Measurement Criteria for Neurophysiological Monitoring and Neuroadaptive Interfaces

Felix Schroeder MSc.

A thesis submitted in partial fulfilment of the requirements of Liverpool John Moores University for the degree of Doctor of Philosophy

August 2025

Table of Contents

List of Tables IV

List of Figures V

List of Abbreviations VII

Abstract IX

Declaration X

Published Papers Included in this Thesis XI

Conference Posters and Presentations XI

Acknowledgements XII

1 Introduction 13

1.1 Background 16
1.1.1 What are mental states? 17
1.1.2 Ways to measure mental states 20
1.1.3 Definitions of Workload 25
1.1.4 Measuring Mental Workload 29
Self-report measures. 30
Performance measures. 31
Physiological measures. 32
1.1.5 Classifying workload 39

1.2 Aims and Objectives 41

1.3 Thesis Structure 42

2 General Methods 43

2.1 Participants 43

2.2 Task and Experimental Design 43
2.2.1 Multi-Attribute Test Battery 49
2.2.2 N-back 49

2.3 Sensors 50
2.3.1 EEG 51
2.3.2 fNIRS 52
2.3.3 ECG 53
2.3.4 Photodiode 53

2.4 Task-irrelevant probes 54
2.4.1 Auditory Probes 54
2.4.2 Visual Probe 54

2.5 Subjective Ratings 56

2.6 Preprocessing 56
2.6.1 ECG 56
2.6.2 EEG 58
2.6.3 fNIRS 60

2.7 Classifiers *61*
2.7.1 Shrinkage Linear Discriminant Analysis 61
2.7.2 Random Forest 62
2.7.3 Riemannian Geometry 63

3 Group-level Analysis of Mental Workload Metrics **65**

3.1 Subjective and Performance Metrics *65*
3.1.1 Subjective ratings 66
3.1.2 MATB Performance 69
3.1.3 N-back Performance 71
3.1.4 Interim Summary 73

3.2 ECG Metrics *74*
3.2.1 Heart-rate 74
3.2.2 Heart Rate Variability 77
3.2.3 Exploratory Rest/Task Analysis 80
3.2.4 Interim Summary 81

3.3 EEG Power Spectrum *82*
3.3.1 Delta effects 84
3.3.2 Theta effects 84
3.3.3 Alpha effects 85
3.3.4 Beta effects 86
3.3.5 Aperiodic effects 86
3.3.6 Time-on-task effects 88
3.3.7 Interim Summary 89

3.4 Task-irrelevant probes *91*
3.4.1 Auditory Probes 92
3.4.2 Interim Summary 96
3.4.3 Visual Probes 97
3.4.4 Interim Summary 106

3.5 Discussion *107*

4 Evaluation of Bias in Cross-Validation Methods for Passive BCIs **110**

4.1 Methods *114*
4.1.1 Dataset Descriptions 115
4.1.2 Dealing with the differences between datasets 116
4.1.3 Classification approaches 120
4.1.4 Cross-validation strategies 121
4.1.5 Statistical Analysis 123

4.2 Results *125*
4.2.1 Impact of Cross-Validation Choices Across Datasets 127
4.2.2 Pair-wise comparisons across datasets 127

4.3 Discussion *129*

5 Continuous Workload Monitoring **133**

5.1 Methods *135*
5.1.1 Feature Extraction 136
5.1.2 Hyperparameter tuning 137
5.1.3 Cross-subject classification 139
5.1.4 Temporal Aggregation 141

5.1.5 Statistics 141

5.2 Results *142*
5.2.1 Lab-grade vs Wearable 143
5.2.2 Temporal Aggregation 144

5.3 Discussion *150*

6 Multimodal Decision Fusion **152**

6.1 Methods *154*
6.1.1 Stacked Ensemble 155
6.1.2 FNIRs Classifiers 156
6.1.3 EEG Classifiers 156
6.1.4 Statistics 157

6.2 Results *157*
6.2.1 Meta-learner vs Single Classifiers 158
6.2.2 Aggregation results 164

6.3 Discussion *166*

7 General Discussion **170**

7.1 Main Findings *170*
7.1.1 Wearable vs Lab-grade EEG 170
7.1.2 Task-irrelevant probes 171
7.1.3 Multimodal Workload Monitoring 172
7.1.4 Reproducible pBCI Research 173

7.2 Neuroadaptive interfaces – a wider scope *174*
7.2.1 Consequences for the Workplace 175

7.3 Limitations *176*

7.4 Future Research *177*

7.5 Conclusion *178*

8 References **181**

List of Tables

Table 1. Overview of the Experiments 45
Table 2. Pilot MATB Settings 48
Table 3. Updated MATB Settings 48
Table 4. fNIRS Channels 53
Table 5. Channels Removed per Participant 59
Table 6. Pilot Data RSME Model 66
Table 7. Lab-grade Data RSME Model 67
Table 8. Wearable Data RSME Model 67
Table 9. Multimodal Data RSME Model 68
Table 10. MATB Performance Results 70
Table 11. N-back Performance Results 72
Table 12. Pilot Heart-Rate Results 75
Table 13. Lab-grade Heart-Rate Results 76
Table 14. Wearable Heart-Rate Results 77
Table 15. Multimodal Heart-Rate Results 77
Table 16. Pilot RMSSD Results 78
Table 17. Lab-grade RMSSD Results 79
Table 18. Wearable RMSSD Results 79
Table 19. Multimodal RMSSD Results 80
Table 20. Average Epochs per Participant (After Epoch Rejection) 93
Table 21. 15Hz Lab-grade (Model C) 105
Table 22. 30Hz Lab-grade (Model B) 105
Table 23. 15Hz Wearable (Model D) 105
Table 24. 30Hz Wearable (Model D) 106
Table 25. Channel Removal Across Datasets 117
Table 26. Average train/test Sample Sizes 123
Table 27. Average Accuracies and Cross-Validation Differences 126
Table 28. Classifiers with Best Accuracy/Responsiveness Trade-Off (Within-Subject) 149
Table 29. Classifiers with Best Accuracy/Responsiveness Trade-Off (Cross-Subject) 149
Table 30. MATB High Class-Separability First- and Second-level Comparisons 158
Table 31. MATB Low Class-Separability First- and Second-level Comparisons 160
Table 32. N-back High Class-Separability First- and Second-level Comparisons 162
Table 33. N-back Low Class-Separability First- and Second-level Comparisons 163
Table 34. Meta-Learner with Best Accuracy/Responsiveness Trade-Off (Within-Subject) 165
Table 35. Meta-Learner with Best Accuracy/Responsiveness Trade-Off (Cross-Subject) 165

List of Figures

Figure 1. Schematic of a Closed-Loop 16
Figure 2. Mental States in the pBCI Literature 18
Figure 3. Mental Workload Across Task Demands 26
Figure 4. The Multiple Resource Model 27
Figure 5. Neuroergonomics Model of Mental Workload 28
Figure 6. EEG's Canonical Band Power Ranges 36
Figure 7. Schematic of the Pilot Experiment 46
Figure 8. Updated Experiment Schema 46
Figure 9. MATB Instructions at a Glance 47
Figure 10. N-back Schematic 50
Figure 11. Sensor Montages 51
Figure 12. fNIRS Sensitivity Profile 52
Figure 13. SSVEP Participant Comfort Pilot 55
Figure 14. Rating Scale of Mental Effort 56
Figure 15. Discarded ECG Data Due to Noise 57
Figure 16. Full Preprocessing Pipeline 58
Figure 17. Schematic of an LDA's Projection Vector 61
Figure 18. Schematic of a Random Forest Model 63
Figure 19. Schematic of Riemannian Classification 63
Figure 20. RSME Results 69
Figure 21. MATB Tracking Performance 71
Figure 22. N-back Performance Results 73
Figure 23. Heart-Rate Results 75
Figure 24. RMSSD Results 78
Figure 25. Delta Power Effects 84
Figure 26. Theta Power Effects 85
Figure 27. Low Alpha Power Effects 86
Figure 28. High Alpha Power Effects 87
Figure 29. Beta Power Effects 87
Figure 30. Aperiodic Slope Effects 88
Figure 31. PSD Time-on-Task Effects 89
Figure 32. PSD Effect Size Summary 90
Figure 33. Pilot Auditory ERP Results 94
Figure 34. Lab-grade Auditory ERP Results 95
Figure 35. Wearable Auditory ERP Results 96
Figure 36. Average RESS Components 101
Figure 37. Sensor-Space SSVEP Results 103
Figure 38. N-Back RESS SNRs 104
Figure 39. Schematic of Multivariate Temporal Dependencies 111
Figure 40. Combined Preprocessing for All Datasets 117
Figure 41. Block Structures and Splitting Procedures 119
Figure 42. Visualisation of Cross-Validation Strategies 123
Figure 43. Classification Accuracy Results 125
Figure 44. Classifier Comparisons 128
Figure 45. Within-Subject Classification Pipeline 135
Figure 46. Hyperparamter Overview 138

Figure 47.Cross-Subject Classification Pipeline 139
Figure 48. Within-Subject Montage Comparisons 143
Figure 49. Cross-Subject Montage Comparisons 144
Figure 50. Within-Subject Extraction Window and Aggregation Results 146
Figure 51. Cross-Subject Extraction Window and Aggregation Results 147
Figure 52. Stacked Ensemble Schematic 155
Figure 53. Multimodal MATB Classification Results (High Class-Separability) 158
Figure 54. Multimodal MATB Classification Results (Low Class-Separability) 159
Figure 55. Multimodal MATB Meta-Learner Weights 161
Figure 56. Multimodal N-back Classification Results (High Class-Separability) 161
Figure 57. Multimodal N-back Classification Results (Low Class-Separability) 163
Figure 58. Multimodal N-back Meta-Learner Weights 164
Figure 59. Meta-Learner Aggregation Results (MATB) 166
Figure 60. Meta-Leaner Aggregation Results (N-back) 166
Figure 61. Per-Subject Variability 167
Figure 62. Average Classifier Rankings 169

List of Abbreviations

ASR	Artifact Subspace Reconstruction
BCI	Brain-Computer Interface
BH	Benjamini-Hochberg
BOLD	Blood-oxygenation-level-dependent
dB	Decibel
DoF	Degrees of Freedom
ECG	Electrocardiogram
ECoG	Electrocorticography
EEG	Electroencephalogram
EMG	Electromyography
FIR	Finite Impulse Response
fMRI	Functional Magnetic Resonance Imaging
fNIRS	Functional Near Infrared Spectroscopy
FOOOF	Fitting Oscillations & one over F
FWHM	Full Width Half Maximum
HbO	Oxyhaemoglobin
HbR	Deoxyhaemoglobin
HR	Heart Rate
HRV	Heart Rate Variability
Hz	Hertz
ICA	Independent Component Analysis
KDE	Kernel Density Estimate
LDA	Linear Discriminant Analysis
Lsl	Labstreaminglayer
MATB	Multi-Attribute Test Battery
MEG	Magnetoencephalography
MRT	Multiple Resource Theory
NASA-TLX	NASA Task-load Index
OPM	Optically Pumped Magnetometers
pBCI	Passive Brain Computer Interface

PCA	Principal Component Analysis
PET	Positron Emission Tomography
PSD	Power Spectral Density
RANSAC	Random Sample Consensus
RESS	Rhythmic Entrainment Sources Separation
RF	Random Forest
RMDM	Riemann Minium Distance to Mean
RMSSD	Root Mean Square of Successive Differences
RSME	Rating Scale of Mental Effort
sLDA	Shrinkage Linear Discriminant Analysis
SNR	Signal to Noise Ratio
SPD	Symmetric Positive Definite
SSC	Short-Separation Channels
SSVEP	Steady State Visual Evoked Potential
STE	Sensor & Trigger Extension
SWAT	Subjective Workload Assessment Technique

Abstract

Human Factors and Psychology have long been interested in the neurophysiological basis of mental states such as mental workload or fatigue due to their relevance to safety and well-being at work. With the recent emergence of Neuroergonomics, research on the human mind in applied settings is becoming more commonplace in an attempt to improve ecological validity and to address the limitations of artificial laboratory environments. This thesis dealt with the applied use of neurophysiological metrics for continuous mental state monitoring, a rapidly growing area thanks to affordable consumer sensing technologies such as smartwatches and eye-trackers. However, neuroimaging techniques like Electroencephalography (EEG) and functional near-infrared spectroscopy (fNIRS) remain largely confined to research contexts. A key question this thesis aimed to answer was the viability of wearable neuroimaging sensors in comparison to lab-grade devices for mental state monitoring. Secondary aims concerned the exploration of novel task-irrelevant probing techniques for mental workload monitoring and ensemble machine learning strategies to overcome the challenge of cross-subject generalisability in mental state monitoring. To achieve this, a two-day experimental paradigm was designed, collecting multimodal data from 80 participants across four datasets using various sensor configurations across three levels of task load in two different tasks. The first was an artificial working memory task called the n-back, and the second was a more ecologically valid multi-tasking paradigm called the Multi-Attribute-Test-Battery. The results revealed task-dependent variations in traditional neurophysiological workload metrics and highlighted the impact of aperiodic contributions to canonical EEG band power metrics, demonstrating varying effects on different frequency ranges, which could clarify some conflicting findings in the existing literature. Additionally, the thesis identified and estimated biases in passive Brain-Computer Interface evaluation methods, which complicate reproducibility in the field. Most importantly, it provided evidence that 64-channel lab-grade EEG and sparse seven-channel sponge-based EEG can achieve comparable mental workload classification performance, along with strategies to address accuracy differences when they occur. However, across the four datasets collected, subtle differences in mental workload - such as between a 0-back and a 1-back task, or the low-effort operation of the MATB compared to medium effort scenarios - proved difficult to distinguish with state-of-the-art methods, with accuracy levels around 60% rather than the 80% - 90% for more extreme workload differences, regardless of sensing hardware. Overall, the results offered evidence for the potential of lightweight, user-friendly EEG devices for neuroadaptive technologies and provided several future research directions towards robust cross-subject mental workload monitoring.

Declaration

I declare that no portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

In addition, I acknowledge the use of generative AI for stylistic suggestions and as a technical reference for debugging. The use of these tools did not extend to the generation of original text or code, which was authored solely by me.

This research was funded by the Defence Science and Technology Laboratory (DSTL) via DSTLX1000156119.

Published Papers Included in this Thesis

Schroeder, F., Fairclough, S., Dehais, F., & Richins, M. (2025). The Impact of Cross-Validation Choices on pBCI Classification Metrics: Lessons for Transparent Reporting. Frontiers in Neuroergonomics, 6. https://doi.org/10.3389/fnrgo.2025.1582724

Conference Posters and Presentations

Schroeder, F., Fairclough, S., Dehais, F., & Richins, M. (2024, July 8-12). *Rapid auditory probes for EEG workload monitoring* [Talk]. 5th Neuroergonomics Conference, Bordeaux, France.

Schroeder, F., Fairclough, S., Dehais, F., & Richins, M. (2023, November 7). *Use of Implicit Stimuli for Neurophysiological Assessment of Mental Workload* [Poster presentation]. Human Augmentation conference, Salisbury, UK.

Schroeder, F., Fairclough, S., Dehais, F., & Richins, M. (2023, July 3-14). *Exploring the Potential of Steady-State Visually Evoked Potentials for Mental Workload Estimation Across Two Task Paradigms* [Poster presentation]. International Summer School on Harmonic and Multifractal Analyses in Neuroscience. Montreal, Canada.

Schroeder, F., Fairclough, S., Dehais, F., & Richins, M. (2023, June 27). *Examining the Feasibility of Rapid Task-Irrelevant Auditory Probes for Mental Workload Monitoring: Insights from EEG-based Classification* [Talk]. Liverpool EEG Day, Liverpool, UK.

Schroeder, F., Fairclough, S., Dehais, F., & Richins, M. (2022, April 20-22*). Dynamics of ECG and EEG-Based Workload Measures Across Tasks* [Talk]. Human Factors and Ergonomics Society Europe Chapter Annual Meeting, Liverpool, UK.

Acknowledgements

Over the past four years, I have had the incredible opportunity to immerse myself in the research questions that sparked my curiosity during my undergraduate and master's studies. For the nearly boundless freedom to pursue these interests, I owe my primary thanks to my supervisors: Stephen Fairclough, Frederic Dehais, and Matt Richins.

Stephen allowed me to enter the field of Neuroergonomics on my own terms while ensuring I remained grounded in the foundational work of the past decades. Having a supervisor I could rely on for deep dives into obscure rabbit holes, and who could pull me back when I veered too far off the deep end, created a great balance between fundamental science and state-of-the-art experimentation. I am equally grateful to Fred for welcoming me so warmly at ISAE-SUPAERO. Getting to work alongside a such a talented and like-minded group was truly inspiring. Finally, Matt made the administrative side of this project an absolute breeze. His support throughout these tumultuous times was always on point, particularly his guidance in disseminating our results at various conferences. I also wish to thank the DSTL for their funding and their trust in our research.

Beyond my supervisors, I was fortunate to work with an exceptional technical team at the tail end of a global pandemic. Getting in-person experimentation up and running so quickly, especially with such a complex paradigm, was largely thanks to Russell and Adarsh. I can only apologize profusely for the endless beeping noises I emanated through the lab hallways for years on end. This apology extends to all the other PGRs in our office. I also want to thank you all—and Shaunna especially—for making Liverpool feel like a true home.

I am also super grateful to my examiners, Fabien Lotte and Nika Adamian, for a viva experience that, while challenging, turned out to be a really insightful and fun discussion.

Finally, while often a rather solitary experience, the final stretch of this journey was made possible by the unwavering support of Joshua and Kim. Your compassion and genuine concern when things got tough meant the world to me. The same goes for family. To my brother, Jari, who has been a lifelong inspiration, and to my parents, Frank and Heike: thank you for never doubting my path. I am incredibly lucky to have you behind me and I simply would not have reached this point without you.

1 Introduction

Modern workplaces are growing ever more complex. Monitoring multiple sources of information while keeping track of environmental conditions and possible extraneous events in maritime (Hanzu-Pazara et al., 2008; Tzannatos, 2010), aviation (Pape et al., 2001), and other heavy-machinery contexts (Chenarboo et al., 2022) has long been known to strain the attentive capacities of even the most senior operators. Kept unchecked, high mental workload, fatigue, or absent-mindedness can lead to human error (Chenarboo et al., 2022; Mehta & Parasuraman, 2013; Pape et al., 2001; Tzannatos, 2010), causing financial and, in the worst case, bodily harm. Human error is not confined to heavy-machinery, and prolonged periods of high mental or physical workload may also drive accidents in the operating room (Wallston et al., 2014; Weigl et al., 2016), as well as errors at desk-based jobs such as air traffic control (Pape et al., 2001) and other, perhaps more mundane, occupations (Balfe et al., 2017; Brookhuis et al., 2003; Di Stasi et al., 2011; Gao et al., 2013). While human error should not be used to explain away systemic sources of accidents (Reason, 1997), its contribution to workplace accidents is still significant (Bliss, 2003; Jorna, 1991; Nakamura et al., 2004). Hence, research into mitigating human error has garnered considerable interest in the past century in the forms of human factors, ergonomics, or industrial-organisational psychology (Lee et al., 2024). A rather recent addition to these efforts, the field of neuroergonomics, seeks to utilise advances in portable neuroimaging technologies as well as insights from cognitive neuroscience to better understand the human mind at the workplace and in everyday settings (Gramann et al., 2017; Parasuraman, 2003).

Textbox 1. Fictional scenario

Andy is an experienced air traffic controller at Frankfurt Airport. Recent staff shortages have resulted in bottlenecks in the coverage of ongoing traffic, and during peak hours, the more senior staff are expected to pick up the slack. Prior to their adoption of novel neuroadaptive headsets, days like these would leave Andy exhausted and overworked. The new headset passively monitors Andy's cognitive states. As incoming traffic grew heavy, the system silently noted his increasing stress. Recognising that he was approaching his limit, it automatically reassigned some less critical tasks to a colleague. Andy barely noticed a slight ease in workload, but it was enough to lower his stress levels and maintain optimal performance during peak hours.

Stories like the scenario in Textbox 1 may soon no longer be science fiction. Neuroadaptive technologies are context-aware computing devices that aim to incorporate implicit user-state information into human-machine interactions by monitoring neurophysiological signals (Fairclough, 2022; Hettinger et al., 2003; Zander et al., 2016). Similar ideas have been studied for the better part of the last 3 decades under various names, like biocybernetic loop (Pope et al., 1995) or physiological computing (Fairclough, 2009). The disparate efforts are united by their use of brain-computer interface methods to passively monitor various neurophysiological processes and translate them into control signals.

Brain-computer interfaces (BCIs) refer to systems that translate brain activity into control signals, an idea that was first formalised in 1973 (Vidal, 1973). Since then, scientific debates regarding technical challenges concerning sensor design or signal processing techniques (Dehais et al., 2020; Fairclough & Lotte, 2020; Lotte et al., 2018), as well as societal challenges (Lebedev & Nicolelis, 2006; Martinez et al., 2022; Muhl, 2024; Vaadia & Birbaumer, 2009), have continued in earnest. BCIs may be grouped into invasive and non-invasive technologies, i.e., those requiring invasive surgery and those utilising sensors on the scalp's surface. Furthermore, BCIs can be grouped into active, reactive, and passive systems (Zander & Kothe, 2011).

- Active BCI. Active BCIs involve a conscious effort to evoke control signals, as is the case for applications in which brain activity is utilised to innervate an artificial limb (Jia et al., 2023; Saha & Baumert, 2020).
- Reactive BCI. Reactive BCIs derive control signals from involuntary brain responses to system-relevant events (external stimulation). Prominent examples of such systems are input/spelling systems that operate on brain activations in response to sensory stimulation utilising steady state visually evoked potentials (Middendorf et al., 2000) or the event related p300 component (Farwell & Donchin, 1988).
- Passive BCI. Passive BCIs extract information on operator state without voluntary control, with or without additional external stimulation. Such systems may be aimed at continuously monitoring the level of vigilance or mental workload of the operator (Aricò et al., 2018).

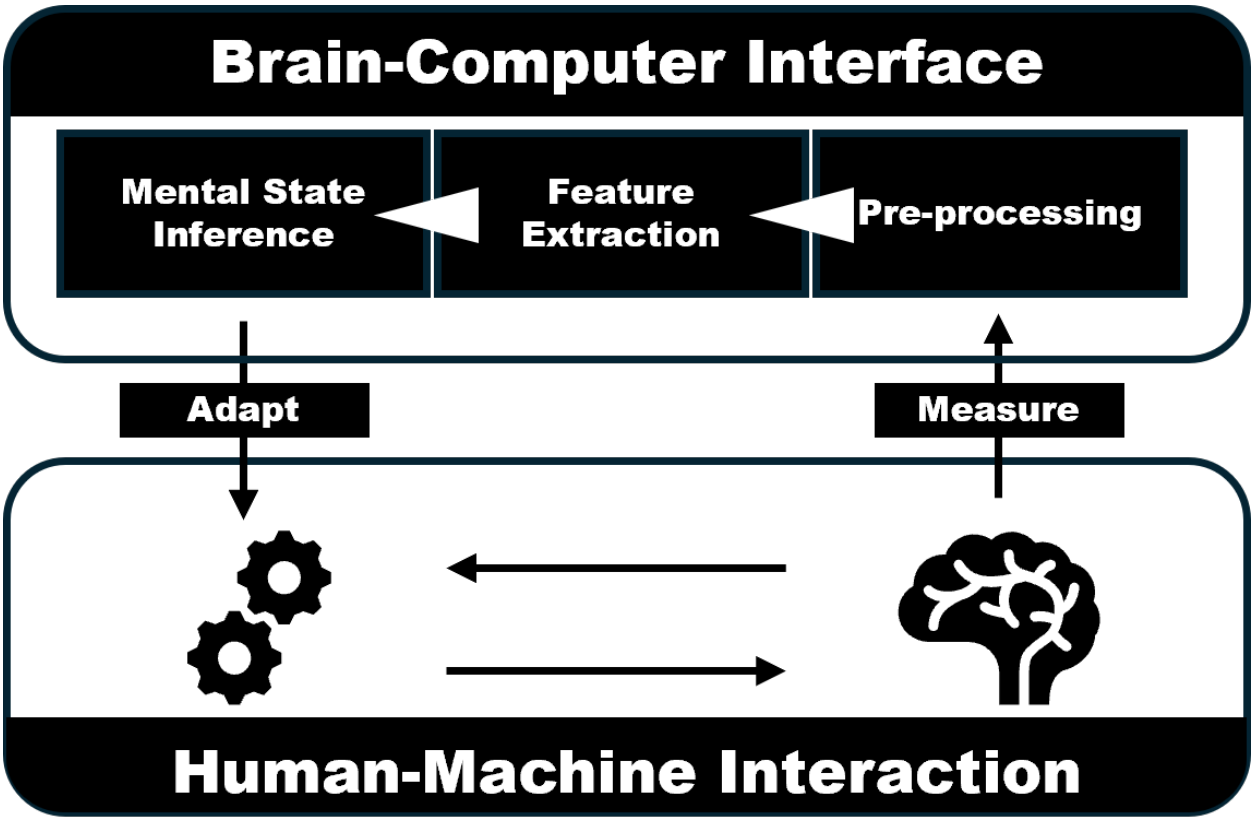
Much of the literature on BCIs concerns active and reactive BCIs in patient populations (Gu et al., 2021; Lotte et al., 2018; Vaadia & Birbaumer, 2009). In the case of active BCIs, the user attempts to actively produce neurological signals, like imagined movements or speech, which can be translated into control signals to operate wheelchairs, prosthetics, communication devices or other software (Thomas et al., 2013). Reactive BCIs have also been developed as a means of communicating with computers without

the use of traditional keyboard and mouse controls (İşcan & Nikulin, 2018; Kalunga et al., 2016). Here, externally induced changes to neurological signals, such as marking specific buttons with different flickering stimuli, can be used to distinguish relevant brain activation from the ongoing background activity (Cabrera et al., 2023a; Dehais et al., 2024).

In the case of neuroadaptive interfaces, there is instead a focus on supporting healthy individuals in working environments with control signals stemming from mostly passive BCI systems (pBCI). Research-grade neuroadaptive technologies have been applied to training (Zammouri et al., 2024), air-traffic control (Aricò et al., 2016), pain management (Fairclough et al., 2023), and various other contexts (Dehais et al., 2022; Gerjets et al., 2014; Karran et al., 2019). By estimating levels of fatigue or mental workload, these applications either gauge the level of understanding in the case of training, the need for added automation in the control of air-traffic control systems, or the level of game demand to optimally distract from pain.

Designing a neuroadaptive interface requires multidisciplinary expertise (Brouwer et al., 2015). The design process includes decision about the goal of the system (i.e. how does it support the operator effectively), the brain processes of interest (i.e. what are we trying to measure), the appropriate sensor modality and form factor (i.e. how can we measure the process of interest), reliable signal processing techniques (i.e. how to assure the signal can be recovered), and accurate classification strategies (i.e. how to produce reliable output). By combining these individual components, a neuroadaptive interface integrates implicit information about the operator into a system's logic, thereby creating an open or closed-loop (Figure 1). In an open-loop system, the inference about the operator state is not expected to affect the state itself, such as merely monitoring mental states without providing feedback (Pan et al., 2022). In a closed-loop system, the operator may receive feedback or the system can be adapted to accommodate changes in user state, affecting the user state in turn (Fairclough, 2009; Hettinger et al., 2003; Krol & Zander, 2017; Zander et al., 2016).

Figure 1. Schematic of a Closed-Loop



Note. Schematic detailing the individual components necessary for a neuroadaptive system.

This thesis aims to contribute to the development of neuroadaptive technologies by testing and validating traditional and state-of-the-art mental state detection techniques using three different mobile neuroimaging montages.

1.1 Background

The following sections of this chapter explore the individual decision-making processes involved in developing neuroadaptive technology. We will clarify the concept of mental states, using mental workload as a specific example, review neurophysiological methods for measuring them, and finally outline the most common metrics used to measure and ultimately classify them. A thorough discussion of the signal processing pipelines will subsequently be presented in the General Methods section and in the individual studies.

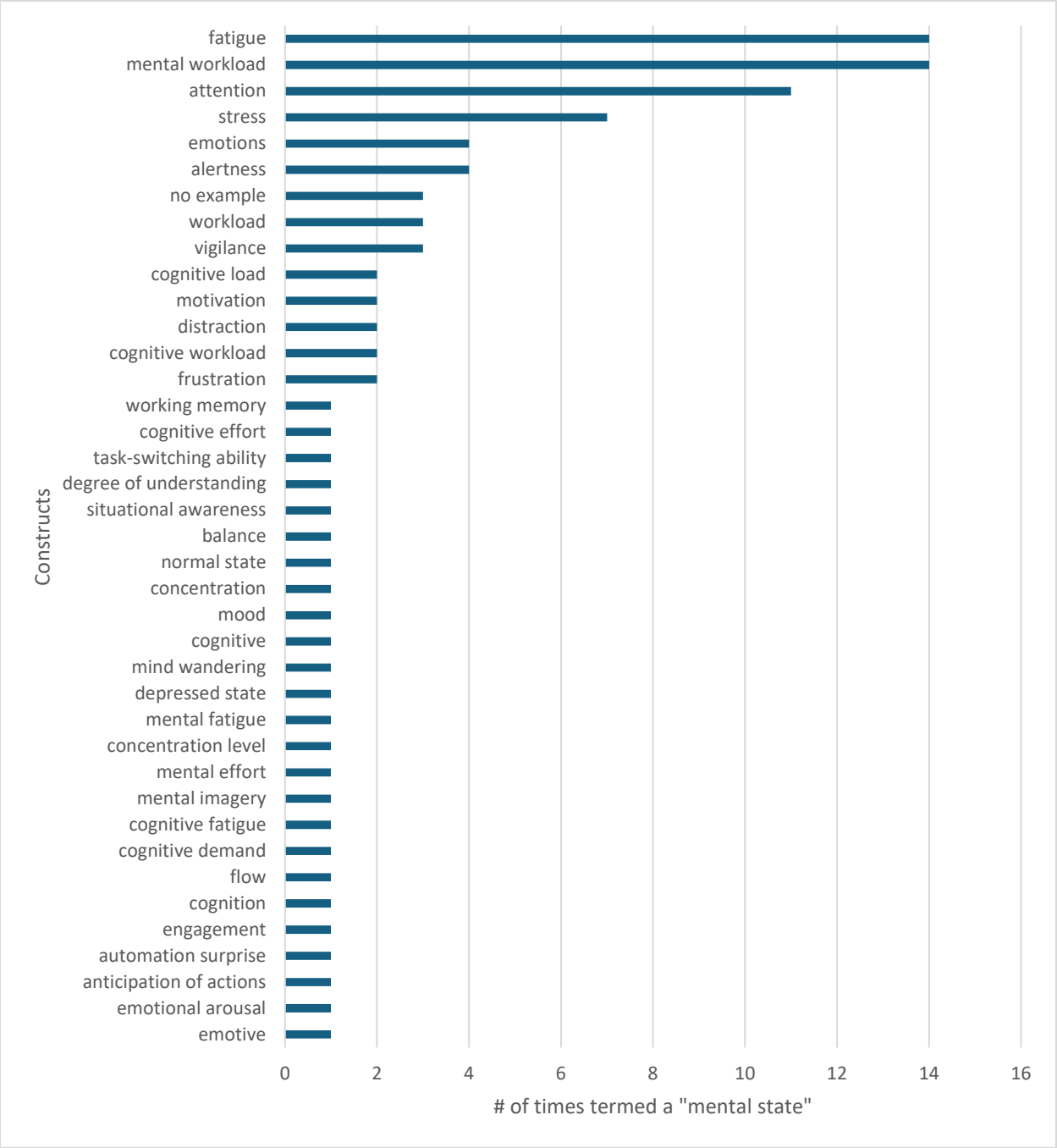
1.1.1 What are mental states?

