer trainable parameters and optimized in nearly half the time as the EEGNetSSVEP network, with the entire optimization study requiring 36.42 hours to complete, operating at an average of 218 minutes per subject. This boost in performance is observed uniformly across 8 subjects excluding Subject 3 which demonstrated a marginal increase ($+$ 0.9%) and Subject 5 which showed a substantial decrease in accuracy ($-$ 4%). The effect is most pronounced in Subjects 2 ($+$ 3.5%), 7 ($+$ 3.4%) and 9 ($+$ 4.7%). Crucially, these subjects demonstrate significant variance in AoC performance, suggesting that the enhancement effects of the optimization process are not dependent on subject data quality.

Both the low-pass and high-pass optimization parameter search patterns reveal that the models successfully converged around unique and narrow frequency ranges. As seen in Figure 6.8 (left plot), a similar parameter value selection pattern emerges for the high-pass cutoff near the 9 Hz boundary. Further, the low-pass selections (right plot) demonstrate a weaker overall trend towards the 15-25 Hz filter range. Interestingly, one subject (Subject 1) reported near-identical parameter selections (high-pass $=$ 9 Hz, low-pass $=$ 29.75 Hz) to the original values. Notably, the increase in performance reported ($+$ 1.8%) is highly marginal. Despite this, the number of epochs used to train each model differs greatly as the Median Pruner optimized network was terminated at just 76 epochs, compared to the 500 epochs used in the training of the original fixed parameter network. These findings suggest aggressive early stopping mechanisms have significant potential to enhance optimization efficiencies in this analysis context.

6.8.3.4 Median Pruner | ShallowConvNet

The increase in mean, cross-subject performance (+ 6.73%) for the ShallowConvNet model following the optimization process is highly significant ($p < 0.005$) (see, Tables, 6.4 & 6.8). This boost in classification accuracies still led to mean network performance (43%) well below functional usage (70%) [250]. There do exist potential applications for these models in SSVEP-based bio-signal classification in lower complexity tasks. This could manifest as an adapted version of the Emoji-Based BCI speller introduced in the first half of this thesis. The system could be deployed using relatively low-cost hardware and as shown in these results, performance could be boosted via the subject-specific optimization of cross-subject data. The network could initially be trained using available data repositories [49, 52, 180] via a transfer learning protocol and finally tuned on subject-specific data as per [215].

Notably, the same positive correlation between network complexity and bandpass filter width is replicated again in these model results. The author asserts that the reduced number of CNN layers in the ShallowConvNet led to a smaller low-pass cutoff value (15.4 Hz, see, Table 6.8), as compared to the deeper DeepConvNet (25.65 Hz, see, Table 6.7). This pattern is also reflected in comparisons between the EEGNetSSVEP, which demonstrates a mean optimized low-pass cutoff of 54.3 Hz, as opposed to the EEGNet reporting a lower 18.93 Hz optimized filter metric. The aforementioned networks possess the same architectural framework, only differing in terms of kernel sizes and the number of convolutional filters per layer. These findings suggest network depth and convolutional filter volume both increase the tendency for the optimization process to favour higher low-pass filter cutoff values.

6.8.4 Pruner Assessments

Here are present all conclusions and interpretations relating to the analyses conducted to compare the three pruner methods assessed in terms of end-point AoC, parameter search behaviour and optimizer computation durations.

6.8.4.1 Classification Accuracy

As shown in Table 6.9, a high degree of consistency in classification accuracies across the different pruners was reported. This suggests that for the pruner configurations as tested, there is no advantage in terms of end-point performance offered by any of the individual search algorithms. Despite this, the highest mean (compound) accuracy across all networks was shown to be the Median pruner.

6.8.4.2 Optimized Filter-Cutoffs

For all pruners and networks assessed a strong trend towards the 9 Hz upper high-pass frequency bound is shown. Crucially, the degree of adherence to this pattern is closely related to network power. In other words, as network complexity decreases, optimized high-pass filter cutoff values increase. This suggests a preference for more heavily pre-processed data in models with fewer trainable parameters. Irrespective, the author asserts that the removal of all data at or immediately below the lowest target SSVEP frequency propagated is arguably the most sound means of data preparation.

The same relatively consistent pattern of optimized frequency ranges is not replicated for the optimized low-pass filter cutoff values. A significant degree of variance between subjects trained via the same pruner algorithm is notable in nearly all instances. This contrasts against the substantially lower within-subject mean differences observed across the different pruners for the same subject. Further, the results corroborate the previous assertions that as network power increases, so does the optimized low-pass filter cutoff value. This is seen when comparing the EEGNetSSVEP model optimized mean low-pass cutoff value of 55.83 Hz against the ShallowConvNet model optimized value of just 19.92 Hz (see, Table 6.11). This suggests that higher-complexity networks are utilizing target SSVEP harmonic information in the upper-frequency range ($>$ 15 Hz) to perform the classification task. In comparison, the lower complexity models demonstrate a preference for more heavily pre-processed data. This is likely owing to a decrease in capacity for parsing these latent embedded waveform components from background noise.

In sum, these findings indicate that there are present individual differences in the expression of the target SSVEP waveforms that can be exploited to boost classification performance via the development of bespoke signal pre-processing stages, namely unique low-pass frequency filters. Extending these findings into, for example, tailored electrode arrangements could lead to further enhancements in classification performance and ultimately BCI speller functionality for end-point users.

6.8.4.3 Optimizer Study Durations

The analyses reported significant differences in the duration of optimization studies across the three pruners assessed (Table, 6.12). These differences in duration show no clear relationship with end-point classification accuracies (see, Tabel 5.10). The Median Pruner required over 50 additional hours of processing time as compared to the other methods tested. As this

offered no additional improvements in AoCs the authors can not recommend implementing this pruner without significant adaptations to the configuration. Further, as seen in Table 6.11, the Successive Halving Pruner consistently produced the highest mean between-subject differences of all pruners assessed. This is principally owing to the relatively higher incidence of outliers. In consideration of all these factors, the author recommends utilizing the Percentile pruner as the baseline method for future optimization studies.

The three different pruners all required significantly different run times per optimization study. This is principally related to the aggressiveness of the pruning protocol. Further, the higher the pruner tendency for early stopping, the fewer results are generated. This increases the likelihood of producing networks that have not fully converged before pausing training. It could be argued that early stopping is preferable to over-training as this can introduce overfitting. Despite this, the author believes that integrating a Patient-Pruner wrapper (hard-coded minimum number of epochs per trial), removing the independent early stopping feature, or modifying the Percentile and Successive Halving Pruner parameters (for example minimum early stopping rate or number of minimum trials) could also help alleviate these issues. Clearly, a balance between under and over-training produces the most desirable conditions for effective network optimization.

6.8.5 Summary

The performance comparisons between fixed and optimized parameters for the EEGNet show a significant decrease in classification performance following the implementation of the optimization method outlined here. Further, marginal and marginally significant improvements (trend towards increased AoCs) were observed for both the EEGNetSSVEP and the DeepConvNet as well as a significant increase in performance for the ShallowConvNet seen at the single and cross-subject level. The EEGNetSSVEP demonstrated the greatest deviance in low-pass and high-pass parameter selections of all models assessed, suggesting the higher network complexity benefitted from the increased availability of SSVEP harmonic information embedded in the frequency space above the 15 Hz target stimulus rate. Despite this, the network did not produce significantly higher or lower AoCs as compared with the fixed parameter model variant. The model is likely underpowered to truly capitalize on the increase in harmonic information via the larger low-pass cutoff value. The inclusion of more convolutional filters or the introduction of an Inception module, as seen in recent work would likely remedy this [263, 278], albeit at a considerable additional network resource cost.

The most promising results suggest that the models assessed herein do not need to be trained to 500 epochs to achieve similar results to fixed parameter assessments. Further, low-power networks can be boosted via subject-specific filter optimization to observe significant increases at the cross-subject level. Moreover, the implementation of parameter optimization can lead to significant increases for relatively more complex networks at the single-subject level. Crucially, these increases are not restricted exclusively to high or low-performing individuals. In some instances, these classification models will eventually have to be deployed in non-ideal lab conditions with clinical end-point patient users. Attenuated versions of these methods to rapidly optimize the selection of bandpass filter ranges could boost classification performance above functional thresholds. In cases of low viability, a tighter range between low-pass and high-pass filters could be deployed. It could be argued that in these situations it would simply be preferable to implement alternative methods will a higher degree of robustness to noisy EEG characteristics, for example, the FBCCA method.

Note, that none of the results collected at the subject or model level showed any significant increase in performance compared to the original Combined FBCCA implemented for SSVEP classification in the corresponding online SSVEP repo article [180] (see, Table 2.1). This pattern of FBCCA-based method supremacy over contemporary CNN techniques in terms of accuracy and information transfer rates is replicated in the surrounding cutting-edge literature. The highest-performing analysis for this SSVEP repository (Mean AoC = 90.2%) listed to date remains the Task Related Component Analysis outlined in [55] (bpm = 352.3). This is a highly modified version of the FBCCA method used in [56, 180] which features substantial pre-trial tuning of subject-specifc EEG spatial filters before classification. Additionally, the system integrates a Bayesian dynamic stopping method to compute the optimal length of data required for effective target prediction while maximally reducing data capture windows to boost ITRs. The greatest advances in CNN-based methods such as those outlined in this thesis are typically related to integrating these elements into the stimulus design, data pre-processing and classification techniques.

As seen in [50], the implementation of a dynamic stopping protocol for the EEGNet architecture dramatically increased accuracies (+ 28.2%) and ITR values (147.6 bpm) from baseline. Further, arguably the second-highest-performing CNN-based SSVEP network, TRCA-Net [54] (bpm = 235.21 bpm), integrates the TRCA method as a data pre-processing stage. Additionally, the most effective CNN-based classifier to date, outlined in [57] (bpm = 318.41), explicitly organises subject data according to filter-bank-based methods. This is done by aggregating the original data alongside banks of the same pre-filtered samples for numerous filter

bounds to amplify the presence of SSVEP harmonic information in the frequency space above the target stimuli utilized.

It must be stated that all of these studies deploy highly similar 3-4 layer CNN model designs. The most significant advances in CNN-based classifier performance are principally related to data handling and preparation, as opposed to neural network architecture modifications. Crucially, the optimization methods detailed in this thesis demonstrate some key findings concerning the relationships between low-pass filter cutoffs, network complexity, individual differences in subject data and SSVEP harmonic information utilization. Despite this, the most important contribution of the research conducted herein relates to the formalization of a method for the automated evaluation and optimisation of data pre-processing stages across numerous models of differing complexity levels. Along these very same lines, this method could be extended to assist researchers in refining currently employed techniques. This could manifest as an exploration into the optimal number of pre-filtered data aggregations, as mentioned above, for a range of different network configurations. Alternatively, the method could be implemented in the validation of completely novel data pre-processing stages, such as the tuning of cross-subject aggregated data for individual differences in SSVEP phase profiles.

6.8.6 Future Work

It could be argued that the sub-FBCCA (see, Tables 6.2, 6.6 & 6.7) classification performance demonstrated by the DeepConvNet and EEGNetSSVEP could be owing to the ratio of network size *vs*. training data volume. The dramatic increase in computational power required to train these models in comparison to the alternative architectures evaluated (EEGNet & ShallowConvNet) likely demands a commensurate increase in training samples. As seen in the article introducing the ShallowConvNet [61], the authors employed the practice of data cropping as inspired by the original computer vision research that the CNN analysis technique emerged from and significantly enhanced model performance. This involves parsing samples of the extracted data segments into windowed overlapping chunks. The author did not implement this here as the differences in target class signal phases would significantly increase the complexity of the windowing operation. The degree of overlap would have to be calibrated to ensure each windowed data chunk presented with the same phase angle degree. Any errors concerning this process would lead to the comingling of signals containing the same frequency and different phase angles in the same class bin.

Ultimately, the labelling of two signal variants with highly uncorrelated signal properties as one class would effectively undo all the efforts taken to increase the discriminability between

all target classes assessed. Note, the author does not infer from these considerations that the process is unviable, only that considerable care must taken for these data preparation techniques and such explorations lay outside the scope of this research. Crucially, this complexity was avoided in these assessments via the use of strict 1-second neighbouring data segregations from the original 6-second trial samples. By subsampling the original trial data at 1-second intervals, the same phase angle is retained across the classes.

An alternative method to the issue of low training samples is the use of data augmentation methods, specifically generative-adversarial networks (GAN) to produce robust imitations of human-derived SSVEPs [298]. This involves training a GAN using a subset of data, either cross-subject or subject-specific, on high-quality class samples, validating them using a baseline network and aggregating the generated samples for integration into the model training scheme. Crucially, the same considerations concerning the phase angle mentioned above must be taken for successful implementation.

Note, if the phase angle complication cannot be fully overcome, the cropped or generated samples could simply be utilized as a first stage in a transfer learning approach, followed by network exposure to genuine samples. Traditionally, the methods for BCI speller classification are restricted to developing either a highly bespoke method trained exclusively using a single subject's data or a highly generic plug-and-play method. In contrast, the notion of exposing a classifier to numerous different EEG profiles from a range of subjects to boost model robustness has been well-established in the surrounding BCI literature for SSVEP waveform classification. Recently, in the context of SSVEP-based BCI spellers, this has been extended to pre-training networks and performing batched transfer learning across multiple different SSVEP datasets [215]. This functions by initializing the weights extensively via exposure to non-end-point target SSVEP waveforms and later tuning these flexible representations to the finalized state immediately before deployment. A systematized chain of pre-training and transfer learning with cropped and augmented SSVEP datasets, for an Inception-style SSVEP-based deep convolutional neural networks, optimized for both signal pre-processing and model hyper-parameters could prove the most effective means of classifying these bio-signals to date.

Overall, the results herein reinforce the general notion that higher complexity or relative power in network design correlates with increased classification accuracies. The limitations imposed by the data modality (downsampled time series) and current hardware reduce the viability of simply increasing network depth (number of layers) or the number of convolutional filters to improve classifier functionality. In contrast, further advances in brain-based bio-signal net-

work performance have been achieved by integrating advancements originally defined in the computer vision literature. This relates to the integration of so-called Inception modules in model architectures. Recent networks utilizing these design features have reported impressive results for similar online SSVEP datasets. Crucially, these networks have managed to retain waveform representations while scaling via the inclusion of parallel convolutional blocks that share connections across layers [263]. Further, the increased implementation of effective data preparation techniques established in the FBCCA and TRCA-based SSVEP classification literature would undoubtedly prove highly fruitful in boosting BCI speller performance.

Chapter 7

Conclusion

7.1 Project Trajectory

The initial aims of the thesis centred on the development of an emoji-based P300 BCI speller system. These investigations were intended to assess the efficacy of emoji-based stimuli for the propagation of viable ERP waveforms, as has previously been explored for more traditional speller array targets, namely letters, numbers, and characters. The rationale behind these assessments was principally couched on the premise that some individuals presenting with the most severe forms of paralysis (Incomplete Locked-In Syndrome) could experience rapid fatigue effects during the use of more commonly utilized 6×6 alphanumeric arrays. In response to these concerns, the reduced density 7 emoji stimulus task variant was introduced to provide a simplified communication format that maximised emotional expressivity. Owing to previous research utilizing icons and device command instructions [44] as stimulus targets, as well as literature demonstrating the P300 peak boosting effects of human face images as augmentation overlay stimuli [153] it was predicted that these emoji targets would perform well in this role. The validation of these speller experiments as robust communication platforms would enable the collection of a large, high-quality data repository necessary for the training of a bespoke CNN-based classifier pipeline. The use of these prediction methods has been adopted widely in the processing of SSVEP-based speller data (see, subsection 2.6.3) and the author aimed to extend the research of the comparatively less developed CNN-based classification of P300 signals.

Further, the author notes that the availability of online bio-signal repositories has dramatically improved the capacity of researchers to collaborate globally on the development of novel bio-signal classification analysis methods. Along these very same lines, the author intended to host all data collected to aid in this process and crucially, upload the corresponding stimulus code.

The issue of analysis replication arguably presents with fewer obstacles owing to the use of Git repository cloning, as opposed to data replication which often requires the implementation of the same hardware for stimulus presentation via computer monitor and respective graphical processing unit in addition to the same data acquisition hardware. These systems are far less flexible; however, efforts must be made to standardize the methods of replication for this key aspect of BCI research.

Continuing this project pathway, the verification of emoji as a viable ERP stimulus would validate the development of an integrated alpha-numeral-emoji array. As noted in the introduction, emojis are typically utilized specifically for text embellishment and serve to enhance communication clarity and specificity. The positioning of these targets within an integrated array was predicted to enable users to maximally harness the expressive capacity of text-based communications. The final iteration of the experimental series was intended to assess the performance of an integrated emoji speller with real-time classification via CNN and post-prediction user feedback.

As is clear from the thesis composition and initial introduction Positioning Statement (see subsection 1.1), the trajectory of this project was altered dramatically owing to the fallout of the COVID-19 pandemic. The P300 emoji-speller project pathway was interrupted and a new area of research was explored that was conducive to the offline restrictions imposed. These comprise the investigations into the methods for optimizing subject-specific pre-processing parameters to boost end-point model classification accuracies across networks of different complexity levels. Note that the experimental aims previously outlined in the P300 speller domain still constitute a fruitful line of research and recent studies aligning with these goals provide evidence to support these assertions [299–301]. This cutting-edge literature and future research considerations for both fields will be discussed following an evaluation of the contributions relating to each respective project defined herein.

7.2 P300 Experimental Series Contributions

The progression of experiments detailed throughout Chapters 3, 4 and 5 demonstrate the substantial efforts employed by the author to explore the initial aims of the thesis. Specifically, these investigations aimed to probe the efficacy of emojis as stimulus targets in an emotion-communication BCI speller context. Following Experiment 1, the author identified the need to modify data buffering protocols, explore impedance-based channel rejection methods, address subject training concerns and identify the need to investigate stimulus array density and

salience effects. The completion of Experiment 2 led to the author developing: an active impedance monitoring system, a localizer data pre-screening and LDA tuning system and a real-time classification and user target prediction feedback method. Overall, these modifications aimed to address issues relating to data quality, LDA model overfitting and subject vigilance, training and fatigue considerations.

The final iteration of the emoji-speller system outlined in 5.4.5 combines all of the adaptations explored in the previous chapters. As shown in the results section relating to the LOCRT experimental variant, 5.5.3 suggested that the method was viable for a single subject (Subject 1) of the 3 assessed and marginal for another individual tested (Subject 3). This LOCRT variant involved the training of LDA models using localizer data before the 7 emoji main experiment and is positioned as the only true real-time classification method explored herein. As noted above and in the corresponding subsection 5.6.1, the low sample size precludes the authors from asserting that these methods are fully validated as reliable and robust in terms of P300 waveform elicitation for the respective emoji-speller context.

Despite these promising findings, the re-analysis of these data via the Pipeline 2 method revealed substantial flaws in the data organisation, pre-processing and analysis associated with Experiments 1, 2 and 3. These were addressed via the implementation of cross-validation procedures, the SMOTE oversampling method and improved baselining and high-pass filtering methods. The offline results produced following these steps produced significantly higher mean classification accuracies across all subjects tested and near-maximal classification of the P300 target samples.

The corresponding code features EEG stream handling functions, data organisation and preprocessing stages, LDA classifier options, the stimulus presentation programmes for the localizer and main experiment as well as tools for the active monitoring of impedance quality during live data acquisition. In addition, ancillary software for data quality assessment via grand average plotting, spectrogram analysis and subject fatigue estimation via Alpha-Theta ratio computations are also included. The author predicts that the reimplementation and adaptation of these methods could assist future researchers by providing a standardized method for capturing, processing and classifying P300-based emoji speller data. In line with these intentions, all aforementioned code will be made available via the corresponding GitHub repository: https://github.com/JoshPod93/EEG_Emoji.

Note, that some efforts to automate experimental features across different systems have been made. This relates to the active detection of presentation monitor dimensions and relative adjustments to the emoji stimulus sizings. Additionally, the respective data-handling functions possess some flexibility in target channel parsing and reference assignment. For the best results, adherence to the system and package requirements is recommended. Further, it must be understood that some bespoke adaptations to the code across platforms are unavoidable.

7.2.1 Future Research

Numerous enhancements are available to improve the final version of the P300 speller defined herein. Firstly, an increase in the number of sequences presented per trial in both the localizer and main experiment would likely enhance the discriminability of P300 and Non-P300 targets by improving the quality of cross-trial average signals. These adaptations were not implemented here due to concerns surrounding an increase in end-point ITR values. This operational performance aspect of the speller was not directly aligned with the key initial goal of the thesis, namely, the assessment of emojis as a viable stimulus for P300 elicitation. The experiment should have been focused primarily on attaining these initial goals before attending to the optimization of BCI functionality metrics.

Despite the predictions that an increase in sequences per trial would have likely improved classifier accuracies, the limited, one-dimensional valance scale here employed would unavoidably restrict the emotional expressivity of the endpoint user. The pleasure-centric Likert scale was adopted principally as a means of providing operators with a clear and simple interface to communicate emotional state information with emojis arranged from agreeable to disagreeable. This experimental user interface was influenced by the affective slider method outlined in [159], however only represents half of the original affective rating instrument, as a secondary scale was also deployed to allow for the reporting of arousal state from low to high. The authors argued that this combination of scale values provided a broad and efficient means of capturing emotional judgements from users.

Along these very same lines, it could be argued that the BCI emoji speller defined herein could be enhanced by adopting a hybrid design in which a pleasure scale trial is followed by an arousal scale trial. This later trial would present a different series of emoji onscreen varying in levels of low to high intensity, as per the arrangement of the current array. Note that, the increase in clarity of emotional state information would also lead to a contingent increase in time spent per communication delivered and, an increase in computational complexity given the need for twice as many classification predictions as compared to the current version.

An alternative method of presentation is suggested by the author that boosts emotional communication specificity, increases the number of affective state targets, and crucially reduces the number of sequences per trial required to differentiate all targets. This involves the adoption of a stimulus array composed of emojis and icons embedded in a 3×3 numpad-style matrix augmented according to a randomised row/column flash protocol (see, Figure 6.1). For this design, only 6 augmentation sequences (3 rows and 3 columns) are needed to generate a unique pairing of P300 propagation instances to identify the target emoji/icon intended for selection. This change in stimulus array design boosts the number of emoji/icons on the screen (7 to 9) while reducing the number of augmentations required per trial (7 to 6). Further, the use of dedicated 'yes' (thumbs up) and 'no' (thumbs down) icons provides greater flexibility for communications in real-time and provides the speller system additional functionality beyond emotional expression. Moreover, the adoption of these methods would both increase the information transfer rate upper limits while also widening the scope of emotional expressivity from a bi-directional agreeable-disagreeable valence scale to 7 of the 8 key facial expression categories (excluding contempt).