Before we can start to measure or classify different mental states, we need to clearly delineate what we mean when we speak of mental states. Firstly, a state refers to a mode or condition of being. According to the National Institute of Mental Health, mental states refer *"to intentions, beliefs, desires, and emotions.".* This definition emphasises introspective constructs that can be subjectively reported. A much wider definition was put forward by Salzman and Fusi (2010), who understand mental states "*as a disposition to action—i.e., every aspect of an organism's inner state that could contribute to its behavior or other responses—which may comprise all the thoughts, feelings, beliefs, intentions, active memories, and perceptions, etc., that are present at a given moment.".* Responses in their framework may be of cognitive (e.g., decision making, memory), behavioural (e.g., Approach or avoidance, reaction times), or physiological (e.g., heart rate, skin conductance) nature.

To understand how researchers in pBCI technology conceptualise mental states, literature from the past 5 years (2020 – 2025) from the Web of Science database, which used the term passive Brain Computer Interface or its acronym, was reviewed. Of the 65 articles identified, only those 42 explicitly mentioning 'mental state' or 'cognitive state' were further analysed to identify common constructs associated with mental states in the field. Figure 2 summarises the frequency of these constructs.

Due to idiosyncrasies in how constructs were named, there were a number of constructs with only a single occurrence. Summarising similar constructs together we find that **mental workload** (also including cognitive demand, mental effort, cognitive effort, cognitive workload, cognitive load, workload), **fatigue** (also including cognitive fatigue, mental fatigue), **emotion** (also including emotive, emotional arousal, depressed state, mood, frustration) and **attention** (also including engagement, concentration level, mind wandering, concentration, situational awareness, distraction, vigilance, alertness) were the dominant constructs of interest to pBCI researchers when they spoke of mental states in the past five years.

Figure 2. Mental States in the pBCI Literature



Note. Counts of individual constructs named in relation to the terms “mental state” or “cognitive state” in pBCI studies of the past 5 years.

From this, it appears that mental states tend to have a transient nature with global effects on cognition and behaviour. They are transient because the constructs described are context-dependent and not trait-like. They affect cognition globally because they describe no single behavioural outcome but rather propensities for positive and negative effects for various behaviours. The latter point on global effects

matches Salzman & Fuzi's (2010) description of mental states being "dispositions to actions", while the transient nature was included implicitly by focusing their definition on "the present moment". However, by incorporating beliefs, thoughts, and memories—elements absent from our survey of the pBCI literature—Salzman and Fuzi's definition may have a scope that is significantly broader than needed for discussions centred on neuroadaptive technologies.

For the purpose of developing pBCIs, a mental state's relationship to neurophysiology is of utmost interest, as a pBCI needs to infer the mental state from neurophysiological variables. To this end, it helps to operationalise mental state in relation to other psychometrically relevant concepts with conscious experience on one end (mental states) and (neuro) physiological processes on the other. Mental states (including affective constructs) as used in the pBCI literature represent subjectively reportable constructs (e.g., "How much effort did this task take?", "How tired do you feel?", "How angry do you feel?", "Did you experience mind wandering?"). Neurophysiological processes, on the other hand, refer to changes in sympathetic and parasympathetic functions, which may not always be reportable. The concept of brain states attempts to capture these large-scale dynamics that recur, continuously evolve, and are behaviourally relevant (Marzetti et al., 2024). Others define brain states as neurophysiological processes that, through internal feedback-loops, give rise to or emerge from mental states (Greene et al., 2023). Their behavioural impact is ultimately the reason why human-machine interaction may be improved through the integration of pBCIs. On the other hand, their complex and dynamical nature may be the reason for some of pBCIs grand challenges (Fairclough & Lotte, 2020).

Importantly, there is likely no one-to-one mapping or simple causal chain between either states or intermediate metrics. They are deeply interlinked with each other, so that physiological changes may affect mental states (as is case with generalized anxiety disorder, where shallow breathing may come before feeling anxious), mental states may affect cognition (as is the case with motivated cognition where performance in visual search rises with increased internal or external motivators), and cognition is deeply interlinked with physiological processes according to embodied cognition (Critchley & Garfinkel, 2018; Harris et al., 2015; Pramme et al., 2016). The exact nature of possible feedback loops remains undetermined, but together they may offer rich descriptions of the antecedents to human-machine interaction relevant behaviours.

Current trends in neuroscience acknowledge brain states to be in constant non-random flux (McCormick et al., 2020) best described in terms of complex systems (Bassett & Gazzaniga, 2011; Mainzer, 2007; Scharfen & Memmert, 2024), leading to an intractable high number of variables necessary to paint a

complete picture of brain states (Jonas & Kording, 2017; Mehler & Kording, 2018). Considering the brain's complexity, it is likely that no single neuroimaging modality would be able to capture the entire spatio-temporal dynamic that underlie mental state related changes to brain states. However, as the next section will explore, not all neuroimaging modalities suit pBCI applications as their hardware may require extensive cooling, their slow temporal resolution may average over important events, or their low spatial resolution may blur disparate dynamics of neighbouring regions of interest.

The next section provides an overview of the most popular neuroimaging techniques used for human experimentation and how they may fit into pBCI applications.

1.1.2 Ways to measure mental states

Information regarding ongoing mental processes is not easily accessible. Various neuroimaging modalities exist, each with its advantages and disadvantages when applied in real-world settings. Generally, the choice of modality is influenced by trade-offs concerning

A. Usability from the operator's perspective
    a. How invasive is the equipment?
    b. Is the apparatus heavy and stationary or lightweight and mobile
B. Temporal resolution
    a. How quickly must the system respond?
    b. Over what temporal scales do the processes of interest occur?
C. Spatial resolution
    a. Is the process of interest measurable in very specific regions of the outer cortex, or does monitoring require a distributed array of sensors or information from deeper brain structures not readily accessible at the outer cortex?

From this outset, methods with the highest temporal and spatial resolution appear best suited for the development of passive BCIs. Possible candidates with the highest spatial resolution include microelectrode arrays that can be implanted directly into the cortex or Electrocorticography (ECoG) devices placed on the arachnoidal surface outside the cortex, beneath the skull. While these options undeniably provide superior spatial resolution, their use is highly invasive. Both ECoG and microelectrode arrays necessitate surgical intervention, typically reserved for clinical populations. Their implantation carries a risk of damaging the surrounding tissue (Edell et al., 1992; Wang et al., 2023) and may lead to biological rejection (Miller et al., 2020). Furthermore, long-term signal quality degradation

may occur, although reports indicate that implants could produce high-quality data for up to ten years (Sponheim et al., 2021). Together, these risks will likely mean that invasive brain imaging sensors will remain a special case for clinical populations for the foreseeable future.

Electroencephalography (EEG) and Magnetoencephalography (MEG) both offer the temporal resolution of the previous invasive sensors, without the need for surgical intervention. Both can pick up on post-synaptic discharge of synchronised neural populations in the cortex and sub-cortical structures (Beniczky & Schomer, 2020; Piastra et al., 2020; Seeber et al., 2019). However, EEG suffers from poor spatial resolution because its sensors measure electric potentials on the scalp's surface. The loss of spatial resolution occurs due to volume conduction, which describes the mixing and diffusion of electric potentials while they travel from their source through several substrates like brain tissue and the skull (Luck, 2005; Nunez & Srinivasan, 2006). MEG avoids this issue by measuring the corresponding magnetic fields of the electric activity, which are not susceptible to the same physical interferences (Hämäläinen, 1992). Due to MEG's costly cooling systems that require their use to be fully stationary (including a fully stationary head) their usage in Neuroergonomics research is limited. Recently, optically pumped magnetometers (OPMs) have emerged as a less stationary MEG alternative. Their use still requires perfectly electronically shielded environments (Brickwedde et al., 2024), and their nascent nature likely requires many years of fundamental research before applied work could harness their benefits. Hence, EEG represents the most popular modality for passive BCI studies (Chevallier et al., 2024a; Fairclough & Lotte, 2020), as its usage benefits from a rich literature (Mushtaq et al., 2024), with devices also becoming increasingly portable over the past decade (Niso et al., 2023).

EEG offers high temporal resolution, allowing for the capture of rapid changes in mental states on a millisecond scale. Furthermore, numerous portable systems are available today compared to a decade ago, when mobile EEG measurements still required complex engineering to accommodate all the cables, processing units, and power supplies for a participant (Gramann et al., 2014). Lastly, EEG is non-invasive, making it suitable for a wide range of applications with minimal discomfort for the participant. The form factor of modern EEG systems, along with their non-invasive nature, gives EEG great potential for high operator acceptability. While, EEG's low spatial resolution tends to be cited for as its predominant disadvantage over alternative imaging modalities, its susceptibility to various forms of noise, such as environmental electromagnetic interference as well as sensor and muscle-related artefacts, likely poses the biggest challenge for its usage in applied settings (Gramann et al., 2014; Makeig et al., 2009; Urigüen & Garcia-Zapirain, 2015).

So far, the previous imaging modalities considered the electromagnetic footprint of the brain. Alternatively, neurovascular coupling can also be utilised for BCI purposes (Ferrari et al., 2004; Ferrari & Quaresima, 2012; Weiskopf et al., 2007). Neurovascular coupling describes the physiological relationship between cerebral blood flow and neuronal activity, in which stronger neural activity generally requires larger amounts of oxygenated blood – i.e., increased blood flow (Buxton, 2012). As such, it is an indirect measure of neural activation as opposed to the direct measurement of post-synaptic activations and action potentials, which the aforementioned M/EEG systems pick up on (Thio & Grill, 2023). Measures of local changes in metabolic responses include functional magnetic resonance imaging (fMRI), positron emission tomography (PET) and functional near-infrared spectroscopy (fNIRS).

In PET, an unstable radioisotope is injected into the bloodstream. During a blocked experiment, the most actively engaged brain areas have the highest metabolic needs, meaning that more blood is pumped towards these areas, resulting in an accumulation of the injected tracer in the tissue (Shukla & Kumar, 2006). By recording and reconstructing the tracer's reaction with the brain tissue in 3D space, an activation map can be created, averaged over the course of about 1 minute. This limitation confines the application of this technique to addressing the question of 'where' neural activity changes, leaving the question of 'when' unanswered. Although novel approaches to PET scanning with increased temporal resolution exist, they have not been commonly used in the past literature (Morigi et al., 2022; Wang, 2019).

Slight increases in temporal resolution, alongside similar or even better spatial resolution, are achievable with fMRI. In this modality, the electromagnetic properties of protons are utilised to differentiate various kinds of molecules, thereby identifying all types of tissues and fluids in the human body. Early studies using magnetic resonance imaging to investigate the blood oxygenation level-dependent (BOLD) effect were limited to a spatial resolution of 3 mm^3 that could be acquired every 2-3 seconds (Bollmann & Barth, 2021; Menon et al., 1997). While blood oxygenation is not equivalent to neural firing, it serves as a decent marker for the former since neurons are energy-hungry cells. The more dynamic nature of the fMRI contrasts allows for event-related designs that can provide insights into the "when" question of cognitive processes (Buckner, 1998). The somewhat slow acquisition time is regarded as sufficiently fast due to the slow peak times of the hemodynamic response (Norris, 2006), which describes the body's adjustment of blood flow in response to neural activity. Similar to the advancements surrounding faster PET methods, the fMRI field has been pushing its technology towards higher temporal and spatial resolutions (Dumoulin et al., 2018; Wiggins et al., 2019). It is now possible to record voxels that are 216

times smaller than those in earlier MRI studies (Seidel et al., 2020). However, a recent review of 80 human fMRI studies revealed that most still conduct their scans using the 3 mm^3 voxel size (Bollmann & Barth, 2021), likely because reducing voxel volume necessitates a significant reduction in the total scan area (Seidel et al., 2020). Promises for wider application of more advanced techniques were made 15 years ago (Logothetis, 2008), yet their adoption remains challenging due to the compromises required to make them viable (Bollmann & Barth, 2021).

While real-time fMRI's application for BCI purposes has been experimented with (Taylor & Martz, 2023; Weiskopf et al., 2007), with enduring uses in the field of neurofeedback (Dewiputri & Auer, 2013; Pindi et al., 2022; Taylor & Martz, 2023), its high cost and immobile apparatus make it an unlikely candidate for neuroadaptive purposes (Klein et al., 2024).

A low-cost and light-weight alternative for BCI applications is fNIRS (Naseer & Hong, 2015). In its most commonplace continuous-wave form (from here on just referred to as fNIRS), fNIRS provides a spatial resolution of 1 cm^3 (Quaresima & Ferrari, 2019) by placing light sources and detectors roughly 3 cm apart across the scalp (Gratton et al., 2006). With this optimal source-detector separation (different recommendations for children and adult-sized heads exist), information on neurovascular coupling is limited to the outer layer of the cortex, with exact spatial specificity being hard to determine due to inter-individual differences in brain size and skull thickness (Chen et al., 2020). Generally, data on oxygenated and deoxygenated haemoglobin (HbO and HbR) is recorded on a curved path between source and detector at sub-centimetre depth into the outer cortex. FNIRS tends to be reported as being less susceptible to muscle artefacts than EEG and easier to set up than wet EEG systems. However, fNIRS is far from free of non-neural noise that needs to be carefully considered to avoid erroneous conclusions (Tachtsidis & Scholkmann, 2016).

Both EEG and fNIRS benefit from denser spatial coverage. Denser and more numerous coverage of EEG electrodes allows for a more exact estimation of (sub-) cortical sources (Michel & Brunet, 2019). FNIRS devices with dense spatial coverage, referred to as (ultra) high-density diffuse optical tomography systems (Scholkmann et al., 2014), further allow for 3-dimensional estimation of the haemoglobin concentration of the outer cortex (Chitnis et al., 2016; Markow et al., 2025; O'Brien et al., 2024). For the sake of user acceptance of applied neuroadaptive systems, however, there are good reasons to scale these systems down to their absolute minimum spatial coverage. Using fewer EEG electrodes or fNIRS channels leads to quicker setup times and less bulky cap designs. Moreover, decreasing the energy consumption of the recording device enables longer battery life or the use of smaller batteries, thereby

lowering the overall weight of the system. This is vital when considering operator acceptability in the development of neuroadaptive systems intended for use throughout the entire workday.

Portable EEG devices have recently seen a great increase in popularity and availability, with cognitive monitoring being the most often cited research focus (Niso et al., 2023). Niso and colleagues' comprehensive review of EEG headsets suitable for mobile use presents a wide array of montage choices and electrode designs, but highlights the challenges of assessing signal quality variations between these choices due to significant variability and gaps in the reported methodologies and processing pipelines (Niso et al., 2023). However, a general rule of thumb to go off is that the type of electrode (gel, wet, semi-dry, dry) influences signal quality regardless of experimental paradigm (Li et al., 2020). Gel-based electrodes utilise a conductive gel between the scalp and the electrode. They are generally considered to produce the highest signal-to-noise ratios, but they also require the longest setup time. Wet electrodes, here referring to saline solution-soaked sponges, are generally faster to set up (Günther et al., 2023), but in turn require top-ups of the saline solution after about 30-60 minutes to maintain optimal signal quality. The lack of gel residue after recordings is generally well received by participants when trying both gel-based and wet systems (Williams et al., 2020), indicating higher operator acceptability for wet systems.

Lastly, semi-dry and dry electrodes use pressure to connect the electrode to the scalp surface. Semi-dry systems apply electrolyte solutions at the micro scale through porous materials using different mechanisms. Dry systems, on the other hand, utilise the head’s sweat and surrounding moisture to a similar effect, albeit with much larger impedance measures (Li et al., 2020). The large variability in materials used in dry systems makes general statements about their signal quality difficult. The necessary pressure needed to achieve low-impedance recordings (Xing et al., 2018), however, renders them ill-suited for longer recording durations (Mathewson et al., 2017). In one particular study, participants reported preferring the dry headset over the gel-based headset, reporting the absence of a chin-strap and the lack of cables tying them to an amplifier (i.e. less applicable to the all-wireless systems we will focus on in this thesis) (Hinrichs et al., 2020).

Due to their more established nature and presumably higher participant comfort than dry electrodes, this thesis utilised and compared traditional Ag/AgCl gel-based electrodes and a low-density wet-electrode system.

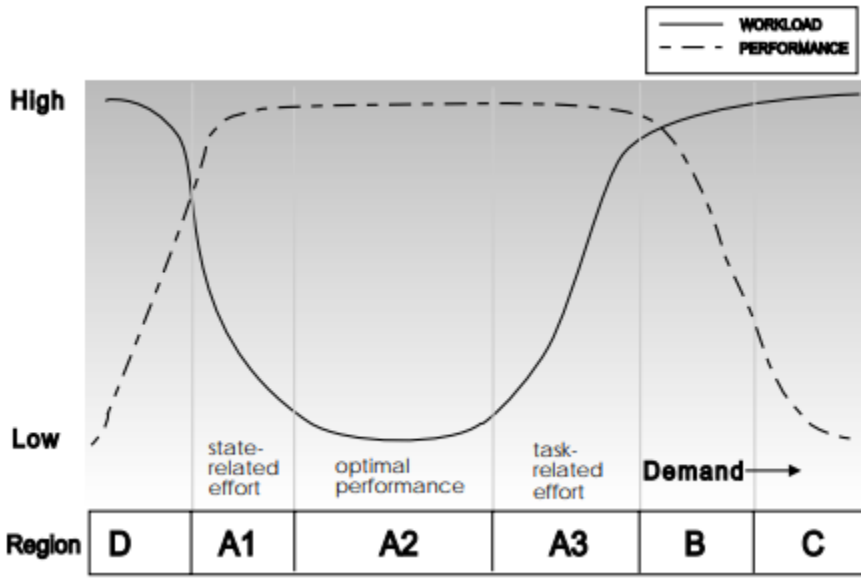
The choices of wireless fNIRS systems are not as numerous, but initial studies validating scaled-down wearable fNIRS systems showed congruence with more expensive and primarily stationary wired solutions (Boere et al., 2024; Friesen et al., 2022; von Lühmann et al., 2021). One hindrance to the development of smaller-scale fNIRS systems is their greater spatial specificity compared to EEG, requiring the accurate placement of optodes over regions of interest (Klein et al., 2024). Therefore, a research-grade wireless system employing the minimal number of optodes positioned over previously identified regions of interest was utilised in this thesis.

1.1.3 Definitions of Workload

"*the portion of an individual's limited capacity that is actually required by task demands*" (O'Donnell & Eggemeier, 1986, p42-2)

Mental workload has been of great interest to human factors and ergonomic research due to the need to consider an operator's capabilities and limitations when designing complex and safety-critical human-machine interactions (Afzal et al., 2022; Young et al., 2015; Dehais et al., 2020; O'Donnell & Eggemeier, 1986; Wickens, 2008; Wilson et al., 2011; Flemisch & Onken, 2002). Generally, it is accepted that suboptimal levels of mental workload lead to reductions in task performance (Figure 3). Identifying the antecedents and causes of suboptimal mental workload presents an opportunity for uncovering design improvements that prevent degraded performance and enhance job satisfaction (Moray, 1979). However, the theoretical construct of mental workload, although omnipresent throughout human factors and ergonomics, evades a concise and agreed-upon definition to this day (Van Acker et al., 2018; Young et al., 2015).

Figure 3. Mental Workload Across Task Demands

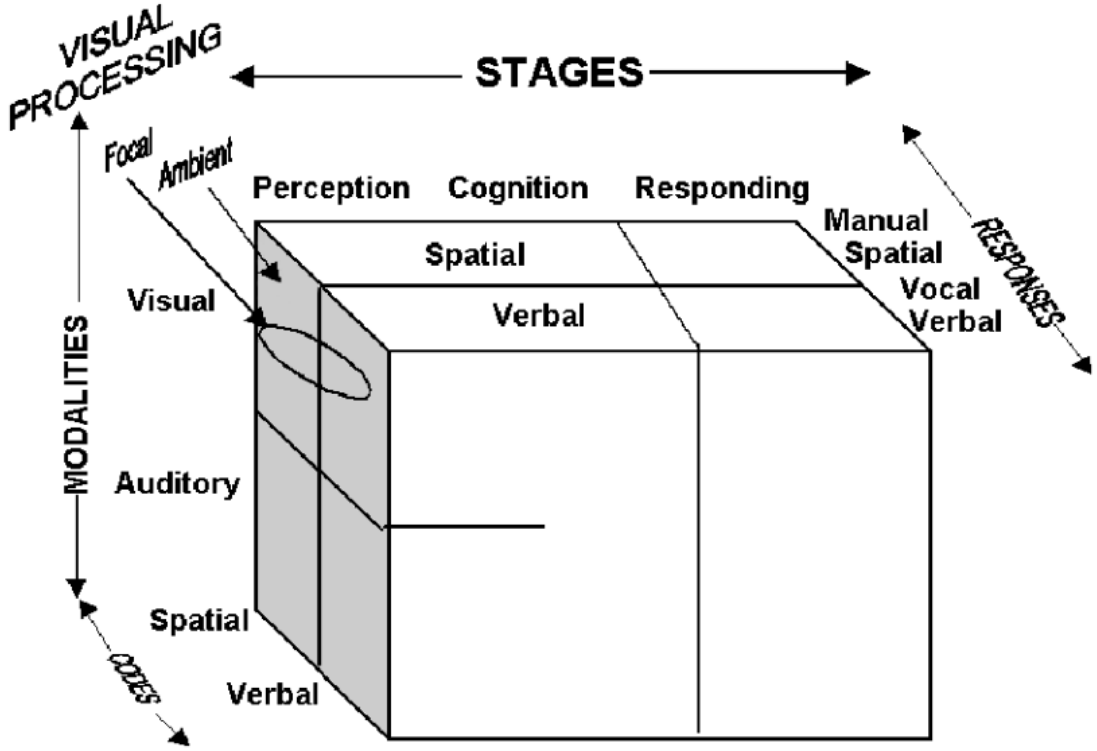


Note. Figure from Waard (1996) - Workload and performance inspired by the Yerkes-Dodson law across 6 task demand regions. In region D (D for deactivation), the operator is bored from engaging with prolonged monotonous tasks. In region A2, performance is optimal, and the operator can easily meet task requirements. In the regions A1 and A3, the operator has to exert effort to preserve performance. In region B, task demands exceed the operator's capacities and performance declines. In region C, performance is at a minimum level: the operator is overloaded

Inspired by models of limited information processing capacity (Kahneman, 1973), many definitions tend to emphasise abstract capacity limits of an operator to accommodate rising task demands, as is visible in the quoted definition by O'Donnell and Eggemeier (1986) at the beginning of this section. In the proceedings of the NATO Symposium on Theory and Measurement of Mental Workload (Moray, 1979), organised due to the construct's lack of an agreed-upon definition, all subfields introduce their general chapter on defining workload by referring to the intensity of effort as well as either fractions of attention or processing capacity in one way or another.

Wickens further formalised and extended this notion of limited resources (Figure 4) by assigning separate pools of resources across three dimensions (Wickens, 2002, 2008). The *stages of processing dimension* divide separate sources of capacity for perceptual, cognitive, and response-related processes. The *codes of processing dimension* differentiate between spatial and verbal activity. The *modality dimension*, specific to the processing stage of perception, posits separate resources for auditory and visual activity. Lastly, in the case of the perceptual processing pool, he further differentiates between focal and ambient channels. Wicken's multiple resource theory (MRT) can account for interference effects where performance degrades when multiple tasks compete for the same resource (Wickens, 2002), making it a useful tool for design decisions in the field of ergonomics.

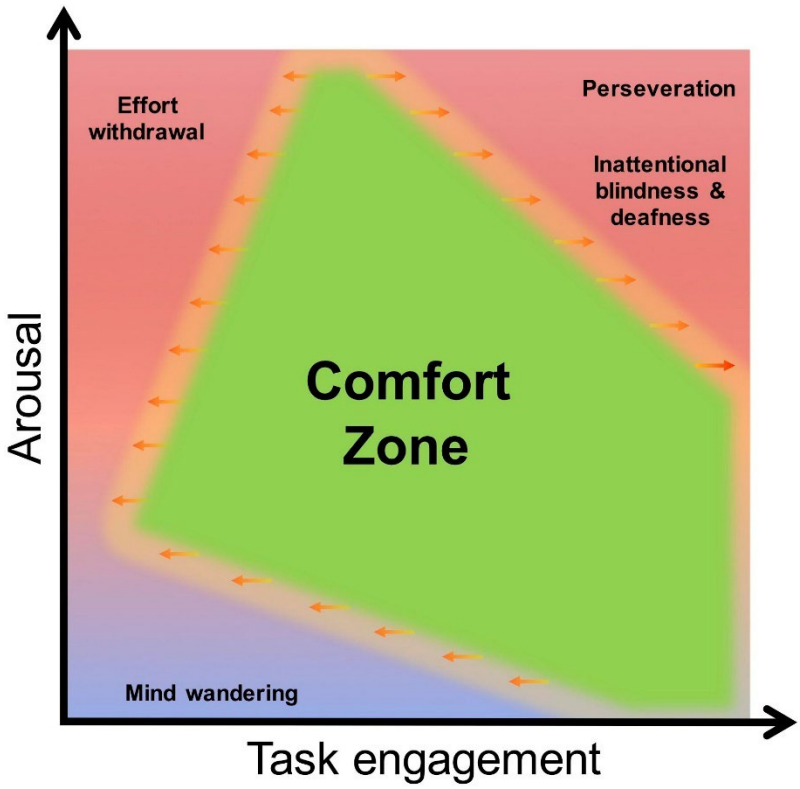
Figure 4. The Multiple Resource Model



Note. from (Wickens, 2008). The Multiple Resource Theory of mental workload splits capacity across several domains, which can be loaded concurrently without degrading performance. Overload occurs when a single domain is taxed beyond its capacity.

While useful, models relying on finite resources cannot account for various sources of variance observed in subjective ratings and performance metrics. These include, among others, the effects of expertise, age, and motivation (Dehais et al., 2020; Van Acker et al., 2018; Young et al., 2015). Empirical data suggests that performance is stable when a given individual's mental workload is neither too high nor too low (as visible in Figure 3). Performance degradation also occurring at underloaded states is at odds with definitions focusing on limited resources, which expect performance to suffer only once resources become scarce. To meet these limitations, the field has turned to operationalised definitions, in which performance-relevant psychological phenomena are linked to (neuro) physiology.

Figure 5. Neuroergonomics Model of Mental Workload



Note. from (Dehais et al., 2020). The neurergonomics model of mental workload situates states of degraded task performance in a 2-dimensional space of arousal and engagement. Only within the comfort zone can an operator preserve performance. Errors due to underload occur at low arousal and engagement levels, while overloaded states are characterised by high arousal levels paired with suboptimal engagement levels.

The neuroergonomics approach to mental workload does not rely on a unitary resource explanation. It links behavioural and neurophysiological metrics to suboptimal states of degraded performance (Dehais et al., 2020). These suboptimal states were mapped within a two-dimensional space, with task engagement on one axis and arousal on the other (Figure 5). Low task engagement and low arousal give rise to mind wandering in one corner of the space. Low task engagement, with excessive arousal levels, may, on the other hand, lead to effort withdrawal, a phenomenon in which disproportional task demands result in disengagement from the task, akin to giving up (Darvishi-Bayazi et al., 2023). In situations where both task engagement and arousal are high, operators may demonstrate degraded problem-solving flexibility, known as perseveration (Dehais, Hodgetts, et al., 2019), or miss crucial visual and auditory alarms due to inattentional blindness and deafness (Dehais et al., 2018; Durantin et al., 2017). Between these three suboptimal zones, they define a sort of Goldilocks zone for performance, similar to Waard's performance plateau (regions A1- A3 in Figure 3). By extending the inquiry beyond

global conceptualisations of mental workload, and instead problematising relationship between behavioural phenomena and physiological processes, the neurergonomics approach may help reframe several dissociate relationships between concurrent operationalisations of mental workload (Hancock & Matthews, 2019) – Dissociation referring to results in which subjective, performance, and or physiological data disagree, providing conflicting information on operator workload (Matthews et al., 2015).

Agreements among the myriad operationalised measures of mental workload are seldom investigated. In cases where multiple measures are compared, it is often their disagreement that is commonly discovered (Matthews et al., 2015), complicating their individual contributions to advancing our understanding of mental workload. The next section aims to provide an overview of how mental workload tends to be measured in research, highlighting inconsistencies which complicate the development of parsimonious definitions of mental workload. While acknowledging the advantages of the neuroergonomics approach to mental workload, most of the literature is based on conceptualisations revolving around limited resources. This is why the current thesis will mostly adopt the latter definition for consistency's sake. When appropriate, links to the neuroergonomics model will be drawn.

1.1.4 Measuring Mental Workload

Since conceptual definitions of workload tend to struggle with fully accounting for the multi-dimensionality of mental workload, defining mental workload by its consequences for human neurophysiology is seen as a viable alternative. Mental workload assessment can be grouped into one of three categories – subjective reports, performance based metrics, and (neuro)physiological metrics (O'Donnell & Eggemeier, 1986). All three have their distinct advantages and drawbacks (Paxion et al., 2014). To assess their relative qualities, seven criteria were identified (de Waard, 1996; Eggemeier, 1988).

1. Sensitivity: Does this metric detect changes in workload?
    - This is dependent on the region of task-load in which the metric is evaluated. Many metrics may be sensitive to changes from region D to C in de Waard's model, but distinguishing between A1, A2, and A3 requires more sensitivity (Figure 3).
2. Diagnosticity: Is this metric sensitive to a specific cognitive resource?
    - This is related to Wicken's model and asks whether a metric changes in response to global workload or only to specific components, like auditory processing (Figure 4).

3. Selectivity: Does this metric only change in response to mental workload or does it fluctuate with workload-unrelated processes?
    - A metric that decreases with increasing workload is not selective if it also decreases with time-of-day or other psychological constructs.
4. Intrusiveness: Does this metric interfere with the primary task?
    - Asking operators for their current mental workload may be sensitive and, depending on the length of the questionnaire, also diagnostic; however, interrupting the primary task with questionnaires could degrade performance itself.
5. Reliability: Does this metric's sensitivity hold across tasks? Also referred to as transferability (Wierwille & Eggemeier, 1993)
    - While task-specific metrics may already provide value, research on mental workload to this day aims to identify generalisable metrics.
6. Implementation requirements: What are the time, training, hardware and software requirements of the metric?
    - A metric that can be attained without fragile sensors, error prone software, or expert supervision would be preferable for applied mental workload monitoring.
7. Operator acceptability: Does the operator find the metric to be useful /worthwhile?
    - A metric that is not reliable or sensitive would likely not provide much value and may even hinder efficient human-machine interactions. Consequently, such metrics may not be deemed acceptable by the operator, especially if implementation requirements are high.

These seven criteria interact with each other. The implementation requirements and intrusiveness of a given metric will inform the operator acceptability. If a metric is highly sensitive and selective, providing valuable information to a neuroadaptive interface, the operator might tolerate higher levels of implementation requirements and intrusiveness. Whereas a metric with limited sensitivity and lacking selectivity would likely be deemed unacceptable if it were to place any additional strain on the operator through its implementation requirements.

Self-report measures. Subjective experiences of workload are not clearly defined in the limited resource views of mental workload. Prompting an operator to reflect on their experienced mental workload sets off a self and situation-appraisal process in which the work demands, estimated performance as well as experienced physiological and affective state are probed (Van Acker et al., 2018). Subjective rating scales

can be roughly divided into multidimensional (Hart, 2006; Reid & Nygren, 1988; Tsang & Velazquez, 1996) and unidimensional scales (Roscoe & Ellis, 1990; Zijlstra & Doorn, 1985). Multidimensional scales may offer higher diagnostic value as they split subjective workload into several subcategories relating to Wicken's MRT. Unidimensional scales, on the other hand, are less intrusive as they can be administered quickly by asking participants to reflect on a single issue. Even though they only require a fraction of the time and deliberation required by the multidimensional scales, their scores have been shown to be in agreement with those of the more complex alternatives (Ghanbary Sartang et al., 2016)

Regardless of their dimensionality, subjective ratings of mental workload tend to suffer from common biases found across subjective reporting literature. Firstly, context plays a big role. Providing operators with a reduced range of task demands still sees full use of the available range of workload ratings, making absolute comparisons across contexts and operators difficult (Colle & Reid, 1998). Secondly, due to response heuristics or attribute substitution, the tendency to substitute difficult questions unwittingly with easier questions to lessen the load on memory storage and retrieval (Kahneman & Frederick, 2002), it has been suggested that operators report their mental workload inversely to their estimates of self-performance when performance cues are readily available (Moore & Picou, 2018). Such response heuristics could introduce biases in studies comparing subjective ratings of mental workload across tasks. Lastly, subjective report questionnaires would be highly intrusive if administered during task performance. Their post-hoc nature makes them a valuable tool for developing neuroadaptive interfaces, by offering training labels or experimental manipulation checks, but not so much for their application.

Performance measures. Performance-based metrics of mental workload generally capture standard metrics like accuracy and speed of response (de Waard, 1996; O'Donnell & Eggemeier, 1986). Speed-based metrics in laboratory settings tend to be reaction or task-completion times, while accuracy-based metrics are derived from error-rates or continuous deviations from a target (tracking error in aviation, or lane deviation in driving contexts). The advantage of incorporating performance metrics into neuroadaptive interfaces is that they are highly sensitive to performance degradation, since this is exactly what they measure. However, as Figure 3 demonstrates, performance metrics would be insensitive to changes in workload during the performance plateau (Louis et al., 2023), during which operators may increase their invested effort to meet increasing task-demands within their capacity (Yeh & Wickens, 1988). Looking at the neuroergonomics model of mental workload (Figure 5), another shortcoming of performance-based metrics is that they may also not be selective as to whether the

performance degradation is due to low or high arousal-related states (e.g. is an error due to mind-wandering or effort withdrawal).

Consequently, by themselves, performance metrics do not suffice for a sensitive mental workload monitoring system. Like subjective ratings, they do, however, provide valuable information for the development of neuroadaptive interfaces as they provide context that informs manipulation checks and "ground-truth" labelling.

Physiological measures. As previously touched upon, mental states like mental workload tend to be influenced by - and exert influence on - the physiology of the brain and body. Inferring mental workload from changes in (neuro)physiology allows for continuous and implicit assessment.

A considerable number of recent review articles have been published regarding the various effects of mental workload on neurophysiological variables (Borghini et al., 2014; Charles & Nixon, 2019; Dehais et al., 2020; Hughes et al., 2019; Luzzani et al., 2024; Tao et al., 2019), which is why we will limit the following discussion to the effects most pertinent to the current thesis. Furthermore, effects from rest vs. active task periods will not be discussed, as we can assume most physiological processes will be sensitive enough to exhibit measurable changes between these extremes. While their analyses have great merit for our understanding of the brain on and off tasks, they provide little value for the development of neuroadaptive technologies, where the main interest lies in differentiating between often more subtle physiological changes when an operator deals with varying task demands.

Electrocardiography. The most extensively studied sensor modality in the mental workload literature is the electrocardiogram (ECG). ECG's popularity in the study of human mental workload likely stems from its comparatively low implementation requirements and low cost (e.g., three-lead ECG or smartwatch-integrated sensors), which allowed for applied experimentation using ECG as early as the late 20th century (Mulder, 1992) when the topic of workload gained momentum. The two main metrics used to study mental workload are heart-rate and heart-rate variability, which can be assessed via the most prominent peak in the ECG's waveform – the R-peak (Gacek & Pedrycz, 2011). Both are derived from the interval between successive R-peaks (RR-interval, normal-to-normal interval or inter-beat interval).

When studying the effects of workload on ECG-based metrics, the sensors' low intrusiveness and implementation requirements are their clearest benefits. The sensitivity of ECG-based metrics has been reported to be lacking, however, as significant differences between conditions are often task-dependent (Nickel & Nachreiner, 2003) and often only between the lowest and highest levels of task demand, with

intermediate levels usually not producing significant changes (Fournier et al., 1999; Splawn & Miller, 2013), making it an all-or-nothing metric. This may also relate to the heart rate's intricate link to general arousal levels, which, as described by the neuroergonomics model of mental workload, does not fully explain the behavioural consequences of high mental workload (Figure 5). Furthermore, during longer experiments, a participant's heart-rate exhibits strong time-on-task effects. This may a) be visible as initial task anxiety, which only subsides after participants become accustomed to the task and laboratory environment, or b) show as gradual increases in HRV over prolonged task duration (Fairclough et al., 2005; O'Hanlon, 1972), meaning designs lacking randomised condition orders may report time-on-task rather than condition-specific effects.

Ocular. Ocular measures of mental workload have become a lot more accessible over the past 3 decades and even surpassed ECG-based metrics in frequency in a recent review of physiological research on mental workload (Charles & Nixon, 2019). Simple metrics regarding blink rates or durations may be extracted from electrooculograms (EOG), but the adoption of eye-tracking devices further allows for the calculation of pupil size, fixation, and saccade-related metrics. While the effects of workload on ocular measures will not be covered in this thesis, some recent findings may still be of interest for the development of neuroadaptive interfaces. Furthermore, workload-related changes to ocular metrics also carry importance for EEG-based workload assessments, due to the strong influence of ocular motion on the EEG signal (Dimigen, 2020; Luck, 2005).

Blink rates seem to decrease with increasing workload (Fournier et al., 1999; Veltman & Gaillard, 1996), although null results are still common (Ahlstrom & Friedman-Berg, 2006) and early research was plagued by inconsistent findings (Kramer, 1991; Marquart et al., 2015). However, Bendetto and colleagues (2011) found that dividing blinks into short (<100ms), long (>171ms) and medium blink durations showed that short-duration blinks actually increase in frequency with increasing workload. Benedetto's results could explain the mixed effects from previous studies, as well as why the average blink duration tends to decrease with increasing workload (Ahlstrom & Friedman-Berg, 2006; Fournier et al., 1999; Veltman & Gaillard, 1996).

Pupil size and the fixation behaviour of the eyes are less commonly investigated than blink-related metrics, as they require the use of an eye-tracker as opposed to EOG electrodes. Studies in the field of aviation tend to find reduced fixation durations (De Rivecourt et al., 2008; Di Stasi et al., 2011; Matthews et al., 2015) and dwell times (De Rivecourt et al., 2008) on points of interest at higher workload levels. Furthermore, some studies found pupil diameter effects (Ahlstrom & Friedman-Berg, 2006; Benedetto et

al., 2011) while others did not (Di Stasi et al., 2011; Matthews et al., 2015). Lastly, the addition of eye-tracking allows for the investigation of changes to fixation dynamics under increasing workload. On average, the distance covered by successive saccades seems to increase with increasing workload (Ahlstrom & Friedman-Berg, 2006; Di Stasi et al., 2010, 2011).

Electroencephalography. A number of comprehensive reviews for the effects of workload on EEG-based metrics have been conducted in the recent past (Borghini et al., 2014; Charles & Nixon, 2019; Ghani et al., 2020; Ismail & Karwowski, 2020; Tao et al., 2019) In general, measures of workload in the realm of EEG can be divided into frequency, temporal, and information domain metrics.

In the temporal domain, EEG analyses focus on latency and amplitude differences in event-related potentials (ERPs) between experimental conditions (Ghani et al., 2020; Luck, 2005). In the realm of mental workload research, one can differentiate between task-dependent ERPs and task-independent ERPs, where task-dependent ERPs are time-locked to task-specific events and task-independent ERPs are time-locked to additional stimuli explicitly employed to probe the attentional resources of the operator (Ghani et al., 2020; Kramer, 1991; Papanicolaou & Johnstone, 1984). The prior have the disadvantage of not being generalisable across tasks. Furthermore, as opposed to laboratory experiments, applied settings likely lack clearly defined events. Only a few studies investigated task-dependent ERPs for workload-related changes. Stimulus locked ERPs in the n-back working memory paradigm (Kirchner, 1958) sometimes show load effects across several ERP components (Shalchy et al., 2020) and sometimes only in the P3-related and later components (Brouwer et al., 2012; Ren et al., 2023). A recent study on noise-induced workload during a mental arithmetic task with auditory delivered stimuli only reported a late, possibly P3-related, effect (Pieper et al., 2021). Due to their flexible nature, task-independent ERPs tend to be much more commonly studied. While an early review by Papanicolaou & Johnstone (1984) on task-independent probing ERPs still included examples of visual, auditory, and tactile probes, the current literature is largely focused on the use of additional auditory stimulation (Ghani et al., 2020).

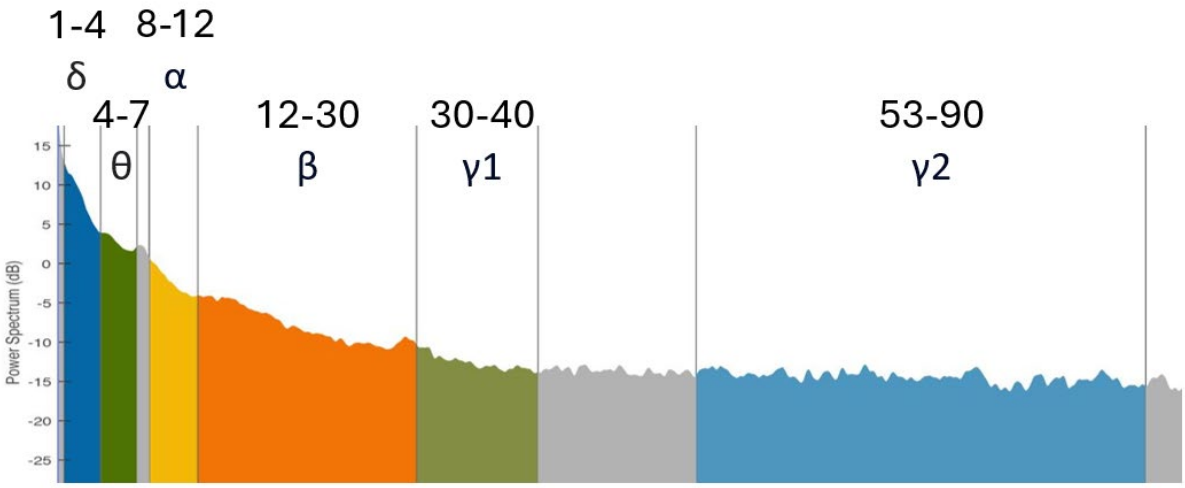
Auditory probes can easily be employed across contexts as they merely require the operator to wear an earpiece in addition to the EEG sensors. They are often reported to show workload related effects in both active (Solís-Marcos & Kircher, 2019; Van Der Heiden et al., 2022) as well as passive designs (Allison & Polich, 2008; Dehais, Duprès, et al., 2019; Dyke et al., 2015; Kramer et al., 1995; Miller et al., 2011; Ullsperger et al., 2001). Active auditory probe approaches make use of oddball designs that present many standard and rare deviant stimuli, with participants having to react to or count the occurrences of deviant stimuli. Here, the p300 ERP component tends to show the strongest effects of mental workload

and is commonly interpreted to relate to a stimulus classification and reorienting response (Polich, 2007), which diminishes in intensity when attentive capacities are directed to the primary-task. However, since active approaches constitute a secondary task, they score high on primary-task intrusion and do not lend themselves well for pBCI applications.

When no explicit attention to the sounds is required, auditory probing paradigms are often denoted as passive or task-irrelevant. Passive oddball paradigms still evoke the p300 response (Mertens & Polich, 1997), albeit at smaller amplitudes (Polich, 2007). Further factors impacting the size of the p300 components are the spacing of deviant stimuli and the type of stimuli. Dyke and colleagues (2015) built on previous work to show that complex sounds with inherent meaning produce a pronounced reorienting response in a passive paradigm (Miller et al., 2011), which decreased monotonically in amplitude from rest to an easy and to a hard task-load level. In the same study, a separate group received the same stimulation using novel pure tones instead, which produced much smaller p300 components and smaller effect sizes, suggesting the attention capturing properties of complex or deviant stimuli are necessary for sensitive workload discrimination (Dyke, et al., 2015). Importantly, since ample spacing between stimuli is required to elicit pronounced reorienting responses (Steiner et al., 2013; Kamp, 2020), relying on the p300 for continuous workload monitoring would result in slow monitoring system in the pBCI context. Furthermore, increased workload seems to subdue the reorienting response in passive oddball designs to a degree where standard could no longer be reliably distinguished from deviant tones (Ladouce, et al., 2023). Thus, a parallel inquiry into the effects of mental workload on ERP components preceding the p300 may be of greater interest for pBCIs.

Frequency domain analysis of workload effects are more easily applied across task settings. Their continuous nature makes frequency domain features ideally suited for continuous workload monitoring. In the past century, a number of canonical band power ranges have been defined (Figure 6), which tend to show functional distinct dynamics (Donoghue et al., 2022). Most prominently, workload-related effects can be found in the theta and alpha ranges of the EEG spectrum, with increased workload often being accompanied by frontal theta increases (Dehais, Duprès, et al., 2019; Fournier et al., 1999; Ke et al., 2021; Roy et al., 2013; Smith et al., 2001) and posterior alpha suppression (Fournier et al., 1999; Roy et al., 2013; Shaw et al., 2018; Smith et al., 2001). However, null-effects are also commonly reported for either spectral range (Baldwin & Penaranda, 2012; Dehais, Duprès, et al., 2019; Ryu & Myung, 2005; Shaw et al., 2018; Smith & Gevins, 2005; Verdière et al., 2019), casting their robustness for applied settings in doubt.

Figure 6. EEG's Canonical Band Power Ranges



Note. Exemplary EEG power spectrum with coloured canonical band power ranges (in Hz) used in the remainder of the thesis.

The delta band (<4 Hz) is often linked to effects of fatigue (Lal & Craig, 2002; G. Li et al., 2020) but rarely reported in relation to workload (Borghini et al., 2014; Chikhi et al., 2022). Some papers found significant increases in delta power comparing baseline to flight checklist checking periods (Dussault et al., 2005) while others found both negative and positive effects (depending on channel location) stemming from computer programming workload (Kosti et al., 2018). The theta band (4-8 Hz) is much more commonly reported in relation to mental workload, showing frontal medial increases with increasing task demand (Gevins et al., 1997; Smit et al., 2005; Zakrzewska & Brzezicka, 2014). Theta activation has been found to be linked to cognitive control signals in the face of novelty, conflict, or errors (Cavanagh et al., 2012; Cavanagh & Frank, 2014), making it a reasonable candidate for task-load induced changes.

The alpha band (8-13Hz) is commonly known for the “Berger effect”, which describes an increase in its power when eyes are closed. Changes to the alpha band have been linked to attention (Foxe & Snyder, 2011; Lobier et al., 2018), working memory (Jokisch & Jensen, 2007; Tuladhar et al., 2007), long-term memory (Hanslmayr et al., 2016), and several clinical outcomes (Fernández-Palleiro et al., 2020; Kim et al., 2022; Zhang et al., 2021). In the field of mental workload, it has been found to reduce in power with increasing workload (Fournier et al., 1999; Roy et al., 2013; Shaw et al., 2018; M. E. Smith et al., 2001). However, increases in alpha with increasing mental workload have also been observed (Borghini et al., 2014; Puma et al., 2018).

The beta band (13-30Hz) has been linked to language processing (Weiss & Mueller, 2012), reward processing (Marco-Pallarés et al., 2015), as well as working and long-term memory (Chen & Huang,

2016; Deiber et al., 2007; Hanslmayr et al., 2014; Weiss & Mueller, 2012). Furthermore, the beta band is generally linked to motor-related processes (Weiss & Mueller, 2012), making it particularly interesting for motor imagery-based BCIs (Bai et al., 2007; Gruenwald et al., 2017). In mental workload research the beta band has been reported to exhibit power increases with increasing workload (Gong et al., 2019; Morales et al., 2019). Lastly, the gamma band (30-100Hz), while interesting to research on human cognition (Herrmann et al., 2010), tends to be rife with EMG-related activations, which, in the case of mental workload research using complex tasks, may carry confounding information that is more closely related to physical rather than mental workload (Brouwer et al., 2015).

Bandpower-metrics have been shown to be affected by time-on-task (Chikhi et al., 2022). One possible explanation for the time-on task effects could lie in the slow rotation of the aperiodic background activity (Donoghue et al., 2022; Gao, 2016), which can be estimated and subtracted from the raw PSD estimates. A first study utilising a parametrised approach to quantify the band-specific power differences after accounting for changes to the aperiodic activation found workload-related effects to be more consistent across subjects when accounting for the aperiodic activity, and even discovered workload-related increases in the exponent describing the slope of the aperiodic activation (Ke et al., 2023). The effects they described from participants carrying out two levels of the MATB task showed the expected effect directions, with increases in theta power, decreases in alpha power with increasing workload. However, the effect of workload on beta power was reversed, increasing with workload in raw estimates but decreasing with workload on aperiodic-free estimates.

Ratio measures of different frequency bands have also been investigated and may be related to the estimation of the slope of the aperiodic activation. One of the ratio measures, termed the engagement index (Pope et al., 1995), is calculated as the ratio of beta/(theta+alpha), which would offer a rough estimation of the spectrum's slope in log-log space and has also been reported to increase with increasing workloads (Raufi & Longo, 2022) or decreases in surgeon performing complicated tasks (Wu et al., 2021).

Another downside to the bandpower metrics is that, as alluded to above, they tend to be involved in numerous processes in the brain. As such, changes in bandpower may not always follow predictable patterns. Previously, workload effects on specifically the alpha band have been shown to disappear under sleep-deprivation (Smith & Gevins, 2005) or longer time on task durations (Fairclough et al., 2005; Roy et al., 2013). Furthermore, the physical changes to the operator (standing vs. walking) also affect the same frequency bands regardless of workload condition (Shaw et al., 2018). Together, these confounds

may render bandpower-based metrics less reliable for applied workload monitoring in uncontrolled settings.

Functional near-infrared spectroscopy. Recent reviews on physiological measures of mental workload tend to either not mention or only briefly mention fNIRS studies (Charles & Nixon, 2019; Luzzani et al., 2024; Pütz et al., 2024; Tao et al., 2019). A recent review of research involving the MATB also did not include fNIRS studies (Prasetyo, 2024). However, the use of a headband-like frontal fNIRS montage has been reported in a number of different mental workload-related studies. In an air-traffic control task, average HbO concentrations over the prefrontal cortex exhibited increases with increasing traffic density (Ayaz et al., 2012). A similar effect (isolated to a single optode) could also be found contrasting 0-back and 3-back conditions in the same study. Concurrent performance of a working memory task with a simulated aviation task resulted in similar results, however, the n-back related effect showed an interaction with the aviation task's difficulty, resulting in opposite effect directions for the n-back in easy and hard aviation conditions (Durantin et al., 2014). Void of additional task-load manipulations, increasing n-back levels seem to be accompanied by HbO increases (Boere et al., 2024; Fishburn et al., 2014) and HbR decreases (Herff et al., 2014; Hirshfield et al., 2024). Furthermore, inducing stress in a virtual reality simulation of shutting down an industrial control system led to increased functional connectivity within a fronto-central fNIRS montage (Shi et al., 2020) . The same study reported a significant correlation between the performance accuracy and the average HbO concentration at various sites across the frontal cortex.

Using an fNIRS montage including parietal regions, in addition to the frontal montage, shows working memory as well as visual load-dependent concentration changes over the parietal regions (Hirshfield et al., 2024) as well as the dorsolateral prefrontal cortex (Fishburn et al., 2014). Additionally, functional connectivity changes between parietal and prefrontal optodes with increasing working memory load have been reported in the past (Fishburn et al., 2014).

The effects reviewed above were all garnered from population averages. To implement a neuroadaptive system, these results need to be reliable across participants and not merely survive the averaging across them. Furthermore, the measures taken for these studies are often computed on participant-wise averages over repeated condition presentations. A neuroadaptive system, on the other hand, will have to work with "single-trial" data – single-trial in quotation marks as we are not necessarily dealing with active or reactive BCIs which need to function off the data from a single sub-second data window. Instead, neuroadaptive systems may work at the pace deemed necessary for their proposed goal. If fast

spikes in workload are expected, the classification may need to work with similarly scarce, and more importantly, noisy information as an active BCI. In cases where workload changes more gradually, the classification could likely operate over longer time frames to come to a more stable, but less immediate result.

1.1.5 Classifying workload

The mental state classification at the core of pBCIs differs from classic neuroimaging analyses discussed before in a number of important ways. The signal-to-noise ratio (SNR) is significantly reduced since we no longer average over trials, blocks, and subjects of a full experiment. From this follows that any given mental workload metric needs to behave reliably / show consistency within a participant, ideally across participants, to facilitate generalisability. Since the necessary signal processing needs to be carried out online, automated signal-processing pipelines that operate without human supervision and careful trial rejection, common to classic physiological research, need to be validated. Computational resources may also be more limited in these online applications, limiting complex computations to a bare minimum.

A recent BCI competition challenged researchers across the globe to submit their best efforts to classify mental workload variations of three multi-tasking load conditions across separate recording sessions (Roy et al., 2022). The teams had access to training and validation data from two of the three recording sessions for this within-participant classification problem. The test data that each team's classifier was evaluated on stemmed from the third session and was withheld from the model training. The top-performing models did not surpass 55% accuracy in this three-class problem. While the misclassification of nearly half the samples appears problematic, the goal of classifying a human-made construct like mental workload using 2-second-wide EEG data samples across recording sessions is far from trivial (Fairclough & Lotte, 2020), and above-chance classification results tend to offer important information for future research.

The top performing models in the competition utilised state-of-the-art classification methods, including Riemannian geometry (Congedo et al., 2017; Yger et al., 2017), deep learning (Roy et al., 2019), and random forests (Breiman, 2001).

- Deep-learning methods comprise a large swath of possible architectures, generally designed as a cascade of non-linear transformations through trainable feature extractor modules (Lotte et al., 2018). Their appeal lies in simplifying the design of pBCI processing pipelines, as their complex architectures allow for end-to-end learning of preprocessing, feature extraction, and

classification (Roy et al., 2019). This may also be rephrased as a disadvantage, however, since the complexity of deep learning methods tends to result in “black-box models” that are difficult to interpret, though efforts towards making deep learning model weights interpretable exist (Samek & Müller, 2019). However, deep learning approaches tend to require large amounts of training data, which is usually not available in neurophysiological experiments (Roy et al., 2019).

- Random Forests are ensemble classifiers capable of handling high-dimensional feature sets by uniting multiple decision trees operating on subsets of the available feature set and combining their outputs for the final classification (Breiman, 2001). They have been shown to operate effectively with little training data and have previously outperformed traditional linear discriminant analysis-based classification for active BCI applications (Lotte et al., 2018)
- Riemannian geometry-based BCIs operate under the assumption that neural time-series are better represented on a curved mathematical space rather than a traditional flat Euclidean space. Here, spatial covariations between sensors tend to be used directly as features in lieu of extensive feature extraction algorithms. Riemannian approaches may hit a sweet spot between the aforementioned competition winners as they require minimal feature extraction procedures and perform well, even on small training sets (Congedo et al., 2017). Their application will play a central role in chapters 4-6.

Deep-learning techniques hold potential, but their need for large training datasets as well as their tendency for carbon emissions several magnitudes larger than those of Riemannian or traditional linear methods (Chevallier et al., 2024a) make them ill-suited for pBCI experimentation at this point in time, where more efficient alternatives exist. If approaches that are more readily explainable and energy efficient can outperform them, these simpler approaches are preferable for BCI technologies (Brouwer et al., 2015; Lemm et al., 2011). According to “the largest EEG-based BCI reproducibility study”, Riemannian classification on spatial covariance matrices can currently be considered one of the most accurate approaches for (re)active BCIs, including the classification of motor imagery, p300 detection, and steady-state visually evoked potential (SSVEP) classification (Chevallier et al., 2024b). However, the no free lunch theorem (Wolpert, 1996b, 1996a) states that no single machine learning model is likely to dominate over alternative models when averaged over all possible datasets. Consequently, whether the dominance of Riemannian-based classifiers holds in passive BCI problems, like mental workload monitoring, warrants further investigation.

The Mother of All BCI Benchmarks (MOABB) project is likely the currently largest effort to centralise and organise open-access EEG datasets for the sake of facilitating offline BCI developments (Chevallier et al., 2024b; Jayaram & Barachant, 2018). However, their focus lies with (re)active BCI paradigms. A recent review found that open-access mental workload-related datasets are rare to non-existent (Hinss et al., 2021). As the team behind MOABB pointed out aptly in their recent reproducibility study, comparing BCI results across a vast and dense literature is a difficult task, due to the tendency for employing idiosyncratic data processing, feature extraction and model evaluation strategies (Chevallier et al., 2024b; Demirezen et al., 2024). An open debate around standardised evaluation strategies would likely facilitate progress in the pBCI realm. For this to occur, openly accessible datasets and cross-dataset model evaluation studies mark an important stepping stone for facilitating the creation of such community wide standards.

The evaluation, comparison and ultimately ensemble-based fusion of various classification techniques across three montage designs will be the main outcome of the current thesis. Rather than focusing on developing ever-new approaches, this thesis focuses on assessing the reliability of popular signal processing, feature extraction, and classification techniques – Riemannian classification now counts as part of this "popularity" definition, given their success in the pBCI challenge (Roy et al., 2022) and the MOABB replication study (Chevallier et al., 2024b).

1.2 Aims and Objectives

Zooming out, the current thesis aims to inform possible avenues for developing reliable and generalisable mental workload detection techniques for incorporation into neuroadaptive technologies. To facilitate this, a paradigm reflective of traditional mental workload research was designed and employed across four experiments utilising different mobile EEG sensor designs and montages, as well as a final multi-modal setup combining frontoparietal fNIRS with lab-grade EEG.

The specific aims of the thesis included:

a) Conducting a comprehensive comparison of pBCI classifiers between a lab-grade and a modern wearable EEG headset.

b) Evaluating novel task-irrelevant auditory and visual probing methodologies.

c) Identifying optimal feature-class dependent extraction parameters.

Finally, the systematic variation of sensor and probing stimuli across the four datasets was designed to fill the gap of open-access datasets for mental workload classification and provide the pBCI community with varied benchmarking data to facilitate the development and validation of new methods.

1.3 Thesis Structure

Chapter 2 (General Methods) will provide an overview of the task paradigms, experimental structure, and participant demographics of the four datasets collected for the thesis. It will also delve into general preprocessing techniques applied to the EEG data for the analyses presented in chapters 3 to 6.

Chapter 3 (Group-level Analyses of Mental Workload Metrics) will aim to characterise the four datasets and their workload manipulations using the subjective, performance, and physiological metrics described in 1.1.4. It will also explore the sensitivity of a novel rapid task-irrelevant auditory probing paradigm, as well as a low-contrast SSVEP for mental workload estimation.

Chapter 4 (Evaluation of Bias in Cross-Validation Methods for Passive BCIs) will investigate biases common to pBCI model evaluation that complicate comparisons of results with the broader literature to establish a standard for the rest of the thesis.

Chapter 5 (Continuous Workload Monitoring) will focus on defining optimal extraction parameters for various machine learning features extracted from EEG and conduct a comprehensive comparison between lab-grade and wearable EEG sensors.