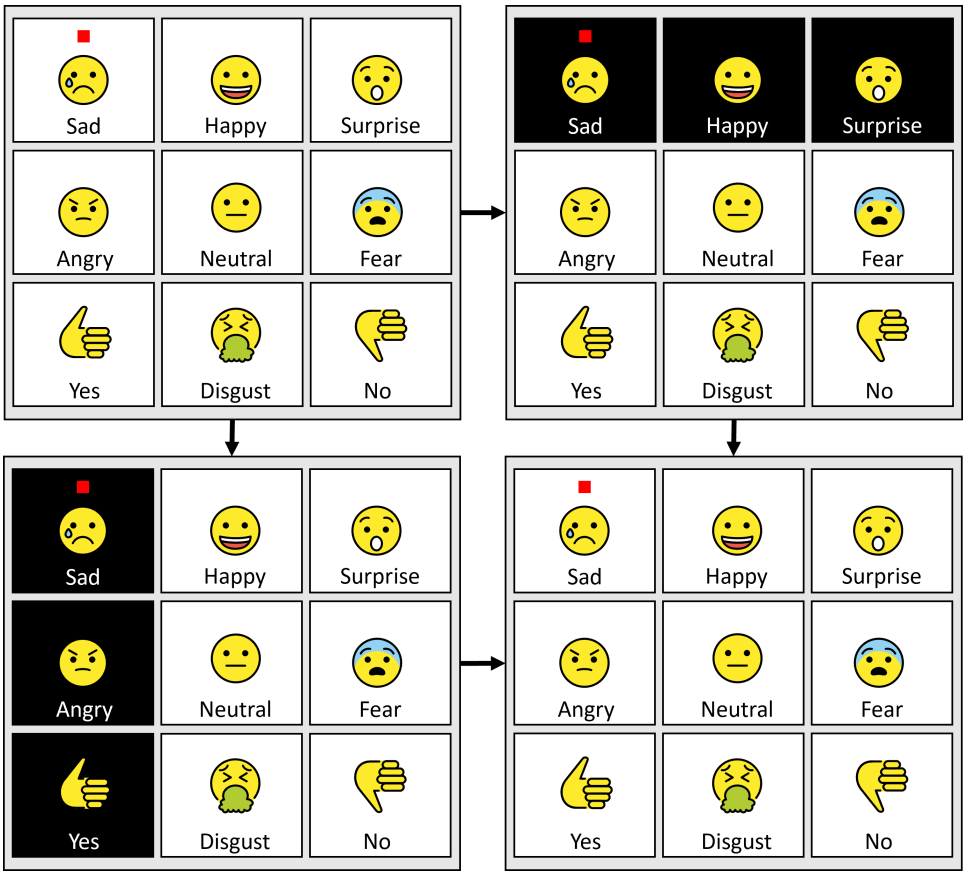

Figure 7.1: Here is presented a novel emoji-matrix speller format for the communication of emotional state information and basic yes/no responses in a traditional row/column paradigm design. The top right quadrant shows a screenshot of the initial stimulus phase displaying a red cueing square over the target 'Sad' emoji in the upper left. The top right and bottom left quadrants show the flash-based stimulus method for the row, and column sequence augmentations respectively. The lower right quadrant shows the default screen presented after each augmentation event.

The application of these foundational affective indicators increases the ease of tuning the stimuli with user-specific data. Recent studies have developed methods for the modification of facial expressions in profile images via the use of generative adversarial networks (GANs) such as the StarGAN [302] and GANimation models [303]. These architectures could be utilized to generate the facial expression variants from a single neutral image input of the lab participants or real-world users to harness the known P300 peaking boost associated with self-face integrated speller arrays [153]. This would expand the original intention of this project, which originally involved validating the efficacy of emoji stimuli in P300 speller contexts, to a paradigm operating as a true emotional expressivity tool.

As noted earlier, emoji are primarily used in tandem with text for embellishment and clarification purposes and the use of these stimuli in an emotion expression paradigm is purely as a standardized and universally recognisable substitute for actual facial expressions. This adapted method suggested here likely fulfils the aims of an emotional expression assistive device more adequately and would likely lead to a significant increase in performance as compared to the stimulus designs noted herein.

If this communication technology were implemented in patient populations, the desire to communicate a specific emotional state using an emoji, such as anger, would more than likely indicate that the individual is actively experiencing that emotion. In other words, the act of selecting an emoji associated with anger presupposes that during the selection process, the user is experiencing the related negative emotional state. There are reasonable concerns regarding the interaction between negative emotional states, such as fear, anger, or disgust, and the effective propagation of the P300 waveform. Previous research has demonstrated that emotionally evocative stimuli can enhance associated P300 signals [304–306]. This effect can be understood as an exogenous cue response to external stimuli.

During the operation of a P300-based BCI system the unintended occurrence of a negative emotion in response to such an exogenous cue could induce an unwanted P300 waveform. Here, the unexpected or surprising event acts as the low-probability oddball event and would undoubtedly confound the trial currently being processed. Group discussions with ALS patients and caregivers have revealed that distraction during the task easily leads to unintended selections, further complicating the accuracy and reliability of the P300-based BCI system [17]. This is particularly concerning when negative emotions like fear, anger, or disgust are involved, as they could unintentionally trigger a P300 response, interfering with the system's ability to correctly interpret the user's intended communication. As a result, the presence of these emotional states could significantly impact the system's performance, making it challenging to distinguish between intentional and unintentional responses, ultimately reducing the effectiveness of the BCI technology in real-world applications.

Moreover, research has shown that negative emotional states often lead to a restriction of attentional mechanisms and cognitive processes. It is therefore reasonable to assume that these effects would result in a relative drop in P300 amplitude and an increase in waveform latency, reducing the accuracy of classifiers designed to interpret P300 signals associated with negative emotions. Research conducted regarding working memory has indicated that exposure to positive or negative emotions can affect P300 responses differently depending on the

individual's working memory capacity. For instance, [307] found that individuals with lower working memory capacity exhibited reduced P300 amplitudes following exposure to negative emotions, while those with higher working memory capacity showed better P300 responses under the same conditions. Given the high incidence of reduced attentional capacity for BCI system target patient populations [308] these findings may indicate that negative emotional states could interact negatively with the classification process by reducing P300 waveform components. Similarly, stress has been shown to affect P300 responses in varied ways, as demonstrated by [309], who observed reduced P300 amplitudes in individuals highly sensitive to stress during a virtual reality simulation of heights, while less stress-sensitive individuals exhibited increased P300 responses.

Arguably the most direct investigation of these concerns was conducted by [310]. Here, audio cues were used to induce emotional states during online P300-speller operation. The effects of emotional stimulation on the online operation were shown to be highly inconsistent and impose no significant impact on the accuracy or efficiency of the standard 6 x 6 BCI speller system employed. These findings suggest that while emotion can influence P300 generation in certain contexts, this influence may not substantially affect P300-based BCI operations. However, given that emotional states are represented cortically alongside the corresponding P300 waveform, this type of system could potentially be used to gather longitudinal data on each emotional state. Over time, the cortical representation of emotional states could be leveraged to predict the emoji intended for selection in a BCI, in conjunction with the P300-predicted emoji, functioning similarly to a predictive text system that maps cortical activity.

7.3 Network Optimization Contributions

The experimental investigations discussed herein contribute to the current literature principally by corroborating the predicted theoretical limits and potential of CNN analysis methods, adding clarity to the process of subject-specific signal processing parameter optimization and outlining potential fruitful avenues for future research. Firstly, the CNNs assessed in this study demonstrate that network complexity did not account for any variance in the classification performance when training on near raw data (see, Table 6.3). These results suggest that fundamental signal processing stages, namely low-pass and high-pass filtering are still relevant for the effective classification of SSVEPs via CNNs. Further, the fixed parameter assessments (see, Table 6.4) show that network complexity in terms of convolutional filter count, as opposed to network depth, arguably leads to more effective model classification results as seen by the higher performance metrics attained by the EEGNetSSVEP as compared to the

DeepConvNet.

The results of the optimization studies performed across subjects for all models evaluated demonstrate a clear positive effect for increased network complexity and larger optimized low-pass cutoff values that align well with the current literature (see, Tables 6.5, 6.6, 6.7 & 6.8). Arguably, the larger networks are not simply more resistant to noise in the high-frequency range, it appears that the adoption of higher low-pass frequency cutoffs also enables these networks to extract more task-relevant harmonic features to boost classification performance. Crucially, the findings indicate that the use of subject-specific optimization for signal preprocessing parameters can enhance classifier performance, independent of subject data quality, especially for lower power network configurations. Note that this research can serve as a starting point for the investigation of alternative parameter optimization studies.

Moreover, the comparison of pruner methods (see subsection 6.8.4) outlines useful base settings for the rapid deployment of these tools such as the optimal pruner algorithm, namely the Percentile variant, and provides a guide for expected optimization durations for the respective data and models evaluated herein. Additionally, methods for increasing the efficiency of these optimization studies while enabling metric validation are outlined via the implementation of the epoch-threshold *k*-fold network assessments (see subsection 6.6.6.2).

Note that all associated scripts, functions and additional software developed for these optimization studies will be hosted on GitHub at: https://github.com/JoshPod93/EEG_Optim. This includes code relating to: data handling and downloads, data preparation, pre-processing and organisation, model loading, optimization wrapper positioning information, results storage, classification performance metric assessments and plotting tools.

7.3.1 Future Research

As discussed in Chapter 6 Subsection, 6.8.5, the enhanced network complexity and continued reimplementation of Filter-Bank Canonical Correlation Analysis methods in a CNN context have significantly boosted classifier performance for SSVEP detection. These methods utilize ultra-wide bandpass filters up to 90 Hz to ensure the extraction of all relevant harmonic SSVEP data. Along these lines, the highest-performing DCNN method for SSVEP classification, [57], avoids applying the strict band limits on the input data, opting for a filter width of 8-90 Hz. These findings corroborate those observed herein and suggest that networks expressing increased depth and complexity do not require the same degree of input pre-processing.

Note that, the results reported herein do not suggest that the use of purely unprocessed data is the ultimate end goal of these investigations. As seen in Table 6.3, the removal of powerline noise and signal frequencies below the minimum target flicker rate is necessary for effective training. Any extension of these signal pre-processing optimization assessments outlined in this thesis would likely be more successfully reimplemented for alternative parameters, this includes electrode arrangement selections, ground and reference sensor assignments and training data phase realignment. As shown in [311], each subject was presented with 10, 12 and 15 Hz SSVEP stimuli at a 0° phase angle. The actual phase of oscillatory components collected in subjects presented with standard deviations ranging between 15-18° for each respective target stimulus. These results suggest that the offset between presentation stimuli control signals and the exogenous SSVEP waveforms can vary dramatically between subjects.

This could explain the reduced capacity of lower-complexity networks for the accurate classification of SSVEP bio-signals described in Chapter 6. A higher number of convolutional filters could enable the networks to adjust for these variances in the training data. A hard-coded solution to enhance data applicability could involve adjusting the phase angle of all training subject data to the unique phase offsets expressed in the target subject. This would increase the coherency between the training and validation data and likely enhance end-point accuracies. These stages are likely being performed by the convolutional filters in the higher-complexity networks. Additional investigations into the phasic waveform representations embedded in convolutional filters, as explored in [60] for stimulus frequency profiles, could help clarify these claims.

Additionally, the author asserts that the application of these signal-preprocessing optimization methods would likely benefit from the expansion of available online SSVEP repo datasets utilized as seen in [49, 52, 56, 180], the inclusion of alternative CNN models [54, 57] and the exploration of additional optimization methods, namely, the Hyperband and Threshold pruners [271]. The integration of these data and methods into the optimization tools defined herein would greatly enhance the ability of researchers to compare the performance of networks in a truly like-for-like format.

References

[1] Antje A Seeber, A Jeannette Pols, Albert Hijdra, Hepke F Grupstra, Dick L Willems, and Marianne de Visser. Advance care planning in progressive neurological diseases lessons from als. *BMC palliative care*, 18:1–10, 2019.

[2] Eran Klein, Betts Peters, and Matt Higger. Ethical considerations in ending exploratory brain computer interface research studies in locked-in syndrome. *Cambridge Quarterly of Healthcare Ethics*, 27:660–674, 2018.

[3] Jane E Huggins, Patricia A Wren, and Kirsten L Gruis. What would brain-computer interface users want? opinions and priorities of potential users with amyotrophic lateral sclerosis. *Amyotrophic Lateral Sclerosis*, 12:318–324, 2011.

[4] Jane E Huggins, Aisha A Moinuddin, Anthony E Chiodo, and Patricia A Wren. What would brain computer interface users want: opinions and priorities of potential users with spinal cord injury. *Archives of physical medicine and rehabilitation*, 96:S38–S45, 2015.

[5] Ralf Stutzki, Markus Weber, Stella Reiter-Theil, Urs Simmen, Gian Domenico Borasio, and Ralf J Jox. Attitudes towards hastened death in als: a prospective study of patients and family caregivers. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, 15:68–76, 2014.

[6] Peter Bede, Rangariroyashe H Chipika, Eoin Finegan, Stacey Li Hi Shing, Mark A Doherty, Jennifer C Hengeveld, Alice Vajda, Siobhan Hutchinson, Colette Donaghy, and Russell L McLaughlin. Brainstem pathology in amyotrophic lateral sclerosis and primary lateral sclerosis: a longitudinal neuroimaging study. *NeuroImage: Clinical*, 24:102054, 2019.

[7] Sonja Korner, Michael Siniawski, Katja Kollewe, Klaus Jan Rath, Klaus Krampfl, Antonia Zapf, Reinhard Dengler, and Susanne Petri. Speech therapy and communication

device: impact on quality of life and mood in patients with amyotrophic lateral sclerosis. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, 14:20–25, 2013.

[8] Marco Caligari, Marco Godi, Simone Guglielmetti, Franco Franchignoni, and Antonio Nardone. Eye tracking communication devices in amyotrophic lateral sclerosis: impact on disability and quality of life. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, 14:546–552, 2013.

[9] Eric W Sellers, Theresa M Vaughan, and Jonathan R Wolpaw. A brain-computer interface for long-term independent home use. *Amyotrophic lateral sclerosis*, 11:449–455, 2010.

[10] Emily M Mugler, Carolin A Ruf, Sebastian Halder, Michael Bensch, and Andrea Kubler. Design and implementation of a p300-based brain-computer interface for controlling an internet browser. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 18:599–609, 2010.

[11] Erwin J O Kompanje, Inez D de Beaufort, and Jan Bakker. Euthanasia in intensive care: a 56-year-old man with a pontine hemorrhage resulting in a locked-in syndrome. *Critical care medicine*, 35:2428–2430, 2007.

[12] James R Patterson and Martin Grabois. Locked-in syndrome: A review of 139 cases. *Stroke*, 17, 1986.

[13] Eimear Smith and Mark Delargy. Locked-in syndrome. *BMJ*, 330:406–409, 2005.

[14] C Reigada, O Mendes, C Paiva, M Tavares, and E Goncalves. Als patients in locked-in syndrome: A systematic review of the literature. *Palliative Medicine*, 28, 2014.

[15] Kunal Khanna, Ajit Verma, and Bella Richard. The locked-in syndrome: Can it be unlocked? *Journal of Clinical Gerontology and Geriatrics*, 2, 2011.

[16] Emanuela Casanova, Rosa E Lazzari, Sergio Lotta, and Anna Mazzucchi. Locked-in syndrome: Improvement in the prognosis after an early intensive multidisciplinary rehabilitation. *Archives of Physical Medicine and Rehabilitation*, 84, 2003.

[17] Stefanie Blain-Moraes, Riley Schaff, Kirsten L Gruis, Jane E Huggins, and Patricia A Wren. Barriers to and mediators of brain-computer interface user acceptance: focus group findings. *Ergonomics*, 55:516–525, 2012.

[18] Zulay R Lugo, Marie-Aurelie Bruno, Olivia Gosseries, Athena Demertzi, Lizette Heine, Marie Thonnard, Veronique Blandin, Frederic Pellas, and Steven Laureys. Beyond the gaze: communicating in chronic locked-in syndrome. *Brain injury*, 29:1056–1061, 2015.

[19] Reza Fazel-Rezai, Scott Gavett, Waqas Ahmad, Ahmed Rabbi, and Eric Schneider. A comparison among several p300 brain-computer interface speller paradigms. *Clinical EEG and neuroscience*, 42:209–213, 2011.

[20] Jiahui Pan, Yuanqing Li, Zhenghui Gu, and Zhuliang Yu. A comparison study of two p300 speller paradigms for brain computer interface. *Cognitive neurodynamics*, 7:523–529, 2013.

[21] Cleiton Eduardo Saturno, Alejandro Rafael Garcia Ramirez, Mauro Jose Conte, Misia Farhat, and Elaine Carmelita Piucco. An augmentative and alternative communication tool for children and adolescents with cerebral palsy. *Behaviour and Information Technology*, 34:632–645, 2015.

[22] Jessica Brown and Amber Thiessen. Using images with individuals with aphasia: Current research and clinical trends. *American Journal of Speech-Language Pathology*, 27:504–515, 2018.

[23] Taejun Lee, Minju Kim, and Sung-Phil Kim. Improvement of p300-based brain-computer interfaces for home appliances control by data balancing techniques. *Sensors*, 20:5576, 2020.

[24] Praveen Kumar Shukla, Rahul Kumar Chaurasiya, Shrish Verma, and G R Sinha. A thresholding-free state detection approach for home appliance control using p300-based bci. *IEEE Sensors Journal*, 21:16927–16936, 2021.

[25] Faraz Akram, Ahmed Alwakeel, Mohammed Alwakeel, Mohammad Hijji, and Usman Masud. A symbols based bci paradigm for intelligent home control using p300 event-related potentials. *Sensors*, 22:10000, 2022.

[26] Xiaoke Chai, Zhimin Zhang, Kai Guan, Yangting Lu, Guitong Liu, Tengyu Zhang, and Haijun Niu. A hybrid bci-controlled smart home system combining ssvep and emg for individuals with paralysis. *Biomedical Signal Processing and Control*, 56:101687, 2020.

[27] Ron D Hays, Jakob B Bjorner, Dennis A Revicki, Karen L Spritzer, and David Cella. Development of physical and mental health summary scores from the patient-reported

outcomes measurement information system (promis) global items. *Quality of life Research*, 18:873–880, 2009.

[28] Kathryn Yorkston, Carolyn Baylor, and Helen Mach. Factors associated with communicative participation in amyotrophic lateral sclerosis. *Journal of Speech, Language, and Hearing Research*, 60:1791–1797, 2017.

[29] Mariana P Branco, Elmar G M Pels, Femke Nijboer, Nick F Ramsey, and Mariska J Vansteensel. Brain-computer interfaces for communication: preferences of individuals with locked-in syndrome, caregivers and researchers. *Disability and Rehabilitation: Assistive Technology*, 18:963–973, 2023.

[30] X Lu, W Ai, X Liu, Q Li, N Wang, G Huang, and Q Mei. Learning from the ubiquitous language: An empirical analysis of emoji usage of smartphone users. *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 770–780, 2016.

[31] Rui Zhou, Jasmine Hentschel, and Neha Kumar. Goodbye text, hello emoji: Mobile communication on wechat in china. *Conference on Human Factors in Computing Systems - Proceedings*, 2017-May, 2017.

[32] Samantha Thomson, Emily Kluftinger, and Jocelyn Wentland. Are you fluent in sexual emoji?: Exploring the use of emoji in romantic and sexual contexts. *Canadian Journal of Human Sexuality*, 27, 2018.

[33] Gabriele Esposito, Penelope Hernandez, Rene Van Bavel, and Jose Vila. Nudging to prevent the purchase of incompatible digital products online: An experimental study. *PLoS ONE*, 12, 2017.

[34] Joanna C Dunlap, Devshikha Bose, Patrick R Lowenthal, Cindy S York, Michael Atkinson, and Jim Murtagh. *What Sunshine Is to Flowers: A Literature Review on the Use of Emoticons to Support Online Learning*. 2015.

[35] Francesco Barbieri, German Kruszewski, Francesco Ronzano, and Horacio Saggion. How cosmopolitan are emojis? exploring emojis usage and meaning over different languages with distributional semantics. *MM 2016 - Proceedings of the 2016 ACM Multimedia Conference*, 2016.

[36] Monica A Riordan. The communicative role of non-face emojis: Affect and disambiguation. *Computers in Human Behavior*, 76:75–86, 2017.

[37] David Rodrigues, Marilia Prada, Rui Gaspar, Margarida V Garrido, and Diniz Lopes. Lisbon emoji and emoticon database (leed): Norms for emoji and emoticons in seven evaluative dimensions. *Behavior Research Methods*, 50:392–405, 2018.

[38] Tina Ganster, Sabrina C Eimler, and Nicole C Kramer. Same same but different!? the differential influence of smilies and emoticons on person perception. *Cyberpsychology, Behavior, and Social Networking*, 15, 2012.

[39] Qiyu Bai, Qi Dan, Zhe Mu, and Maokun Yang. A systematic review of emoji: Current research and future perspectives. *Frontiers in Psychology*, 10, 2019.

[40] Ben Eisner, Tim Rocktaschel, Isabelle Augenstein, Matko Bosnjak, and Sebastian Riedel. emoji2vec: Learning emoji representations from their description. *arXiv preprint arXiv:1609.08359*, 2016.

[41] Mayank Kejriwal, Qile Wang, Hongyu Li, and Lu Wang. An empirical study of emoji usage on twitter in linguistic and national contexts. *Online Social Networks and Media*, 24, 2021.

[42] Marilia Prada, David L Rodrigues, Margarida V Garrido, Diniz Lopes, Bernardo Cavalheiro, and Rui Gaspar. Motives, frequency and attitudes toward emoji and emoticon use. *Telematics and Informatics*, 35, 2018.

[43] Susan C Herring and Ashley R Dainas. Gender and age influences on interpretation of emoji functions. *ACM Transactions on Social Computing*, 3, 2020.

[44] Alexandra Comaniciu and Laleh Najafizadeh. Enabling communication for locked-in syndrome patients using deep learning and an emoji-based brain computer interface. *2018 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, pages 1–4, 2018.