Chapter 6 (Multimodal Decision Fusion) Here, the focus will lie with fusing multiple classifiers into a large ensemble that is tuned to specific participants. The addition of fNIRS-based features will allow for the assessment of their incremental value above and beyond EEG-based mental workload classification.

Finally, Chapter 7 (General Discussion) summarises the main findings and links them to the open challenges in pBCI research.

2 General Methods

This chapter provides a high-level overview of data, signal processing, and classification techniques that will be utilised throughout this thesis. Four datasets resulting from four experiments have been collected for this thesis. Each experiment followed the same procedure, but they differed in the sensor and/or task-irrelevant probing techniques utilised. Furthermore, of the four experiments, the first differed from the remaining experiments in three critical ways. These differences pertain to the training regime participants underwent on a prior day to the experiment, the task-load manipulations in both tasks that were employed, and differences in the timing accuracy of the task-irrelevant auditory probes. The reasons and nature of the changes will be highlighted in this chapter. Table 1 provides an overview of the main differences and commonalities between the four experiments.

2.1 Participants

All participants signed an informed consent form after being briefed about the entire nature of the experiment. Participants were included if they were 16-50 years old, had normal or corrected-to-normal vision, and had no history of epilepsy. Participants who attempted to complete the experiment were compensated for their time with shopping vouchers (£25). The studies' experimental protocol complied with the Declaration of Helsinki and was approved by the British Ministry of Defence research ethics council (MODREC Reference: 2143/MODREC/22).

2.2 Task and Experimental Design

The purpose of all studies conducted during the thesis was to manipulate mental workload through changes in task-load. Task-load was manipulated at three levels, with the design choices being primarily guided by the need to minimise confounding variables. Confounding variables, here, refer to factors that could aid the identification of mental workload through other means than changes in brain states, such as differences in motor-affordances, or short-term trends in the data uniquely identifying blocks belonging to a certain condition. To mitigate the latter factor, conditions were presented in a pseudo-randomised order. Across the recording, participants completed three sets of blocks containing a randomised order of all conditions per task (see Figure 7 and Figure 8). To mitigate the influence of electromyographic (EMG) information from muscle activity in the EEG, the experimental paradigm employed no "rest" condition and all movements required to succeed in the hardest task-load level were necessary for the easiest level as well.

Furthermore, the experimental session (from now on termed recording session) was preceded by a training session on a prior day. This was done to reduce learning effects in the recording session and to stabilise performance and subjective mental workload experiences throughout the recording session.

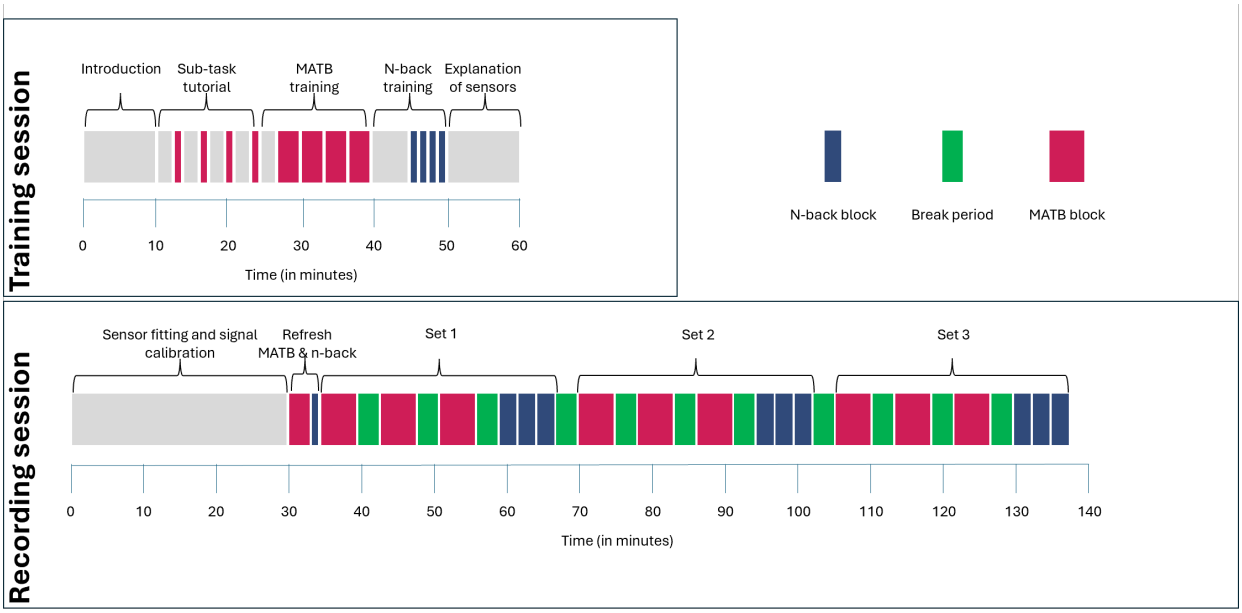
Two tasks were employed - (1) the n-back task (Kirchner, 1958), in which mental workload was manipulated via working memory load, and (2) the micro-world simulation environment called the Multi-Attribute Test Battery (Santiago-Espada, 2011), a multitasking paradigm, in which task-load was manipulated via changes to the density of events that required attention. Each task was presented at three task-load levels to investigate differences in sensitivity of neurophysiological mental workload metrics.

After completion of the pilot study, many participants were observed struggling with the hardest n-back level and some with the easiest MATB level. Subsequently, the task-load settings of both tasks were amended (see Table 2 and Table 3) to make the easy and medium task levels easier. Furthermore, as the hardest n-back level caused lots of frustration, we also opted to change it. To further help low performers reach performance levels comparable to their high-performing counterparts, the training regime was also amended to offer more MATB training time and dedicated time for feedback on mistakes and strategies.

Table 1. Overview of the Experiments

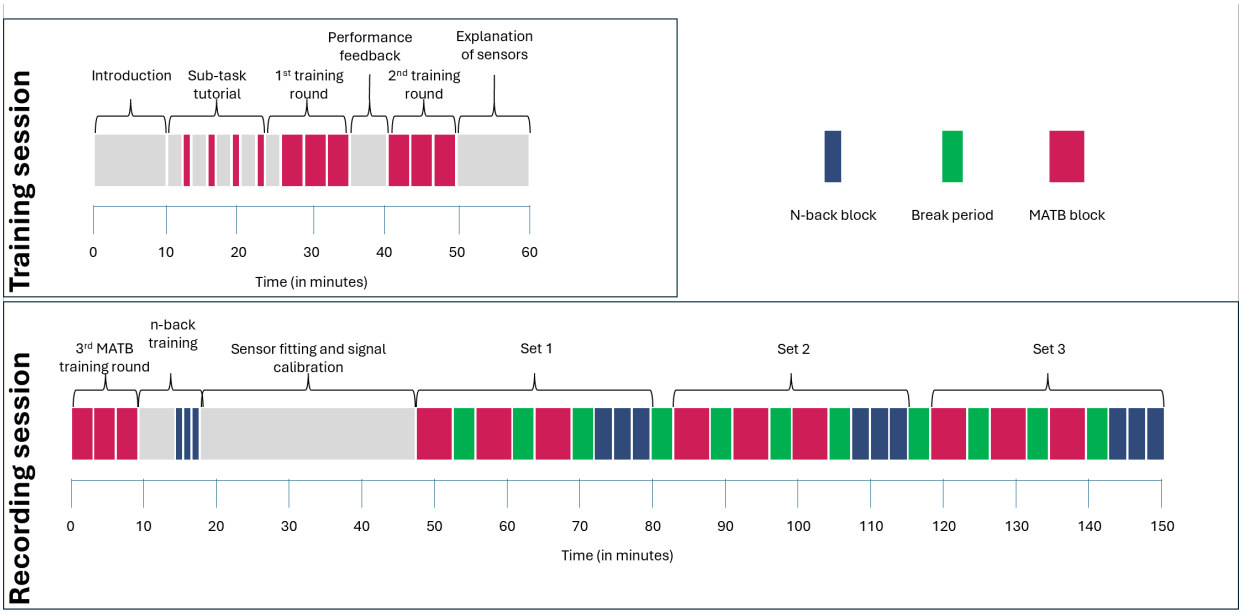
Datasets	**Pilot**	**Lab-grade**	**Wearable**	**Multi-modal**
Sensors	64-channel Gel-based LiveAmp, Three-lead ECG	64-channel Gel-based LiveAmp, Three-lead ECG, Photodiode	7-channel wet-electrodes X.on, Three-lead ECG, Photodiode	32-channel Gel-based LiveAmp, 17-channel Frontoparietal NirsSport2 + 8 short-channels, Three-lead ECG
Screens	22” LED monitor (iiyama ProLite B2283HS at 1920x1080@60Hz)	27” OLED monitor (LG 27GR95QE at 1920x1200@240Hz)	27” OLED monitor (LG 27GR95QE at 1920x1200@240Hz)	27” OLED monitor (LG 27GR95QE at 1920x1200@240Hz)
Auditory probes	Pure-tones ISI = 400-800ms Presented on same machine as the task using SoundPYO engine	Pure-tones ISI = 400-800ms Presented on a dedicated Linux machine using Psychtoolbox engine	Pure-tones ISI = 400-800ms Presented on a dedicated Linux machine using Psychtoolbox engine	-
Visual Probes	-	15Hz Sinusoidal full-screen stimulation at 10% contrast via FlickersOnTop	15Hz Sinusoidal full-screen stimulation at 10% contrast via FlickersOnTop	-
n-back conditions	1-back; 3-back; 6-back	0-back;1-back;3-back	0-back;1-back;3-back	0-back;1-back;3-back
MATB settings	See Table 2	See Table 3	See Table 3	See Table 3
Participants	20 (14 female, 19 right-handed, age: M = 26.7, SD = 7.8, Range = 18 – 49)	20 (12 female, 18 right-handed, age: M = 20.8, SD = 3.18, Range = 18 – 28)	21 (12 female, 19 right-handed, age: M = 23.8, SD = 6.16, Range = 18 – 38)	19 (9 female, 14 right-handed, age: M = 26.6, SD = 7.44, Range = 18 – 40)

Figure 7. Schematic of the Pilot Experiment



Note. Schema of the 60 minute training and ~2.5hour recording session. The recording session was split into three sets, each containing one block per task x task-load combination

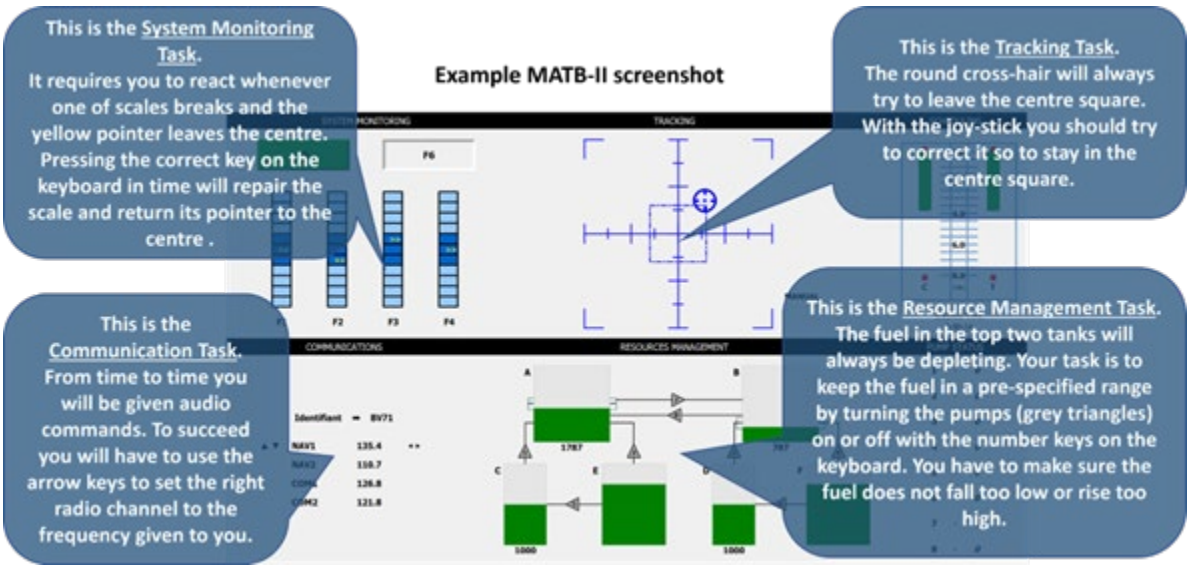
Figure 8. Updated Experiment Schema



Note. Schema of the 60 minute training and ~2.5hour recording session. The training regime now contained an extra round of MATB scenarios. A third round of MATB scenarios and the n-back training was added to the start of the recording session. The recording session was split into three sets, each containing one block per task x task-load combination

The training session was also adapted after after the Pilot study. For the Pilot study, participants were first instructed to perform each MATB sub-task in isolation before completing two 3-minute scenarios generated with the easy settings, and one 3-minute scenario each for the medium and hard level. Afterwards, they were instructed on the n-back and completed two 1-minute 1-back rounds and 1 minute of the 3-back and 6-back each. After each trial, the n-back training rounds presented feedback on whether the response was correct or incorrect to help participants understand the task better. The updated training regime for the latter three experiments spent more time on the MATB and moved the n-back training to the recording day (Figure 7), as participants generally struggled more with the MATB task. To remedy this, the training regime now included a first round with one 3-minute scenario per task-load condition, followed by a debrief in which the experimenter and the participant inspected plots detailing the errors in the individual subtasks for each task-load level. Afterwards, participants could apply the feedback to a second round of all three task-load conditions. Finally, upon returning for the recording session, participants could inspect their performance from the training session before completing a third round containing all three task-load settings.

Figure 9. MATB Instructions at a Glance



Note. Screenshot of the OpenMATB interface that was shown to the participants before the training session started.

Table 2. Pilot MATB Settings

	Resource Management						Communications		Tracking	System Monitoring	
task-load condition	Tank A & B loss	Pump 1-6 flow	pump 7&8 flow	Tank tolerance	Pumps that fail	N pump failures	N prompts	Own prompts	Track radius	Scales that fail	N scale failures
Easy	1000	1000	500	600	1 to 8	0	5	50%	0.1	3 & 4	10
Medium	1300	1000	500	400	1 to 6	20	5	50%	0.1	1 to 4	20
Hard	1300	1000	1000	350	1 to 6	30	5	50%	0.1	1 to 4	30

Table 3. Updated MATB Settings

	Resource Management						Communications		Tracking	System Monitoring	
task-load condition	Tank A & B loss	Pump 1-6 flow	pump 7&8 flow	Tank tolerance	Pumps that fail	N pump failures	N prompts	Own prompts	Track radius	Scales that fail	N scale failures
Easy	1000	1000	500	800	1 to 8	0	5	1	0.2	3 & 4	10
Medium	1000	1000	500	600	1 to 8	15	5	2	0.2	3 & 4	20
Hard	1300	1000	1000	350	1 to 6	30	5	4	0.2	1 to 4	30

2.2.1 *Multi-Attribute Test Battery*

The open-source python implementation of the Multi-Attribute Task Battery-II (MATB-II - Santiago-Espada, 2011), named OpenMATB (v1.1 - Cegarra et al., 2020), offers a modern and customisable version of the multi-tasking paradigm widely used for studying mental workload (see Pontiggia et al., 2024 for a review). For the current study, a custom script was used to generate scenarios with three varying levels of task load. All subtasks (see Figure 9) were presented at once in all three task load conditions, as opposed to increasing the number of subtasks to increase task-load. This was done to avoid introducing new regions of interest that could have resulted in fundamentally different fixation dynamics across conditions. Task load was instead manipulated by increasing the number of required interactions with the system monitoring and resource management systems while keeping the communications and tracking tasks constant across conditions (see Table 2 and Table 3 for a detailed list of task load parameter changes). The Communications and Tracking systems were kept at a fixed setting to keep muscle-related artefacts from the joystick and auditory interference from the radio commands at similar levels across conditions.

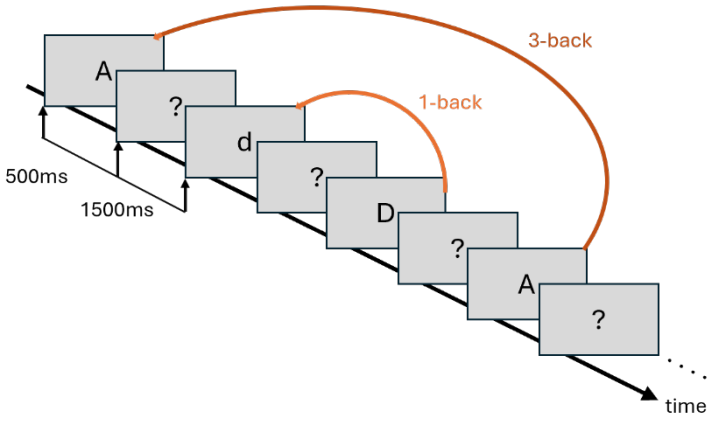
While each task load condition had a fixed number of events (e.g., pump failures, radio prompts, etc.), their order and timing were randomly generated for each scenario. Events targeting the same variable were separated from each other by at least 2 seconds; 15 seconds for the communications task. This meant a single scale in the system monitoring task could only fail once within 2 seconds, but theoretically all four scales could fail within the same 2-second window. For the communications task, however, only a single audio command could be given every 15 seconds.

2.2.2 *N-back*

The n-back was initially designed to study mechanistic explanations of short-term / working memory (Gajewski et al., 2018; Kirchner, 1958). Since the difficulty of the task (here, memory load) can be easily manipulated, it is commonly used in the workload literature (Ahonen et al., 2021; Dai et al., 2017; Devos et al., 2022; Roy et al., 2013). In the classic n-back task, the participant sits in front of a screen displaying a series of individual letters. The participant must decide with the press of a button if each letter is a target or not. What constitutes a target is defined by the ‘n’ in n-back. In a 1-back, targets are letters that repeat twice in a row – the participant needs to remember the last letter and compare it to the current letter. In a 3-back, targets are letters that match the letter three trials before, requiring constant updating and retention of three letter identities as well as their temporal order. Additionally, a simpler 0-back condition is often employed in the literature, which removes the working memory aspect. In our 0-back condition, only the letter ‘X’ constituted a target, removing the need to retain information from previous trials. The Lab-grade, Wearable, Multi-modal

experiments utilised the 0-back, 1-back and 3-back, whereas the Pilot experiment utilised a 1-back, 3-back and 6-back condition. The initial choice of including a 6-back was based on the rationale that to induce overload, the condition should be nearly impossible to perform. The manipulation turned out to be too difficult, as many participants seemingly disengaged from the task prematurely – i.e. participants would not even attempt to succeed at the 6-back. This resulted in us removing the 6-back for the following studies and making the low task-load condition easier by including a 0-back.

Figure 10. N-back Schematic



Note. Schematic of the n-back paradigm. Participants were instructed respond to both targets and non-targets.

The task and stimuli were generated with a custom script using PsychoPy2 (Peirce et al., 2019). Participants were presented with 70 letters per block, each randomly selected from 12 visually distinct letters (B, F, G, H, K, M, P, R, S, T, X, and Z). They were presented in black, centred on a grey screen for 500ms, and subtended approximately 1.17^{o} of visual angle, given a fixed viewing distance of 50 cm (the approximate the distance of the participants' eyes to the screen in this experiment). After the letter was presented, a question mark was displayed in the centre of the screen, prompting participants for a response if they hadn't yet responded. The question mark was displayed for 1500ms +/- 100 ms, giving participants about 2 seconds to make their responses. Participants could respond with the 'z' key for targets and 'm' key for non-targets. Each stream consisted of 21 randomly placed targets and 49 non-targets. The first 6 letters never included targets and were not included in any of the following analyses. In the training n-backs, participants additionally received visual feedback after each trial in the form of a happy smiley face for correct responses, or a sad smiley face in case of a wrong or no response.

2.3 Sensors

One aim of the current thesis was to compare pBCI classifiers across lab-grade and wearable neurophysiological sensors. To this end, we selected three devices that could be used outside

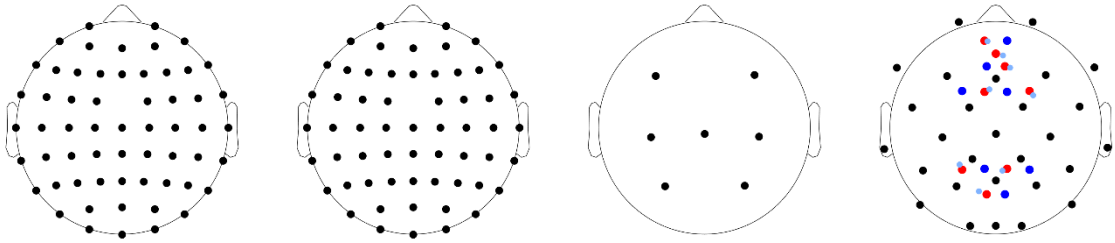
laboratory environments and differed in their spatial coverage and signal quality, as well as operator acceptance factors such as setup time, comfort, and any residual materials left upon removal. Furthermore, all used devices were selected for their compatibility with Labstreaminglayer (lsl - Stenner et al., 2022), such that all data recording and synchronisation could be carried out with lsl's open source LabRecorder[1].

2.3.1 EEG

EEG data was recorded with two different systems: the LiveAmp (Brain Products GmbH, Germany) and the X.on (Brain Products GmbH, Gilching, Germany). The LiveAmp was used with either 64 or 32 actiCAP snap electrodes (Brain Products GmbH, Gilching, Germany). These are active Ag/AgCl gel-based and shielded electrodes positioned according to the international 10-20 system. Data was sampled at 500Hz and recorded using the LiveAmp-lsl connector[2]. Impedance checks were carried out using the proprietary BrainVision Recorder, as no open-source solution is currently available for this. The battery power of the LiveAmp is advertised as being 4 hours, but this was not tested in our experiment, as we utilised an additional power bar to ensure no potential data loss.

The X.on system employs 7 wet passive electrodes in fixed 10-20 positions. The headset does not feature shielded leads and employs a default online 0.1Hz highpass filter. Data was sampled at 250Hz, even though 500Hz was technically possible. The lower sampling frequency was opted for due to battery concerns, as the advertised 5-hour battery life was not observed in our testing. Instead, the inbuilt battery delivered approximately 3 hours under typical in-lab use, possibly due to frequent impedance checks that may have accelerated battery depletion. Both impedance checks and data recording were carried out using the X.on LSL-connector[3] (version: 1.0.7).

Figure 11. Sensor Montages



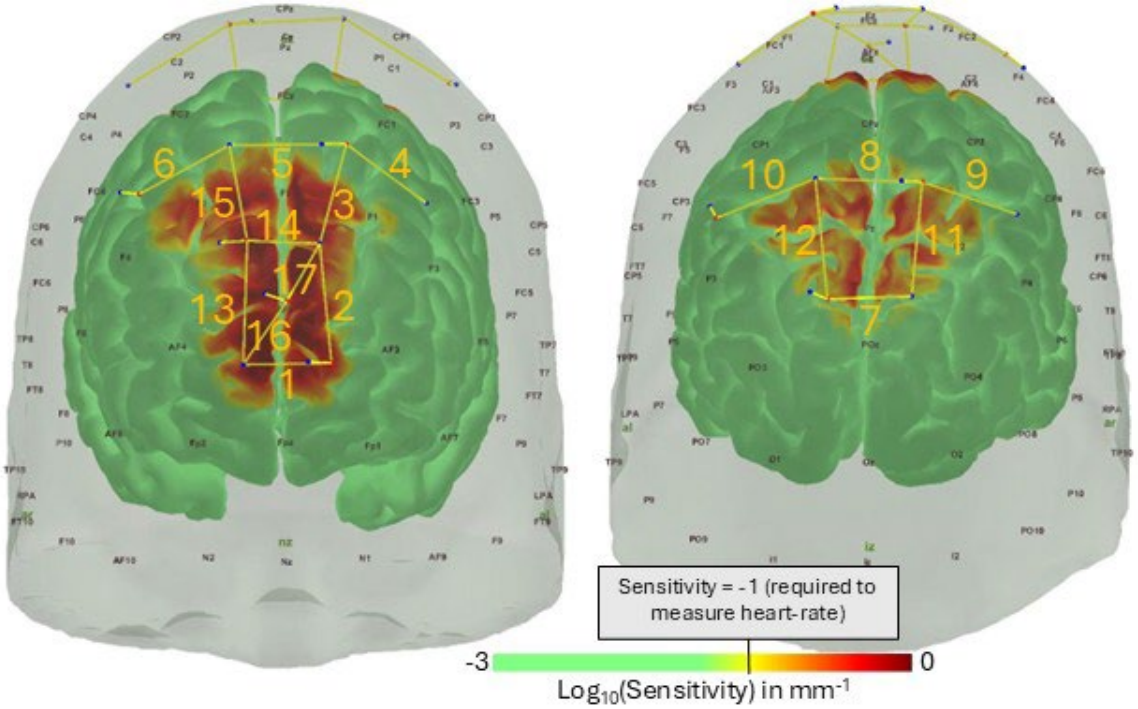
Note. The Pilot and Lab-grade montages (to the left) used the same 64-channel layout. The Wearable montage (centre-right) used 7 channels. The Multi-modal montage (right-most) used 32 EEG channels in combination with 17 fNIRS channels (8 sources in red, 7

[1] https://github.com/labstreaminglayer/App-LabRecorder
[2] https://github.com/brain-products/LSL-LiveAmp
[3] Only available through a password protected repository

detectors in dark blue, and an additional 8 short separation channels in light blue). EEG channels were arranged according to the 10-20 system; fNIRS probes were placed in 5-10 locations to not obstruct the EEG montage.

2.3.2 *fNIRS*

Figure 12. fNIRS Sensitivity Profile



Note. Sensitivity profile of the fNIRS montage (left for anterioir view and right for posterior view) exported from AtlasViewer with channel numbers additionally added (matching entries in table 4).

Neurovascular coupling was measured in the Multimodal experiment using the NIRSport2 (NIRx Medical Technologies LLC, NY, USA). The NIRSport 2 offers a lightweight and easy-to-set-up sensor design that some consider a wearable headset (Vidal-Rosas et al., 2023). Here, we used an 8x7 source-detector (dual-tip) montage with short channels (8mm distance) placed underneath each source diode (the 8th detector collected the short separation channel data). Data was sampled at 10.2 Hz and recorded with Aurora. Aurora also streamed the data via LSL to synchronise with the EEG stream.

The montage was selected to cover frontal and parietal areas, inspired by previous applied work, utilising sparse fNIRS montages for mental state monitoring (Fairclough et al., 2023; Fishburn et al., 2014). Channels were selected to cover Broadman areas 5,7,8,9,10, roughly corresponding to the Somatosensory association cortex, frontal eye fields, dorsolateral prefrontal cortex, and anterior prefrontal cortex across the left and right hemispheres (Table 4). The montage design was achieved via the fNIRS Optodes' Location Decider (fOld) toolbox (Zimeo Morais et al., 2018), which uses photon transport simulation through five types of tissue in the Colin27 head atlas to optimise

channel placements covering user-selected regions of interest. Additional photon transport simulations were carried out in AtlasViewer (Aasted et al., 2015) to produce the sensitivity profile in Figure 12.

Table 4. fNIRS Channels

Channel	Optode Pair (Source-Detector)	Rough Brain Area (Brodmann Area)
1	S1 (AFp1) - D1 (AFp2)	Medial Anterior Prefrontal Cortex (BA 10)
2	S1 (AFp1) - D2 (AFF1h)	Left Anterior Prefrontal Cortex (BA 10)
3	S2 (FFC1h) - D2 (AFF1h)	Left Dorsolateral Prefrontal Cortex (BA 9)
4	S2 (FFC1h) - D3 (FFC3h)	Left Frontal Eye Fields / DLPFC (BA 8/9)
5	S2 (FFC1h) - D4 (FFC2h)	Medial Dorsolateral Prefrontal Cortex (BA 9)
6	S3 (FFC4h) - D4 (FFC2h)	Right Dorsolateral Prefrontal Cortex (BA 9)
7	S4 (CPP2h) - D5 (PPO2h)	Right Somatosensory Association Cortex (BA 5, 7)
8	S4 (CPP2h) - D6 (CPP1h)	Medial Somatosensory Association Cortex (BA 5, 7)
9	S4 (CPP2h) - D7 (CPP4h)	Right Somatosensory Association Cortex (BA 5, 7)
10	S5 (CPP3h) - D6 (CPP1h)	Left Somatosensory Association Cortex (BA 5, 7)
11	S6 (PPO1h) - D5 (PPO2h)	Medial Somatosensory Association Cortex (BA 5, 7)
12	S6 (PPO1h) - D6 (CPP1h)	Left Somatosensory Association Cortex (BA 5, 7)
13	S7 (AFF2h) - D1 (AFp2)	Right Anterior Prefrontal Cortex (BA 10)
14	S7 (AFF2h) - D2 (AFF1h)	Medial Anterior Prefrontal Cortex (BA 10)
15	S7 (AFF2h) - D4 (FFC2h)	Right Frontal Eye Fields / DLPFC (BA 8/9)
16	S8 (AFz) - D1 (AFp2)	Right Medial Anterior Prefrontal Cortex (BA 10)
17	S8 (AFz) - D2 (AFF1h)	Left Medial Anterior Prefrontal Cortex (BA 10)

2.3.3 ECG

ECG data for the Pilot, Lab-grade, and Multimodal experiments was recorded using a 3-lead chest configuration with two electrodes placed above the left and right clavicle bones and the ground electrode on the left hip. Data were recorded using the same LSL connection as the EEG data via the Sensor & Trigger Extension (STE) for the LiveAmp amplifier and sampled at 500 Hz. For the Wearable experiment, ECG data was recorded using the Aux input of the X.on headset (250Hz) with the ground electrode placed on the right earlobe.

2.3.4 Photodiode

The Lab-grade and Wearable experiments included the visual probe. To assure its accurate display, an additional photodiode was included in the experiments, which was attached to the bottom left

corner of the screen to measure the 15 Hz flicker. The photodiode was recorded using a LiveAmp STE and sampled at 500Hz.

2.4 Task-irrelevant probes

2.4.1 Auditory Probes

While performing the two tasks, participants in Pilot, Lab-grade and Wearable experiments were instructed to ignore the rapidly presented pure tones (500, 600, 700, 800, 900, 1000, 1100, 1200, 1300, 1400, 1500 or 1600Hz). The selected tones and their presentation rate were chosen to conceptually replicate the long-variable condition reported on by Sugimoto and colleagues (2022), meaning the interstimulus interval (onset to onset) was uniformly sampled from a range of 400 – 800ms (mean = 600ms). All tones were presented at maximally 75 dB/SPL, calibrated before each session with a sound level meter (Tenma 72-860 – Farnell, US) held at the head level of the participant. In the Pilot study, tones were generated with the pyo python library[4]. However, its timing accuracy was deemed insufficient, and in the Lab-grade and Wearable experiments, all tones were instead generated by the Psychtoolbox sound engine[5] (Peirce et al., 2019).

2.4.2 Visual Probe

While performing the two tasks and hearing the task-irrelevant auditory probes, participants in the Lab-grade and Wearable experiments were instructed to also ignore a full-screen low-contrast sinusoidal 15Hz flicker overlaid over both tasks. The flicker was generated using a modified version of the FlickersOnTop software (v0.1.2 - Darmet, et al., 2022), with the contrast set to 10% (resulting in a measured modulation-depth of 10.5% - minimum brightness: 128 lux; maximum brightness: 158 lux) to reduce visual discomfort (Ladouce et al., 2022). The maximum and minimum brightness were measured using a Lux Meter (Tenma 72-6693 – Farnell, US). While providing a convenient measure of the light falling on the sensor, it is important to note that lux is a measure of illuminance, not luminance, which is the more appropriate unit for describing the brightness of a light-emitting source like a computer monitor. However, a luminance meter was not available at the time of writing.

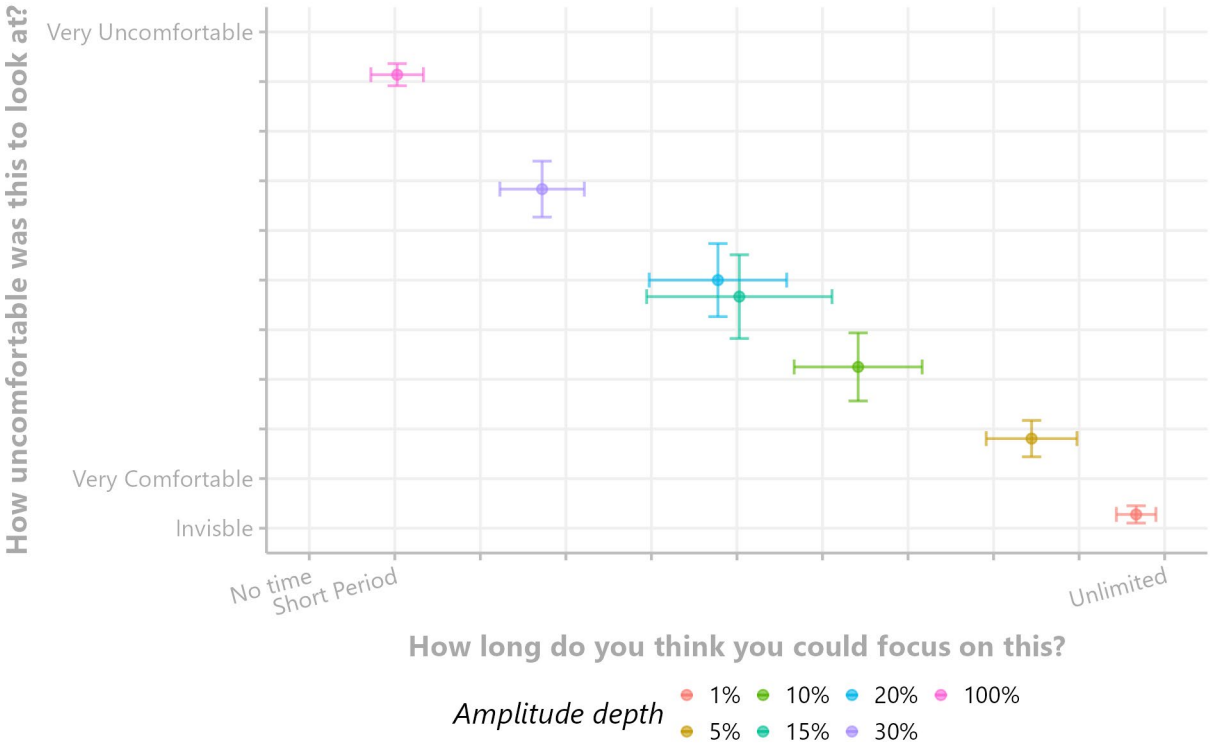
The 15Hz frequency for the sinusoidal flicker was decided upon as it still offered high signal-to-noise ratios (Ladouce et al., 2022) while not overlapping with the alpha band. The contrast of 10% was decided upon after a brief pilot study in which 10 volunteers were asked to focus on a fixation cross for 10 seconds while either a 1%, 5%, 10%, 15%, 20%, 30%, or 100% flicker was overlaid over the grey background (Hex code: #D3D3D3; same background as for the n-back and MATB) followed by

[4] https://pypi.org/project/pyo/
[5] https://www.psychopy.org/api/sound/playback.html#psychopy.sound.backend_ptb.SoundPTB

two Likert scale ratings. The first asked for the level of visual comfort on an 11-point scale ranging from 0 to 10, with 0 labelled as “invisible”, 1 as “very comfortable”, and 10 as “very uncomfortable”. The second scale asked the participants to estimate how long they could focus on the presented fixation cross with the flicker overlay. This was, again, an 11-point Likert scale, this time with 0 labelled as “No time”, 1 as “Short period” and 10 as “Unlimited”. Each contrast setting was presented 3 times in a pseudo-random order, and individual trials were interleaved by 20 seconds, in which the screen was left blank and participants were instructed to keep their eyes closed. Each trial commenced with an auditory alarm that signalled participants to reopen their eyes.

Since the 10% contrast setting seemed separated from the higher contrasts by a good margin in both comfort ratings (Figure 13), we opted for 10% over even lower contrast settings to ensure a high enough SNR for the SSVEP-based analyses.

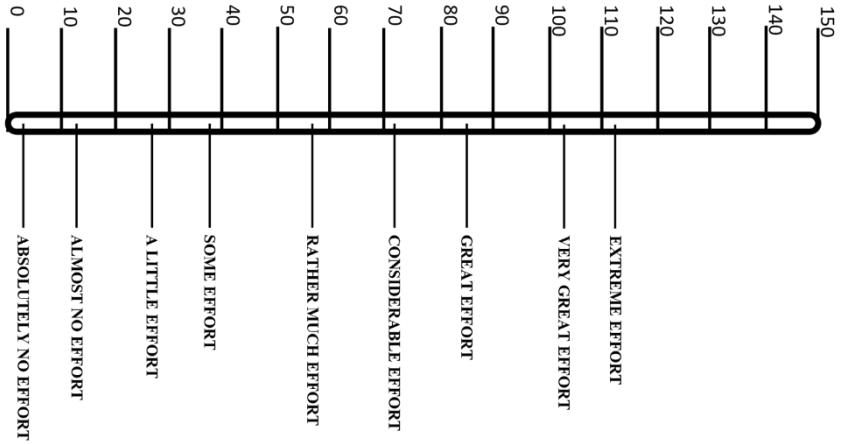
Figure 13. SSVEP Participant Comfort Pilot



Note. Average comfort ratings per contrast setting with error bars denoting the standard error of the mean.

2.5 Subjective Ratings

Figure 14. Rating Scale of Mental Effort



Note. The RSME scale recreated from the original study (Zijlstra & Doorn, 1985)

Subjective ratings of mental workload were collected using the Rating Scale of Mental effort (RSME). The RSME (Figure 14) is a continuous scale from 0 to 150 and is presented to participants with nine verbal anchor points ("Absolutely No Effort" to "Extreme Effort"). For all four experiments, we printed the scale vertically on pieces of paper (A4) to span 15cm. We asked participants to give their rating by drawing a line on the scale, as this was the method of delivery in its original validation studies (Zijlstra & Doorn, 1985). Ratings were digitised using a simple ruler.

2.6 Preprocessing

2.6.1 ECG

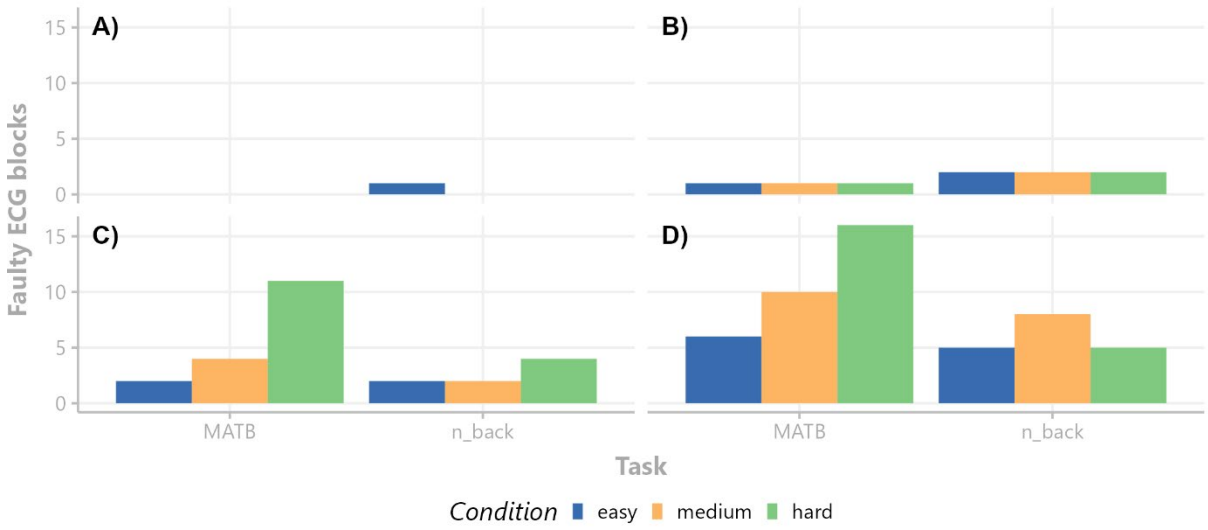
ECG preprocessing was carried out in MATLAB using an adapted version of the R-peak detection algorithm of the biosignal-specific processing toolbox (Nabian et al., 2018). Their algorithm was built on top of the popular Pan-Tompkins algorithm (J. Pan & Tompkins, 1985). Nabian and colleagues (2018) reported various bandpass filter designs that achieve comparable R-peak detection results. Here, the signal was filtered between 8 Hz and 20 Hz (Elgendi et al., 2010) using a zero-phase Butterworth bandpass filter (stopbands = 7.5 and 30Hz, stopband attenuation = 10dB and 40dB, maximum passband ripple = 1db). The filtered signal was subsequently squared to emphasise the R-peaks.

For the R-peak detection, the adapted Pan-Tompkins algorithm uses a 400-ms sliding window that scans the entire time series in 1-sample steps; if the global maximum is at the centre of the window, it is marked as a potential R-peak. Next, an amplitude threshold is set to one-third of the maximum amplitude of the first two seconds of the signal, and R-peaks below this threshold are removed. This

threshold is later updated to .75 of the mean of the previous 8 R-peaks. Next, false negatives (missed R-peaks) are detected by checking whether any RR interval exceeds a predefined threshold (initially set to the previous RR interval multiplied by 1.66 and later adjusted to the average of the past 8 RR intervals multiplied by 1.66). Missing R-peaks are placed at the global maximum 200ms after the last and 200ms before the following R-peak.

Since the MATB's joystick caused many high-amplitude artifacts in the ECG trace, peak detection was manually inspected for missed or erroneously placed peaks. In case of suboptimal peak detection, the squared signal was clamped to a maximum amplitude threshold. This threshold was iteratively decreased to 5% of the maximum amplitude of the time series. R-peak detection was performed on a block-by-block basis. If the R-peak detection after clamping still exhibited misplaced R-peaks, the participant's block was marked as faulty and excluded from further analysis.

Figure 15. Discarded ECG Data Due to Noise



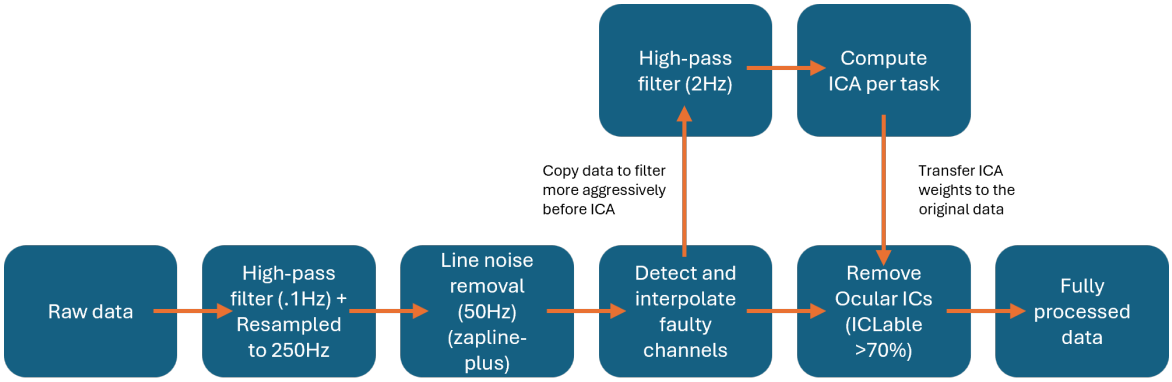
Note. Total number of blocks in A) Pilot, B) Lab-grade, C) Wearable, and D) Multimodal datasets considered too noisy for accurately estimating RMSSD.

Faulty R-peak detection may be particularly harmful to HRV analyses using RMSSD, as a single misplaced peak can significantly increase the measured variance. In the case that joystick motion was significantly different between MATB conditions, a less stringent R-peak detection procedure could lead to misleading results. Figure 15 shows how many blocks across participants were deemed too noisy for RMSSD estimation. Especially the Wearable and Multimodal datasets exhibited increasingly noisy ECG traces with increasing MATB difficulty, but also generally more noisy ECG traces.

2.6.2 *EEG*

Preprocessing was carried out with custom EEGLAB (version: 2021.1) scripts (Delorme & Makeig, 2004). All but the Wearable EEG data were preprocessed in the same manner to yield two processed EEG files per participant. The first file was heavily processed for traditional EEG analyses (Figure 16). Line noise was removed with zapline-plus (De Cheveigné, 2020; Klug & Kloosterman, 2022), Idependent Component Analysis (ICA) was used to subtract ocular artifacts, and faulty channels were interpolated after performing a correlation-based detection. These sets will be referred to as the *fully processed data*. The second file represents minimally processed EEG data, as would likely be used in a real-time neuroadaptive interface. They will be referred to as *minimally processed data*. The 7-channel EEG data collected for the Wearable dataset were only *minimally processed*, as methods based on spatial filters like Zapline-plus or ICA require enough degrees of freedom to fully separate artifactual from brain-related activations (Troller-Renfree et al., 2021). Artifact Subspace Reconstruction (ASR - Blum et al., 2019; Mullen et al., 2013) has been suggested as an alternative, but previous studies did not show convincing SNR increases using ASR correction on low-density headsets (Kumaravel et al., 2021). ASR could also not noticeably improve SNR in a 3-electrode setting (Yang & Lin, 2023). So far, no research into ASR's application to low-density EEG systems could provide convincing evidence that the method is not also removing brain-related information when reconstructing artifact laden segments, which is why we decided not to process the Wearable data to the same extent as the higher density datasets.

Figure 16. Full Preprocessing Pipeline



Note. Signal processing pipeline used for analyses without classification.

The .xdf files from LSL's LabRecorder software were imported using EEGLAB's load_xdf wrapper. Breaks between experimental blocks were removed to avoid excessive EMG artefacts impacting the ICA and faulty channel detection later on. Next, the data was high-pass filtered at 0.1 Hz using EEGLAB's default FIR filter (zero-phase filter with -6dB cutoff at 0.05Hz) and resampled down to 250 Hz (preceded by a low-pass filter at 125 Hz to avoid aliasing effects) for more efficient storage and

processing. For the fully processed data, line noise was attenuated using Zapline-plus (Klug & Kloosterman, 2022), an algorithm that detects periods with stable noise characteristics and adaptively removes line noise by combining spectral and spatial filters to only remove noise-related components in the specified line-noise band (and its harmonics). Afterwards, the "cleaned" line-noise band-ranges are added back to the rest of the signal, resulting in no data-rank loss and no "notch" in the power-spectrum. Using the now mostly line-noise-free data, EEGLAB's clean_rawdata function's correlation-based bad-channel detection was used with a cut-off of 70%. In it, bad channels are detected using random sample consensus (RANSAC). If the median correlation with 50 interpolated alternatives computed from random subsets consisting of 25% of the other channels fell below the cut-off, the channel was removed from the participant's data and interpolated using EEGLAB's spherical spline interpolation. Table 4 contains information on how many channels needed to be interpolated in the fully processed files.

Table 5. Channels Removed per Participant

Dataset	Participant-ID	Channels interpolated
Pilot	11	AF3, FT7, C1
	13	FT10
	18	C1
Lab-grade	4	C1,Po3
	5	FT9
	10	C1
Multi-modal	5	Fp1
	17	Fp1

Once faulty channels were interpolated, the data was referenced to the common average, and the original reference channel (FCz) was reconstructed (Kim et al., 2023).

The next step was the computation of Independent components using EEGLAB's extended infomax ICA implementation with Principal Component Analysis (PCA) dimensionality reduction to avoid rank-deficiency issues due to channel interpolation. Data was separately high-pass filtered at 2Hz before ICA to maximise the number and quality of ocular components (Dimigen, 2020). The weights of the heavily high-pass filtered data were then transferred to the original data and IClabel's (Pion-Tonachini et al., 2019) default algorithm was used to detect ocular ICs. Any ICs with a >70% chance of being

classed ocular were removed from the data. For the MATB, the Pilot and Lab-grade datasets had 4 ocular ICs removed on average and the Multi-modal dataset 2. For the n-back, the Pilot and Lab-grade datasets had 3 ICs removed on average, and the Multi-modal 1.

2.6.3 fNIRS

The fNIRS data was preprocessed using custom MATLAB scripts leveraging functions from Homer3 (v1.87.0) and EEGLAB (v2021.1). The preprocessing pipeline was designed to convert raw light intensity data into haemoglobin concentration changes, correct for systemic artifacts, and segment the data into experimental blocks.

The initial step involved loading the fNIRS data and synchronised event markers, which were previously saved in an EEGLAB .set file format from the original .xdf recordings. The raw intensity data was then loaded into a .snirf file format using the SnirfClass object in Homer3. Time periods corresponding to inter-block rest intervals were identified using event markers and were excluded from subsequent processing steps to prevent them from influencing data quality assessments.

A channel pruning algorithm (hmrR_PruneChannels) was applied to identify channels with an excessively low signal-to-noise ratio (SNR < 2), which were stored for diagnostic purposes but not removed. Raw light intensity data was then converted to changes in optical density using hmrR_Intensity2OD. Subsequently, the modified Beer-Lambert law was applied via the hmrR_OD2Conc function to convert optical density changes into relative concentration changes of oxyhemoglobin (HbO) and deoxyhemoglobin (HbR). A differential pathlength factor of 1.0 was used for all wavelengths.

To correct for systemic physiological artifacts originating from the scalp, a short-separation channel (SSC) regression was performed. For each source, the signal from its corresponding SSC was regressed from the signal of each associated long-separation channel using a general linear model (GLM) approach (Brigadoi & Cooper, 2015; Novi et al., 2023). Following this correction, the SSCs were removed from the data matrix. The resulting concentration data was then bandpass filtered using a zero-phase filter with cutoff frequencies of 0.01 Hz and 0.4 Hz.

Finally, the continuous data was epoched into the 18 experimental blocks. The data within each block was resampled to a new sampling rate of 10 Hz using linear interpolation to standardize the time series length across blocks and subjects for subsequent analysis / classification.

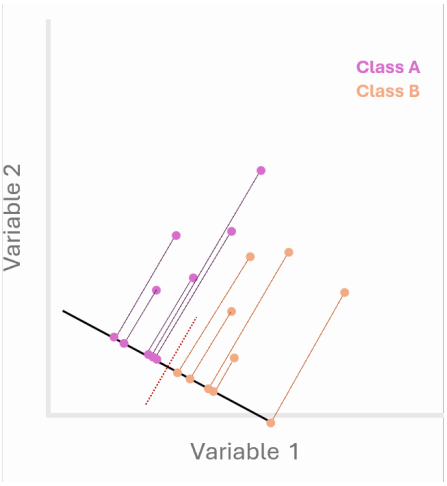
2.7 Classifiers

Throughout the classification chapters of this thesis, three distinct supervised classification approaches were utilised, which have been shown to work well with limited amounts of training data (Lotte et al., 2018). Their respective theoretical underpinnings as well as their relevance to the BCI literature are described below.

2.7.1 Shrinkage Linear Discriminant Analysis

Linear discriminant analysis (LDA) models are considered a simple type of classifier due to their linear nature and their interpretable projection vector. In a binary classification problem, LDAs project multiple features onto a single dimension that optimally separates the two classes (Figure 17). LDAs are popular in neuroimaging classification due to their interpretability (uninformative features get near-zeros weights in the projection) and lower tendency to overfit on limited training data, compared to more complex models (Lemm et al., 2011). Shrinkage LDA (sLDA) is a variant of LDA that adds regularisation through automated Ledoit-Wolf shrinkage, which helps to prevent noisy estimates from biasing the classification. Ledoit-Wolf shrinkage will be utilised throughout this thesis when sample covariance matrices need to be estimated. It computes an optimal shrinkage parameter by means of the distance between the, possibly very noisy, sample covariance matrix and an identity matrix scaled by the data, requiring no manual tuning and resulting in more robust and well-conditioned covariance estimates (Ledoit & Wolf, 2003).

Figure 17. Schematic of an LDA's Projection Vector



Note. Two variables which, by themselves, cannot be used to clearly separate data from two classes. A linear combination of the two variables allows for projecting the datapoints onto a more discriminative axis.

While more complex models tend to make the titles of new publications, sLDAs are still widely used (Lotte et al., 2018). They offer comparable performance to more complex approaches (Simões et al.,

2020), especially when paired with efficient dimensionality reduction techniques (Ang et al., 2012; Blankertz et al., 2008; Rivet et al., 2011).

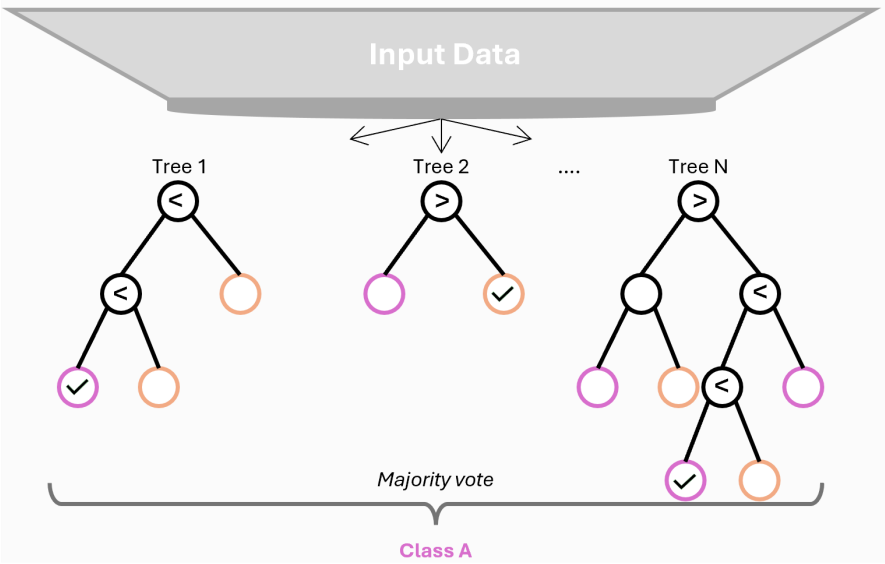
2.7.2 *Random Forest*

Random Forests are an ensemble of classifiers, which, via a combined majority vote, produce accurate predictions in many different applications (Fife & D'Onofrio, 2023; Loef et al., 2022). The individual classifiers are known as decision trees, which learn a number of simple if/else decision rules (Figure 18). They split the data iteratively into two subsets through their decision nodes to reach a prediction at the terminal nodes (i.e. "leafs"), which represent the classes (Barnova et al., 2023). A single decision tree, without limits to its depth, will create enough decision nodes to perfectly classify every single training datapoint (i.e. it badly overfits on the training data), which is why decision trees, unlike sLDAs, require either a number of a priori decisions about sensible hyperparameters (e.g. maximum depth of trees, minimum number of samples in a leaf node, etc.) or require separate hyperparameter tuning using training data.

A Random Forest classifier is based on the principle of bagging (bootstrap aggregating), in which the available training data is randomly subsampled (with replacement) to create numerous training sets for the individual decision trees. The available features are also subsampled, and by allowing individual trees to see only a subset of the available features, the Random Forest becomes more resilient to single domineering features by decorrelating the individual trees. Each tree may overfit; however, by combining hundreds of decision trees, the Random Forest classifier becomes more resilient to overfitting and robust against noise and outliers (Fawagreh et al., 2014).

As opposed to LDAs, Random Forests can additionally learn non-linear decision rules. A single decision tree may split data iteratively by using "rectangular" (in the case of two features) regions, in which the dominant class within the final regions represents the prediction of a single tree. When combining multiple of these decision rules, highly complex decision boundaries are formed that go beyond the linear hyperplanes of LDAs. Random Forests have performed in the top 10 of the recent cross-session pBCI challenge (Roy et al., 2022)

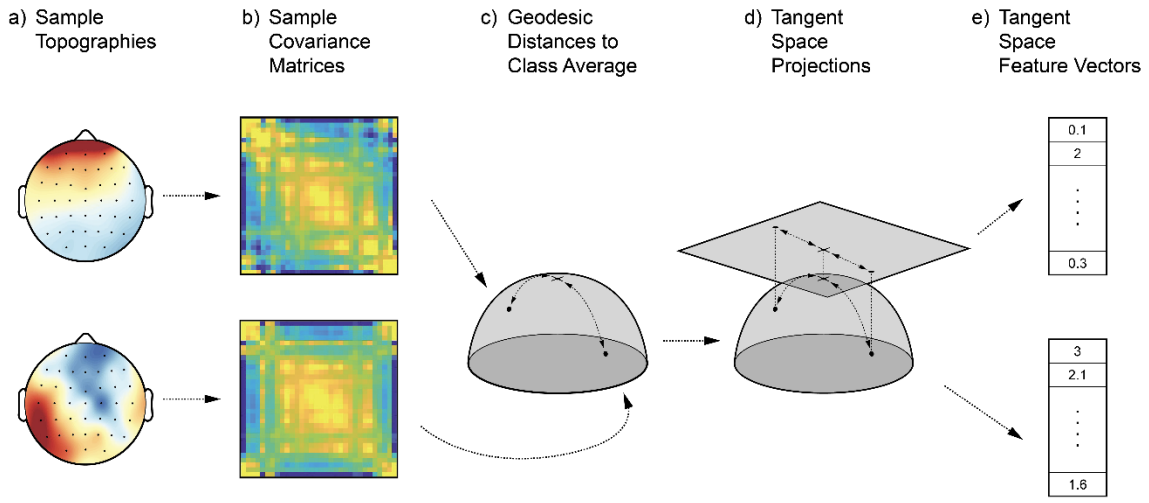
Figure 18. Schematic of a Random Forest Model



Note. Example of how data is funnelled through individual decision trees of a random forest to yield a prediction.

2.7.3 *Riemannian Geometry*

Figure 19. Schematic of Riemannian Classification



Note. Schematic of how N by N sensor-space covariance gets mapped onto a convex Riemannian space, and subsequently mapped onto the tangent space of a previously computed Riemannian average, resulting in feature vectors of size N*(N+1) / 2 that maintain the curved distance relationships to be used with traditional Euclidean classification models.

Riemannian geometry-based classifiers have become a staple in the BCI literature in recent years and performed well in a number of BCI competitions and benchmarking studies (Chevallier et al., 2024a; Congedo et al., 2017; Lotte et al., 2018; Roy et al., 2022; Yger et al., 2017). A major benefit of Riemannian classification is its robustness when dealing with non-normal or noisy data distributions (Congedo et al., 2017). The minimum distance between sensor-space covariance matrices for BCI

purposes represents the simplest Riemannian classifier. Covariance matrices are symmetric positive definite (SPD) and reside on a differentiable Riemannian manifold. This space is inherently curved, and two points on this surface (for example, two covariance matrices) can be connected by an optimal curved line named a geodesic. Their Riemannian mean lies in the centre of this geodesic. While minimum distance-based classification using the geodesic distances to compute class means and assign class labels by a simple nearest-neighbour (i.e., by finding which class's mean is closest to a new sample; see Figure 19, c.) is already a very effective BCI approach (Barachant et al., 2010; Barachant & Congedo, 2014), it cannot leverage the full discriminative power of more complex models. Tangen Space mapping allows for the projection of the data from a curved surface onto a flat, Euclidean one (Figure 19, d.). Here, a reference point (like the Riemannian mean of the reference class) is used to project all points on the manifold back onto a flat plane. This "flattening" of the manifold at a single point linearises the problem, allowing the use of simpler, well-known Euclidean classifiers. The resulting feature vectors preserve the advantageous distance relationships of the convex space (Congedo et al., 2017) but now allow for more complex decision functions like those from the Random Forest classifier.

The simplest Riemannian classifier for BCI purposes is based on the minimum distance between sensor-space covariance matrices, but more advanced approaches have also been designed to contain phase differences in event-related responses (Congedo et al., 2013; Ladouce et al., 2024), cross-frequency coupling information (Yamamoto et al., 2023), or relationships between HbO and HbR measures of neurovascular coupling (Näher et al., 2024).

The flexibility and robustness of Riemannian geometry methods extend beyond classification, making them a powerful and versatile toolset for BCI data analysis, from feature engineering to artefact management. Distance and dispersion metrics of covariance matrices on the Riemannian manifold have been shown to be efficient artifact rejection (Barthelemy et al., 2019; Blum et al., 2019) and channel/feature selection tools (Barachant & Bonnet, 2011; Lotte & Jeunet, 2018; Roy et al., 2014; Yamamoto et al., 2022).