[45] Honglin Hu, Zhenyu Wang, Xi Zhao, Ruxue Li, Ang Li, Yuan Si, Jiaheng Wang, Ting Zhou, and Tianheng Xu. A survey on brain-computer interface-inspired communications: Opportunities and challenges. *IEEE Communications Surveys and Tutorials*, 2024.

[46] Anthony M Norcia, L G Gregory Appelbaum, J M Justin M Ales, B R Benoit R Cottereau, and Bruno Rossion. The steady-state visual evoked potential in vision research: a review. *Journal of Vision*, 15:4, 2015.

[47] Yonghao Chen, Chen Yang, Xiaochen Ye, Xiaogang Chen, Yijun Wang, and Xiaorong Gao. Implementing a calibration-free ssvep-based bci system with 160 targets. *Journal of Neural Engineering*, 18, 2021.

[48] Guangyu Bin, Xiaorong Gao, Zheng Yan, Bo Hong, and Shangkai Gao. An online multi-channel ssvep-based brain-computer interface using a canonical correlation analysis method. *Journal of Neural Engineering*, 6, 2009.

[49] Bingchuan Liu, Xiaoshan Huang, Yijun Wang, Xiaogang Chen, and Xiaorong Gao. Beta: A large benchmark database toward ssvep-bci application. *Frontiers in Neuroscience*, 14, 2020.

[50] Weizhi Zhou, Aiping Liu, and Xun Chen. Compact cnn with dynamic window for ssvep-based bcis. *Chinese Control Conference, CCC*, 2022-July, 2022.

[51] Ke Liu, Zhaolin Yao, Li Zheng, Qingguo Wei, Weihua Pei, Xiaorong Gao, and Yijun Wang. A high-frequency ssvep-bci system based on a 360 hz refresh rate. *Journal of Neural Engineering*, 20:46042, 2023.

[52] Yijun Wang, Xiaogang Chen, Xiaorong Gao, and Shangkai Gao. A benchmark dataset for ssvep-based brain-computer interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25:1746–1752, 2017.

[53] Masaki Nakanishi, Yijun Wang, Xiaogang Chen, Yu-Te Wang, Xiaorong Gao, and Tzyy-Ping Jung. Enhancing detection of ssveps for a high-speed brain speller using task-related component analysis. *IEEE Transactions on Biomedical Engineering*, page 1, 2017.

[54] Yang Deng, Qingyu Sun, Ce Wang, Yijun Wang, and S Kevin Zhou. Trca-net: using trca filters to boost the ssvep classification with convolutional neural network. *Journal of Neural Engineering*, 20:46005, 2023.

[55] Jing Jiang, Erwei Yin, Chunhui Wang, Minpeng Xu, and Dong Ming. Incorporation of dynamic stopping strategy into the high-speed ssvep-based bcis. *Journal of Neural Engineering*, 15, 2018.

[56] Xiaogang Chen, Yijun Wang, Masaki Nakanishi, Xiaorong Gao, Tzyy-Ping Jung, and Shangkai Gao. High-speed spelling with a noninvasive brain-computer interface. *Proceedings of the National Academy of Sciences*, 112:E6058–E6067, 2015.

[57] Osman Berke Guney, Muhtasham Oblokulov, and Huseyin Ozkan. A deep neural network for ssvep-based brain-computer interfaces. *IEEE transactions on biomedical engineering*, 69:932–944, 2021.

[58] Hubert Cecotti. A time frequency convolutional neural network for the offline classification of steady-state visual evoked potential responses. *Pattern Recognition Letters*, 32:1145–1153, 2011.

[59] Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. Eegnet: a compact convolutional neural network for eeg-based brain-computer interfaces. *Journal of neural engineering*, 15:56013, 2018.

[60] Nicholas Waytowich, Vernon J Lawhern, Javier O Garcia, Jennifer Cummings, Josef Faller, Paul Sajda, and Jean M Vettel. Compact convolutional neural networks for classification of asynchronous steady-state visual evoked potentials. *Journal of Neural Engineering*, 15, 2018.

[61] Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for eeg decoding and visualization. *Human Brain Mapping*, 38:5391–5420, 2017.

[62] Piotr Olejniczak. Neurophysiologic basis of eeg. *Journal of Clinical Neurophysiology*, 23:186–189, 2006.

[63] Kevin M Spencer, Margaret A Niznikiewicz, Martha E Shenton, and Robert W McCarley. Sensory-evoked gamma oscillations in chronic schizophrenia. *Biological Psychiatry*, 63, 2008.

[64] Jack Solomon, Shaun Boe, and Timothy Bardouille. Reliability for non-invasive somatosensory cortex localization: Implications for pre-surgical mapping. *Clinical Neurology and Neurosurgery*, 139, 2015.

[65] Raheel Ahmed and James T Rutka. The role of meg in pre-surgical evaluation of epilepsy: current use and future directions. *Expert Review of Neurotherapeutics*, 16, 2016.

[66] Kenan Alkhalili, Ajay Niranjan, and Johnathan Engh. Preoperative magnetoencephalography improves tumor resection safety in awake craniotomy: Our initial experience. *Journal of Neurological Surgery Part B: Skull Base*, 77, 2016.

[67] Robert C Knowlton. Can magnetoencephalography aid epilepsy surgery? *Epilepsy Currents*, 8, 2008.

[68] Robert C Knowlton, Rotem A Elgavish, Al Bartolucci, Buddhiwardhan Ojha, Nita Limdi, Jeffrey Blount, Jorge G Burneo, Lawrence Ver Hoef, Lebron Paige, Edward Faught, Pongkiat Kankirawatana, Kristen Riley, and Ruben Kuzniecky. Functional imaging: Ii. prediction of epilepsy surgery outcome. *Annals of Neurology*, 64, 2008.

[69] Iris B Mauss and Michael D Robinson. Measures of emotion: A review. *Cognition and Emotion*, 23:209–237, 2009.

[70] Aashit K Shah and Sandeep Mittal. Invasive electroencephalography monitoring: Indications and presurgical planning. *Annals of Indian Academy of Neurology*, 17, 2014.

[71] Emily S Kappenman and Steven J Luck. The effects of electrode impedance on data quality and statistical significance in erp recordings. *Psychophysiology*, 47:888–904, 2010.

[72] Yu Mike Chi, Yu Te Wang, Yijun Wang, Christoph Maier, Tzyy Ping Jung, and Gert Cauwenberghs. Dry and noncontact eeg sensors for mobile brain-computer interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 20:228–235, 2012.

[73] Alexandra-Maria Tautan, Vojkan Mihajlovic, Yun-Hsuan Chen, Bernard Grundlehner, Julien Penders, and Wouter Serdijn. Signal quality in dry electrode eeg and the relation to skin-electrode contact impedance magnitude. *Proceedings of the International Conference on Biomedical Electronics and Devices (BIOSTEC 2014)*, pages 12–22, 2014.

[74] F Freire, M Becchi, S Ponti, E Miraldi, and A Strigazzi. Impedance spectroscopy of conductive commercial hydrogels for electromyography and electroencephalography. *Physiological Measurement*, 31:S157, 2010.

[75] Marc Nuwer. Assessment of digital eeg, quantitative eeg, and eeg brain mapping: Report of the american academy of neurology and the american clinical neurophysiology society. *Neurology*, 49:277–292, 1997.

[76] Saurabh R Sinha, Lucy Sullivan, Dragos Sabau, Daniel San-Juan, Keith E Dombrowski, Jonathan J Halford, Abeer J Hani, Frank W Drislane, and Mark M Stecker. American clinical neurophysiology society guideline 1: minimum technical requirements for performing clinical electroencephalography. *Journal of Clinical Neurophysiology*, 33:303–307, 2016.

[77] Francesco Marini, Clement Lee, Johanna Wagner, Scott Makeig, and Mateusz Gola. A comparative evaluation of signal quality between a research-grade and a wireless dry-electrode mobile eeg system. *Journal of neural engineering*, 16:54001, 2019.

[78] Gianluca Di Flumeri, Pietro Arico, Gianluca Borghini, Nicolina Sciaraffa, Antonello Di Florio, and Fabio Babiloni. The dry revolution: Evaluation of three different eeg dry electrode types in terms of signal spectral features, mental states classification and usability. *Sensors*, 19:1365, 2019.

[79] Benedict Shien Wei Ng, Nikos K Logothetis, and Christoph Kayser. Eeg phase patterns reflect the selectivity of neural firing. *Cerebral Cortex*, 23, 2013.

[80] S P Layne, G Mayer-Kress, and J Holzfuss. *Problems Associated with Dimensional Analysis of Electroencephalogram Data*. 1986.

[81] S Blanco, H Garcia, R Quian Quiroga, L Romanelli, and O A Rosso. Stationarity of the eeg series. *IEEE Engineering in medicine and biology Magazine*, 14:395–399, 1995.

[82] Swati Vaid, Preeti Singh, and Chamandeep Kaur. Eeg signal analysis for bci interface: A review. *International Conference on Advanced Computing and Communication Technologies, ACCT*, 2015-April, 2015.

[83] Benedetto Falsini and Vittorio Porciatti. The temporal frequency response function of pattern erg and vep: Changes in optic neuritis. *Electroencephalography and Clinical Neurophysiology - Evoked Potentials*, 100, 1996.

[84] Francesco DiRusso and Donatella Spinelli. Electrophysiological evidence for an early attentional mechanism in visual processing in humans. *Vision research*, 39:2975–2985, 1999.

[85] B Johansson and P Jakobsson. Fourier analysis of steady-state visual evoked potentials in subjects with normal and defective stereo vision. *Doc Ophthalmol*, 101:233–246, 2000.

[86] Gabriela Alcaraz and Pirjo Manninen. Intraoperative electrocorticography. *Journal of Neuroanaesthesiology and Critical Care*, 04, 2017.

[87] Fedor Panov, Emily Levin, Coralie De Hemptinne, Nicole C Swann, Salman Qasim, Svjetlana Miocinovic, Jill L Ostrem, and Philip A Starr. Intraoperative electrocorticography for physiological research in movement disorders: Principles and experience in 200 cases. *Journal of Neurosurgery*, 126, 2017.

[88] Niels Birbaumer. Breaking the silence: Brain-computer interfaces (bci) for communication and motor control. *Psychophysiology*, 43, 2006.

[89] J J Vidal. Toward direct brain-computer communication. *Annual review of biophysics and bioengineering*, 2, 1973.

[90] Jonathan R Wolpaw, Niels Birbaumer, Dennis J McFarland, Gert Pfurtscheller, and Theresa M Vaughan. Brain-computer interfaces for communication and control. *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology*, 113:767–791, 2002.

[91] Benjamin Blankertz, Guido Dornhege, Matthias Krauledat, Michael Schroder, John Williamson, Roderick Murray-Smith, and Klaus-Robert Muller. The berlin brain-computer interface presents the novel mental typewriter hex-o-spell. 2006.

[92] Scott Mackenzie and Behrooz Ashtiani. Blinkwrite efficient text entry using eye blinks. *Universal Access in the Information Society*, 10:69–80, 2011.

[93] Mindaugas Vasiljevas, Turcinas Rutenis, and Damasevicius Robertas. Emg speller with adaptive stimulus rate and dictionary support. *2014 Federated Conference on Computer Science and Information Systems IEEE*, 2:227–234, 2014.

[94] Shiv Kumar Mudgal, Suresh K Sharma, Jitender Chaturvedi, and Anil Sharma. Brain computer interface advancement in neurosciences: Applications and issues. *Interdisciplinary Neurosurgery: Advanced Techniques and Case Management*, 20, 2020.

[95] Fakhreddine Karray, Milad Alemzadeh, Jamil Abou Saleh, and Mo Nours Arab. Human-computer interaction: Overview on state of the art. *International journal on smart sensing and intelligent systems*, 1:137, 2008.

[96] Tristan Joseph C Limchesing, Arianna Elise C Chua, Christalline Jhine L Shi, Renann G Baldovino, Francisco Emmanuel T Munsayac, and Nilo T Bugtai. A review on recent applications of eeg-based bci in wheelchairs and other assistive devices. *2021 IEEE 13th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM)*, pages 1–6, 2021.

[97] Annalisa Colucci, Mareike Vermehren, Alessia Cavallo, Cornelius Angerhofer, Niels Peekhaus, Loredana Zollo, Won-Seok Kim, Nam-Jong Paik, and Surjo R Soekadar. Brain-computer interface-controlled exoskeletons in clinical neurorehabilitation: ready or not? *Neurorehabilitation and Neural Repair*, 36:747–756, 2022.

[98] Amin Hekmatmanesh, Pedro H J Nardelli, and Heikki Handroos. Review of the state-of-the-art of brain-controlled vehicles. *IEEE Access*, 9:110173–110193, 2021.

[99] Muhammad Ahmed Khan, Rig Das, Helle K Iversen, and Sadasivan Puthusserypady. Review on motor imagery based bci systems for upper limb post-stroke neurorehabilitation: From designing to application. *Computers in biology and medicine*, 123:103843, 2020.

[100] Paul Dominick E Baniqued, Emily C Stanyer, Muhammad Awais, Ali Alazmani, Andrew E Jackson, Mark A Mon-Williams, Faisal Mushtaq, and Raymond J Holt. Brain-computer interface robotics for hand rehabilitation after stroke: A systematic review. *Journal of neuroengineering and rehabilitation*, 18:1–25, 2021.

[101] S M Riazul Islam, Daehan Kwak, M D Humaun Kabir, Mahmud Hossain, and Kyung-Sup Kwak. The internet of things for health care: a comprehensive survey. *IEEE access*, 3:678–708, 2015.

[102] Pierce Stegman, Chris S Crawford, Marvin Andujar, Anton Nijholt, and Juan E Gilbert. Brain-computer interface software: A review and discussion. *IEEE Transactions on Human-Machine Systems*, 50:101–115, 2020.

[103] Francisco Javier Velasco-Alvarez, Alvaro Fernandez-Rodriguez, and Ricardo Ron-Angevin. Brain computer interface control of smartphone messaging applications. 2021.

[104] Francisco Velasco-Alvarez, Alvaro Fernandez-Rodriguez, Francisco-Javier Vizcaino-Martin, Antonio Diaz-Estrella, and Ricardo Ron-Angevin. Brain computer interface (bci) control of a virtual assistant in a smartphone to manage messaging applications. *Sensors*, 21:3716, 2021.

[105] Gabriel Alves Mendes Vasiljevic and Leonardo Cunha De Miranda. Brain-computer interface games based on consumer-grade eeg devices: A systematic literature review. *International Journal of Human-Computer Interaction*, 36:105–142, 2020.

[106] Eduardo R Miranda. Plymouth brain-computer music interfacing project: from eeg audio mixers to composition informed by cognitive neuroscience. *International Journal of Arts and Technology*, 3:154–176, 2010.

[107] Duncan A H Williams and Eduardo R Miranda. *BCI for music making: then, now, and next*, pages 193–206. CRC Press, 2018.

[108] Aleksandra Kawala-Sterniuk, Natalia Browarska, Amir Al-Bakri, Mariusz Pelc, Jaroslaw Zygarlicki, Michaela Sidikova, Radek Martinek, and Edward Jacek Gorzelanczyk. Summary of over fifty years with brain computer interfaces a review. *Brain Sciences*, 11:43, 2021.

[109] J R Wolpaw, H Ramoser, D J McFarland, and G Pfurtscheller. Eeg-based communication: Improved accuracy by response verification. *IEEE Transactions on Rehabilitation Engineering*, 6:326–333, 1998.

[110] Pasquale Arpaia, Antonio Esposito, Angela Natalizio, and Marco Parvis. How to successfully classify eeg in motor imagery bci: A metrological analysis of the state of the art. *Journal of Neural Engineering*, 19:31002, 2022.

[111] Hongtao Wang, Fan Yan, Tao Xu, Haojun Yin, Peng Chen, Hongwei Yue, Chuangquan Chen, Hongfei Zhang, Linfeng Xu, Yuebang He, and Anastasios Bezerianos. Brain-controlled wheelchair review: From wet electrode to dry electrode, from single modal to hybrid modal, from synchronous to asynchronous. *IEEE Access*, 9, 2021.

[112] Ioulietta Lazarou, Spiros Nikolopoulos, Panagiotis C Petrantonakis, Ioannis Kompatsiaris, and Magda Tsolaki. Eeg-based brain-computer interfaces for communication and rehabilitation of people with motor impairment: A novel approach of the 21st century. *Frontiers in Human Neuroscience*, 12, 2018.

[113] Pablo F Diez, Vicente A Mut, Enrique M Avila Perona, and Eric Laciar Leber. Asynchronous bci control using high-frequency ssvep. *Journal of NeuroEngineering and Rehabilitation*, 8, 2011.

[114] Tanja Krumpe, Carina Walter, Wolfgang Rosenstiel, and Martin Spuler. Asynchronous p300 classification in a reactive brain-computer interface during an outlier detection task. *Journal of Neural Engineering*, 13, 2016.

[115] Thorsten O Zander, Christian Kothe, Sabine Jatzev, and Matti Gaertner. *Enhancing Human-Computer Interaction withÂ Input from Active and Passive Brain-Computer Interfaces*. 2010.

[116] Hubert Cecotti. Spelling with non-invasive brain computer interfaces current and future trends. *Journal of Physiology-Paris*, 105:106–114, 2011.

[117] Jesus Minguillon, M Angel Lopez-Gordo, and Francisco Pelayo. Trends in eeg-bci for daily-life: Requirements for artifact removal. *Biomedical Signal Processing and Control*, 31:407–418, 2017.

[118] Samuel Sutton, Magery Braren, Joseph Zubin, and E R John. Evoked-potential correlates of stimulus uncertainty. *Science*, 1965.

[119] Shravani Sur and V K Sinha. Event-related potential: An overview. *Industrial Psychiatry Journal*, 18:70, 2009.

[120] Michael G H Coles, Henderikus G O M Smid, Marten K Scheffers, and Leun J Otten. *Mental chronometry and the study of human information processing*, volume 25, pages 86–131. 1995.

[121] Rolf Verleger. Event-related potentials and cognition: A critique of the context updating hypothesis and an alternative interpretation of p3. *Behavioral and Brain Sciences*, 11:343–356, 1988.

[122] John Polich. Updating p300: An integrative theory of p3a and p3b. *Clinical Neurophysiology*, 118:2128–2148, 2007.

[123] Eric Courchesne, Steven A Hillyard, and Robert Galambos. Stimulus novelty, task relevance and the visual evoked potential in man. *Electroencephalography and Clinical Neurophysiology*, 39:131–143, 1975.

[124] Nancy K Squires, Kenneth C Squires, and Steven A Hillyard. Two varieties of long-latency positive waves evoked by unpredictable auditory stimuli in man. *Electroencephalography and Clinical Neurophysiology*, 38:387–401, 1975.

[125] Hadar Levi-Aharoni, Oren Shriki, and Naftali Tishby. Surprise response as a probe for compressed memory states. *PLOS Computational Biology*, 16:e1007065, 2020.

[126] S Sutton. *P300 - Thirteen years later*, pages 107–126. 1979.

[127] Ray Johnson. On the neural generators of the p300 component of the event-related potential. *Psychophysiology*, 30, 1993.

[128] Craig Gonsalvez and John Polich. P300 amplitude is determined by target-to-target interval. *Psychophysiology*, 39:388–396, 2002.

[129] P Brunner, S Joshi, S Briskin, J R Wolpaw, H Bischof, and G Schalk. Does the 'p300' speller depend on eye gaze? *Journal of Neural Engineering*, 7, 2010.

[130] Matthias S Treder and Benjamin Blankertz. (c)overt attention and visual speller design in an erp-based brain-computer interface. *Behavioral and Brain Functions*, 6, 2010.

[131] Lynn M McCane, Susan M Heckman, Dennis J McFarland, George Townsend, Joseph N Mak, Eric W Sellers, Debra Zeitlin, Laura M Tenteromano, Jonathan R Wolpaw, and Theresa M Vaughan. P300-based brain-computer interface (bci) event-related potentials (erps): People with amyotrophic lateral sclerosis (als) vs. age-matched controls. *Clinical Neurophysiology*, 126, 2015.

[132] Benjamin Blankertz, Steven Lemm, Matthias Treder, Stefan Haufe, and Klaus-Robert Muller. Single-trial analysis and classification of erp components-a tutorial. *NeuroImage*, 56:814–825, 2011.

[133] Seul Ki Yeom, Siamac Fazli, Klaus Robert Mu Ller, and Seong Whan Lee. An efficient erp-based brain-computer interface using random set presentation and face familiarity. *PLoS ONE*, 9, 2014.

[134] Niels Birbaumer, Ander Ramos Murguialday, Cornelia Weber, and Pedro Montoya. *Chapter 8 Neurofeedback and Brain-Computer Interface. Clinical Applications*, volume 86, pages 107–117. 2009.

[135] D J McFarland and J R Wolpaw. Eeg-based brain-computer interfaces. *Current Opinion in Biomedical Engineering*, 2017.

[136] Fabien Lotte, Marco Congedo, Anatole Lecuyer, Fabrice Lamarche, Bruno Arnaldi, L Anatole, Fabien Lotte, Marco Congedo, L Anatole, Enas Abdulhay, Rami Oweis, Areej Mohammad, Lujain Ahmad, Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, Tonio Ball, Zhicheng Jiao, Xinbo Gao, Ying Wang, Jie Li, Haojun Xu, U Rajendra Acharya, Shu Lih Oh, Yuki Hagiwara, Jen Hong Tan, and Hojjat Adeli. A review of classification algorithms for eeg-based brain-computer interfaces to cite this version : A review of classification algorithms for eeg-based brain-computer interfaces. *Human brain mapping*, 38, 2018.

[137] Patricia Tueting, Samuel Sutton, and Joseph Zubin. Quantitative evoked potential correlates of the probability of events. *Psychophysiology*, 7:385–394, 1970.

[138] John Polich, Patricia Crane Ellerson, and Jill Cohen. P300, stimulus intensity, modality, and probability. *International Journal of Psychophysiology*, 23:55–62, 1996.

[139] Richard L Horst, Ray Johnson, and Emanuel Donchin. Event-related brain potentials and subjective probability in a learning task. *Memory and Cognition*, 8:476–488, 1980.

[140] John Polich. Task difficulty, probability, and inter-stimulus interval as determinants of p300 from auditory stimuli. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section*, 68:311–320, 1987.

[141] M Teresa Medina-Julia, Alvaro Fernandez-Rodriguez, Francisco Velasco-Alvarez, and Ricardo Ron-Angevin. P300-based brain-computer interface speller usability evaluation of three speller sizes by severely motor-disabled patients. *Frontiers in Human Neuroscience*, 14:583358, 2020.

[142] Nitzan S Artzi and Oren Shriki. An analysis of the accuracy of the p300 bci. *Brain-Computer Interfaces*, 5:112–120, 2018.

[143] John Polich and Jose R Criado. Neuropsychology and neuropharmacology of p3a and p3b. *International Journal of Psychophysiology*, 60, 2006.

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

[145] Christoph Guger, Shahab Daban, Eric Sellers, Clemens Holzner, Gunther Krausz, Roberta Carabalona, Furio Gramatica, and Guenter Edlinger. How many people are able to control a p300-based brain-computer interface (bci)? *Neuroscience Letters*, 462:94–98, 2009.

[146] G Townsend and V Platsko. Pushing the p300-based brain-computer interface beyond 100 bpm: Extending performance guided constraints into the temporal domain. *Journal of Neural Engineering*, 13, 2016.

[147] Dean J Krusienski, Eric W Sellers, Francois Cabestaing, Sabri Bayoudh, Dennis J McFarland, Theresa M Vaughan, and Jonathan R Wolpaw. A comparison of classification techniques for the p300 speller. *Journal of Neural Engineering*, 3, 2006.

[148] E Donchin, K M Spencer, and R Wijesinghe. The mental prosthesis: Assessing the speed of a p300-based brain- computer interface. *IEEE Transactions on Rehabilitation Engineering*, 8:174–179, 2000.

[149] Benjamin Blankertz, Gabriel Curio, and Klaus-Robert Muller. Classifying single trial eeg: Towards brain computer interfacing. *Advances in Neural Information Processing Systems*, 1:157–164, 2002.

[150] K R Muller, M Krauledat, G Dornhege, G Curio, and B Blankertz. Machine learning techniques for brain-computer interfaces. *Biomed Tech*, pages 11–24, 2004.

[151] Nikolay V Manyakov, Nikolay Chumerin, Adrien Combaz, and Marc M Van Hulle. Comparison of classification methods for p300 brain-computer interface on disabled subjects. *Computational Intelligence and Neuroscience*, 2011, 2011.

[152] Nikolay V Manyakov, Nikolay Chumerin, Adrien Combaz, Arne Robben, Marijn Van Vliet, and Marc M Van Hulle. Decoding phase-based information from steady-state visual evoked potentials with use of complex-valued neural network. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6936 LNCS:135–143, 2011.

[153] Zhaohua Lu, Qi Li, Ning Gao, and Jingjing Yang. The self-face paradigm improves the performance of the p300-speller system. *Frontiers in computational neuroscience*, 13:93, 2020.

[154] Alvaro Fernandez-Rodriguez, Francisco Velasco-Alvarez, Maria Teresa Medina-Julia, and Ricardo Ron-Angevin. Evaluation of emotional and neutral pictures as flashing stimuli using a p300 brain computer interface speller. *Journal of Neural Engineering*, 16:056024, 2019.

[155] D B Ryan, G Townsend, N A Gates, K Colwell, and E W Sellers. Evaluating brain-computer interface performance using color in the p300 checkerboard speller. *Clinical Neurophysiology*, 2017.

[156] Hesam Moradkhani and Vahid Shalchyan. Using emoji stimuli for checker-board p300 speller. *Iranian Journal of Biomedical Engineering*, 10:325–337, 2016.

[157] Yanghao Lei, Dong Wang, Weizhen Wang, Hao Qu, Jing Wang, and Bin Shi. Improving single-hand open/close motor imagery classification by error-related potentials correction. *Heliyon*, 9, 2023.

[158] A Fernandez-Rodriguez, Francisco Velasco-Alvarez, and Ricardo Ron-Angevin. Evaluation of a p300 brain-computer interface using different sets of flashing stimuli. pages 1–4, 2018.

[159] Alberto Betella and Paul F M J Verschure. The affective slider: A digital self-assessment scale for the measurement of human emotions. *PLoS ONE*, 11, 2016.

[160] C S Herrmann. Human eeg responses to 1-100 hz flicker: Resonance phenomena in visual cortex and their potential correlation to cognitive phenomena. *Experimental Brain Research*, 137, 2001.

[161] Maria A Pastor, Julio Artieda, Javier Arbizu, Miguel Valencia, and Jose C Masdeu. Human cerebral activation during steady-state visual-evoked responses. *Journal of neuroscience*, 23:11621–11627, 2003.

[162] Christoph Guger, Brendan Z Allison, Bernhard Growindhager, Robert Pruckl, Christoph Hintermuller, Christoph Kapeller, Markus Bruckner, Gunther Krausz, and Gunter Edlinger. How many people could use an ssvep bci? *Frontiers in Neuroscience*, 2012.

[163] C Guger, G Edlinger, W Harkam, I Niedermayer, and G Pfurtscheller. How many people are able to operate an eeg-based brain-computer interface (bci)? *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 11, 2003.

[164] Akshay Katyal and Rajesh Singla. A novel hybrid paradigm based on steady state visually evoked potential and p300 to enhance information transfer rate. *Biomedical Signal Processing and Control*, 59:101884, 2020.

[165] Xin Bai, Minglun Li, Shouliang Qi, Anna Ching Mei Ng, Tit Ng, and Wei Qian. A hybrid p300 ssvep brain computer interface speller with a frequency enhanced row and column paradigm. *Frontiers in Neuroscience*, 17:1133933, 2023.

[166] Robert S Fisher, Graham Harding, Giuseppe Erba, Gregory L Barkley, and Arnold Wilkins. Photic- and pattern-induced seizures: A review for the epilepsy foundation of america working group. *Epilepsia*, 46, 2005.

[167] Dong Ok Won, Han Jeong Hwang, Sven Dahne, Klaus Robert Muller, and Seong Whan Lee. Effect of higher frequency on the classification of steady-state visual evoked potentials. *Journal of Neural Engineering*, 13, 2015.

[168] Andrea Kubler, Boris Kotchoubey, Jochen Kaiser, Jonathan R Wolpaw, and Niels Birbaumer. Brain computer communication: Unlocking the locked in. *Psychological bulletin*, 127:358, 2001.

[169] Xiaorong Gao, Dingfeng Xu, Ming Cheng, and Shangkai Gao. A bci-based environmental controller for the motion-disabled. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 11, 2003.

[170] A Riccio, D Mattia, L Simione, M Olivetti, and F Cincotti. Eye-gaze independent eeg-based brain-computer interfaces for communication. *Journal of Neural Engineering*, 9, 2012.

[171] Matthias M Muller, P Malinowski, T Gruber, and S A Hillyard. Sustained division of the attentional spotlight. *Nature*, 424:309–312, 2003.

[172] Brendan Z Allison, Dennis J McFarland, Gerwin Schalk, Shi Dong Zheng, Melody Moore Jackson, and Jonathan R Wolpaw. Towards an independent brain-computer interface using steady state visual evoked potentials. *Clinical Neurophysiology*, 119, 2008.

[173] Simon P Kelly, Edmund C Lalor, Richard B Reilly, and John J Foxe. Visual spatial attention tracking using high-density ssvep data for independent brain-computer communication. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 13, 2005.

[174] Nikolay V Manyakov, Nikolay Chumerin, and Marc M Van Hulle. Multichannel decoding for phase-coded ssvep brain-computer interface. *International Journal of Neural Systems*, 22:1250022, 2012.

[175] Xiaogang Chen, Bingchuan Liu, Yijun Wang, and Xiaorong Gao. A spectrally-dense encoding method for designing a high-speed ssvep-bci with 120 stimuli. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 30, 2022.

[176] Siyu Liu, Deyu Zhang, Ziyu Liu, Mengzhen Liu, Zhiyuan Ming, Tiantian Liu, Dingjie Suo, Shintaro Funahashi, and Tianyi Yan. Review of brain-computer interface based on steady-state visual evoked potential. *Brain Science Advances*, 8, 2022.

[177] Zhonglin Lin, Changshui Zhang, Wei Wu, and Xiaorong Gao. Frequency recognition based on canonical correlation analysis for ssvep-based bcis. *IEEE Transactions on Biomedical Engineering*, 54, 2007.

[178] Yu Zhang, Guoxu Zhou, Qibin Zhao, Akinari Onishi, Jing Jin, Xingyu Wang, and Andrzej Cichocki. Multiway canonical correlation analysis for frequency components recognition in ssvep-based bcis. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7062 LNCS, 2011.

[179] Masaki Nakanishi, Yijun Wang, Yu Te Wang, Yasue Mitsukura, and Tzyy Ping Jung. Generating visual flickers for eliciting robust steady-state visual evoked potentials at flexible frequencies using monitor refresh rate. *PLoS ONE*, 9, 2014.

[180] Masaki Nakanishi, Yijun Wang, Yu Te Wang, and Tzyy Ping Jung. A comparison study of canonical correlation analysis based methods for detecting steady-state visual evoked potentials. *PLoS ONE*, 2015.

[181] Yijun Wang, Masaki Nakanishi, Yu-Te Wang, and Tzyy-Ping Jung. Enhancing detection of steady-state visual evoked potentials using individual training data. pages 3037–3040. Ieee, 2014.

[182] Masaki Nakanishi, Yijun Wang, Yu Te Wang, and Tzyy Ping Jung. Does frequency resolution affect the classification performance of steady-state visual evoked potentials? *International IEEE/EMBS Conference on Neural Engineering, NER*, pages 341–344, 2017.

[183] Akash Goel, Amit Kumar Goel, and Adesh Kumar. The role of artificial neural network and machine learning in utilizing spatial information. *Spatial Information Research*, 31:275–285, 2023.

[184] Luis Fernando Nicolas-Alonso and Jaime Gomez-Gil. Brain computer interfaces, a review. *sensors*, 12:1211–1279, 2012.

[185] Ou Bai, Peter Lin, Sherry Vorbach, Jiang Li, Steve Furlani, and Mark Hallett. Exploration of computational methods for classification of movement intention during human voluntary movement from single trial eeg. *Clinical Neurophysiology*, 118:2637–2655, 2007.

[186] Howida A Shedeed, Mohamed F Issa, and Salah M El-Sayed. Brain eeg signal processing for controlling a robotic arm. *Proceedings - 2013 8th International Conference on Computer Engineering and Systems, ICCES 2013*, pages 152–157, 2013.

[187] Kyoung-Su Oh and Keechul Jung. Gpu implementation of neural networks. *Pattern Recognition*, 37:1311–1314, 2004.

[188] Jurgen Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015.

[189] Edward J Kim and Robert J Brunner. Star-galaxy classification using deep convolutional neural networks. *Monthly Notices of the Royal Astronomical Society*, 464:4463–4475, 2017.

[190] A Kimura, I Takahashi, M Tanaka, N Yasuda, N Ueda, and N Yoshida. Single-epoch supernova classification with deep convolutional neural networks. *Proceedings - IEEE 37th International Conference on Distributed Computing Systems Workshops, ICDCSW 2017*, 2017.

[191] Vitoantonio Bevilacqua, Giuseppe Mastronardi, and Mario Marinelli. *A Neural Network Approach to Medical Image Segmentation and Three-Dimensional Reconstruction*, pages 22–31. 2006.

[192] Q Dou, H Chen, L Yu, L Zhao, J Qin, D Wang, V C Mok, L Shi, and P A Heng. Automatic detection of cerebral microbleeds from mr images via 3d convolutional neural networks. *IEEE Transactions on Medical Imaging*, 35:1182–1195, 2016.

[193] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542:115–118, 2017.

[194] P Sermanet and Y LeCun. Traffic sign recognition with multi-scale convolutional networks. *Neural Networks (IJCNN), The 2011 International Joint Conference on*, pages 2809–2813, 2011.

[195] Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36:193–202, 1980.

[196] Steven J Nowlan and Geoffrey E Hinton. Simplifying neural networks by soft weight-sharing. *Neural Computation*, 4:473–493, 1992.

[197] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. *AISTATS '11: Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, 15:315–323, 2011.

[198] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances In Neural Information Processing Systems*, pages 1–9, 2012.

[199] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.

[200] Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Y Ng. Unsupervised learning of hierarchical representations with convolutional deep belief networks. *Communications of the ACM*, 54:95, 2011.

[201] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *arXiv:1409.4842*, 2014.

[202] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521:436–444, 2015.

[203] A Antoniades, L Spyrou, C C Took, and S Sanei. Deep learning for epileptic intracranial eeg data. *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, 2016.

[204] A R Johansen, J Jin, T Maszczyk, J Dauwels, S S Cash, and M B Westover. Epileptiform spike detection via convolutional neural networks. *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 754–758, 2016.

[205] Yousef Rezaei Tabar and Ugur Halici. A novel deep learning approach for classification of eeg motor imagery signals. *Journal of neural engineering*, 14:016003, 2016.

[206] Ian Walker, Marc Deisenroth, and Aldo Faisal. Deep convolutional neural networks for brain computer interface using motor imagery. *Imperial College of Science, Technology and Medicine Department of Computing*, 2015.

[207] Manfredo Atzori, Matteo Cognolato, and Henning Muller. Deep learning with convolutional neural networks applied to electromyography data: A resource for the classification of movements for prosthetic hands. *Frontiers in Neurorobotics*, 10, 2016.

[208] Ulysse Coteallard, Francois Nougarou, Cheikh Latyr Fall, Philippe Giguere, Clement Gosselin, Francois Laviolette, and Benoit Gosselin. A convolutional neural network for robotic arm guidance using semg based frequency features. pages 2464–2470. IEEE, 2016.

[209] William Prew, Toby Breckon, Magnus Bordewich, and Ulrik Beierholm. Improving robotic grasping on monocular images via multi-task learning and positional loss. *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 9843–9850, 2021.

[210] Alexander Craik, Yongtian He, and Jose L Contreras-Vidal. Deep learning for electroencephalogram (eeg) classification tasks: A review. *Journal of Neural Engineering*, 16, 2019.

[211] Yudong Pan, Jianbo Chen, and Yangsong Zhang. A survey of deep learning-based classification methods for steady-state visual evoked potentials. *Brain-Apparatus Communication: A Journal of Bacomics*, 2, 2023.

[212] Nik Khadijah Nik Aznan, Stephen Bonner, Jason D Connolly, Noura Al Moubayed, and Toby P Breckon. On the classification of ssvep-based dry-eeg signals via convolutional neural networks. *arXiv:1805.04157*, 7 2018.

[213] Joshua J Podmore, Toby P Breckon, Nik K N Aznan, and Jason D Connolly. On the relative contribution of deep convolutional neural networks for ssvep-based bio-signal decoding in bci speller applications. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27:611–618, 2019.

[214] Dongcen Xu, Fengzhen Tang, Yiping Li, Qifeng Zhang, and Xisheng Feng. An analysis of deep learning models in ssvep-based bci: A survey. *Brain Sciences*, 13, 2023.

[215] Elham Rostami, Farnaz Ghassemi, and Zahra Tabanfar. Improving the classification of real-world ssvep data in brain-computer interface speller systems using deep convolutional neural networks. *Frontiers in Biomedical Technologies*, 9:248–254, 2022.

[216] Vinay Jayaram and Alexandre Barachant. Moabb: trustworthy algorithm benchmarking for bcis. *Journal of Neural Engineering*, 15:66011, 2018.

[217] Xia Chen, Xiangbin Teng, Han Chen, Yafeng Pan, and Philipp Geyer. Toward reliable signals decoding for electroencephalogram: A benchmark study to eegnex. *Biomedical Signal Processing and Control*, 87:105475, 2024.

[218] Gernot R Muller-Putz, Reinhold Scherer, Christian Brauneis, and Gert Pfurtscheller. Steady-state visual evoked potential (ssvep)-based communication: Impact of harmonic frequency components. *Journal of Neural Engineering*, 2, 2005.

[219] Murat Kancaoglu and Mehmet Kuntalp. The effect of harmonics count on ssvep-based bci results. *2019 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 1–4, 2019.

[220] Volkan Cetin, Serhat Ozekes, and Huseyin Selcuk Varol. Harmonic analysis of steady-state visual evoked potentials in brain computer interfaces. *Biomedical Signal Processing and Control*, 60:101999, 2020.

[221] Jonathan Peirce, Jeremy R Gray, Sol Simpson, Michael MacAskill, Richard Hochenberger, Hiroyuki Sogo, Erik Kastman, and Jonas Kristoffer Lindelov. Psychopy2 experiments in behavior made easy. *Behavior research methods*, 51:195–203, 2019.

[222] Benedikt Gros and Daniel Utz. Openmoji: https://openmoji.org/library/, retrieved, 10 2023, 2023.

[223] Dennis J McFarland, William A Sarnacki, George Townsend, Theresa Vaughan, and Jonathan R Wolpaw. The p300-based brain-computer interface (bci): Effects of stimulus rate. *Clinical Neurophysiology*, 122:731–737, 2011.

[224] LSL Developers. Github - sccn/labstreaminglayer: Labstreaminglayer super repository comprising submodules for lsl and associated apps: https://github.com/sccn/labstreaminglayer.

[225] Yu M Chi, Yijun Wang, Yu-Te Wang, Tzyy-Ping Jung, Trevor Kerth, and Yuchen Cao. A practical mobile dry eeg system for human computer interfaces. *Foundations of Augmented Cognition: 7th International Conference, AC 2013, Held as Part of HCI International 2013, Las Vegas, NV, USA, July 21-26, 2013. Proceedings 7*, pages 649–655, 2013.

[226] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, and Jonathan Bright. Scipy 1.0 fundamental algorithms for scientific computing in python. *Nature methods*, 17:261–272, 2020.

[227] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, and Vincent Dubourg. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[228] Gerwin Schalk and Jurgen Mellinger. *A practical guide to brain computer interfacing with BCI2000 General-purpose software for brain-computer interface research, data acquisition, stimulus presentation, and brain monitoring*. Springer Science and Business Media, 2010.

[229] Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.

[230] Emanuel Neto, Felix Biessmann, Harald Aurlien, Helge Nordby, and Tom Eichele. Regularized linear discriminant analysis of eeg features in dementia patients. *Frontiers in aging neuroscience*, 8:273, 2016.

[231] Fangzhou Yao, Jeff Coquery, and Kim-Anh Le Cao. Independent principal component analysis for biologically meaningful dimension reduction of large biological data sets. *BMC bioinformatics*, 13:1–15, 2012.

[232] Sungho Tak and Jong Chul Ye. Statistical analysis of fnirs data: a comprehensive review. *Neuroimage*, 85:72–91, 2014.

[233] Gan Huang, Guangquan Liu, Jianjun Meng, Dingguo Zhang, and Xiangyang Zhu. Model based generalization analysis of common spatial pattern in brain computer interfaces. *Cognitive neurodynamics*, 4:217–223, 2010.

[234] Jieping Ye. Least squares linear discriminant analysis. *Proceedings of the 24th international conference on Machine learning*, pages 1087–1093, 2007.

[235] Student. The probable error of a mean. *Biometrika*, pages 1–25, 1908.

[236] Samuel Sanford Shapiro and Martin B Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52:591–611, 1965.

[237] Phillip Good. *Permutation tests: a practical guide to resampling methods for testing hypotheses*. Springer Science and Business Media, 2013.

[238] Anjana Gosain and Saanchi Sardana. Handling class imbalance problem using oversampling techniques a review. pages 79–85. IEEE, 2017.

[239] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[240] Guillaume Lemaztre, Fernando Nogueira, and Christos K Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of machine learning research*, 18:1–5, 2017.

[241] Abraham Savitzky and Marcel J E Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36:1627–1639, 1964.

[242] Rik van Dinteren, Martijn Arns, Marijtje L A Jongsma, and Roy P C Kessels. P300 development across the lifespan a systematic review and meta-analysis. *PloS one*, 9:e87347, 2014.

[243] Mark W Geisler and John Polich. P300 and individual differences: morning/evening activity preference, food, and time-of-day. *Psychophysiology*, 29:86–94, 1992.

[244] Ivo Kathner, Selina C Wriessnegger, Gernot R Muller-Putz, Andrea Kubler, and Sebastian Halder. Effects of mental workload and fatigue on the p300, alpha and theta band power during operation of an erp (p300) brain computer interface. *Biological psychology*, 102:118–129, 2014.

[245] C Guger, G Krausz, and G Edlinger. *Brain-computer interface control with dry EEG electrodes*. Citeseer, 2011.

[246] Hermann Hinrichs, Michael Scholz, Anne Katrin Baum, Julia W Y Kam, Robert T Knight, and Hans-Jochen Heinze. Comparison between a wireless dry electrode eeg system with a conventional wired wet electrode eeg system for clinical applications. *Scientific reports*, 10:5218, 2020.

[247] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, and Lauri Parkkonen. Meg and eeg data analysis with mne-python. *Frontiers in Neuroinformatics*, 7:267, 2013.

[248] James W Covington and John Polich. P300, stimulus intensity, and modality. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section*, 100:579–584, 1996.

[249] H Nolan, R Whelan, and R B Reilly. Faster: Fully automated statistical thresholding for eeg artifact rejection. *Journal of Neuroscience Methods*, 2010.

[250] Brendan Z Allison and Christa Neuper. Could anyone use a bci? *Brain-computer interfaces: Applying our minds to human-computer interaction*, pages 35–54, 2010.

[251] Sebastian Halder, Andreas Pinegger, Ivo Kathner, Selina C Wriessnegger, Josef Faller, Joao B Pires Antunes, Gernot R Muller-Putz, and Andrea Kubler. Brain-controlled applications using dynamic p300 speller matrices. *Artificial intelligence in medicine*, 63:7–17, 2015.

[252] Eric W Sellers, Dean J Krusienski, Dennis J McFarland, Theresa M Vaughan, and Jonathan R Wolpaw. A p300 event-related potential brain computer interface the effects of matrix size and inter stimulus interval on performance. *Biological psychology*, 73:242–252, 2006.

[253] Sandra G Hart and Lowell E Staveland. *Development of NASA-TLX Task Load Index Results of empirical and theoretical research*, volume 52, pages 139–183. Elsevier, 1988.

[254] Charles R Harris, K Jarrod Millman, Stefan J Van Der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, and Nathaniel J Smith. Array programming with numpy. *Nature*, 585:357–362, 2020.

[255] Clemens Brunner, Brendan Z Allison, Dean J Krusienski, Vera Kaiser, Gernot R Muller-Putz, Gert Pfurtscheller, and Christa Neuper. Improved signal processing approaches in an offline simulation of a hybrid brain-computer interface. *Journal of neuroscience methods*, 188:165–173, 2010.

[256] Xiaogang Chen, Zhikai Chen, Shangkai Gao, and Xiaorong Gao. A high-itr ssvep-based bci speller. *Brain-Computer Interfaces*, 1:181–191, 2014.