3 Group-level Analysis of Mental Workload Metrics

This chapter aims to compare the four datasets analysed throughout the thesis with the broader literature on mental workload through a number of group-level analyses. What follows is an overview of commonly studied metrics of mental workload, ranging from subjective to performance to physiological. The reported statistics furthermore offer an opportunity to interrogate the consistency of neurophysiological metrics across EEG montages and provide context for the classification analyses of the later chapters.

In the long history of mental workload research, continuous monitoring can be considered a relatively recent trend. Most research was based on group-level inference, which attempts to estimate population-level effects. This could pose an issue when researchers expect to find one-to-one mappings between widely reported mental workload effects in the literature and subject-level classifier performance (i.e. if the EEG alpha band shows strong workload effects in the population, we expect classifiers based on alpha features to produce high accuracy in a continuous monitoring setting). However, subject-level variance tends to be much higher than what group-level results suggest (Fisher et al., 2018), and averaging over individual differences, especially in neuropsychological measures (Höller et al., 2013), may paint a misleading picture (Höller et al., 2019; Mahini et al., 2024; Wijk et al., 2021) as the idealised “average” brain cannot be found in nature. When it comes to neuroadaptive interfaces, neurophysiological metrics often reported to show strong effects in group-level analyses may not offer the best performance for any given individual operator.

3.1 Subjective and Performance Metrics

Subjective ratings of mental workload and performance metrics are commonplace in the literature. Subjective ratings provide important context to neurophysiological metrics, as the subjective experience of workload could be regarded as the “ground-truth” that other metrics try to approximate. Performance-based metrics, like subjective ratings, provide contextual information about operator workload in pBCI research. However, they could also be included in closed-loop systems themselves, provided a) the system requires frequent input (manual vs automated driving) and b) that performance degradation due to increases in mental workload is not regarded an immediate safety risk itself. More often than providing direct information to the adaptation, performance metrics tend to be used as a means to evaluate the benefits of an adaptive system, as the adaptations are usually aimed at stabilising/improving performance (Grubov et al., 2024; Valeriani et al., 2022; Zander et al., 2016).

In the current chapter, subjective ratings and performance metrics were treated as manipulation checks.

3.1.1 Subjective ratings

All four experiments employed two tasks, each with three levels of task-load. It was expected that subjective effort ratings would increase with increasing task-load. To analyse the effects of our task-load manipulation on subjective workload, a mixed factorial design with both within- and between-subject factors was used. Per dataset, participants were entered into a linear mixed-effect model (lme4 version: 1.1-33).

```
Formula: RSME ~ Task * Condition *  Repetition + (1 | subject)
```

The variables task, condition and repetition (set number) were coded as factors with condition and repetition treated as ordered variables with the easiest task-load level and first set as reference level, respectively. The factor subject (their participant-IDs) was treated as a random intercept to account for repeated measures. Type III sums of squares were used to test fixed effects with Satterthwaite's approximation for degrees of freedom.

Effect sizes were estimated using eta squared. Post hoc pairwise comparisons were conducted using estimated marginal means (EMMs) with Tukey adjustment for multiple comparisons via the emmeans package (version: 1.10.5).

Table 6. Pilot Data RSME Model

Effect	F-statistic	DoF	p-value	η^2_p
Task	0.92	(1, 323)	0.34	<0.01
Condition	92.34	(2, 323)	<0.001***	0.36
Repetition	0.03	(2, 323)	0.98	<0.01
Task x Condition	1.89	(2, 323)	0.15	0.01
Task x Repetition	2.15	(2, 323)	0.12	0.01
Condition x Repetition	0.3	(4, 323)	0.87	<0.01
Task x Condition x Repetition	0.52	(4, 323)	0.72	<0.01

Table 7. Lab-grade Data RSME Model

Effect	F-statistic	DoF	p-value	η^2_p
Task	30.81	(1, 323)	<0.001***	0.09
Condition	175.65	(2, 323)	<0.001***	0.52
Repetition	3.26	(2, 323)	0.04*	0.02
Task x Condition	1.82	(2, 323)	0.16	0.01
Task x Repetition	0.05	(2, 323)	0.95	<0.01
Condition x Repetition	0.18	(4, 323)	0.95	<0.01
Task x Condition x Repetition	0.76	(4, 323)	0.55	<0.01

Table 8. Wearable Data RSME Model

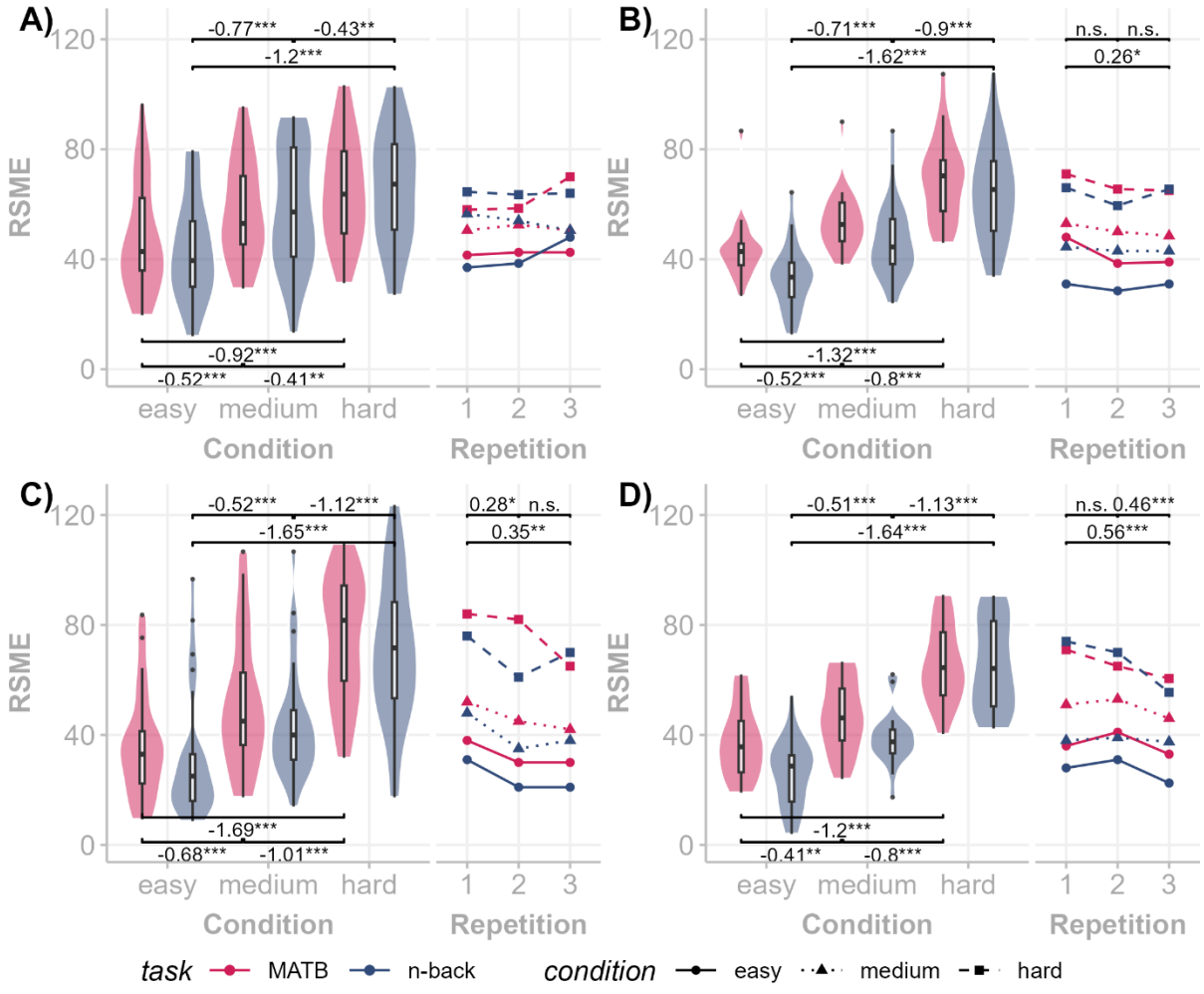
Effect	F-statistic	DoF	p-value	η^2_p
Task	10.86	(1, 340)	0.001**	0.03
Condition	242.73	(2, 340)	<0.001***	0.59
Repetition	5.76	(2, 340)	0.003**	0.03
Task x Condition	0.53	(2, 340)	0.59	<0.01
Task x Repetition	0.07	(2, 340)	0.93	<0.01
Condition x Repetition	0.11	(4, 340)	0.98	<0.01
Task x Condition x Repetition	0.91	(4, 340)	0.91	<0.01

Table 9. Multimodal Data RSME Model

Effect	F-statistic	DoF	p-value	(η^2_p)
Task	16.80	(1, 300)	<0.001***	0.05
Condition	158.23	(2, 300)	<0.001***	0.51
Repetition	13.38	(2, 300)	<0.001***	0.08
Task x Condition	3.93	(2, 300)	0.02*	0.03
Task x Repetition	0.79	(2, 300)	0.45	<0.01
Condition x Repetition	2.01	(4, 300)	0.09	0.03
Task x Condition x Repetition	0.08	(4, 300)	0.98	<0.01

Throughout all four datasets, task-load condition exhibited significant effects on subjective ratings of mental workload, with effect sizes ranging from 36% - 59% of variance being explained by condition labels alone. Task differences were significant in all but the Pilot dataset, likely due to the changes undertaken to the memory-load condition of the n-back after completion of the Pilot data collection. The MATB task was on average rated 6.9 points higher in the Lab-grade ($t(323) = 5.55$, $p < .001$, $d = 0.62$), 4.7 points higher in the Wearable ($t(340) = 2.30$, $p = .001$, $d = 0.36$), and 6.56 points higher in the Multimodal dataset ($t(300) = 3.10$, $p < .001$, $d = 0.47$). Time on task effects (here measured by the factor repetition, which corresponds to the three sets of condition repetitions throughout the 90-minute-long data collection) were also significant in all but the Pilot dataset, albeit while only explaining small fractions of RSME variance. Even though partial eta squared measures were smaller, the Cohen's d of the pair-wise comparisons, comparing the first with the last condition-repetitions, showed moderately sized effects (see Figure 20). Of the tested interaction effects, only the Multi-modal dataset demonstrated a small but significant interaction between task type and task load condition. This interaction was likely influenced by the modest effect size of the difference score between easy and medium task load levels in the MATB. While the n-back showed nearly a linear increase in RSME ratings, the MATB's increase was initially shallow and then became much steeper from medium to hard. No other interaction effects were significant. (see Tables 6-9)

Figure 20. RSME Results



Note. Four plots showing the RSME distribution per task and task-load condition for the pilot (A), the Lab-grade (B), the Wearable (C) and the Multi-modal (D) experiments. Next to the distributions, the median RSME values across participants for the condition repetitions are plotted. Cohen's d effect sizes of pairwise comparisons are reported within tasks for condition differences and across tasks for repetition differences.

3.1.2 MATB Performance

All four experiments employed three levels of MATB task-load. It was expected that performance in the MATB would decline with increasing task-load levels. The performance on the MATB was analysed using the same mixed factorial design as used for the RSME ratings.

To simplify the analysis, MATB performance was estimated using the tracking task alone. This decision was taken because a) the tacking task offered a continuous measure that could be assessed throughout the entire task, and b) tracking lent itself to more straightforward interpretations than the other subtasks – i.e. a reduction in tracking accuracy suggested the operator could not focus on the tracking subtask.

An additional treatment of the reaction times or tendencies for false positives could also be of interest when analysing system-monitoring performance. However, as an exhaustive MATB subtask analysis was not of primary interest to this thesis, we have opted to use the tracking performance as a proxy of general MATB performance.

A linear mixed-effects model was used to evaluate the effects of task-load condition, experiment, and condition repetition on MATB performance measured as the proportion of time the tracking task's crosshair was within the target zone. To assure approximate normality of residuals, the tracking score was logit transformed. The model included fixed effects for task-load `condition` and condition `repetition`, along with their interactions and a random intercept per subject.

$$\texttt{Formula: Logit(Tracking}_{\texttt{onTarget}}\texttt{) ~ condition * repetition + (1 | subject)}$$

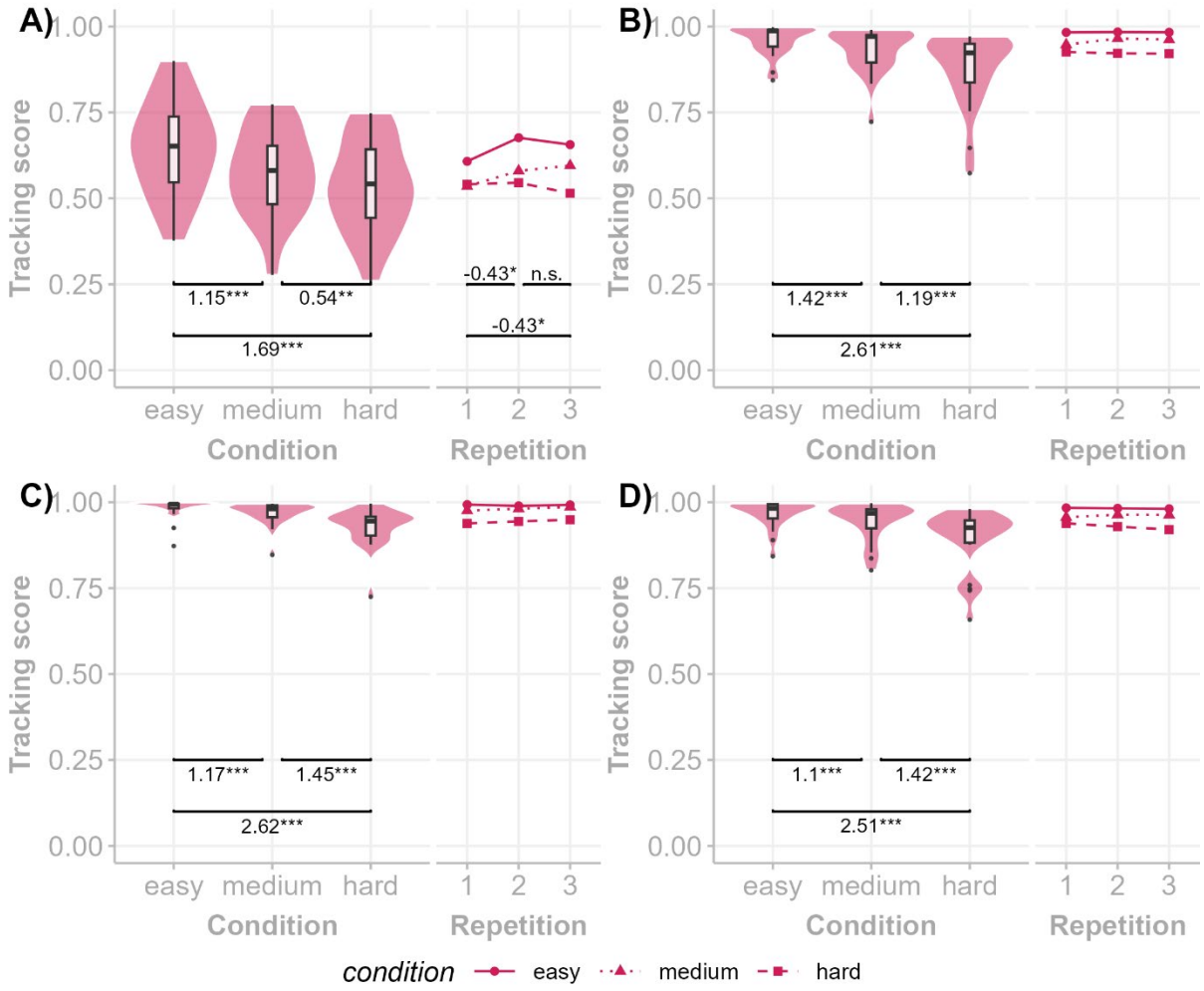
Table 10. MATB Performance Results

Experiment	Effect	F-statistic	DoF	p-value	η^2_p
Pilot	Condition	56.69	(2, 152)	<0.001***	0.43
	Repetition	4.72	(2, 152)	0.01*	0.06
	Condition x Repetition	0.40	(4, 152)	0.81	0.01
Lab-grade	Condition	129.40	(2, 152)	<0.001***	0.63
	Repetition	2.34	(2, 152)	0.1	0.03
	Condition x Repetition	0.59	(4, 152)	0.67	0.02
Wearable	Condition	137.83	(2, 160)	<0.001***	0.63
	Repetition	1.10	(2, 160)	0.34	0.01
	Condition x Repetition	0.25	(4, 160)	0.91	<0.01
Multimodal	Condition	108.58	(2, 137.05)	<0.001***	0.61
	Repetition	2.51	(2, 137.26)	0.08	0.04
	Condition x Repetition	0.91	(4, 137.05)	0.46	0.03

The task-load condition strongly affected tracking accuracy across datasets (Table 10). Comparing the Pilot to the other datasets changes to task-load settings increased the effect size of the condition factor. Additionally, all datasets following the Pilot dataset exhibited more than double the effect size

between the easiest and hardest task-load levels, suggesting the changes to the MATB's task-load settings were successful. Time on task was only significant in the Pilot dataset (Figure 21, A.). The absence of significant repetition effects may have been due to the updated training regime spending more time on the MATB. Subsequently, participants exhibited less of a training effect during data recording. A ceiling effect due to the larger tracking radius may, however, also have been the reason for the reduced time on task effects.

Figure 21. MATB Tracking Performance



Note. Four plots showing the Tracking score distribution per task-load condition for the Pilot (A), the Lab-grade (B), the Wearable (C) and the Multimodal (D) experiments. Next to the distributions, the median tracking scores across participants for the condition repetitions are plotted.

3.1.3 N-back Performance

All four experiments employed three levels of n-back memory load. Performance accuracy was expected to decline with increasing memory load. Performance on the n-back task was analysed using the same mixed factorial design as for the MATB and RSME.

The dependent variable analysed was d-prime, a measure from signal detection theory that assesses the sensitivity in distinguishing between signal (target stimuli) and noise (non-target stimuli). It offers a combined assessment of the participant's hit-rate of true target stimuli and false alarm rate in response to non-target stimuli (Haatveit et al., 2010; Meule, 2017) and is calculated as $d' = Z_{Hit} - Z_{FA,}$ with Hit being the proportion of correctly detected targets and FA the proportion of false alarms in non-target trials (Macmillan & Creelman, 1990). To prevent the z-transforms of extreme hit or FA rates from resulting in infinite values, they were adjusted by replacing 0 with 0.5 / n, and 1 with (n – 0.5) / n (termed the "conventional" fourth approach in Stanislaw & Todorov, 1999)

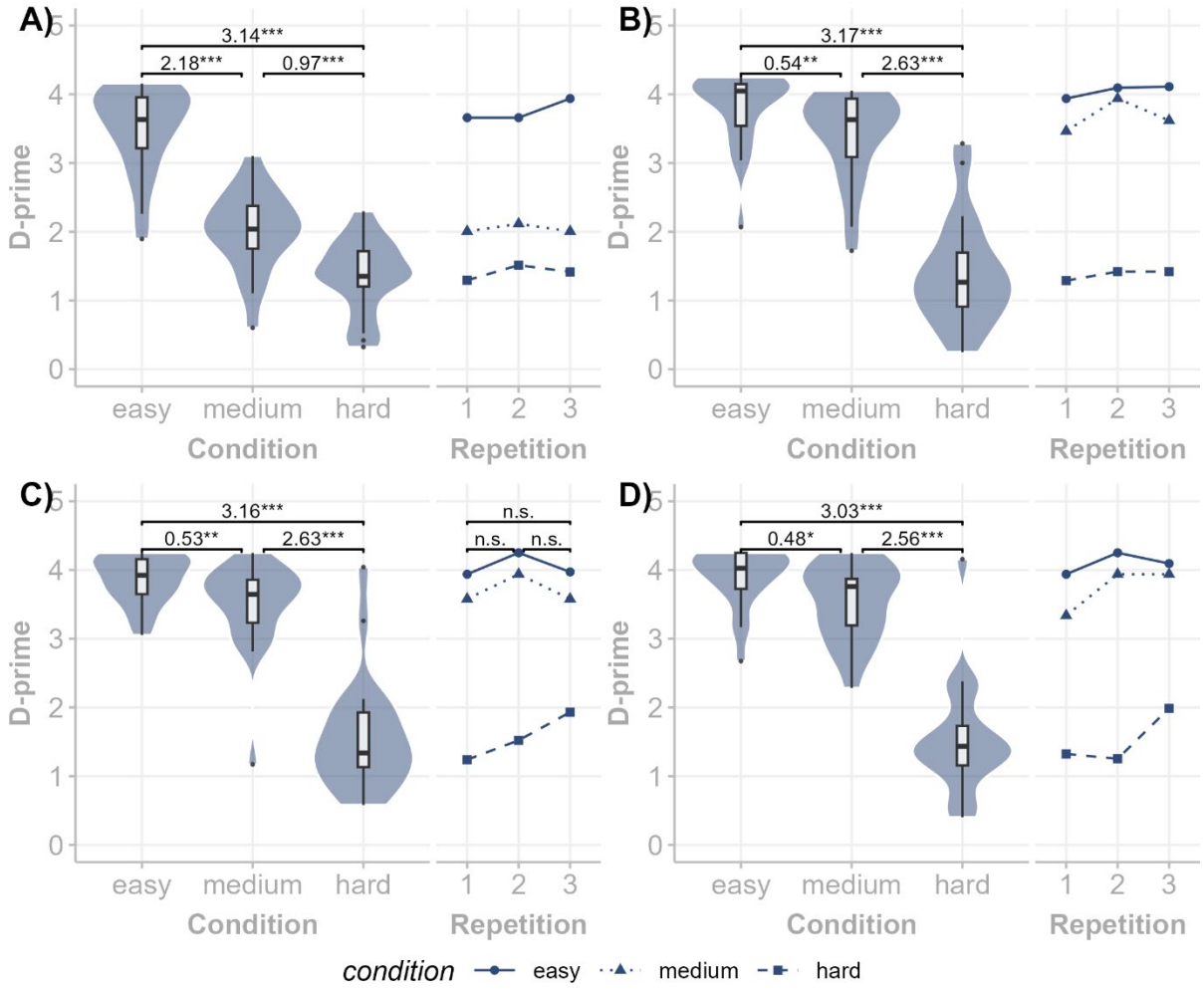
```
Formula: d-prime ~  condition *  repetition + (1 | subject)
```

Table 11. N-back Performance Results

Experiment	Effect	F-statistic	DoF	p-value	η^2_p
Pilot	Condition	71.84	(2, 154.21)	<0.001***	0.72
	Repetition	0.92	(2, 153.51)	0.08	0.03
	Condition x Repetition	0.16	(4, 153.51)	0.78	0.01
Lab-grade	Condition	219.1	(2, 152)	<0.001***	0.74
	Repetition	1.16	(2, 152)	0.32	0.02
	Condition x Repetition	0.31	(4, 152)	0.86	<0.01
Wearable	Condition	229.35	(2, 160)	<0.001***	0.74
	Repetition	3.38	(2, 160)	0.04*	0.04
	Condition x Repetition	0.48	(4, 160)	0.32	0.03
Multimodal	Condition	179.83	(2, 135.54)	<0.001***	0.73
	Repetition	1.30	(2, 137.62)	0.27	0.02
	Condition x Repetition	1.69	(4, 135.54)	0.15	0.05

Task-load condition exhibited strong effects on d-prime in line with previous literature assessing d-prime (Table 11) in the context of n-backs (Haatveit et al., 2010; Meule, 2017). Time-on-task was only significant for the wearable dataset, without any of the post-hoc comparisons exhibiting significant differences between the condition repetitions (see Figure 22 C.). None of the tested interactions between task-load level and condition repetitions were significant.

Figure 22. N-back Performance Results



Note. Four plots showing the Tracking score distribution per task-load condition for the Pilot (A), the Lab-grade (B), the Wearable (C) and the Multimodal (D) experiments. Next to the distributions, the median d-prime across participants for the condition repetitions is plotted.

3.1.4 Interim Summary

The effects of time on task and task-load condition on subjective mental workload ratings and performance measures suggested successful task-load manipulations, showing strong effect sizes of task-load conditions across the three models in all datasets. RSME ratings showed task differences after the changes to the paradigm from the pilot study, with the MATB being rated more difficult than the n-back. The changes also seemed to have increased the effect sizes of task-load condition on subjective mental workload ratings, suggesting the changes to the task-load levels did indeed widen the gap in subjective effort between the lowest and highest task-load levels. Performance in the n-back and MATB significantly decreased with increasing workload across all datasets and condition contrasts.

Time-on-task effects, which could be interpreted as effects of learning or fatigue, were observed in the subjective ratings, which decreased over time. While significant, the small effect sizes on the omnibus tests in all but the RSME ratings of the Multimodal dataset suggested that the influence of time-on-task on subjective ratings and performance metrics was negligible.

No interaction terms of condition and time-on-task were significant, indicating there was no evidence of learning effects across the recording session that influenced specific task-load levels more than others.

3.2 ECG Metrics

The previously extracted R-peaks were used here to test whether the mental workload manipulations sufficed to induce the expected increases in heart rate and decreases in RMSSD. Blocks with inaccurate R-peak detection results due to noisy ECG traces were excluded from the analysis (see Figure 15). To further minimise the impact of suboptimal R-peak detection on the analysis, both dependent variables were calculated using a 10-second sliding window with a 1 R-peak step size (Laborde et al., 2017). Subsequently, outliers were removed with a cut-off of ±1.5 times the interquartile range of the RMSSD values. The final statistics per block were the median heart rate and RMSSD of the remaining measurement windows.

Only 18 of the 20 participants in the Pilot datasets could be included in this analysis, as one participant's recording was missing event markers, and the other's ECG trace was corrupted.

3.2.1 Heart-rate

Per experiment, a linear mixed-effects model was used to evaluate the effects of task type, task-load condition, and condition repetition on the participants' heart rate. The model included their two-way interactions of all fixed factors and a random intercept for subject-ID. Both the condition repetition and the task-load condition were coded as ordered factors.

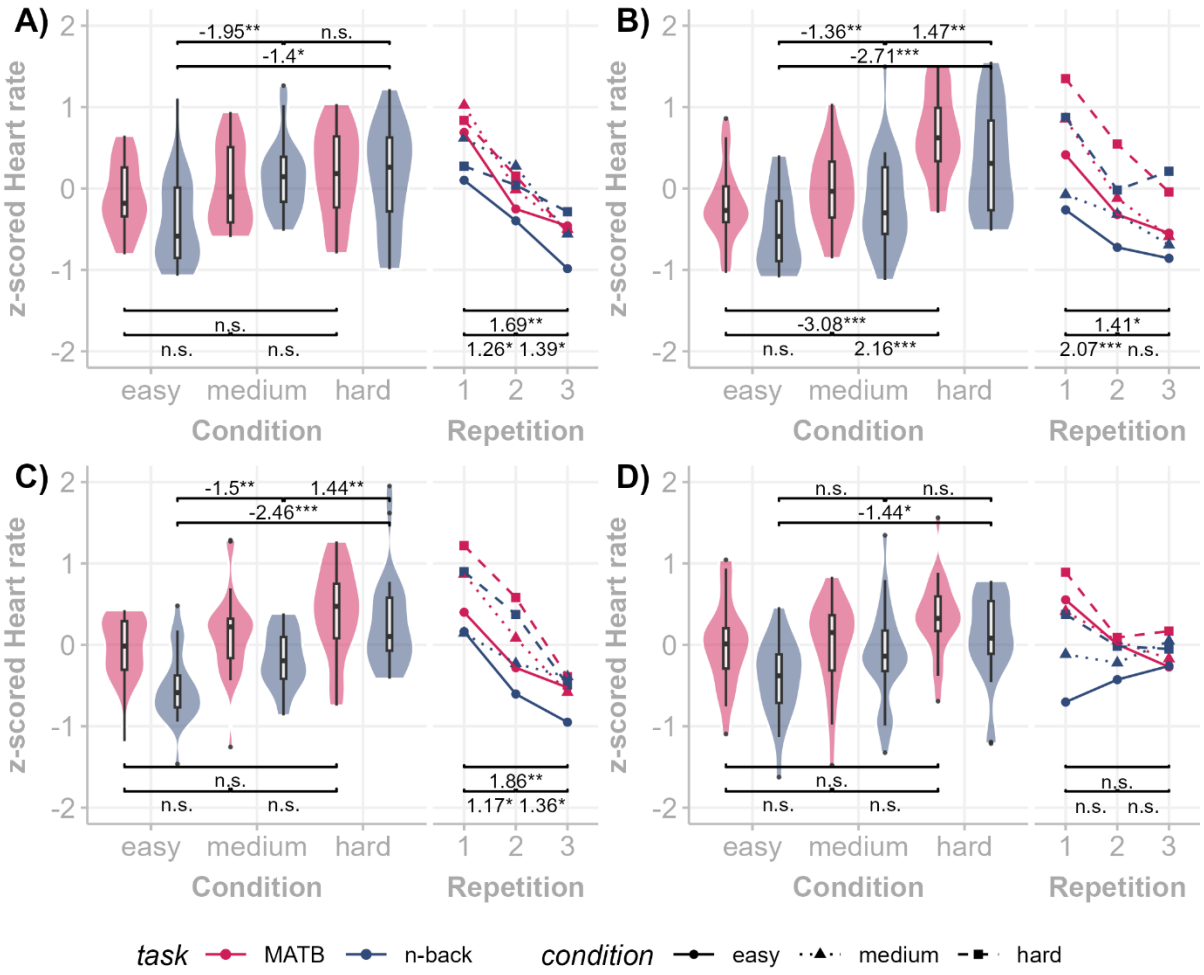
```
Formula: Heart rate ~ (Task + Condition + Repetition)^2 + (1 | subject)
```

Statistical significance was again assessed via Type III sums of squares with Satterthwaite's approximation for degrees of freedom, and partial eta squared effect sizes were computed.

Table 12. Pilot Heart-Rate Results

Effect	F-statistic	DoF	p-value	η^2_p
Task	0.70	(1, 292)	0.40	<0.01
Condition	5.25	(2, 292)	0.006**	0.03
Repetition	27.46	(2, 292)	<0.001***	0.16
Task x Condition	1.19	(2, 292)	0.31	0.01
Task x Repetition	3.02	(2, 292)	0.05	0.03
Condition x Repetition	0.56	(4, 292)	0.69	0.01

Figure 23. Heart-Rate Results



Note. Four plots showing the distribution of within-subject normalised heart-rates per task and task-load condition for the Pilot (A), the Lab-grade (B), the Wearable (C) and the Multimodal (D) experiments. Next to the distributions, the median normalised heart-rate across participants for the condition repetitions are plotted. Cohen's dz effect sizes of pairwise comparisons are reported within tasks for condition differences and across tasks for repetition differences.

In the Pilot dataset Condition effects and Repetition effects were significant (Table 12). However, paired-comparisons exhibited no significant differences between MATB conditions, suggesting the condition effect was driven by the differences between the 1-back and remaining n-back conditions (see Figure 23 panel A). Heart-rate further significantly decreased across both tasks with increasing condition repetition numbers (see Figure 23 panel A).

After adaptation of the task-load condition manipulations, the Lab-grade and Multimodal datasets exhibited additional task differences. Higher heart rates in the MATB (Lab-grade: $t(318.02) = 3.88$ $p < .001$; Multimodal: $t(279.07) = 2.75$, $p = .0064$), in addition to a stronger effect size for the condition level (see Table 13 and Table 15). Further, the interaction of task and condition repetition was significant in both datasets. In both cases, the interaction of the linear contrast of repetition and task suggested a less extreme reduction in heart rate over time in the n-back compared with the MATB (see Figure 23, panels B and D).

In the Wearable dataset, the fixed factors of condition and repetition were significant, while the aforementioned interaction or task differences were not (see Table 14).

Table 13. Lab-grade Heart-Rate Results

Effect	F-statistic	DoF	p-value	η^2_p
Task	16.07	(1, 318.02)	<0.001***	0.05
Condition	22.18	(2, 318)	<0.001***	0.13
Repetition	17.59	(2, 318.05)	<0.001***	0.10
Task x Condition	0.01	(2, 318)	0.16	<0.01
Task x Repetition	5.78	(2, 318.02)	0.003**	0.04
Condition x Repetition	0.46	(4, 318)	0.95	<0.01

Table 14. Wearable Heart-Rate Results

Effect	F-statistic	DoF	p-value	η^2_p
Task	3.25	(1, 325.08)	0.07	<0.01
Condition	12.47	(2, 325.01)	<0.001***	0.07
Repetition	34.85	(2, 325.01)	<0.001***	0.18
Task x Condition	2.72	(2, 325.01)	0.06	0.02
Task x Repetition	2.94	(2, 325.01)	0.05	0.01
Condition x Repetition	0.57	(4, 325.01)	0.68	0.01

Table 15. Multimodal Heart-Rate Results

Effect	F-statistic	DoF	p-value	η^2_p
Task	7.55	(1, 279.07)	0.006**	0.03
Condition	4.46	(2, 279.02)	0.012*	0.03
Repetition	2.43	(2, 279.03)	0.089	0.02
Task x Condition	0.47	(2, 279.02)	0.62	<0.01
Task x Repetition	3.93	(2, 279.03)	0.021*	0.03
Condition x Repetition	0.13	(4, 279.01)	0.97	<0.01

3.2.2 Heart Rate Variability

Per experiment, a linear mixed-effects model was used to evaluate the effects of task type, task-load condition, and condition repetition on the participants' heart rate variability. The model included their two-way interactions of all fixed factors and a random intercept for subject-ID. Both the condition repetition and the task-load condition were coded as ordered factors.

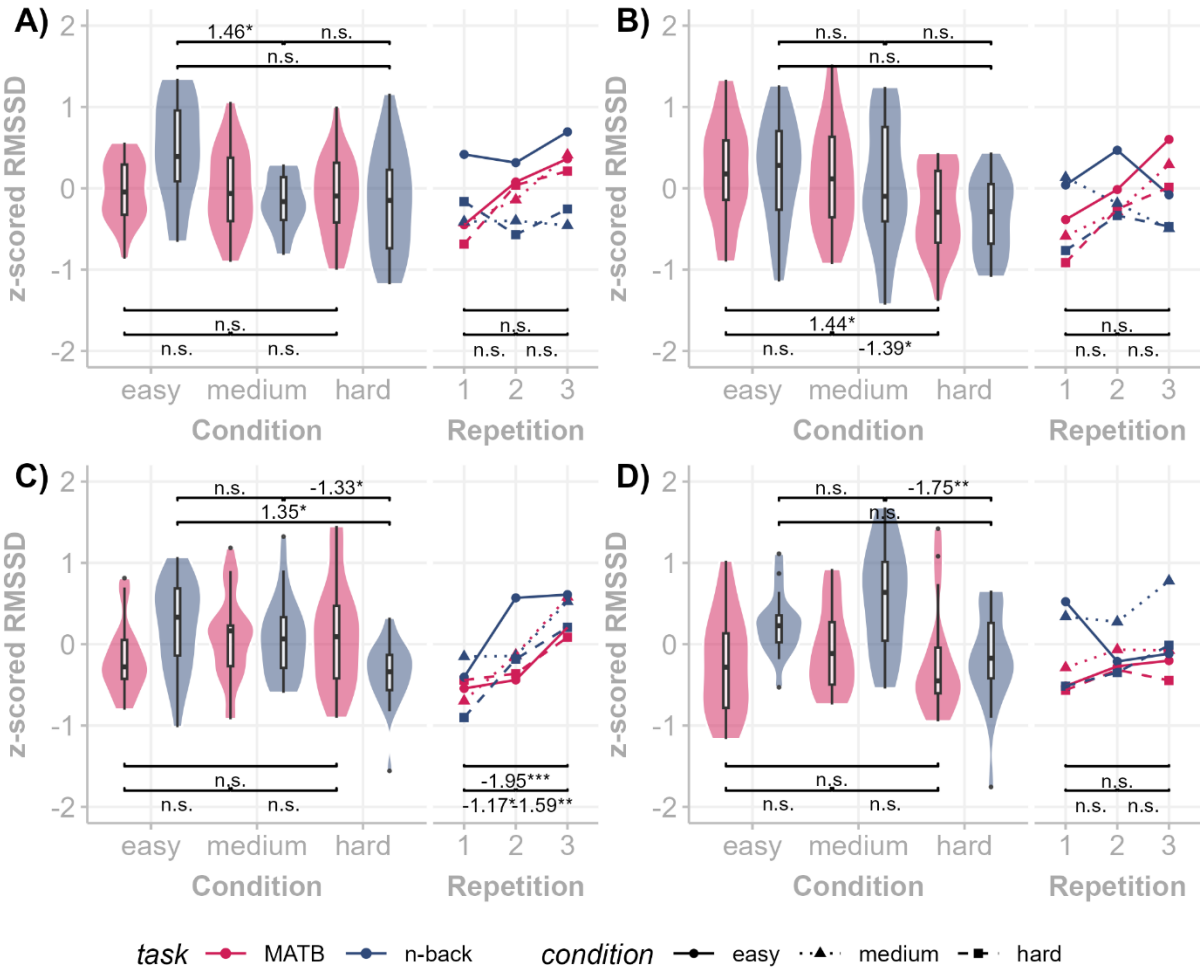
```
Formula: RMSSD ~  (Task + Condition +  Repetition)^2 + (1 | subject)
```

Statistical significance was assessed via Type III sums of squares with Satterthwaite's approximation for degrees of freedom, and partial eta squared effect sizes were computed.

Table 16. Pilot RMSSD Results

Effect	F-statistic	DoF	p-value	η^2_p
Task	1.52	(1, 292)	0.22	<0.01
Condition	1.48	(2, 292)	0.23	0.01
Repetition	7.95	(2, 292)	0.004**	0.05
Task x Condition	1.43	(2, 292)	0.24	<0.01
Task x Repetition	0.03	(2, 292)	0.97	<0.01
Condition x Repetition	0.39	(4, 292)	0.81	<0.01

Figure 24. RMSSD Results



Note. Four plots showing the distribution of within-subject normalised RMSSD scores per task and task-load condition for the Pilot (A), the Lab-grade (B), the Wearable (C) and the Multimodal (D) experiments. Next to the distributions, the median normalised RMSSD score across participants for the condition repetitions are plotted. Cohen's dz effect sizes of pairwise comparisons are reported within tasks for condition differences and across tasks for repetition differences.

In the Pilot dataset, only the repetition effect was significant (Table 16). Paired comparisons exhibited no significant differences between the three repetitions, even though a clear trend was visible in Figure 24 (panel A).

Table 17. Lab-grade RMSSD Results

Effect	F-statistic	DoF	p-value	η^2_p
Task	0.36	(1, 317.97)	0.55	<0.01
Condition	5.97	(2, 317.93)	0.002**	0.04
Repetition	4.89	(2, 318.05)	0.008**	0.03
Task x Condition	0.53	(2, 317.93)	0.59	<0.01
Task x Repetition	4.39	(2, 317.97)	0.013*	0.03
Condition x Repetition	0.26	(4, 317.93)	0.26	<0.01

Table 18. Wearable RMSSD Results

Effect	F-statistic	DoF	p-value	η^2_p
Task	2.47	(1, 325.17)	0.12	<0.01
Condition	4.10	(2, 325.05)	0.017*	0.02
Repetition	23.28	(2, 325.03)	<0.001***	0.13
Task x Condition	1.20	(2, 325.05)	0.30	<0.01
Task x Repetition	1.46	(2, 325.03)	0.23	<0.01
Condition x Repetition	0.93	(4, 325.02)	0.45	0.01

After adaptation of the task-load condition manipulations, the Lab-grade, Wearable, and Multimodal dataset exhibited condition differences (Tables 17-19). Paired comparisons showed different contrasts reaching significance across the experiments. The Lab-grade data contained the only significant MATB contrast comparing the easy with the hard condition. For the Wearable data, the 1 vs 3-back and 0 vs. 3-back contrasts showed a significant reduction in RMSSD (see Figure 24, panels B and C). For the Multimodal dataset, this was only significant between the 1-back and 3-back (see Figure 24, panel D).

The Multimodal data contained an additional significant task difference with the MATB having induced lower RMSSD than the n-back ($t(279.23) = 3.65$, $p < .001$). Furthermore, the Multimodal data was the only dataset without significant repetition effects (Table 19).

Table 19. Multimodal RMSSD Results

Effect	F-statistic	Degrees of freedom	p-value	Partial eta squared (η^2_p)
Task	12.01	(1, 279.21)	0.006**	0.03
Condition	4.53	(2, 279.03)	0.012*	0.03
Repetition	0.08	(2, 279.06)	0.089	0.02
Task x Condition	0.46	(2, 279.04)	0.62	<0.01
Task x Repetition	0.45	(2, 279.05)	0.021*	0.03
Condition x Repetition	0.26	(4, 279.01)	0.97	<0.01

3.2.3 Exploratory Rest/Task Analysis

The reasons for the inconsistent effects observed across various experiments were likely complex and multifaceted. However, an appropriately sensitive metric should not exhibit such variability in response to changes in demographics or environmental conditions. To verify whether the heart-rate and RMSSD metrics truly captured mental workload-related information, and to assess their likely lack of sensitivity in detecting differences between the active task-load conditions used in the four experiments, a comparison of the multimodal dataset's 30-second pre-task rest periods, which were introduced as possible baseline periods for the fNIRS data, and the task-load conditions follows.

For each 30-second rest period, which preceded each active task period, R-peak detection was carried out just like described in 2.4.1. As before, some ECG traces were too noisy for accurate R-peak detection, resulting in the removal of one, two or four rest periods (two, three, and one participant, respectively). Blocks without a matching rest period were removed from the analysis.

If differences between rest and task-load conditions interact significantly, the statistics of interest were the marginal means tests for differences between the rest and active task periods per task and task-load condition. Since the task-load condition should not have affected the rest periods at all, the factor 'Rest' was additionally entered as a random slope to account for these differences in variance.

```
Formular: Metric ~ (Task + Condition +  Repetition + Rest)^2 + (1 + Rest |
subject)
```

For the heart rate test, the expected effect of the fixed factor 'Rest' was significant (F(1,522.01) = 127.64, p < 0.001, η^2_p = .20). Planned pair-wise comparisons for the MATB showed that heart rate during rest periods was on average 4.13 beats per second slower for easy (t(522) =-6.34, p <0.001, d = -0.56), 3.83 for medium (t(522) =-5.75, p <0.001, d = -0.5), and -5.61 for hard blocks (t(522) =-8.1, p < .001, d = -0.71). Planned pair-wise comparisons for the n-back showed that heart rate during rest periods was on average -2.06 beats per second slower for 0-backs (t(522) =4.06, p <0.001, d = -0.35), 2.3 for 1-backs (t(522) =-3.45, p < .001, d = -0.30) and 4.1 for 3-backs (t(522) =-6.18, p < .001, d = -0.54).

For the RMSSD test, the previously reported task effect remained consistent in this updated model (F(1, 522.56) = 8.44, p = .003,η^2_p = .02). More interestingly, the expected effect of the fixed factor 'Rest' was also significant (F(1,521.99) = 79.1, p = .004, η^2_p = .13). Planned pair-wise comparisons for the MATB showed that RMSSD during rest periods was on average 12.7 points higher for easy (t(522) =4.56, p <0.001, d = 0.40), 10 for medium (t(522) =3.5, p <0.001, d = 0.31), and 18.4 for hard blocks (t(522) =6.2, p < .001, d = 0.37). Planned pair-wise comparisons for the n-back showed that RMSSD during rest periods was on average 10.7 points higher for 0-backs (t(522) =3.85, p <0.001, d = .34), 8 for 1-backs (t(522) =2.78, p = .006, d = .07) and 16.3 for 3-backs (t(522) =5.78, p < .001, d = 0.51).

Notably, the effect sizes of the previous tests were Cohen's d, which appear smaller than the reported effect sizes in the earlier figures (Figures 19-23), which were Cohen's dZ computed from paired differences.

3.2.4 Interim Summary

Heart rate and heart rate variability as indices of parasympathetic adaptations to mental workload exhibited task-load condition-related effects across all datasets (with the single exception of RMSSD in the Pilot data). However, unlike heart rate, which increased significantly for most pair-wise n-back contrasts (Figure 22), RMSSD only exhibited very sporadic pair-wise significant differences. Additionally, both metrics appeared insensitive to task-load differences in the MATB (with the exception of the Lab-grade data), indicating that the easy MATB condition already saturated either metric's sensitivity within the laboratory context. The exploratory comparisons between rest and on-task durations of the Multimodal data, where both metrics exhibited significant differences, further supported this statement. While these differences increased with increasing task-load conditions between the rest periods and the on-task periods, the differences between conditions tended to be too small for the statistical power of the tests employed here.

Time-on-task effects were large and may have been caused by a combination of fading nervousness and habituation to the tasks and the laboratory environment. The lack of a significant repetition effect for both metrics in the Multimodal dataset was surprising and may have been caused by the more extensive preparation time as the separate donning of the fNIRS optodes and EEG sensors required multiple location changes (from the prepping table to the task computer and back) and more elaborate explanations (to justify the lengthy prep time and possible uncomfortable optodes springs) This may have reduced nervousness before the recording more effectively compared to the other experiments prepping period in which the participant were solely sitting at the prepping table for an extended duration.

3.3 EEG Power Spectrum

As described in Chapter 1, mental workload affects different frequency ranges in the EEG's power spectrum. However, there are two shortfalls in the literature. Firstly, not all frequency ranges are reported on, suggesting null results tend to be omitted from publications. A recent meta-analysis suspected a file-drawer problem when examining publications investigating mental workload's effects on spectral power (Chikhi et al., 2022). The file-drawer problem describes the trend that null-results are disproportionately missing from the peer-reviewed literature, suggesting their omission rather than nonexistence, thereby painting a skewed picture of true population effects (Rosenthal, 1979). Secondly, traditional frequency power estimates have recently been criticised for their inaccuracies (Donoghue et al., 2022). Donoghue and colleagues (2022) described how effects that are usually interpreted to stem from changes in oscillatory power may also occur due to shifts in the oscillatory centre frequency, changes in broadband power, or changes to aperiodic activations (describing the exponent of the 1/f power distribution) that characterise EEG power spectra (Buzsáki et al., 2012). How these methodological considerations possibly affect commonly accepted workload effects should be investigated in more detail.

Here, we investigated whether a) the task-load conditions affected any of the canonical bandpower ranges, b) time on task effects on oscillatory power, and c) whether accounting for aperiodic activation in the EEG power spectrum changes the commonly reported relationships between workload and oscillatory power.

Beyond possibly improving cross-day classification (Ke et al., 2023), accounting for changes in the aperiodic component of the EEG frequency spectrum may invalidate some previously reported effects, which may have been spuriously ascribed to changes in oscillatory power of periodic components (Donoghue et al., 2022). Several software packages now offer the parameterisation of the power-spectral density function to estimate the 1/f slope in log-log space (Gerster et al., 2022).

While the benefits of accounting for changes in the aperiodic activity have been described aptly by Donoghue and colleagues (2022), discussions about ideal parameters, like frequency ranges, oscillatory bandwidth limits, or the fitting of platues and 'knees' in the EEG's power spectrum are ongoing (Gerster et al., 2022). Furthermore, the functional relevance of the exponent describing the slope of the aperiodic spectrum is currently not clearly defined in cognitive or neurophysiological terms. (Buzsáki et al., 2012; Gao, 2016; Gao et al., 2017; Kramer & Chu, 2023).

In this analysis, the effects of workload on the delta, theta, alpha and beta bands, as well as the aperiodic-slope were tested using the fully processed data. Tests were carried out once on raw PSD estimates computed per experimental block and once more on the same spectrum after the aperiodic component was subtracted. PSD estimates were computed with a 1Hz resolution using the Welch method (nfft = 250, window size = 250, window overlap = 125). The aperiodic component was estimated using the Fitting Oscillations and One-Over F (FOOOF) python implementation (version 1.1). The fitting algorithm was set to cover the 1-40Hz range as this avoided having to fit a "knee" in the log-log power spectrum, which could have resulted in worse overall model fits. Furthermore, the algorithm was set to ignore oscillatory peaks with bandwidth below 2 or above 8Hz and to fit maximally 4 peaks, inspired by previous work (Kałamała et al., 2024). All other settings were kept at their default values.

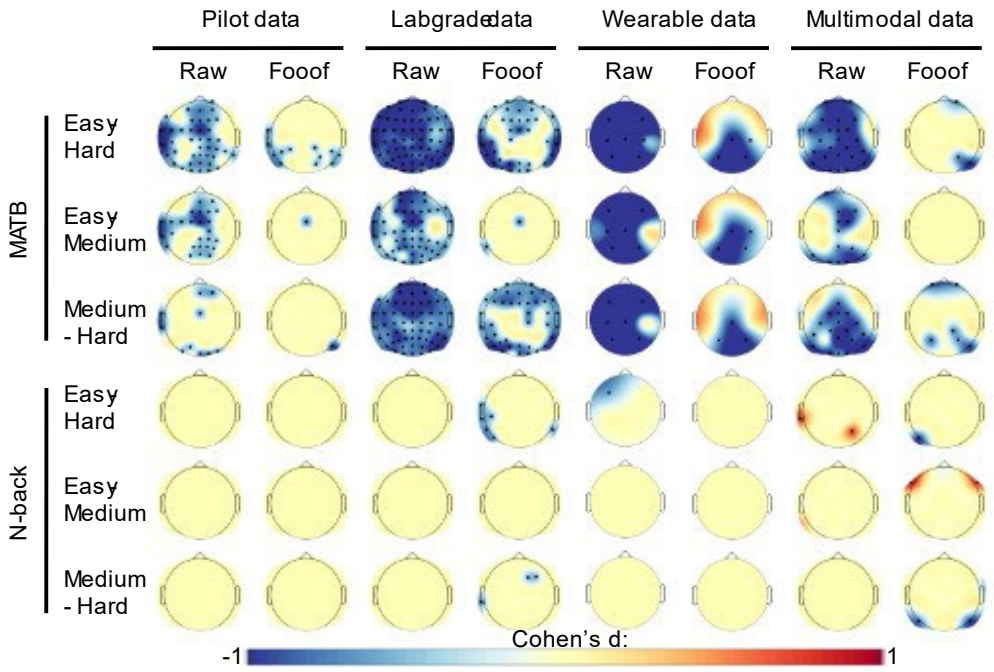
Effects of task load and time on task were tested for using a data-driven detection of significant effects across electrodes. Threshold-Free Cluster Enhancement (TFCE) is a non-parametric, cluster-based statistical method that enhances spatiotemporal differences without the a priori definition of cluster-defining thresholds (Mensen & Khatami, 2013; Smith & Nichols, 2009). Two parameters inform the TFCE's enhancement of signal differences in space and time, and through permutation testing and the Max-T method, the method maintains control over the family-wise error rate. We chose the recommended parameters of 2.0 for cluster height and 0.5 for cluster extent across neighbouring electrodes (Mensen & Khatami, 2013). Neighbouring electrodes were defined as being under 40mm apart in the standard 10-20 template. The tests were carried out using Fieldtrip's (version: 20240111; Oostenveld et al., 2011) ft_timelockedstatistics function, with 10.000 permutations to build the null distribution of TFCE statistics. Significant clusters were defined using an alpha of 5%.

TFCE was carried out per dataset, task, condition contrast, and frequency band. It should be noted that these tests did not include a time dimension and that for the lower density Wearable and Multimodal datasets, the TFCE algorithm operates as a standard permutation paired t-test, as either

no or only a few channels possess neighbours, however, through the use of the Max-T procedure, the family-wise error rate remains controlled.

3.3.1 *Delta effects*

Figure 25. Delta Power Effects



Note. Topoplots of Cohen's D effect sizes per experiment and condition contrast in the delta band. TFCE tests were carried out once using raw PSD estimates and once using PSD estimates after removal of the estimated 1/f slope. Significant channels (p< .05) are displayed as black points.

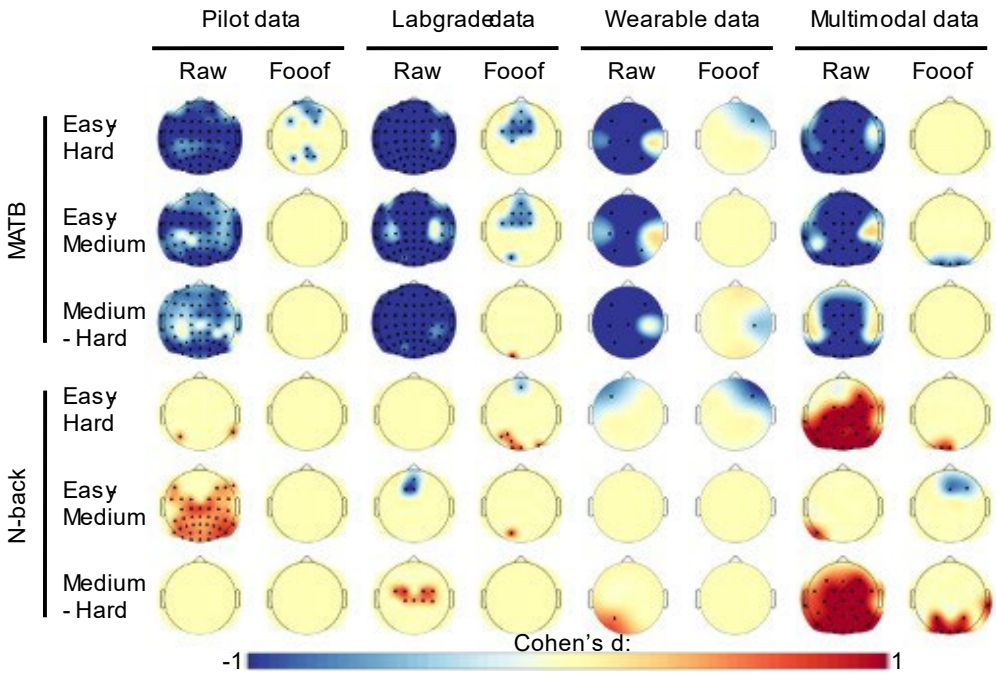
For the MATB, significant clusters in the delta band exhibited increases with increasing task-load and were reduced in their spatial extent when controlling for the aperiodic activity (Figure 25). Effects in the n-back were sparse and inconsistent across the four datasets, with the Pilot data showing no effects, the Lab-grade data showing effects for 0-3 and 1-3 contrasts, the wearable data showing a single significant electrode in the raw estimates, and the Multimodal data showing positive effects (delta reduced in higher task-load n-back conditions) in the 0-3 raw estimate data and 0-1 aperiodic-free estimate (Figure 25).

3.3.2 *Theta effects*

Theta effects showed global increases with increasing MATB task-load in the raw PSD estimates and even more pronounced reductions in spatial extent than the delta effects when moving from raw to aperiodic-free estimates (Figure 26). While the MATB effects were global across all four datasets, the corrected data exhibited small frontal and sporadic parietal (Pilot data) or occipital clusters (Lab-

grade and Multimodal data). Furthermore, posterior effects were positive in many of the raw and some of the aperiodic free n-back contrasts, suggesting a reduction in posterior, and in some cases central theta power at higher n-back loads (Figure 26).

Figure 26. Theta Power Effects

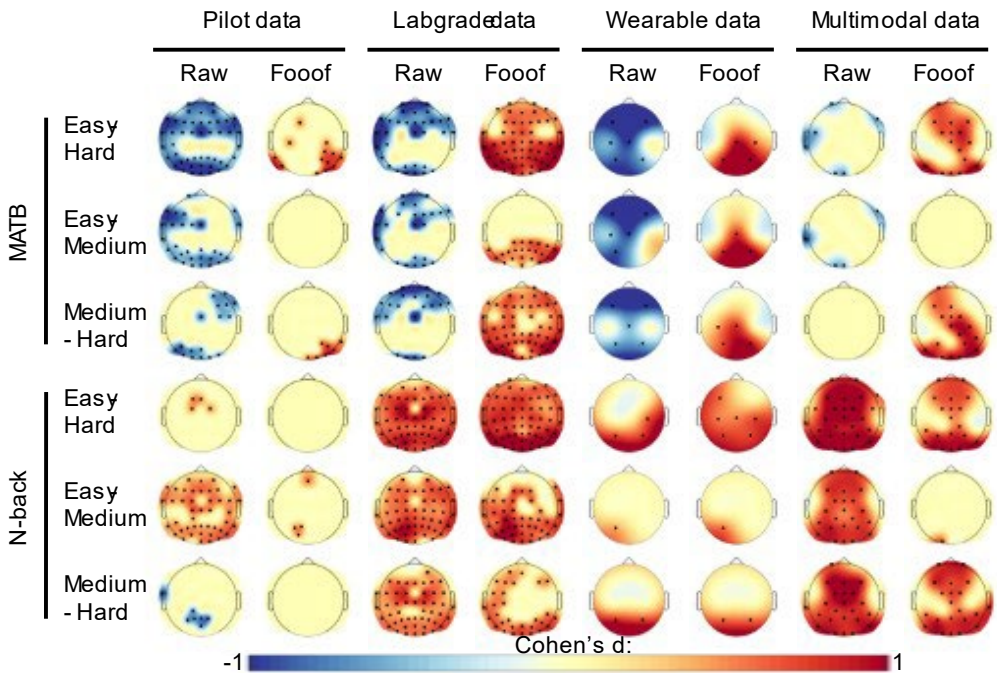


Note. Topoplots of Cohen's D effect sizes per experiment and condition contrast in the theta band. TFCE tests were carried out once using raw PSD estimates and once using PSD estimates after removal of the estimated 1/f slope. Significant channels (p< .05) are displayed as black points.

3.3.3 *Alpha effects*

Alpha effects were tested in two separate sub-bands termed low (8-10Hz) and high (10-13Hz) alpha, due to previous research suggesting low and high Alpha band ranges differing in their effect strengths (Chikhi et al., 2022). Both sub-bands showed significant effects across all condition contrasts (Figures 27 & 28). Interestingly, the effect direction changed from mixed negative and positive effects in the raw estimates to exclusively positive effects in the aperiodic-free estimates in the MATB contrasts (lower alpha power with increasing task-load).

Figure 27. Low Alpha Power Effects



Note. Topoplots of Cohen's D effect sizes per experiment and condition contrast in the low alpha band. TFCE tests were carried out once using raw PSD estimates and once using PSD estimates after removal of the estimated 1/f slope. Significant channels (p< .05) are displayed as black points.

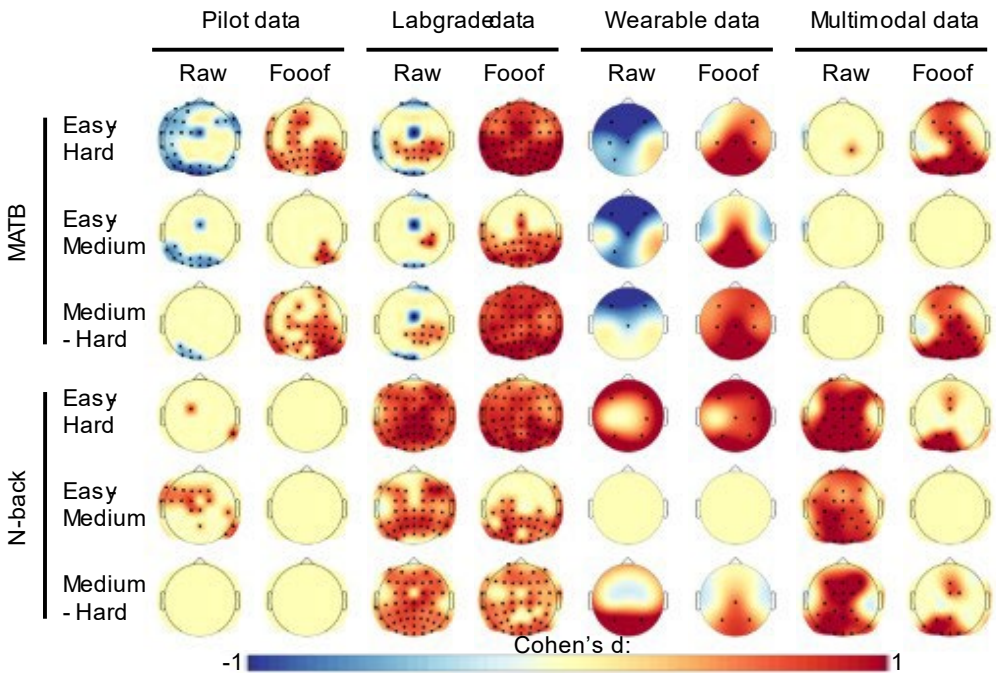
3.3.4 *Beta effects*

Effects in the beta band also switched direction from negative to positive in the MATB contrasts across all four datasets (Figure 29). The aperiodic-free data even exhibited some positive effects, where no effects were present using the raw data in the widest n-back contrast in both the wearable and pilot.