[257] Ulrich Hoffmann, Jean-Marc Vesin, Touradj Ebrahimi, and Karin Diserens. An efficient p300-based brain-computer interface for disabled subjects. *Journal of Neuroscience methods*, 167:115–125, 2008.

[258] Angela Riccio, Francesca Schettini, Valentina Galiotta, Enrico Giraldi, Maria Grazia Grasso, Febo Cincotti, and Donatella Mattia. Usability of a hybrid system combining p300-based brain-computer interface and commercial assistive technologies to enhance communication in people with multiple sclerosis. *Frontiers in Human Neuroscience*, 16:868419, 2022.

[259] Lukas Vareka. Evaluation of convolutional neural networks using a large multi-subject p300 dataset. *Biomedical signal processing and control*, 58:101837, 2020.

[260] Pu Du, Penghai Li, Longlong Cheng, Xueqing Li, and Jianxian Su. Single-trial p300 classification algorithm based on centralized multi-person data fusion cnn. *Frontiers in Neuroscience*, 17:1132290, 2023.

[261] A S Albahri, Z T Al-Qaysi, Laith Alzubaidi, Alhamzah Alnoor, O S Albahri, A H Alamoodi, and Anizah Abu Bakar. A systematic review of using deep learning technology in the steady-state visually evoked potential-based brain-computer interface applications: Current trends and future trust methodology. *International Journal of Telemedicine and Applications*, 2023:7741735, 2023.

[262] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *32nd International Conference on Machine Learning, ICML 2015*, 1, 2015.

[263] Yipeng Du and Jian Liu. Ienet: a robust convolutional neural network for eeg based brain-computer interfaces. *Journal of neural engineering*, 19:036031, 2022.

[264] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[265] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *7th International Conference on Learning Representations, ICLR 2019*, 2019.

[266] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. *13th International Conference on Artificial Intelligence and Statistics*, 2010.

[267] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. pages 1026–1034, 2015.

[268] Tanja Krumpe, Katrin Baumgaertner, Wolfgang Rosenstiel, and Martin Spuler. Non-stationarity and inter-subject variability of eeg characteristics in the context of bci development. Verlag der TU Graz, 2017.

[269] Alexander Kamrud, Brett Borghetti, and Christine Schubert Kabban. The effects of individual differences, non-stationarity, and the importance of data partitioning decisions for training and testing of eeg cross-participant models. *Sensors*, 21:3225, 2021.

[270] Yannick Roy, Hubert Banville, Isabela Albuquerque, Alexandre Gramfort, Tiago H Falk, and Jocelyn Faubert. Deep learning-based electroencephalography analysis: a systematic review. *Journal of neural engineering*, 16:051001, 2019.

[271] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 2623–2631, 2019.

[272] Liam Li, Kevin Jamieson, Afshin Rostamizadeh, Ekaterina Gonina, Jonathan Ben-Tzur, Moritz Hardt, Benjamin Recht, and Ameet Talwalkar. A system for massively parallel hyperparameter tuning. *Proceedings of Machine Learning and Systems*, 2:230–246, 2020.

[273] Irene Winkler, Stefan Debener, Klaus-Robert Muller, and Michael Tangermann. On the influence of high-pass filtering on ica-based artifact reduction in eeg-erp. *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 4101–4105, 2015.

[274] Wenqiang Yan, Bo He, Jin Zhao, Yongcheng Wu, Chenghang Du, and Guanghua Xu. Frequency domain filtering method for ssvep-eeg preprocessing. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2023.

[275] Trung-Hau Nguyen and Wan-Young Chung. A single-channel ssvep-based bci speller using deep learning. *IEEE Access*, 7:1752–1763, 2018.

[276] Zhongke Gao, Tao Yuan, Xinjun Zhou, Chao Ma, Kai Ma, and Pan Hui. A deep learning method for improving the classification accuracy of ssmvep-based bci. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 67:3447–3451, 2020.

[277] Yuanlu Zhu, Ying Li, Jinling Lu, and Pengcheng Li. Eegnet with ensemble learning to improve the cross-session classification of ssvep based bci from ear-eeg. *IEEE Access*, 9:15295–15303, 2021.

[278] Yipeng Du, Mingxi Yin, and Bingli Jiao. Inceptionssvep: a multi-scale convolutional neural network for steady-state visual evoked potential classification. pages 2080–2085. IEEE, 2020.

[279] Dechun Zhao, Tian Wang, Yuanyuan Tian, and Xiaoming Jiang. Filter bank convolutional neural network for ssvep classification. *IEEE Access*, 9:147129–147141, 2021.

[280] Sina Ardabili, Amir Mosavi, and Annamaria R Varkonyi-Koczy. Systematic review of deep learning and machine learning models in biofuels research. *International Conference on Global Research and Education*, pages 19–32, 2019.

[281] Pezhman Taherei Ghazvinei, Hossein Hassanpour Darvishi, Amir Mosavi, Khamaruzaman bin Wan Yusof, Meysam Alizamir, Shahaboddin Shamshirband, and Kwok-Wing Chau. Sugarcane growth prediction based on meteorological parameters using extreme learning machine and artificial neural network. *Engineering Applications of Computational Fluid Mechanics*, 12:738–749, 2018.

[282] Maher G M Abdolrasol, S M Suhail Hussain, Taha Selim Ustun, Mahidur R Sarker, Mahammad A Hannan, Ramizi Mohamed, Jamal Abd Ali, Saad Mekhilef, and Abdalrhman Milad. Artificial neural networks based optimization techniques: A review. *Electronics*, 10:2689, 2021.

[283] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

[284] D Randall Wilson and Tony R Martinez. The need for small learning rates on large problems. *IJCNN International Joint Conference on Neural Networks. Proceedings (Cat. No. 01CH37222)*, 1:115–119, 2001.

[285] Guangyu Bin, Xiaorong Gao, Yijun Wang, Yun Li, Bo Hong, and Shangkai Gao. A high-speed bci based on code modulation vep. *Journal of neural engineering*, 8:25015, 2011.

[286] Vernon J Lawhern. Github - vlawhern/arl-eegmodels: This is the army research laboratory (arl) eegmodels project: A collection of convolutional neural network (cnn) models for eeg signal classification, using keras and tensorflow, retrieved from https://github.com/vlawhern/arl-eegmodels, 2023.

[287] Francois Chollet. Keras: https://github.com/fchollet/keras, 2015.

[288] Yijun Wang, Shangkai Gao, and Xiaorong Gao. Common spatial pattern method for channel selelction in motor imagery based brain-computer interface. pages 5392–5395. IEEE, 2006.

[289] Ioannis Xygonakis, Alkinoos Athanasiou, Niki Pandria, Dimitris Kugiumtzis, and Panagiotis D Bamidis. Decoding motor imagery through common spatial pattern filters at the eeg source space. *Computational intelligence and neuroscience*, 2018, 2018.

[290] Kai Keng Ang, Zheng Yang Chin, Chuanchu Wang, Cuntai Guan, and Haihong Zhang. Filter bank common spatial pattern algorithm on bci competition iv datasets 2a and 2b. *Frontiers in Neuroscience*, 2012.

[291] Michael Tangermann, Klaus-Robert Muller, Ad Aertsen, Niels Birbaumer, Christoph Braun, Clemens Brunner, Robert Leeb, Carsten Mehring, Kai J Miller, and Gernot Mueller-Putz. Review of the bci competition iv. *Frontiers in neuroscience*, page 55, 2012.

[292] Robert Leeb, Felix Lee, Claudia Keinrath, Reinhold Scherer, Horst Bischof, and Gert Pfurtscheller. Brain-computer communication: motivation, aim, and impact of exploring a virtual apartment. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 15:473–482, 2007.

[293] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12, 2011.

[294] Djork Arne Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*, 2016.

[295] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[296] ai neptune. neptune.ai: experiment tracker, 2023.

[297] James Bergstra, Remi Bardenet, Yoshua Bengio, and Balazs Kegl. Algorithms for hyper-parameter optimization. *Advances in neural information processing systems*, 24, 2011.

[298] Nik Khadijah Nik Aznan, Amir Atapour-Abarghouei, Stephen Bonner, Jason D Connolly, Noura Al Moubayed, and Toby P Breckon. Simulating brain signals: Creating synthetic eeg data via neural-based generative models for improved ssvep classification. *2019 International joint conference on neural networks (IJCNN)*, pages 1–8, 2019.

[299] Sourav Kundu and Samit Ari. Mscnn: A deep learning framework for p300-based brain-computer interface speller. *IEEE Transactions on Medical Robotics and Bionics*, 2:86–93, 2019.

[300] Pu Du, Penghai Li, Longlong Cheng, Xueqing Li, and Jianxian Su. Single-trial p300 classification algorithm based on centralized multi-person data fusion cnn. *Frontiers in Neuroscience*, 17:1132290, 2023.

[301] Wei Gao, Weichen Huang, Man Li, Zhenghui Gu, Jiahui Pan, Tianyou Yu, Zhu Liang Yu, and Yuanqing Li. Eliminating or shortening the calibration for a p300 brain-computer interface based on a convolutional neural network and big electroencephalography data: An online study. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:1754–1763, 2023.

[302] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. pages 8789–8797, 2018.

[303] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation: One-shot anatomically consistent facial animation. *International Journal of Computer Vision*, 128:698–713, 2020.

[304] Sarah F Lang, Charles A Nelson, and Paul F Collins. Event-related potentials to emotional and neutral stimuli. *Journal of Clinical and Experimental Neuropsychology*, 12:946–958, 1990.

[305] Edward Bernat, Scott Bunce, and Howard Shevrin. Event-related brain potentials differentiate positive and negative mood adjectives during both supraliminal and subliminal visual processing. *International journal of psychophysiology*, 42:11–34, 2001.

[306] Susan J Thomas, Stuart J Johnstone, and Craig J Gonsalvez. Event-related potentials during an emotional stroop task. *International journal of psychophysiology*, 63:221–231, 2007.

[307] Yuanyuan Zhang, Gaoyan Zhang, and Baolin Liu. Investigation of the influence of emotions on working memory capacity using erp and ersp. *Neuroscience*, 357:338–348, 2017.

[308] Richard A Osborne, Ramnik Sekhon, Wendy Johnston, and Sanjay Kalra. Screening for frontal lobe and general cognitive impairment in patients with amyotrophic lateral sclerosis. *Journal of the neurological sciences*, 336:191–196, 2014.

[309] Howe Yuan Zhu, Hsiang-Ting Chen, and Chin-Teng Lin. Understanding the effects of stress on the p300 response during naturalistic simulation of heights exposure. *Plos one*, 19:e0301052, 2024.

[310] Minju Kim, Jongsu Kim, Dojin Heo, Yunjoo Choi, Taejun Lee, and Sung-Phil Kim. Effects of emotional stimulations on the online operation of a p300-based brain-computer interface. *Frontiers in human neuroscience*, 15:612777, 2021.

[311] Chuan Jia, Xiaorong Gao, Bo Hong, and Shangkai Gao. Frequency and phase mixed coding in ssvep-based brain - computer interface. *IEEE Transactions on Biomedical Engineering*, 58:200–206, 2011.

[312] Lutz Roeder. Netron, visualizer for neural network, deep learning, and machine learning models, 2017.

Appendix A

Appendix

A.1 Experiment 1: Inversion Method Results: No Class-Balancing: Pipeline 1

A.1.1 Pooled-Subject: No Class-Balancing: Pipeline 1

The results reported in this subsection refer to analyses undertaken for the Inversion No Balance data partition (see, Table A.1). The data in question contain all subject EEG time-series collected during the simultaneous visual presentation of the Inversion method of stimulus augmentation (see subsection 3.3.3). The ratio of P300 and Non-P300 events is 1:6, with one target emoji stimulus selected at the start of each trial according to a non-consecutive randomisation protocol.

	Mean Accuracy (%)	P300 Accuracy (%)	Non-P300 Accuracy (%)	Solver	Shrinkage	Num Test Events
Cross Subject	83.38	0	100	lsqr	0.01	679
Subject 1	85.29	0	100	lsqr	0.21	68
Subject 2	85.29	0	100	lsqr	0.01	68
Subject 3	86.96	0	100	lsqr	0.04	69
Subject 4	86.96	0	100	lsqr	0.24	69
Subject 5	83.82	0	98.28	lsqr	0.10	68
Subject 6	86.96	0	100	lsqr	0.11	69
Subject 7	86.96	0	100	lsqr	0.26	69
Subject 8	84.13	0	100	lsqr	0.31	63
Subject 9	85.29	0	100	lsqr	0.08	68
Subject 10	85.29	0	100	lsqr	0.05	68
Single Subject Avg.	85.70	0	99.83	n/a	0.14	67.9
Single Subject Var.	1.57	0	0.86	n/a	0.15	3

Table A.1: A table of classification performance metrics and optimization results from the Inversion No Balance dataset (refer to Table 3.5 for data partition info). For further information on field-headings refer to Table 3.6.

As can be seen in Table A.1 the LDA classifier trained using the pooled-subject data achieved greater than random performance. Mean accuracy is high (83.38%) with a clear imbalance in performance metrics for P300 events (MA= 0%) as compared to Non-P300 events (MA=100%). When applying the grid search technique it was found the optimal combination of solver and shrinkage value to be the Least Squares (lsqr) method at 0.01, respectively.

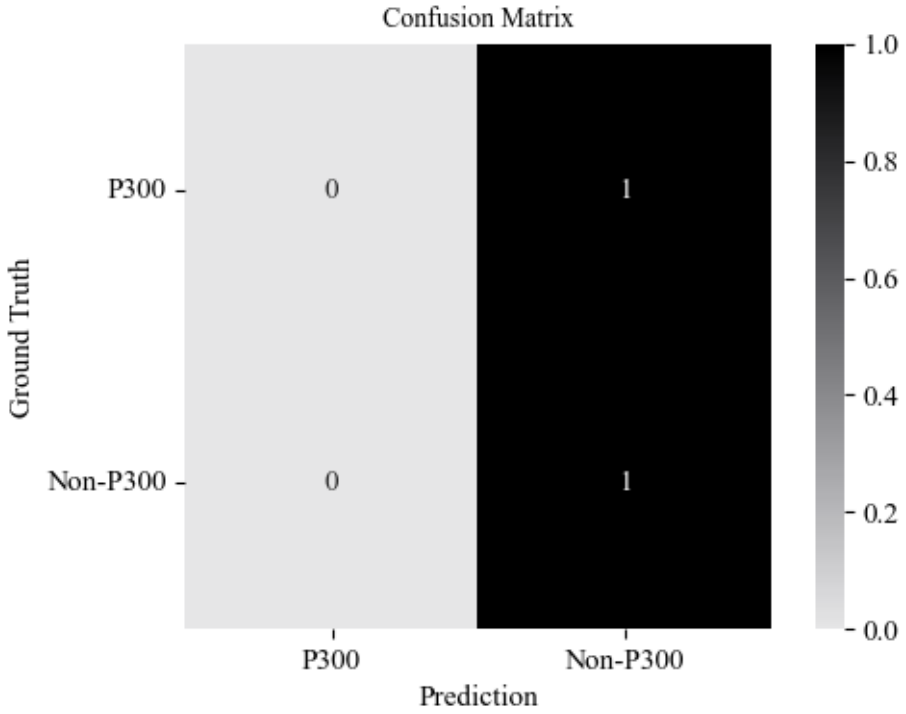

Figure A.1: Displayed here is a confusion matrix generated via the LDA classifier across all subjects for the Inversion No Balance data partition (refer to Table A.1). For more information on interpreting this figure see, Figure 3.10.

In the above figure (Figure A.1) the confusion matrix indicates a selective bias towards the prediction of Non-P300 events, leading to a hit score of 1 for this class (see, bottom right quadrant). The hit score for P300 events was found to be zero, with significant confusion present in the classification of these target waveforms.

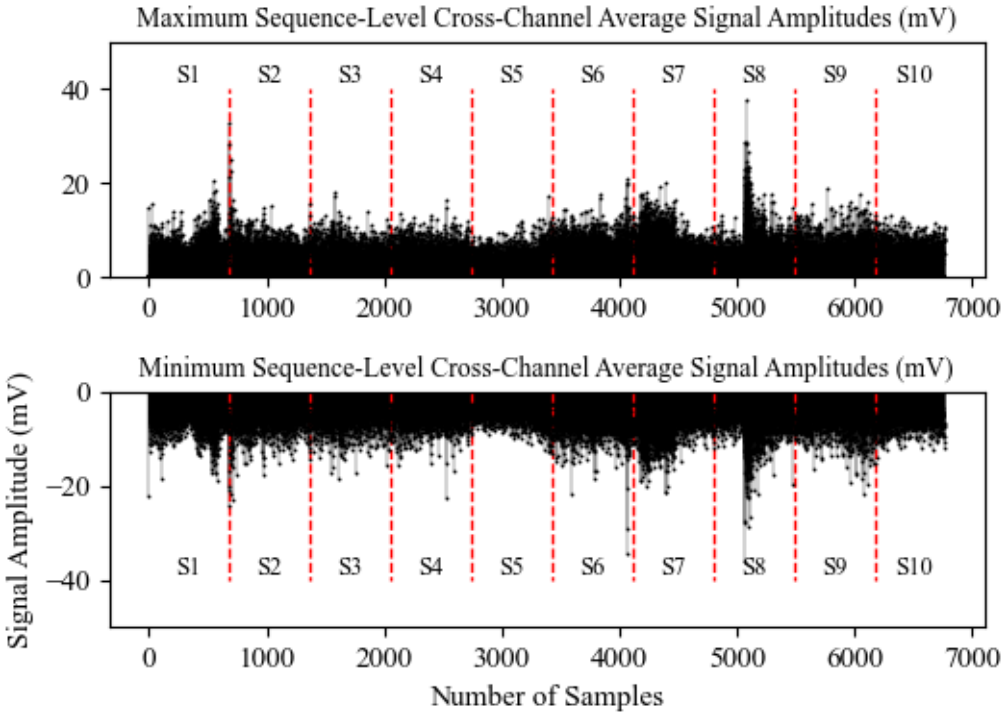

Figure A.2: This shows the distribution of the maximum positive and negative values for every respective event across all subjects for the Inversion No Balance dataset (refer to Table A.1) retained post-processing. For further info on interpretation see, Figure 3.14.

The above figure (Figure A.2) indicates the vast majority of subject events included for analysis, in both positive and negative deflections, remained well within $+/-20$mV. Some select subjects (Subjects 2 & 8) demonstrated significantly higher positive deflections, along with Subjects 7 and 8 demonstrating substantially larger negative deflections, as compared to the average range of events sampled. The relevance of these deflections will be discussed further in the following analysis sections.

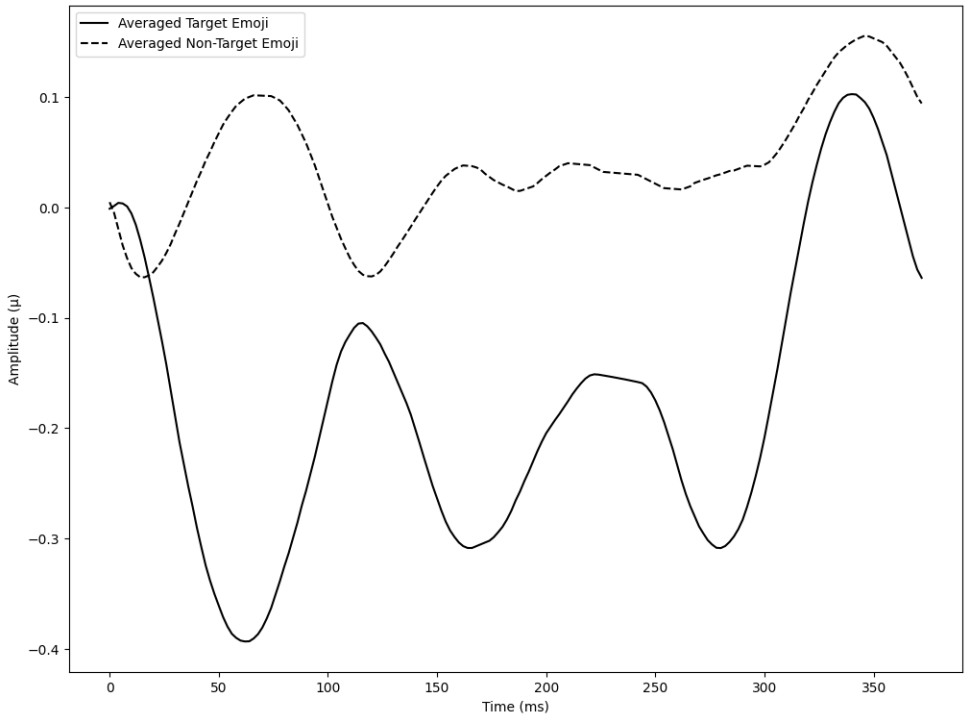

Figure A.3: Here is shown a Cz grand average plot for all trial P300 (solid line) and Non-P300 (dashed line) events respectively collected during the Inversion No Balance data partition (refer to, Table 3.1 for data partition info). Each augmentation event generated a data stream marker used to window the data into 350ms chunks. Given the onset and offset times of the augmentations through each trial sequence, some data is shared across data chunks for different emoji targets. These waveforms are computed by averaging across all P300 (relating to cued target emoji instances) or Non-P300 events (non-cued target emoji instances) and isolated exclusively to the central Cz channel. The averages generated across these classes amplify underlying EEG waveform patterns embedded in the signals. Further, all plots were baselined by computing the average of the first 50ms of the samples collected. Note, that these baselining measures were implemented exclusively for presentation purposes and were not applied during the Pipeline 1 approach to data pre-processing as stated in subsection 3.3.5.3, see table 3.1.

As seen in Figure A.3, the averaged pooled-subject waveforms for the Inversion No Balance data partitions are relatively representative of the typical waveform profiles expected for P300 and Non-P300 signal averages. The upper P300 average plot presents a significant negative component at 50ms, with a large positive deflection around 300-350ms. Further, the Non-P300 waveform average signal (lower plot) does present the correct visual properties, in that the range of values is significantly reduced in comparison to the upper plot. This suggests redundant noise and non-stationary components have converged towards zero over the course of iterative averaging.

A.1.2 Within-Subject: No Class-Balancing: Pipeline 1

As seen in Table A.1 the average within-subject accuracy is 85.70% $+/-1.57\%$, this is significantly above the non-class balanced random performance threshold. The P300 performance levelled out at zero (0%), with Non-P300 accuracy at 99.83% $+/-0.86\%$. The solver selected in all instances is the lsqr method, with an average shrinkage metric of 0.14 $+/-0.15$. Subject 8 returned the least amount of samples post-channel rejection (63 *vs.* the average of 67.9).

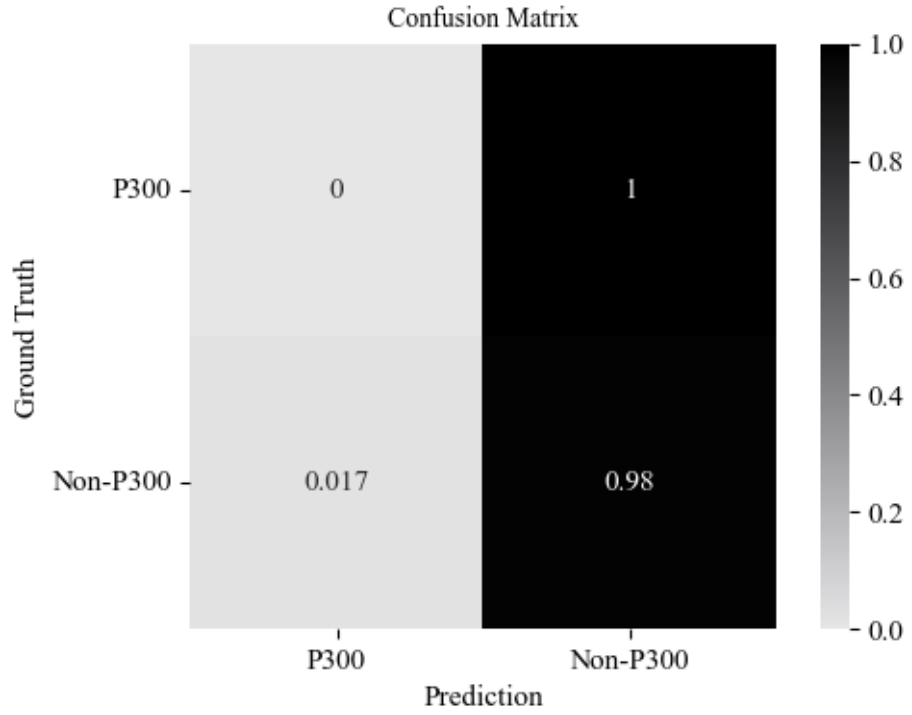

Figure A.4: Displayed here is a confusion matrix generated via the LDA classifier for subject 5 in the Inversion No Balance data partition (refer to Table A.1). For more information on interpretation see, Figure 3.10.

As seen above in Figure A.4, the results at the single-subject level broadly conform with the performance observed at the pooled-subject level. The classification data displayed here represents the only subject with a corresponding LDA classifier that did not produce exclusively Non-P300 predictions. As seen above, one instance is recorded that involved a Non-P300 event being misclassified as a P300 event.

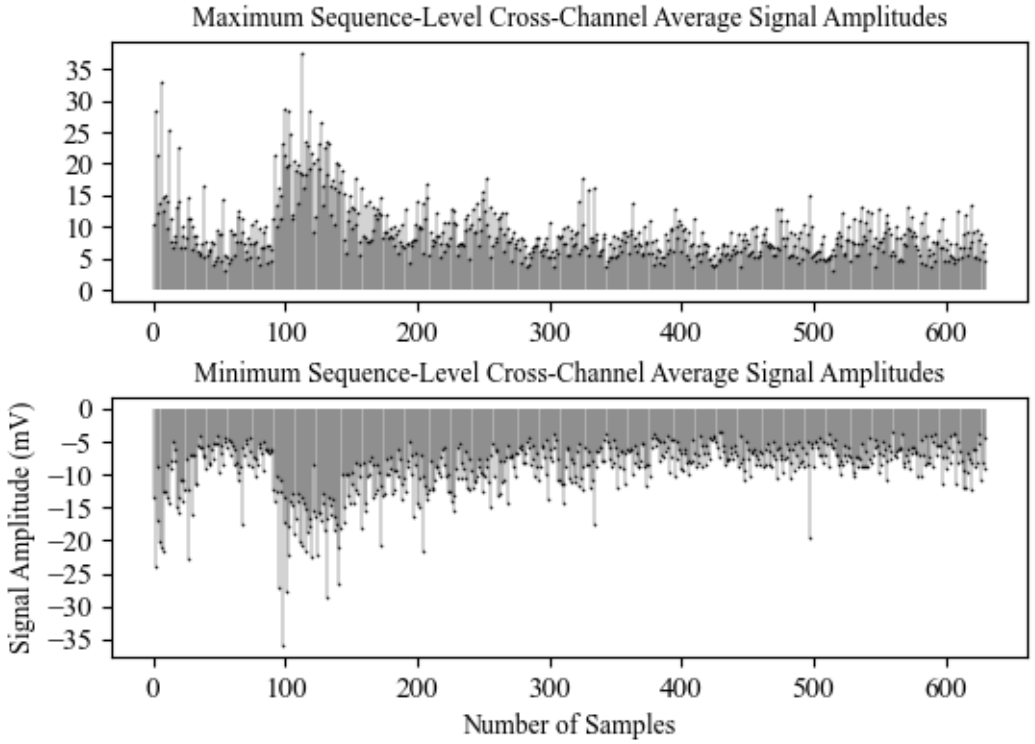

Figure A.5: This shows the distribution of the maximum positive and negative values for every respective event for Subject 8 in the Inversion No Balance dataset (refer to Table A.1) retained post-processing. For further information on interpretation see, Figure 3.14.

In Figure A.5 (above), a distribution of maximum and minimum amplitude values is presented for Subject 8. This subject was previously highlighted in the above pooled-subject section as presenting some of the highest ranges in µV amplitudes. The larger range values are present in the early stages of the experiment and tail off towards the later portion of the experimental session. Less than 5% of all samples included presented with amplitude values exceeding $+/-25$ µV.

A.2 Experiment 1: Combined Method Results: No Class-Balancing: Pipeline 1

After compiling and interpreting the results for the Flash and Inversion augmentation methods data individually it was evident that one of the reasons the analyses may have struggled to accurately distinguish between P300 and Non-P300 events might be related to the volume of examples available to the LDA analyses models employed. To examine the possibility that increasing the volume of samples available to the LDA discriminant function could enhance classification performance the Flash and Inversion method datasets were aggregated into the Combined No Balance dataset (refer to Table 2.1). The basis of each experimental method is

identical, aside from the method of stimulus augmentation. This, in the mind of the author, justifies the combination of these datasets. Again, the data and results discussed in this section are collated from all subjects across both experimental augmentation methods (Flash and Inversion) and possess a P300 to Non-P300 event ratio of 1:6. Note, that amplitude range plots will not be discussed in the aggregate subsections as these data have already been presented piecemeal throughout the text.

	Mean Accuracy (%)	P300 Accuracy (%)	Non-P300 Accuracy (%)	Solver	Shrinkage	Num Test Events
Cross Subject	84.36	0	100	lsqr	0.01	1364
Subject 1	84.67	0	100	lsqr	0.56	137
Subject 2	84.67	0	100	lsqr	0.03	137
Subject 3	85.51	0	100	lsqr	0.01	138
Subject 4	85.51	0	100	lsqr	0.03	138
Subject 5	85.93	0	100	lsqr	0.08	135
Subject 6	85.51	0	100	lsqr	0.10	138
Subject 7	85.51	0	100	lsqr	0.23	138
Subject 8	86.92	0	100	lsqr	0.05	130
Subject 9	86.03	0	100	lsqr	0.08	136
Subject 10	84.67	0	100	lsqr	0.05	137
Single Subject Avg.	85.49	0	100	n/a	0.12	136.4
Single Subject Var.	1.13	0	0	n/a	0.28	4

Table A.2: A table of classification performance metrics and optimization results from the Combined No Balance dataset (refer to Table 2.1). For further information on field-headings refer to Table 2.2

A.2.1 Pooled-Subject: No Class-Balancing: Pipeline 1

As can be seen in Table A.2, the pooled-subject mean accuracy for the Combined No Balance dataset is reported at 84.36%. The P300 Accuracy is reported as 0%, with the LDA classifier demonstrating a clear bias for the Non-P300 class events showing a 100% classification accuracy. The grid search optimisation returned a selection for the lsqr solver with an optimised shrinkage rate of 0.01.

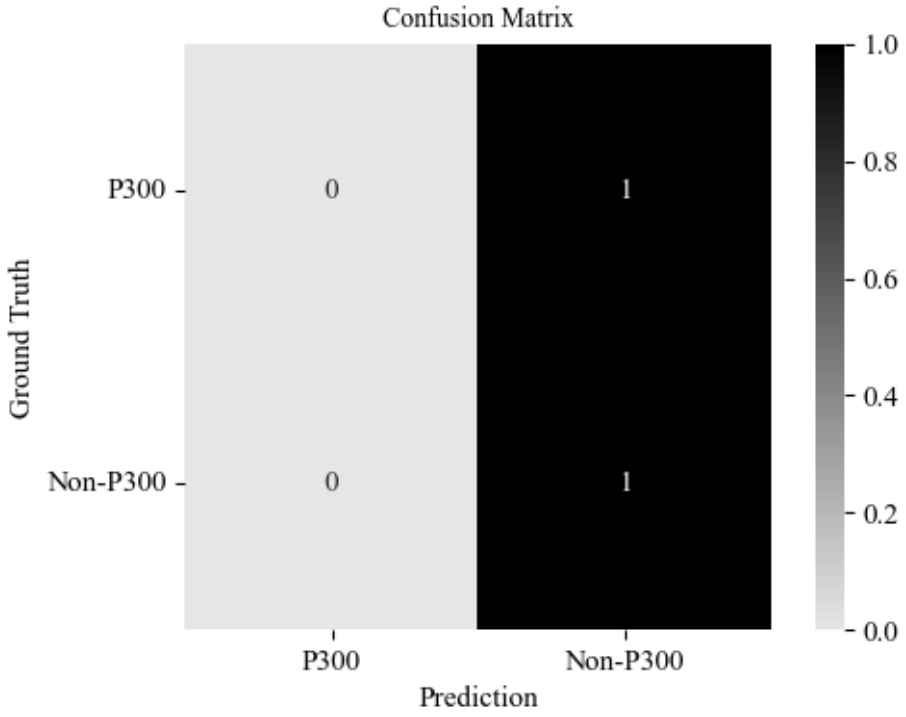

Figure A.6: Confusion matrix displaying the normalized classification performance of trained LDA models of each respective class for the pooled-subject Combined No Balance data partition.

The above confusion matrix figure (Figure A.7) demonstrates the exclusive attribution of all test events to the Non-P300 event class. At no point is the P300 class selected, even in an instance of classifier confusion.

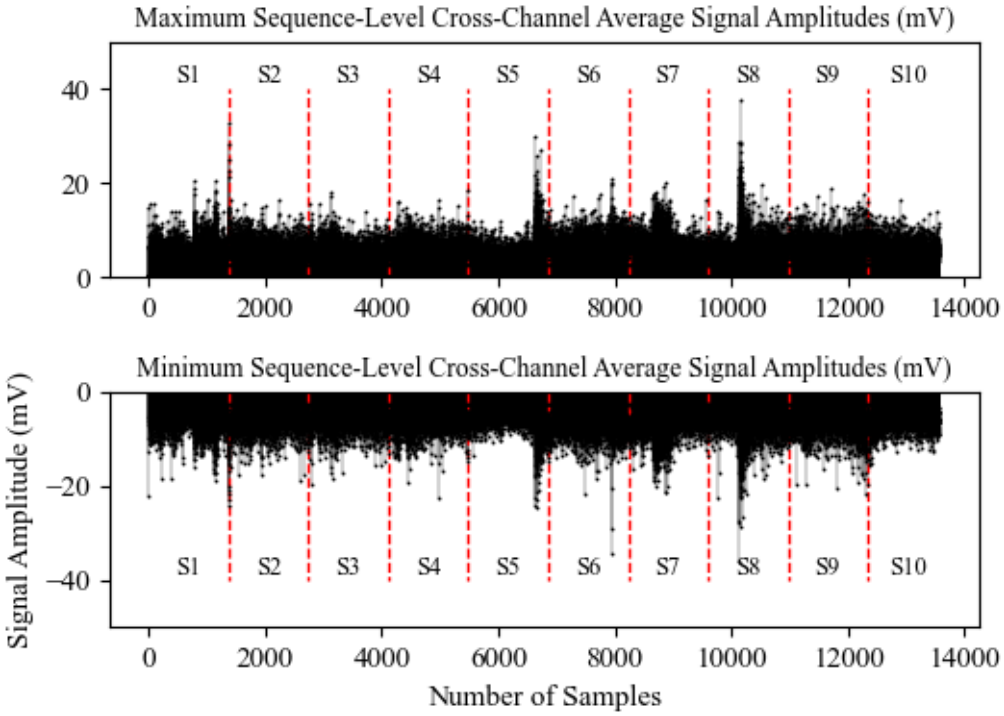

Figure A.7: Here is presented a plot of µV amplitude ranges for all test events in the Combined No Balance data partition (including both P300 and Non-P300 events). The upper plot contains the maximum µV amplitude recorded in each respective event and the low plot displays the minimum µV amplitude sampled in each respective event. The total number of events attainable post-processing amounts to 7 (emoji) $\times$ 49 (trials) $\times$ 4 (blocks) $\times$ 10 (subjects) $=$ 13720 samples.

As can be seen in Figure A.8, the µV amplitude upper plot reveals 3 subjects displaying maximum µV amplitude values outside the typical range of the dataset as a whole, Subjects: 2, 6 and 8. Concerning the lower minimum µV value plots, Subjects 7 and 8 demonstrate the presence of lower minimum amplitudes. It must be noted that the higher overall variance in the lower minimum µV amplitude plot reduces the ability to visually appraise potential outliers.

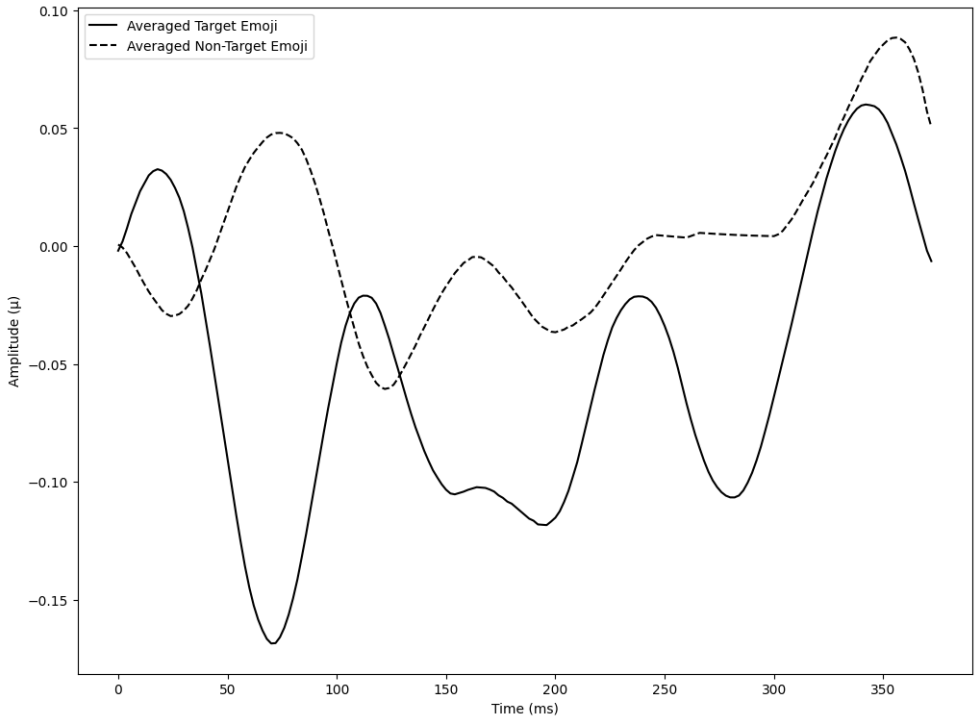

Figure A.8: Here is shown a Cz grand average plot for all trial P300 (solid line) and Non-P300 (dashed line) events respectively collected for the Both No Balance data partition (refer to, Table 3.1 for data partition info). Each augmentation event generated a data stream marker used to window the data into 350ms chunks. Given the onset and offset times of the augmentations through each trial sequence, some data is shared across data chunks for different emoji targets. These waveforms are computed by averaging across all P300 (relating to cued target emoji instances) or Non-P300 events (non-cued target emoji instances) and isolated exclusively to the central Cz channel. The averages generated across these classes amplify underlying EEG waveform patterns embedded in the signals. Further, all plots were baselined by computing the average of the first 50ms of the samples collected. Note, that these baselining measures were implemented exclusively for presentation purposes and were not applied during the Pipeline 1 approach to data pre-processing as stated in subsection 3.3.5.3, see table 3.1.

The figure presented above (Figure A.9) shows the averaged signals taken across both stimulus augmentation methods (Flash and Inversion) for the P300 and Non-P300 classes. The P300 average plot (upper) does not appear to be adequately baselined to zero. Despite this, the presence of a strong negative component at 50-100ms and a robust positive component between 300-350ms does suggest the presence of P300-related components in the signals averaged. The Non-P300 average plot (lower) shows a similar trend in amplitude changes over time, at a reduced scale. It must be noted that these signals present with a significantly lower range of μV amplitude ranges (-0.1 to 0.1mV), owing to the dramatically higher number of samples used and the increased incidence of orthogonal waveforms trending the signal towards the baseline.

A.2.2 Within-Subject: No Class-Balancing: Pipeline 1

As can be seen from Table A.2, the within-subject results are highly uniform, with an average mean classification accuracy of 85.49% $+/-1.13\%$. All subjects evaluated demonstrate a P300 event classification accuracy of 0% and a corresponding 100% accuracy for Non-P300 events. The incidence of event rejection is highest for Subject 8, with 130 (of 196) total events retained post-channel rejection. See previous sections for further discussion. The exclusive selection of the lsqr solver method for the LDA evaluation remains consistent for these results, with a slightly lower variance in overall shrinkage values selected across subjects (0.12 $+/-0.28$).

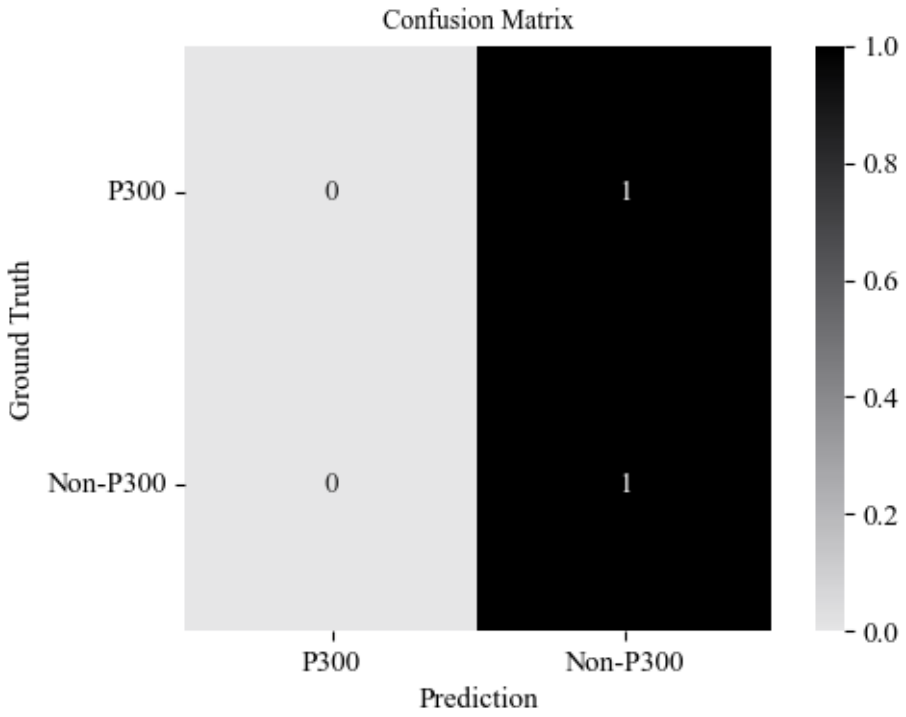

Figure A.9: Here is presented the confusion matrix generated from the results of the LDA analyses conducted using Subject 10 for the Combined No Balance data partition (refer to Table A.2).

The figure positioned above (Figure A10) displays the confusion matrix for Subject 10 and breaks down the prediction preferences demonstrated by the LDA analyses trained using the data in question. This figure illustrates the complete overfitting of the model towards a selection bias for the Non-P300 event data, as all predictions made identified each test event as belonging to the Non-P300 class.

A.3 Experiment 1: Inversion Method Data Partitions: Pipeline 2

	Inversion Collapsed						
	Total Post-Rejection		Test		Train		
Subjects	P300	Non-P300	P300	Non-P300	P300 (Real)	P300 (Synthetic)	Non-P300
Subject 3	93	581	9	58	84	438	522
Subject 5	95	586	9	59	86	441	527
Subject 8	92	584	9	58	83	442	525

Table A.3: Here is presented a table detailing the distribution of sample quantities for the datasets associated with the Inversion Collapsed Pipeline 2 approach. All samples here are composed of signals collected over all 5 sequences of each trial, for more information see subsection 3.3.5.3. For further information on field headings and interpretation please refer to the table above (Table 3.14). Note, that the ratios between Target and Non-Target samples for all datasets listed, including the proportion of Real vs Synthetic P300 instances mirror those in the Non-Collapsed data preparation variant (see table above).

	Inversion Non-Collapsed						
	Total Post-Rejection		Test		Train		
Subjects	P300	Non-P300	P300	Non-P300	P300 (Real)	P300 (Synthetic)	Non-P300
Subject 3	45	289	4	59	41	189	230
Subject 5	46	292	4	59	42	191	233
Subject 8	45	291	4	59	41	191	232

Table A.4: Here is presented a table detailing the distribution of sample quantities for the datasets associated with the Inversion Collapsed Pipeline 2 approach. All samples here are composed of signals from 10 sequences, this was generated by averaging corresponding Target and Non-Target samples for 2 neighbouring trials containing 5 sequences each. For further information on field headings and interpretation please refer to the table above (Table 3.14). Note, that the ratios between Target and Non-Target samples for all datasets listed, including the proportion of Real vs Synthetic P300 instances mirror those in the Non-Collapsed data preparation variant (see table above).

A.4 Experiment 2: Collapsed Data Partitions: Pipeline 2

	3-Emoji Non-Collapsed						
	Total Post-Rejection		Test		Train		
Subjects	P300	Non-P300	P300	Non-P300	P300 (Real)	P300 (Synthetic)	Non-P300
Subject 1	13	28	1	3	12	13	25
Subject 3	11	26	1	3	10	14	24
Subject 5	13	27	1	3	12	12	24

Table A.5: Here is presented a table detailing the distribution of sample quantities for the subject-specific datasets associated with the 3-Emoji (see Figure 4.4), Collapsed Pipeline 2 approach (see subsection 3.3.5.3). All samples here are composed of signals from 10 sequences, this was generated by averaging corresponding Target and Non-Target samples for 2 neighbouring trials containing 5 sequences each. (see subsection 4.3.7). For further information on field headings and interpretation please refer to Table 3.14.

	5-Emoji Non-Collapsed						
	Total Post-Rejection		Test		Train		
Subjects	P300	Non-P300	P300	Non-P300	P300 (Real)	P300 (Synthetic)	Non-P300
Subject 1	11	57	1	6	10	41	51
Subject 3	12	55	1	6	11	39	50
Subject 5	12	56	1	6	11	39	50

Table A.6: Here is presented a table detailing the distribution of sample quantities for the subject-specific datasets associated with the 5-Emoji (see Figure 4.5), Collapsed Pipeline 2 approach (see subsection 3.3.5.3). All samples here are composed of signals from 10 sequences, this was generated by averaging corresponding Target and Non-Target samples for 2 neighbouring trials containing 5 sequences each. (see subsection 4.3.7). For further information on field headings and interpretation please refer to Table 3.14.

	3-Emoji Non-Collapsed						
	Total Post-Rejection		Test		Train		
Subjects	P300	Non-P300	P300	Non-P300	P300 (Real)	P300 (Synthetic)	Non-P300
Subject 1	13	85	1	8	12	65	76
Subject 3	13	86	1	9	12	65	77
Subject 5	13	83	1	8	12	63	75

Table A.7: Here is presented a table detailing the distribution of sample quantities for the subject-specific datasets associated with the 7-Emoji (see Figure 4.42, Collapsed Pipeline 2 approach (see subsection 3.3.5.3). All samples here are composed of signals from 10 sequences, this was generated by averaging corresponding Target and Non-Target samples for 2 neighbouring trials containing 5 sequences each. (see subsection 4.3.7). For further information on field headings and interpretation please refer to Table 3.14.

A.5 Experiment 3: MAINOFF: Pipeline 1

The results herein were computed exclusively using the data gathered during the main experiment. The datasets discussed at both the pooled-subject and single-subject levels were partitioned 9:1 in terms of training and test divisions. At no point was any test data included in the training phase of the LDA classifiers discussed herein. All data comprising the test set consisted of the last 10% of samples gathered for each respective participant. This was done to enhance the ecological validity of the analyses conducted as these methods mirror the conditions observed in a real-world clinical setting. Concerning the pooled-subject dataset, the test dataset was constructed from the final samples acquired across all subjects sampled. This method was employed to ensure an even representation of all subjects across both training and test datasets to avoid any subject-specific biasing. Once partitioned into the respective training and test datasets, all events were randomised to prevent confounding of classifier training via chronological artefacts/order effects. Again, this same process was repeated for the pooled-subject data.

A.5.1 Pooled-Subject

The results described herein relate to the aggregated dataset including samples from all subjects tested (refer to, Tables 4.1 & 4.2 for further details). It must be noted that all pooled-subject data analysis and pre-processing were performed offline.

	Mean Acc (%)	P300 Acc (%)	Non-P300 Acc (%)	Solver	Shrinkage
Pooled-Sub	80.65	0.00	94.34	lsqr	0.36
Subject 1	85.71	0.00	100.00	lsqr	0.08
Subject 2	80.00	0.00	94.12	lsqr	0.09
Subject 3	76.19	0.00	88.89	lsqr	0.31
Sub Avg	80.63	0.00	94.34	n/a	0.16
Sub Var	4.76	0.00	5.56	n/a	0.12

Table A.8: The classification table contains the metrics relevant to the pooled-subject dataset computed without the inclusion of a localizer pre-training stage (MAINOFF).

The classification performance of trained LDA models for this aggregated pooled-subject dataset (see, TableA.3) demonstrates significant biasing towards the Non-P300 (94.34% accuracy), to the detriment of the P300 class (0.00% accuracy). The grid optimization method for computing the ideal solver method led to the selection of the lsqr solver, with a shrinkage of 0.36.

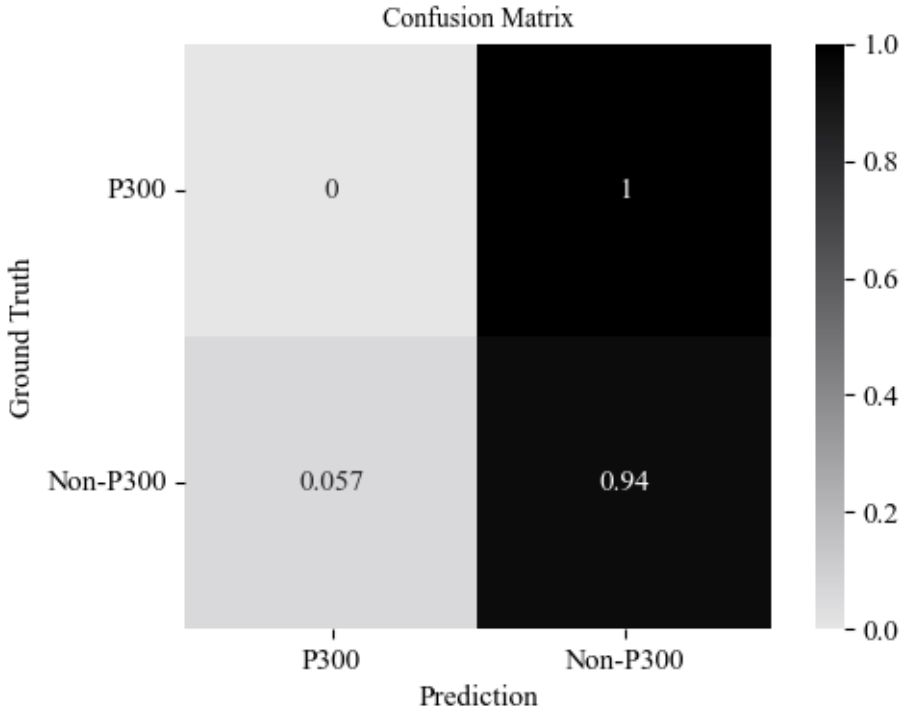

Figure A.10: Here is displayed a normalized confusion matrix reporting the classification performance of the trained LDA model for both P300 and Non-P300 classes relating to the pooled-subject MAIN-OFF analyses block (refer to, Table A.3). Note, that these evaluations were computed using an aggregate dataset comprising all subjects sampled. For further information on the interpretation of this plot refer to, Figure 3.10.

The pattern of performance across classes is illustrated in Figure A.12. This confusion matrix demonstrates that all P300 instances were misclassified as Non-P300 events. The only event preventing complete overfitting was a single Non-P300 event being misclassified as belonging to the P300 class.

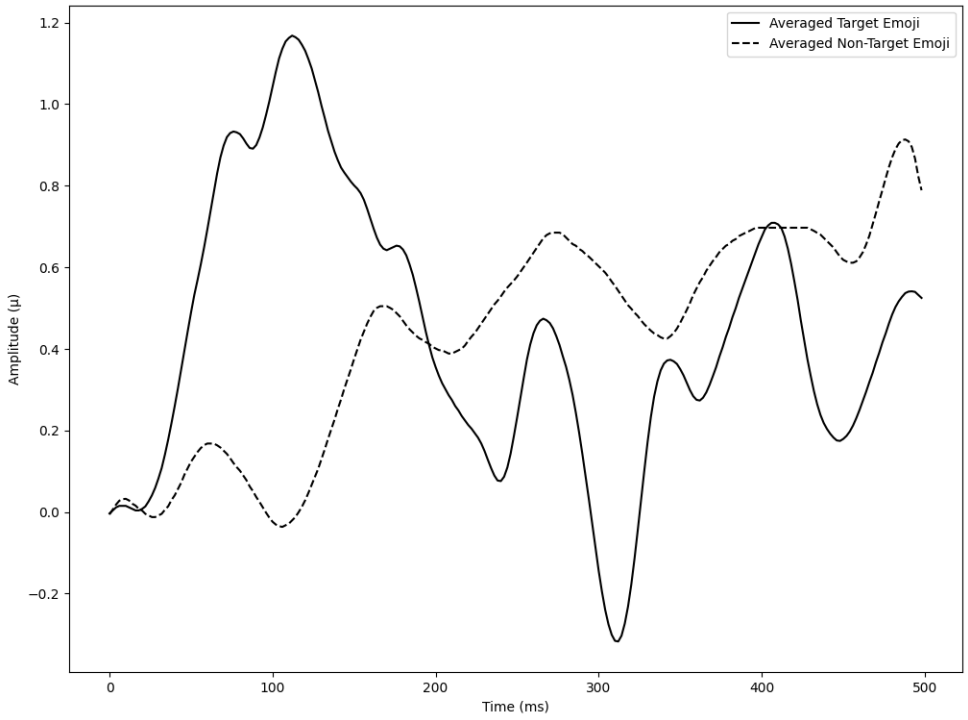

Figure A.11: Here is displayed a Cz grand average plot containing the pooled-trial P300 (upper plot) and non-P300 (lower plot) events for the pooled-subject MAINOFF analysis block (refer to, Table A.8). The x-axis denotes time in milliseconds for the 500ms event data chunk. On the y-axis is amplitude, and displays changes in micro-voltage of the EEG signal. The averages generated across these classes amplify underlying EEG waveform patterns embedded in the signals.

The average plot positioned above (see, Figure A.13) does not provide us with a recognisable example of either the P300 (upper plot) or Non-P300 (lower plot) standard signal. The P300 average signal presents with a large initial crest at 100ms, followed by a drop in µV amplitude and a final positive deflection back to the 0mV baseline. These are not typical waveform characteristics for the expression of a P300 bio-signal. In the lower plot, the signal volatility is comparatively reduced against the upper plot. There are also present some signs of drifting from around 100ms to the end of the 500ms data chunk.

A.5.2 Within-Subject

The single-subject performance for all 3 individuals sampled largely replicates the results observed for the pooled-subject subset. In all instances, significant signs of overfitting are found towards the Non-P300 class, with a universal 0.00% accuracy for the P300 target waveform. The only variance in performance between subjects is the degree of overfitting, with Subject 3 demonstrating the highest incidence of Non-P300 target misclassification at 88.89% accuracy.

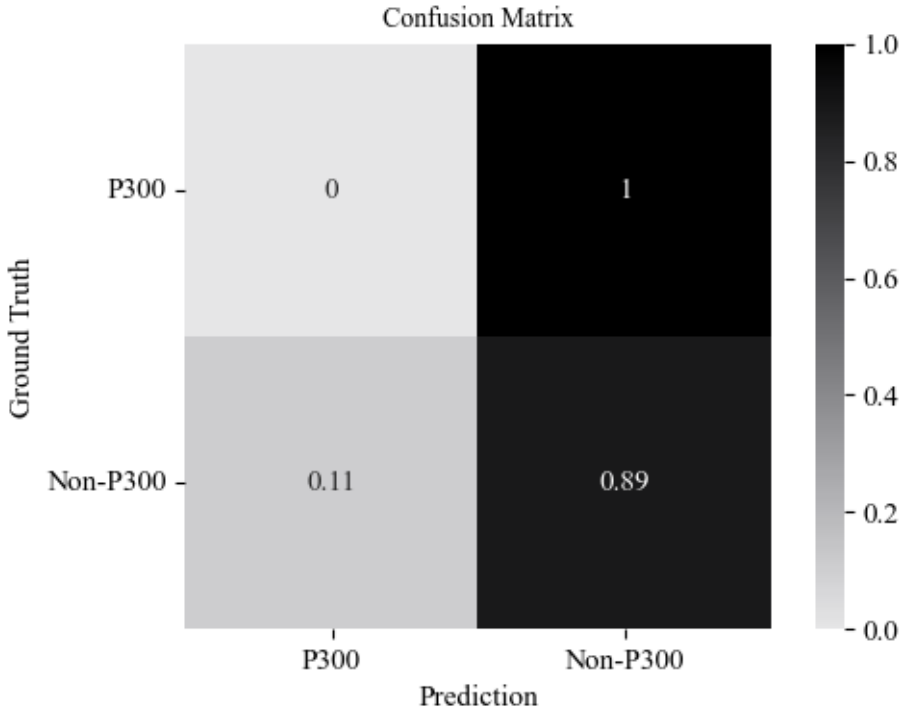

Figure A.12: Here is displayed a normalized confusion matrix reporting the classification performance of the trained LDA model for both P300 and Non-P300 classes relating to Subject 3 in the MAINOFF analyses block (refer to, Table A.8). For further information on the interpretation of this plot refer to, Figure 3.10.

The pattern of Non-P300 misclassification for Subject 3 is presented in the above confusion matrix (see, Figure A.14). The incidence of this phenomenon extends to just over 10% of all Non-P300 events comprising the test dataset. Additionally, all instances of P300 target prediction resulted in misclassifications as Non-P300 waveforms.

A.6 Optimized Network Architectures

A.6.1 ShallowConvNet

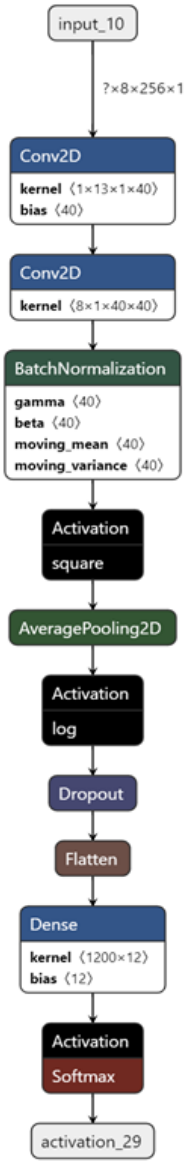

Figure A.13: Here is presented a visualization of the ShallowConvNet [61] architecture utilized throughout the evaluations conducted herein. The input data works through a series of stages primarily composed of 2 convolutional phases along the temporal and spatial dimensions respectively. The visualization app Netron was used to generate this, and all similarly referenced architecture visualizations ([312]).

A.6.2 DeepConvNet

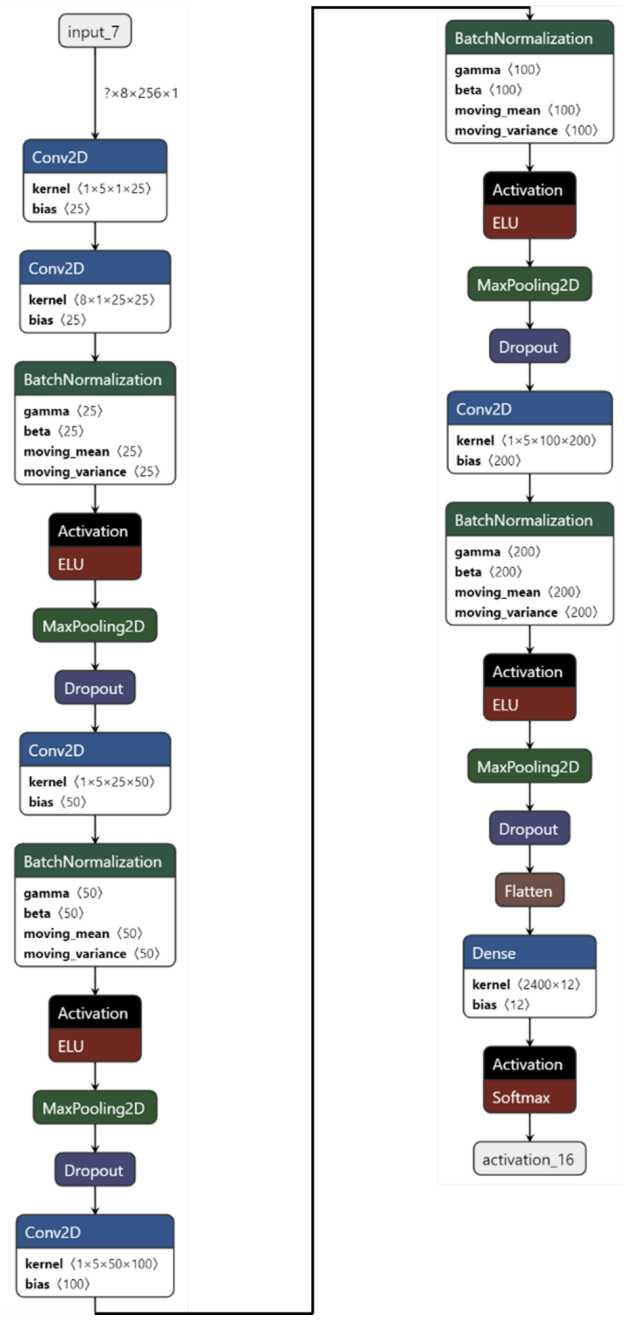

Figure A.14: Here is presented an architecture visualization of the DeepConvNet [61] generated via the Netron app ([312]).

A.6.3 EEGNet

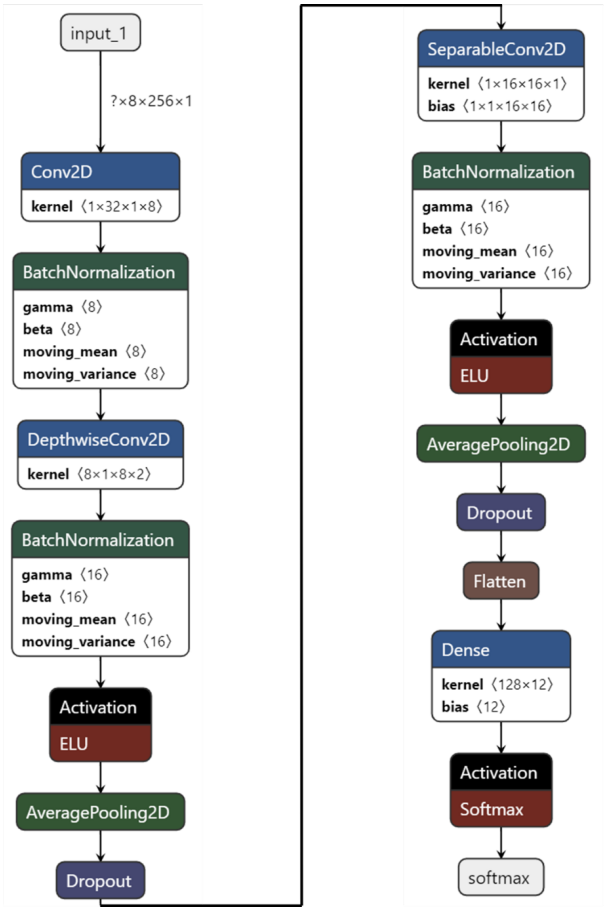

Figure A.15: Here is presented a visualization of the EEGNet [59] architecture generated via the Netron app ([312]).

A.6.4 EEGNetSSVEP

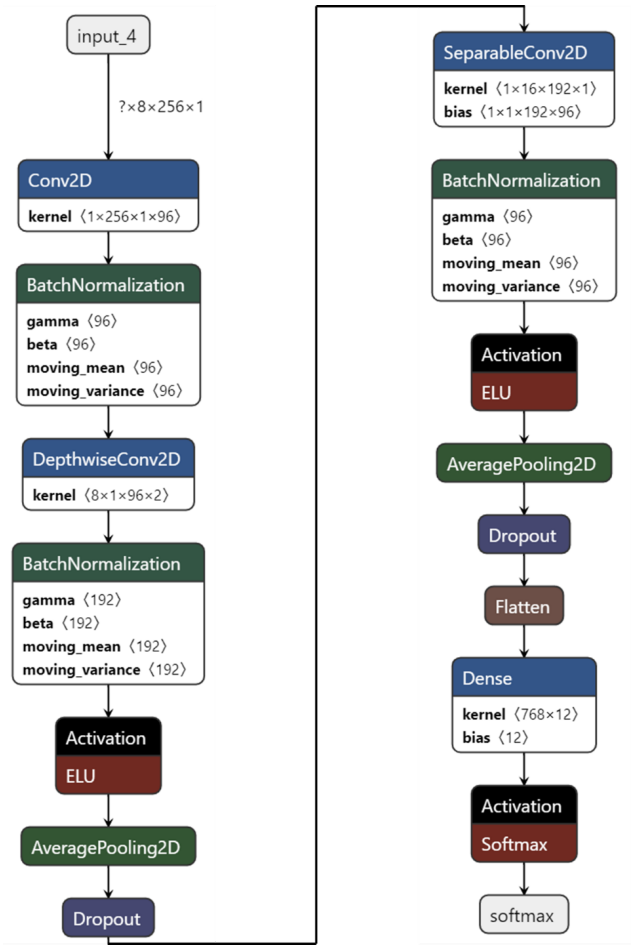

Figure A.16: Here is presented an architecture visualization of the EEGNetSSVEP [60] network generated via the Netron app ([312]).

A.7 Median Pruner Subject-Level Parameter Selection Plots

A.7.1 Subject 2

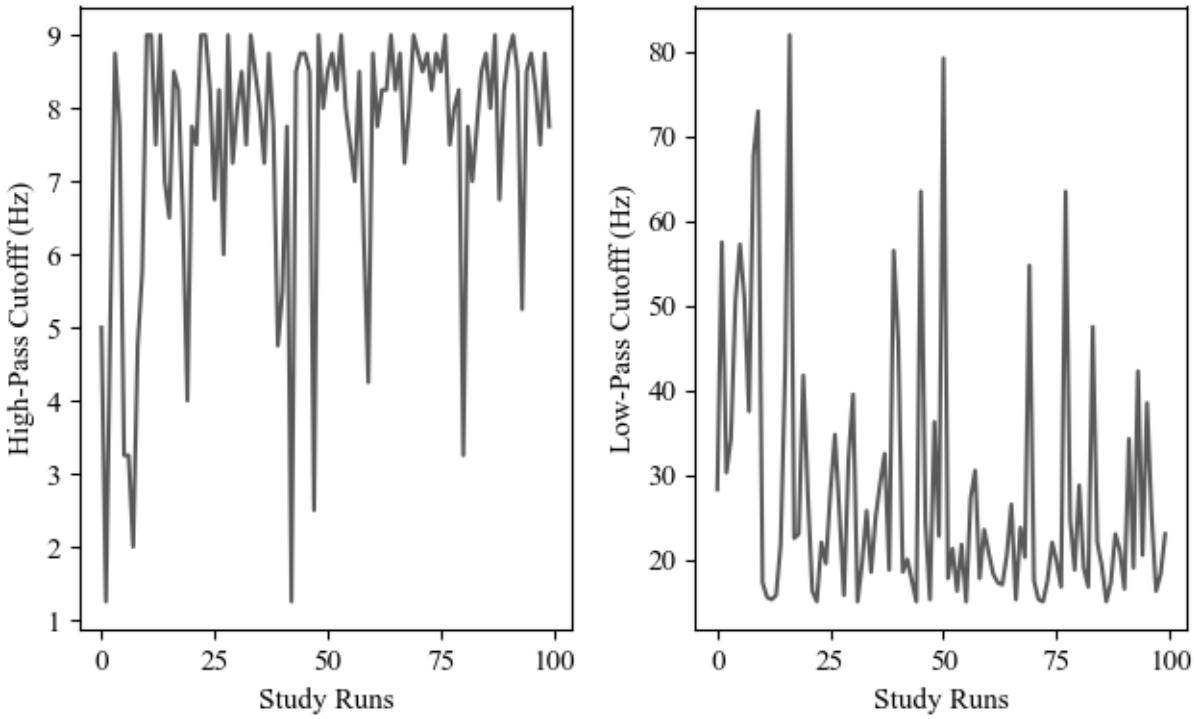

Figure A.17: Here is presented a figure displaying the Median Pruner high and low-pass parameter selections over the course of the EEGNet optimization study (100 trials) for Subject 2.

A.7.2 Subject 5

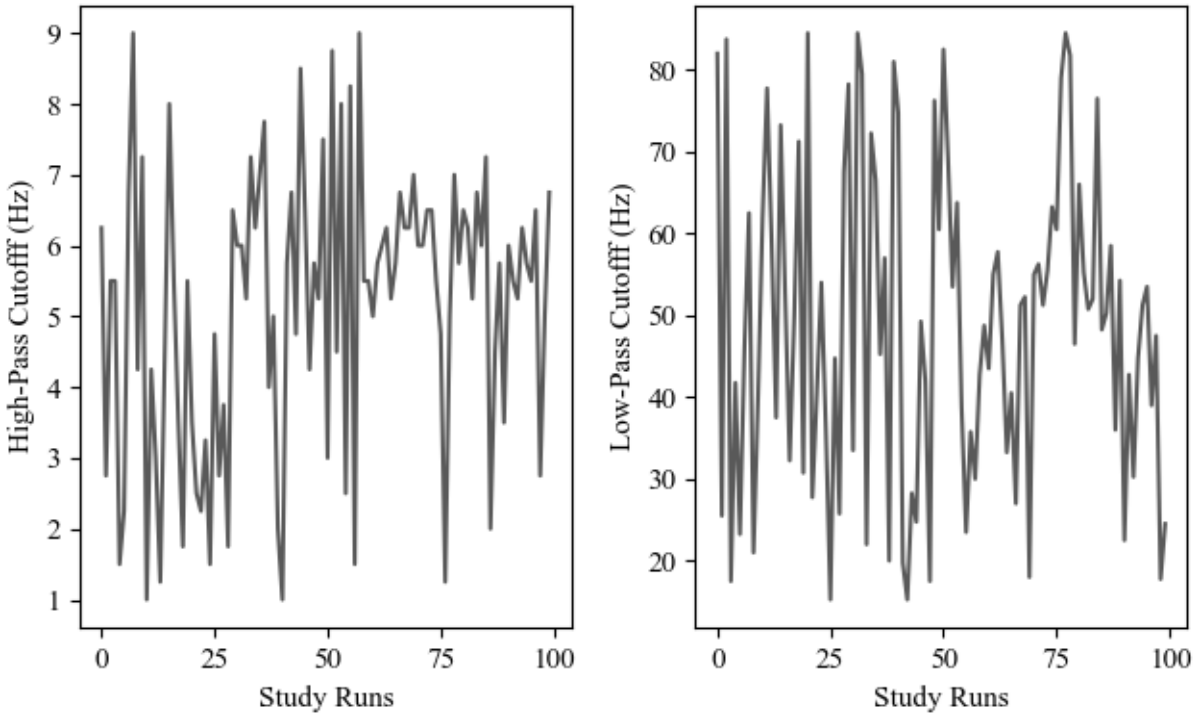

Figure A.18: Here is presented a figure displaying the Median Pruner high and low-pass parameter selections over the course of the EEGNet optimization study (100 trials) for Subject 5.

A.7.3 Subject 7

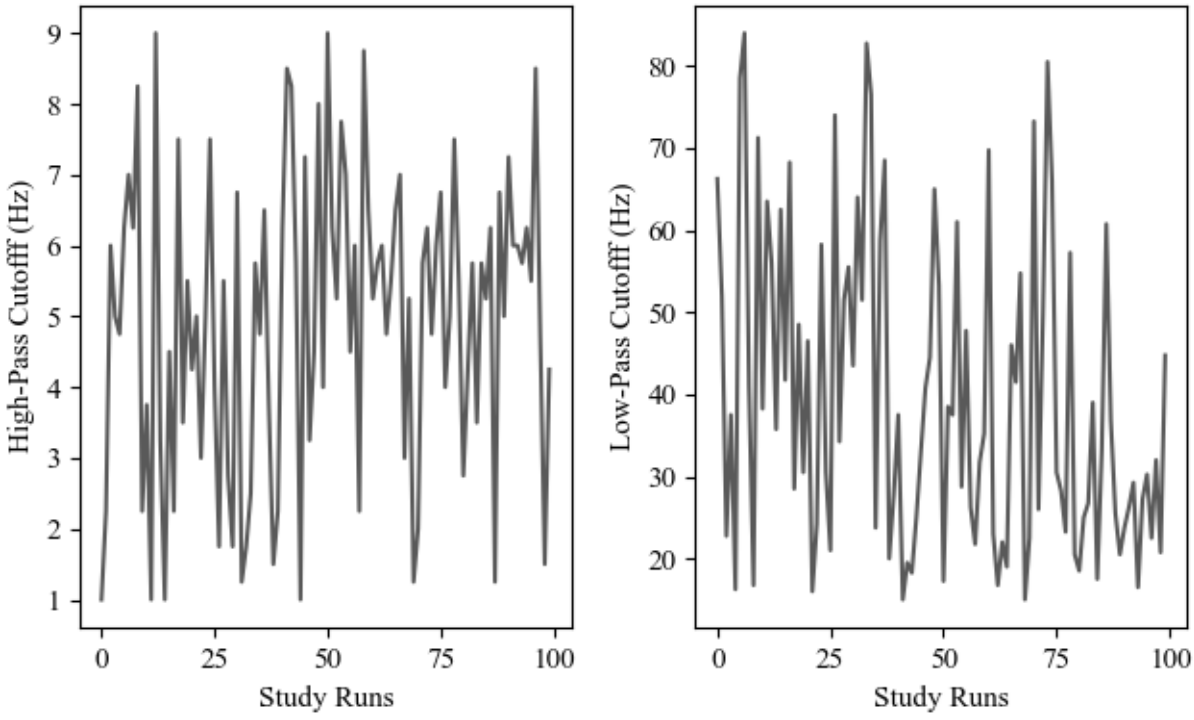

Figure A.19: Here is presented a figure displaying the Median Pruner high and low-pass parameter selections over the course of the EEGNet optimization study (100 trials) for Subject 7.

A.8 Optimizer Loss Profiles

A.8.1 Subject 8

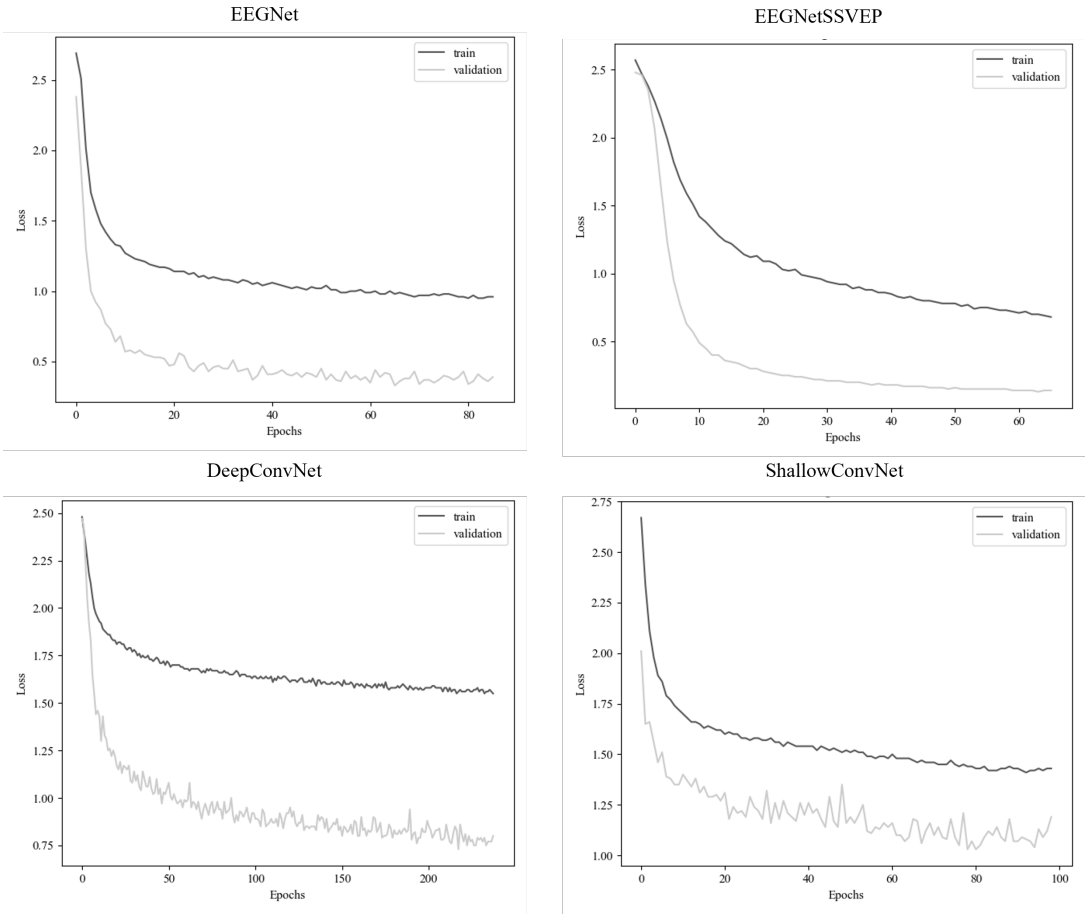

Figure A.20: Here is presented a figure displaying a representative subsample of the loss in network optimization trials for Subject 8 across the 4 models evaluated. These metrics are taken from the highest-performing networks in each respective model optimisation study. The loss values are positioned on the y-axis against the number of epochs on the x-axis. The black line presents the training data loss, and the grey shows the validation loss. As can be seen in the bottom right plot, the variance in the validation data loss values is the highest across all networks assessed. These patterns are highly uniform across subjects and models, with some variation in presentation for extremely low or high relative high-pass and low-pass filter cutoff parameter selections.

A.9 Pruner-Wise Optimized Classification Results

A.9.1 Percentile Pruner: EEGNet

	AoC	Min K	Max K	High-Pass	Low-Pass	Epochs
Sub 1	0.28	0.25	0.34	9	23	171
Sub 2	0.16	0.17	0.20	7.5	15.25	75
Sub 3	0.42	0.38	0.48	9	16.25	177
Sub 4	0.64	0.61	0.68	8.75	15	185
Sub 5	0.53	0.44	0.62	6.25	61.25	154
Sub 6	0.66	0.60	0.74	8	35.75	235
Sub 7	0.53	0.49	0.57	8.75	20.75	180
Sub 8	0.81	0.76	0.86	7.5	16.25	241
Sub 9	0.53	0.50	0.58	8.5	20	147
Sub 10	0.58	0.53	0.63	9	16	112
Mean	0.51	0.48	0.57	8.23	23.95	n/a

Table A.9: Here is shown a table of classification accuracies for subjects (Sub) relating to the optimization studies conducted utilizing the Percentile pruner algorithm for the EEGNet model. For further information on table interpretation see Table 5.6.

A.9.2 Percentile Pruner: EEGNetSSVEP

	AoC	Min K	Max K	High-Pass	Low-Pass	Epochs
Sub 1	0.55	0.53	0.56	8.5	73.25	52
Sub 2	0.31	0.230	0.34	4.75	17	47
Sub 3	0.72	0.70	0.74	8.5	50	62
Sub 4	0.92	0.91	0.93	3.25	43.25	109
Sub 5	0.87	0.86	0.86	8.75	83.25	35
Sub 6	0.92	0.91	0.93	5	54.5	88
Sub 7	0.84	0.83	0.85	9	50.75	78
Sub 8	0.97	0.96	0.98	8.75	68.25	78
Sub 9	0.86	0.85	0.87	8.5	62	141
Sub 10	0.82	0.81	0.82	6	67.75	47
Mean	0.78	0.77	0.79	7.1	57	n/a

Table A.10: Here is shown a table of classification accuracies for subjects (Sub) relating to the optimization studies conducted utilizing the Percentile pruner algorithm for the EEGNetSSVEP model. For further information on table interpretation see Table 5.6.

A.9.3 Percentile Pruner: DeepConvNet

	AoC	Min K	Max K	High-Pass	Low-Pass	Epochs
Sub 1	0.42	0.40	0.45	8.25	78	96
Sub 2	0.26	0.25	0.26	5.5	17.25	57
Sub 3	0.52	0.50	0.55	8.25	46.75	115
Sub 4	0.79	0.79	0.80	9	16.5	91
Sub 5	0.75	0.74	0.77	8.75	32.75	85
Sub 6	0.80	0.80	0.80	8.5	21.25	138
Sub 7	0.69	0.69	0.70	9	40	161
Sub 8	0.89	0.88	0.91	8.5	20	95
Sub 9	0.82	0.80	0.85	8.25	40	88
Sub 10	0.73	0.72	0.75	6.5	74.25	98
Mean	0.67	0.65	0.69	8.05	38.68	n/a

Table A.11: Here is shown a table of classification accuracies for subjects (Sub) relating to the optimization studies conducted utilizing the Percentile pruner algorithm for the DeepConvNet model. For further information on table interpretation see Table 5.6.

A.9.4 Percentile Pruner: ShallowConvNet

	AoC	Min K	Max K	High-Pass	Low-Pass	Epochs
Sub 1	0.25	0.23	0.30	7.75	15.25	129
Sub 2	0.16	0.15	0.17	8.5	72	40
Sub 3	0.32	0.28	0.41	8	15.75	111
Sub 4	0.51	0.47	0.58	9	15	116
Sub 5	0.40	0.35	0.44	8.25	19.25	103
Sub 6	0.58	0.56	0.62	8.75	15	114
Sub 7	0.46	0.42	0.50	9	15	107
Sub 8	0.62	0.59	0.68	8.75	15	165
Sub 9	0.43	0.38	0.51	9	18.5	99
Sub 10	0.510	0.49	0.55	9	16	108
Mean	0.42	0.39	0.48	8.6	21.68	n/a

Table A.12: Here is shown a table of classification accuracies for subjects (Sub) relating to the optimization studies conducted utilizing the Percentile pruner algorithm for the ShallowConvNet model. For further information on table interpretation see Table 5.6.

A.9.5 Successive Halving Pruner: EEGNet

	AoC	Min K	Max K	High-Pass	Low-Pass	Epochs
Sub 1	0.29	0.27	0.32	8.75	20.5	180
Sub 2	0.19	0.17	0.21	8.75	15	107
Sub 3	0.41	0.35	0.50	3.5	16	196
Sub 4	0.57	0.51	0.69	5.75	80.25	342
Sub 5	0.51	0.42	0.65	1.75	22.25	171
Sub 6	0.70	0.66	0.75	8.75	15	124
Sub 7	0.55	0.53	0.59	8.5	15	137
Sub 8	0.75	0.71	0.82	6.75	22.25	193
Sub 9	0.55	0.49	0.65	1.25	48.75	355
Sub 10	0.60	0.57	0.65	9	15.5	160
Mean	0.51	0.47	0.58	6.28	27.05	n/a

Table A.13: Here is shown a table of classification accuracies for subjects (Sub) relating to the optimization studies conducted utilizing the Successive Halving pruner algorithm for the EEGNet model. For further information on table interpretation see Table 5.6.

A.9.6 Successive Halving Pruner: EEGNetSSVEP

	AoC	Min K	Max K	High-Pass	Low-Pass	Epochs
Sub 1	0.55	0.55	0.55	7.75	74	16
Sub 2	0.30	0.29	0.32	3.5	18	39
Sub 3	0.74	0.72	0.46	7.74	83.5	76
Sub 4	0.91	0.90	0.94	5.25	64.75	110
Sub 5	0.83	0.81	0.85	7.5	70.5	42
Sub 6	0.92	0.91	0.93	6	54.5	85
Sub 7	0.83	0.82	0.84	5.25	63.25	73
Sub 8	0.97	0.97	0.97	8.5	47.5	105
Sub 9	0.86	0.54	0.88	7.5	53.25	114
Sub 10	0.82	0.81	0.82	5.75	32.5	42
Mean	0.77	0.73	0.76	6.47	56.18	n/a

Table A.14: Here is shown a table of classification accuracies for subjects (Sub) relating to the optimization studies conducted utilizing the Successive Halving pruner algorithm for the EEGNetSSVEP model. For further information on table interpretation see Table 5.6.

A.9.7 Successive Halving Pruner: DeepConvNet

	AoC	Min K	Max K	High-Pass	Low-Pass	Epochs
Sub 1	0.40	0.38	0.44	9	29.5	59
Sub 2	0.25	0.24	0.25	7.5	16.75	47
Sub 3	0.52	0.51	0.55	8.5	20	112
Sub 4	0.78	0.78	0.78	8.25	19	81
Sub 5	0.73	0.70	0.76	3.5	80.5	161
Sub 6	0.76	0.73	0.79	7.75	17.25	47
Sub 7	0.70	0.69	0.71	8.25	15.25	86
Sub 8	0.89	0.88	0.90	8.5	64.75	96
Sub 9	0.84	0.83	0.84	7.75	46.25	149
Sub 10	0.72	0.69	0.78	7.25	29.5	61
Mean	0.66	0.64	0.68	7.63	33.88	n/a

Table A.15: Here is shown a table of classification accuracies for subjects (Sub) relating to the optimization studies conducted utilizing the Successive Halving pruner algorithm for the DeepConvNet model. For further information on table interpretation see Table 5.6.

A.9.8 Successive Halving Pruner: ShallowConvNet

	AoC	Min K	Max K	High-Pass	Low-Pass	Epochs
Sub 1	0.24	0.21	0.29	8.75	16.75	70
Sub 2	0.15	0.15	0.16	8.5	75.5	43
Sub 3	0.29	0.24	0.37	7.5	16.5	99
Sub 4	0.48	0.45	0.54	9	16	79
Sub 5	0.39	0.37	0.42	8.5	21.75	77
Sub 6	0.57	0.54	0.61	8.75	15.25	97
Sub 7	0.46	0.44	0.48	8.5	15.5	121
Sub 8	0.58	0.55	0.65	9	17	84
Sub 9	0.42	0.38	0.48	9	17.5	88
Sub 10	0.50	0.48	0.53	9	15	46
Mean	0.41	0.38	0.45	8.65	22.68	n/a

Table A.16: Here is shown a table of classification accuracies for subjects (Sub) relating to the optimization studies conducted utilizing the Successive Halving pruner algorithm for the ShallowConvNet model. For further information on table interpretation see Table 5.6.

A.10 Subject-Wise Pruner Computation Durations

A.10.1 Median Pruner

Model	Sub 1	Sub 2	Sub 3	Sub 4	Sub 5	Sub 6	Sub 7	Sub 8	Sub 9	Sub 10
EEGNet	217.00	176.16	200.83	290.17	200.13	260.96	232.17	254.60	238.87	205.23
SSVEP	293.90	294.30	433.17	368.43	390.97	324.57	351.43	612.73	677.00	238.10
DCN	215.10	155.17	216.43	297.03	222.67	194.87	272.37	209.73	220.73	181.30
SCN	218.50	206.63	272.83	263.87	220.63	299.83	320.17	327.07	286.20	365.13
Range (mins)	78.80	139.13	232.34	104.56	190.84	129.70	119.26	403.00	456.27	183.83

Table A.17: Here is presented a table containing the duration of computational time required for the completion of all respective model optimization studies across each subject (Sub) assessed using the Median Pruner algorithm. All duration times are shown in minutes.

A.10.2 Percentile Pruner

Model	Sub 1	Sub 2	Sub 3	Sub 4	Sub 5	Sub 6	Sub 7	Sub 8	Sub 9	Sub 10
EEGNet	165.87	172.13	181.07	237.87	163.33	190.40	175.27	181.23	165.97	172.03
SSVEP	262.93	236.33	289.87	344.20	243.27	288.90	278.40	322.10	359.27	246.80
DCN	156.93	157.50	163.43	187.10	169.87	172.37	193.60	159.77	159.67	162.67
SCN	151.37	214.90	150.90	251.37	163.97	194.67	170.73	191.03	161.74	166.70
Range (mins)	111.56	78.83	138.97	157.10	79.94	116.53	107.67	162.33	199.60	84.13

Table A.18: Here is presented a table containing the duration of computational time required for the completion of all respective model optimization studies across each subject (Sub) assessed using the Percentile Pruner algorithm. All duration times are shown in minutes.

A.10.3 Successive Halving Pruner

Model	Sub 1	Sub 2	Sub 3	Sub 4	Sub 5	Sub 6	Sub 7	Sub 8	Sub 9	Sub 10
EEGNet	149.73	156.30	170.73	161.90	155.87	165.70	195.73	162.93	204.43	179.00
SSVEP	205.46	643.70	283.70	348.03	217.20	233.07	238.53	374.50	328.27	204.70
DCN	145.80	144.13	156.77	153.57	164.20	144.07	173.87	150.57	169.13	150.80
SCN	146.23	141.23	144.37	198.03	149.57	166.20	152.77	165.33	165.00	164.90
Range (mins)	59.66	502.47	139.33	194.46	67.63	89.00	85.76	223.93	163.27	53.90

Table A.19: Here is presented a table containing the duration of computational time required for the completion of all respective model optimization studies across each subject (Sub) assessed using the Successive Halving Pruner algorithm. All duration times are shown in minutes.