3.3.5 *Aperiodic effects*

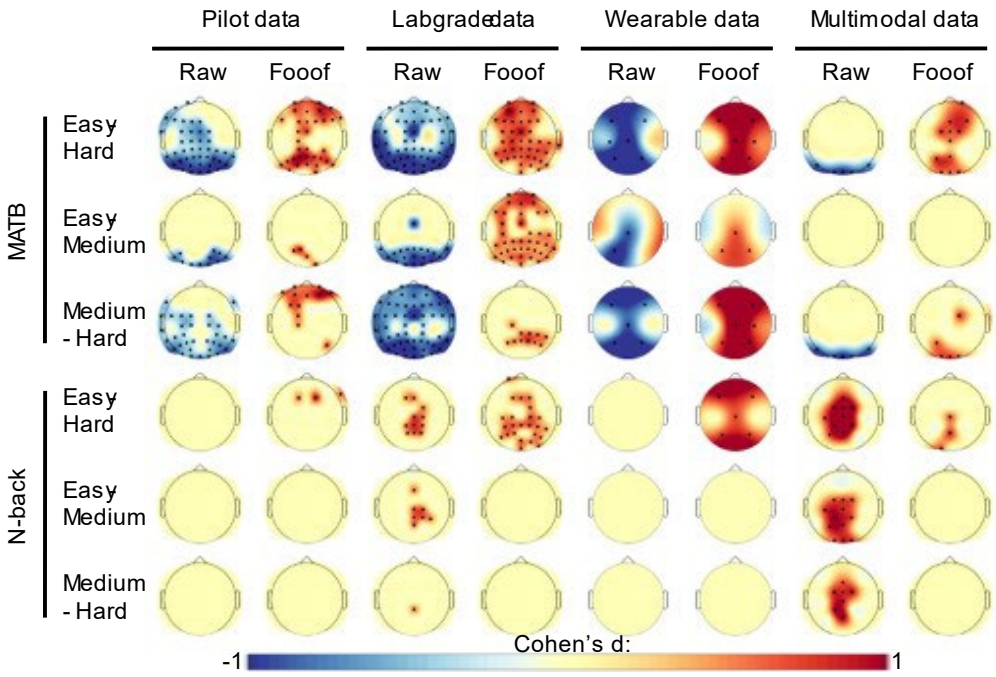
The exponent of the 1/f slope fitted by the Fooof algorithm showed no significant effects in the n-back. For the MATB however, the exponent seemed to have increased (steeper slope) with increasing task load in frontal areas and decreased with increasing task load in occipital areas (Figure 30).

Figure 28. High Alpha Power Effects



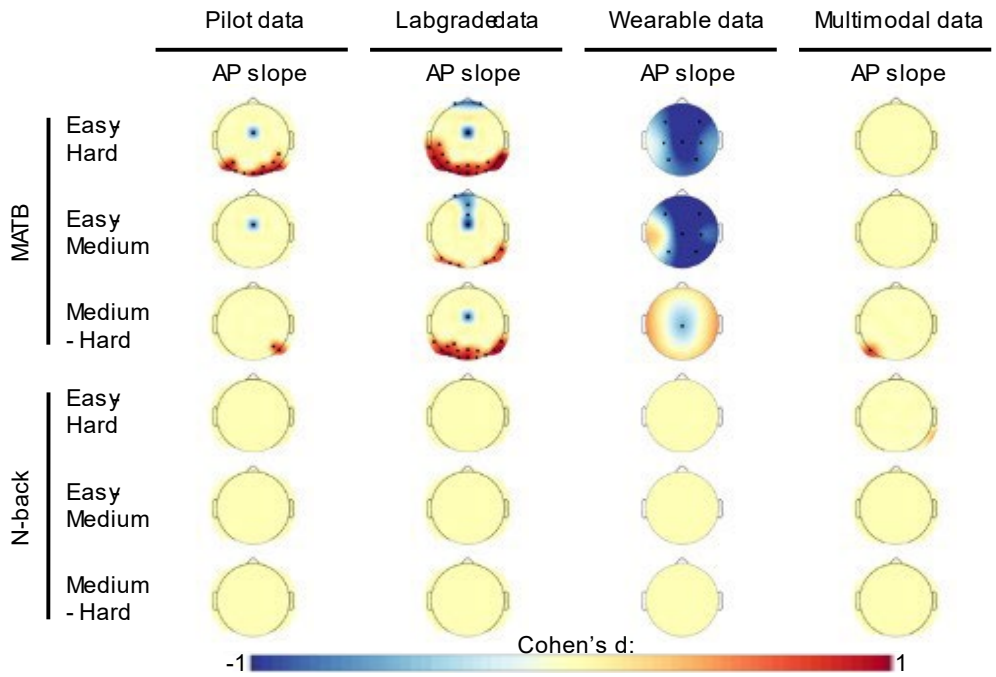
Note. Topoplots of Cohen's D effect sizes per experiment and condition contrast in the high alpha band. TFCE tests were carried out once using raw PSD estimates and once using PSD estimates after removal of the estimated 1/f slope. Significant channels (p< .05) are displayed as black points.

Figure 29. Beta Power Effects



Note. Topoplots of Cohen's D effect sizes per experiment and condition contrast in the beta band. TFCE tests were carried out once using raw PSD estimates and once using PSD estimates after removal of the estimated 1/f slope. Significant channels (p< 0.05) are displayed as black points.

Figure 30. Aperiodic Slope Effects

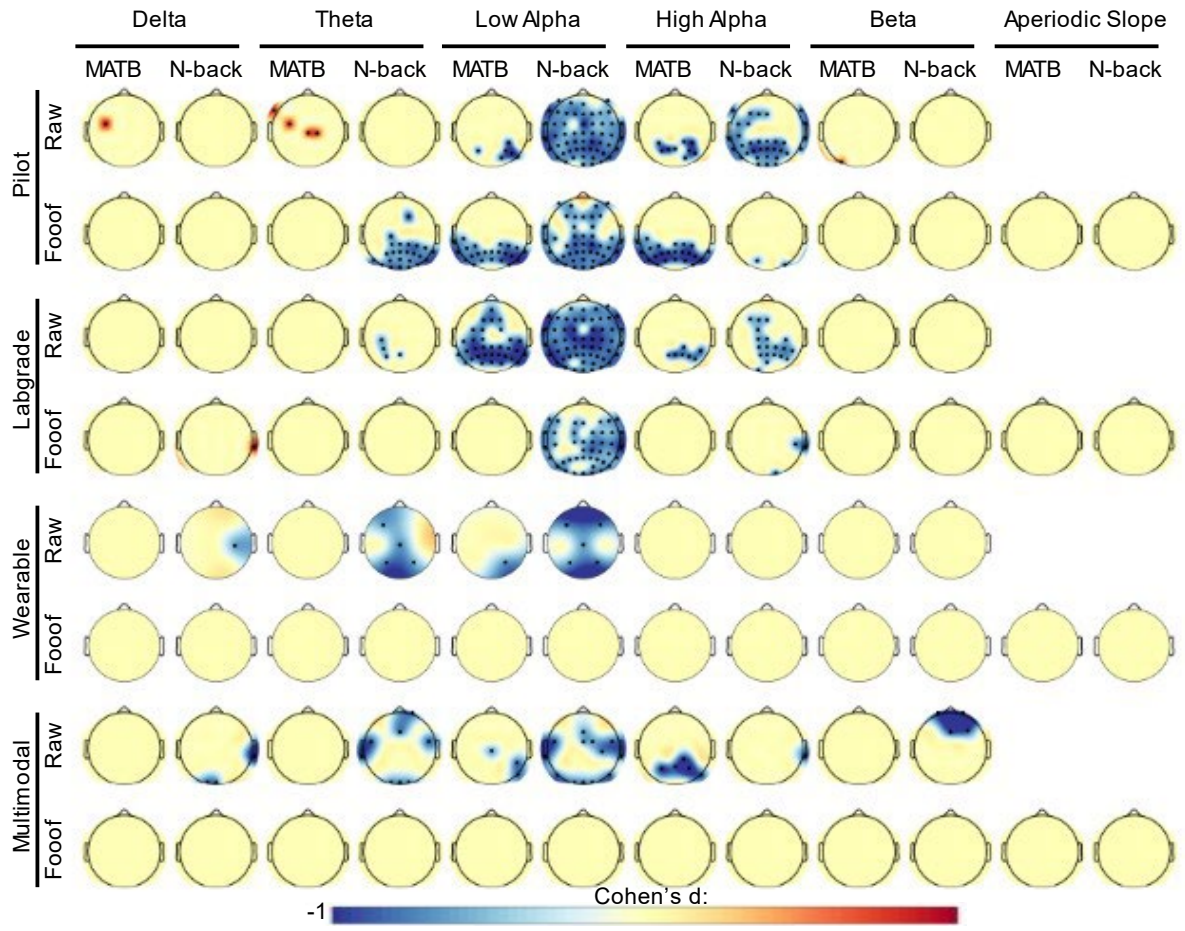


Note. Topoplots of Cohen's D effect sizes per experiment and condition contrast of the 1/f exponent. TFCE tests were carried out once using raw PSD estimates and once using PSD estimates after removal of the estimated 1/f slope. Significant channels (p< .05) are displayed as black points.

3.3.6 *Time-on-task effects*

Next, the effects of time on task were tested by repeating the same TFCE analysis only using the first and third condition repetition as conditions (Figure 31). The data was averaged over task-load conditions in the first and last set of repetitions per task. Time on task showed sporadic effects in the delta band with changing effect direction. The theta band, on the other hand, seemed to increase in power from the first to the last set of condition repetitions in the n-back. However, this effect was also not consistent throughout experiments, as it was sometimes visible in raw and sometimes in aperiodic free estimates. Both alpha sub-bands, especially the low one, showed significant increases over time. Most prominently visible in the n-back. The beta band showed no consistent effect across datasets, with a small positive cluster containing a single electrode in the MATB of the pilot data and a larger negative cluster in the n-back of the Multimodal data. Lastly, the aperiodic slope showed no significant effect across any of the tested datasets.

Figure 31. PSD Time-on-Task Effects



Note. Topoplots of TFCE statistics per experiment and task contrasting the first set of condition repetition and the last. TFCE tests were carried out once using raw PSD estimates and once using PSD estimates after removal of the estimated 1/f slope. Significant channels (p< .05) are displayed as black points.

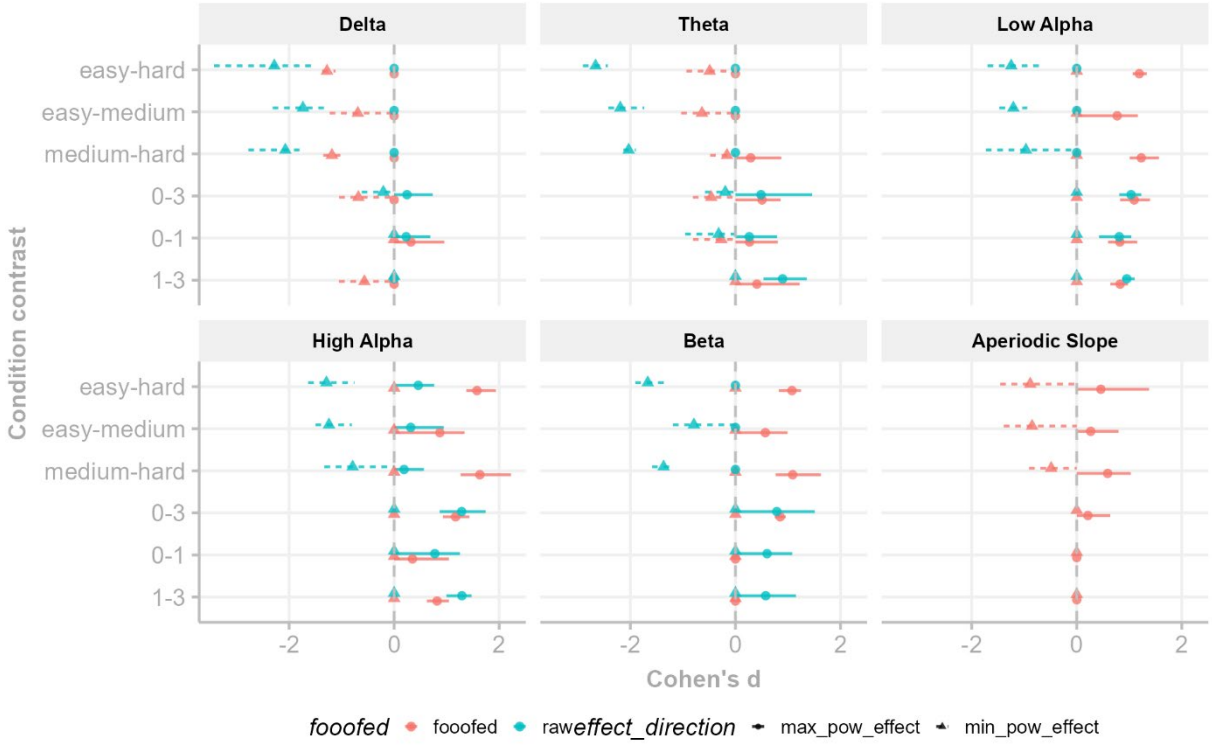
3.3.7 Interim Summary

The PSD analysis had three purposes. It should a) replicate mental workload effects previously described in the literature, b) investigate whether controlling for aperiodic activation changes the observed effects, and c) characterise the four datasets and put them in relation with each other.

In Figure 32, the maximal effect sizes of the Lab-grade, Wearable and Multimodal datasets are summarised. The Pilot dataset was excluded from this summary as the differences in task-load manipulation could have obfuscated trends emerging from the three studies with the exact same experimental paradigm. The effect sizes show that condition contrasts in the MATB exhibited significant differences within all tested frequency band ranges, even in the aperiodic slope parameters. For the delta and theta band, increasing task-load in the MATB caused power increases (negative effects). These effect sizes seem reduced from raw to aperiodic-free estimates, and, in the case of the medium-hard contrast, a positive effect (theta reductions) appeared after removing the

aperiodic activation. In the n-back, effects in the delta and theta bands were more sporadic, with both small negative as well as positive effects in the raw and aperiodic-free estimates that did not occur consistently in all three datasets. The reduction in effect size from raw to aperiodic-free estimates suggested aperiodic activation contributed to the observed differences in the raw-PSD estimates.

Figure 32. PSD Effect Size Summary



Note. Effect sizes (Cohen's d) of the Lab-grade, wearable, and Multi-modal datasets per tested PSD measure. Point ranges describe the median, the lowest and the highest effect size. Effects were separated into positive and negative effects since some frequency ranges exhibit both simultaneously.

The confounding nature of aperiodic activity was even clearer in the two alpha and the beta bands, in which the MATB effect direction changed when comparing the raw and aperiodic-free estimates. Raw effects showed increases in power with increasing task-load. When accounting for the aperiodic activity, the negative effects disappeared, and instead the expected reductions with increasing task-load emerged. The alpha effects in the n-back showed no such direction change, and alpha power appeared to consistently decrease with increasing task load. As with the alpha effects, the n-back showed weaker but consistently positive effects using raw estimates. Only the widest workload contrast produced a positive effect after accounting for aperiodic-free estimates. Lastly, the exponent that described the aperiodic slope exhibited both positive and negative effects with increasing task-load in the MATB. For the n-back, only the widest contrast showed a positive effect in two of three datasets.

With all three datasets and both tasks having exhibited effect size decreases when accounting for aperiodic activity, its influence on PSD-based measures of mental workload seems quite robust. While mostly amplifying oscillatory effects, for the MATB, it also caused sign-changes across several band ranges and electrode locations. This makes aperiodic activation an interesting topic for meta-analytical efforts that try to find explanations for disparate effects in these frequency ranges under mental workload (Borghini et al., 2014; Chikhi et al., 2022), but also for BCI research, as these results may offer a possible explanation for the difficulty of cross-task generalisability.

Time on task effects were inconsistent for the delta, theta, and beta bands. However, the alpha sub-bands showed consistent increases with time-on-task, especially in the low alpha band. The n-back exhibited more global effects in the 64-channel montages, while the MATB's time-on-task effect in the alpha band was more isolated to posterior electrodes. This was likely due to the MATB inducing more pronounced alpha suppression (Foxe & Snyder, 2011) that counteracted time/fatigue-related increases in the alpha band.

The exponent describing the slope of the 1/f spectrum (the aperiodic activity) showed no effect in the n-back. In the MATB, the exponent increased over centro-frontal electrodes in all but the Multimodal dataset and exhibited decreases with increasing task-load in the MATB across Pilot, Lab-grade, and, to a lesser extent, the Multimodal dataset. This was contrary to previous findings that reported only increases with increasing workload (Ke et al., 2023). Task-related differing excitability regime changes across locations of the cortex are less commonly reported on than the global changes following anaesthetics (Gao et al., 2017; Medel et al., 2023) but were previously shown in animal (Boustani et al., 2009) and human recordings (Waschke et al., 2021) during tasks requiring visual attention. The observed flattening of the power spectrum over occipital sites with increasing task-load in the MATB could suggest increases in excitatory and decreases in inhibitory signals (Gao et al., 2017) according to the E:I hypothesis, or an overall increase of activation throughout the visual cortex, as previously found using ECoG recordings (Podvalny et al., 2015).

These results underline the need for controlling for aperiodic activation in EEG research and warrant further research into 1/f activity in relation to task-load-related brain-state changes, as 1/f exponents could offer a potentially robust metric to track increases in visual task-load.

3.4 Task-irrelevant probes

As mentioned in the introduction, transient brain responses to external stimulation often carry workload-related information. In the mental workload literature, probing participants with tactile,

visual, or auditory stimuli has a long history (O'Donnell & Eggemeier, 1986; Papanicolaou & Johnstone, 1984). However, recent work tended to focus on the use of auditory probing in particular.

3.4.1 Auditory Probes

As alluded to in Chapter 1, while the P300 family of ERP components tends to be the main target for studies employing auditory probes for mental workload research, earlier ERP components could also offer valuable insights into mental workload-related changes to brain states. Components like the N1 and P2 are thought to relate to earlier, low-level sensory processing, and these components have previously also been shown to be modulated by differences in task-load (Allison & Polich, 2008; Dyke et al., 2015; Kramer et al., 1995). Disregarding the later p300 components removes the need to space stimuli far apart, instead allowing for much faster probe paradigms, which consequently increases the responsiveness of a pBCI utilising such probes - possibly at the cost of reduced sensitivity.

Recently, Sugimoto (2022) investigated the optimal spacing for such a pure tone paradigm for group-level analysis. Their aim was to accumulate a maximum number of probes in as little time as possible while maintaining the ability to distinguish between two levels of workload. They described significant effects of driving simulator difficulty on the N1 and P2 , with variable ISIs averaging around 0.6 seconds. They reported requiring around 2 minutes of data for a significant difference between workload conditions, whereas previous comparable studies with longer ISIs required around 10 minutes of data (Allisson et al, 2008).

Due to the rapid nature of Sugimoto et al's paradigm, it may lend itself greatly to the pBCI context, as its nearly continuous stream of events allows for the accumulation of many training examples in a short time span. This could allow for fast calibration and, if accurate, also allow for the detection of sudden spikes in workload. A major concern with the use of pure tones in implicit probing paradigms is that participants may habituate to rapid, repetitive auditory stimulation and that this habituation will dampen the sensitivity and statistical power of the approach (Näätänen & Picton, 1987). Furthermore, this reduction in statistical power may be exacerbated when the inter-stimulus interval is short due to previous observations that short ISIs may outpace the required refractory period of the neuronal sources of the ERP components (Budd et al., 1998; Pereira et al., 2014). While the rapid nature of the probing approach presented by Sugimoto and colleagues (2022) promises significant improvements for the responsiveness of pBCI applications, it may also reduce the sensitivity of the ERP to changes in workload over longer recording sessions.

While the primary interest in adopting the rapid task-irrelevant auditory probe method was its use for continuous mental workload assessment, replicating the effect of workload on the subject-wise averaged ERPs forms an important step towards developing faster probing paradigms. The original

publication reported significant effects of driving speed in a simulated driving task on both the average N1 and P2 amplitudes (Sugimoto et al., 2022) using average peak amplitudes extracted from 25ms windows centred around the maximal amplitude of the grand-average ERP.

Here, the analysis was carried out using fieldtrips TFCE implementation with cluster extent now covering time and neighbouring electrodes. The tests were carried out using Fieldtrip's ft_timelockedstatistics function, with 10.000 permutations to build the null-distribution of TFCE statistics from 0ms to 500ms after stimulus onset. Significant clusters were defined using an alpha of 5%.

The analysis was conducted on block-wise ERPs computed per subject. For the Lab-grade and pilot datasets, the fully processed data was re-referenced to the averaged Mastoids (TP9 and TP10). The wearable data did not contain these channels and was analysed using its original right-earlobe online reference. All three datasets were additionally low-pass filtered at 30Hz using EEGLAB's default FIR filter – resulting in the same retained frequency contents as in the original publication (Sugimoto et al., 2022). Epochs with high-amplitude artifacts (-150µV,150µV) were discarded. Table 20 details the average number retained epochs per block. The original publication reported needing 120 seconds of data per participant to detect the reduction in P2 amplitude, and 90s of data for the N1 reduction. Both the 140-second blocks of the n-back and the 300-second blocks of the MATB should thus have provided sufficient data per block.

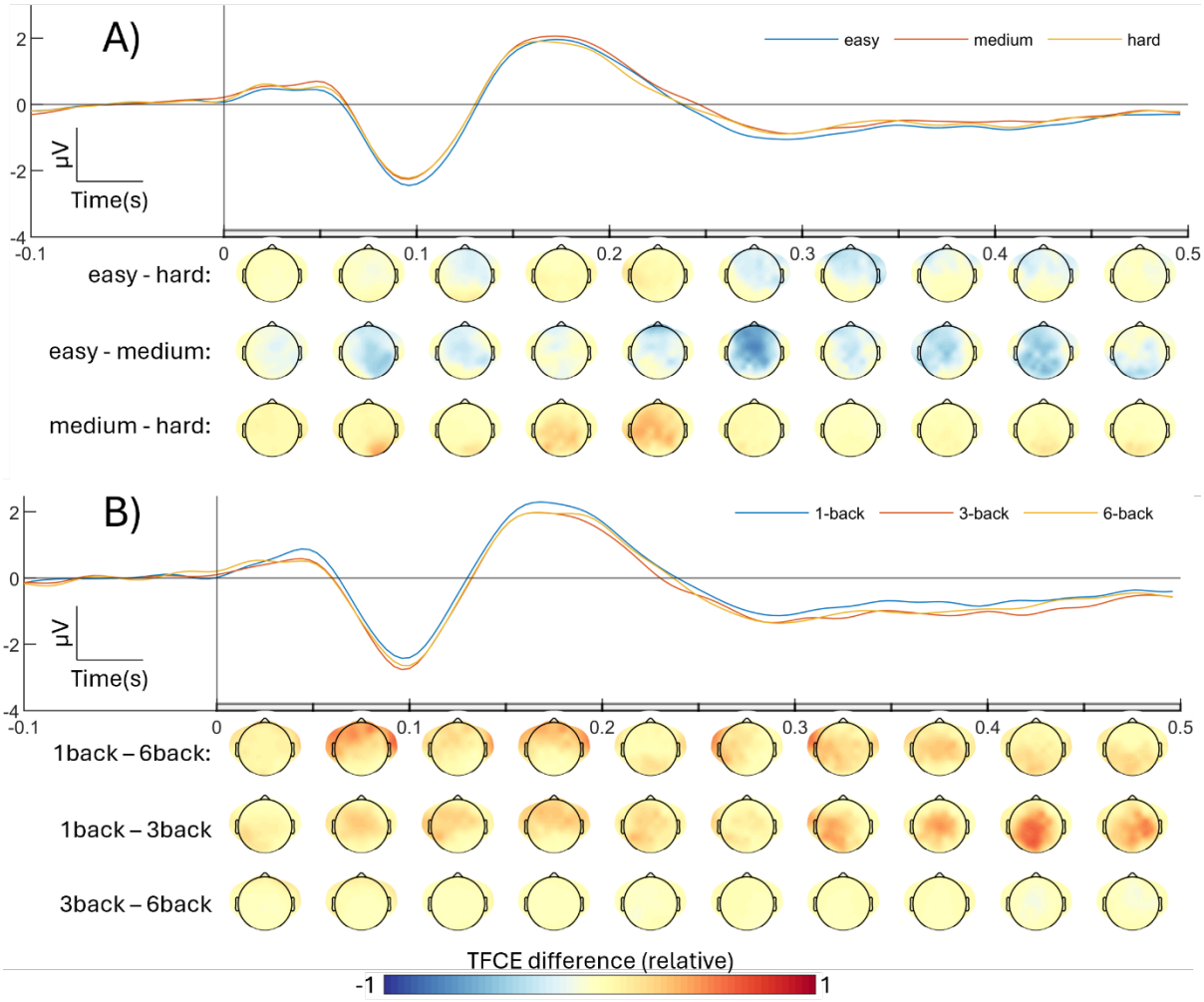
Table 20. Average Epochs per Participant (After Epoch Rejection)

	MATB			N-back		
	Easy	Medium	Hard	Easy	Medium	Hard
Pilot	477.1	474.5	472.25	221.9	219.7	221.8
Lab-grade	404.1	399.1	379.25	207.7	203.7	204.1
Wearable	444.5	440.3	435.8	208.9	207.4	202.6

The Pilot dataset exhibited no significant effects (Figure 33). The Lab-grade data exhibited higher amplitudes around the P2 peak across frontal electrodes in the hardest MATB condition compared to the medium and easy task-load conditions (Figure 34). This effect was the opposite of the results reported in the original study, where increased driving task-load caused a reduction in P2 amplitude at FCz (Sugimoto et al., 2022). The widest MATB contrast exhibited another significant cluster around 350ms after stimulus onset in the Lab-grade data, followed by a larger cluster after 400ms. Both effects were again negative, suggesting larger amplitudes with increased mental workload. In the Wearable dataset, only the widest contrast of the MATB showed significant effects after 300ms,

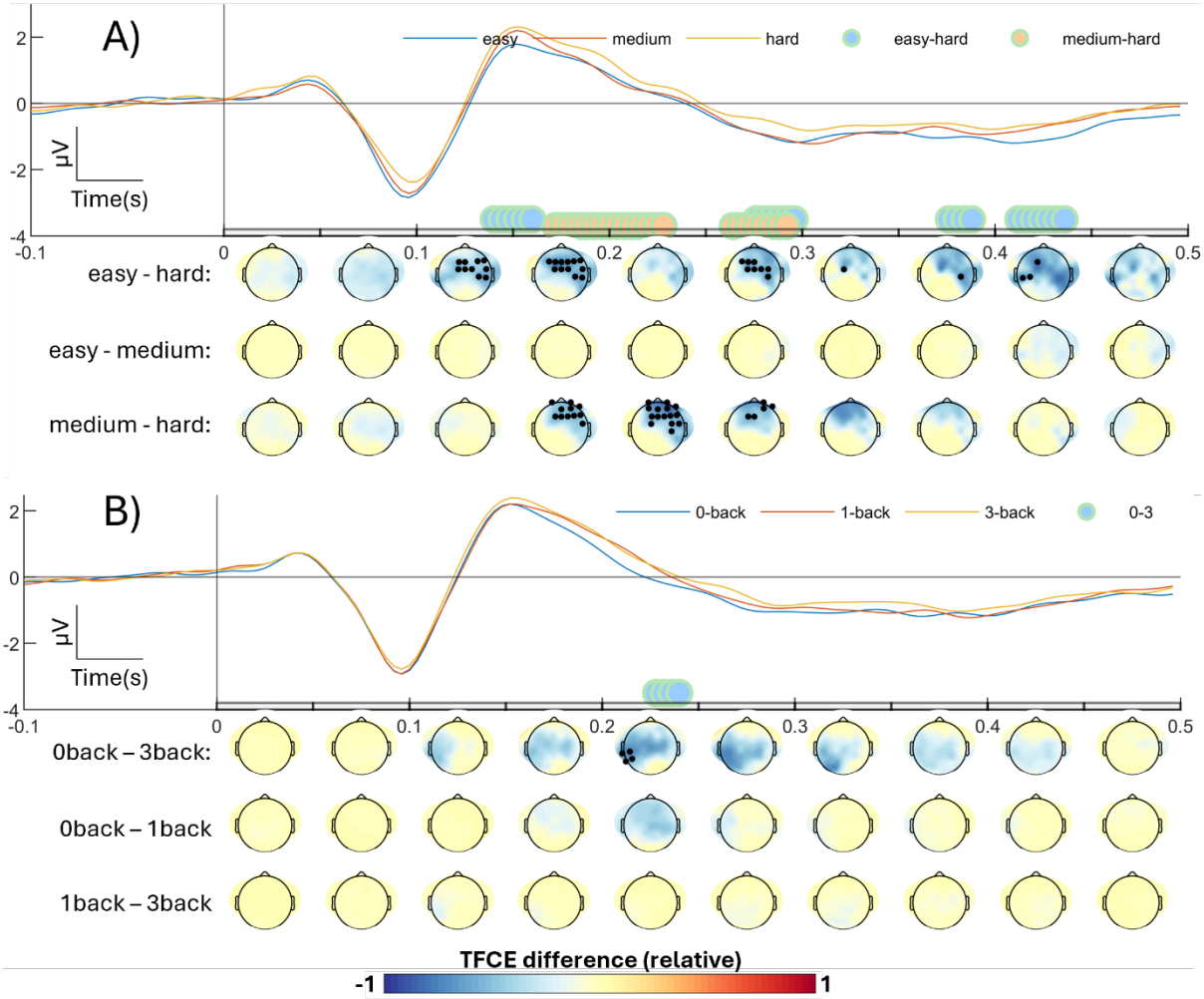
however, here, in the opposite direction compared to the effects observed in the Lab-grade data, with reduced amplitude in the hardest MATB condition (Figure 35).

Figure 33. Pilot Auditory ERP Results



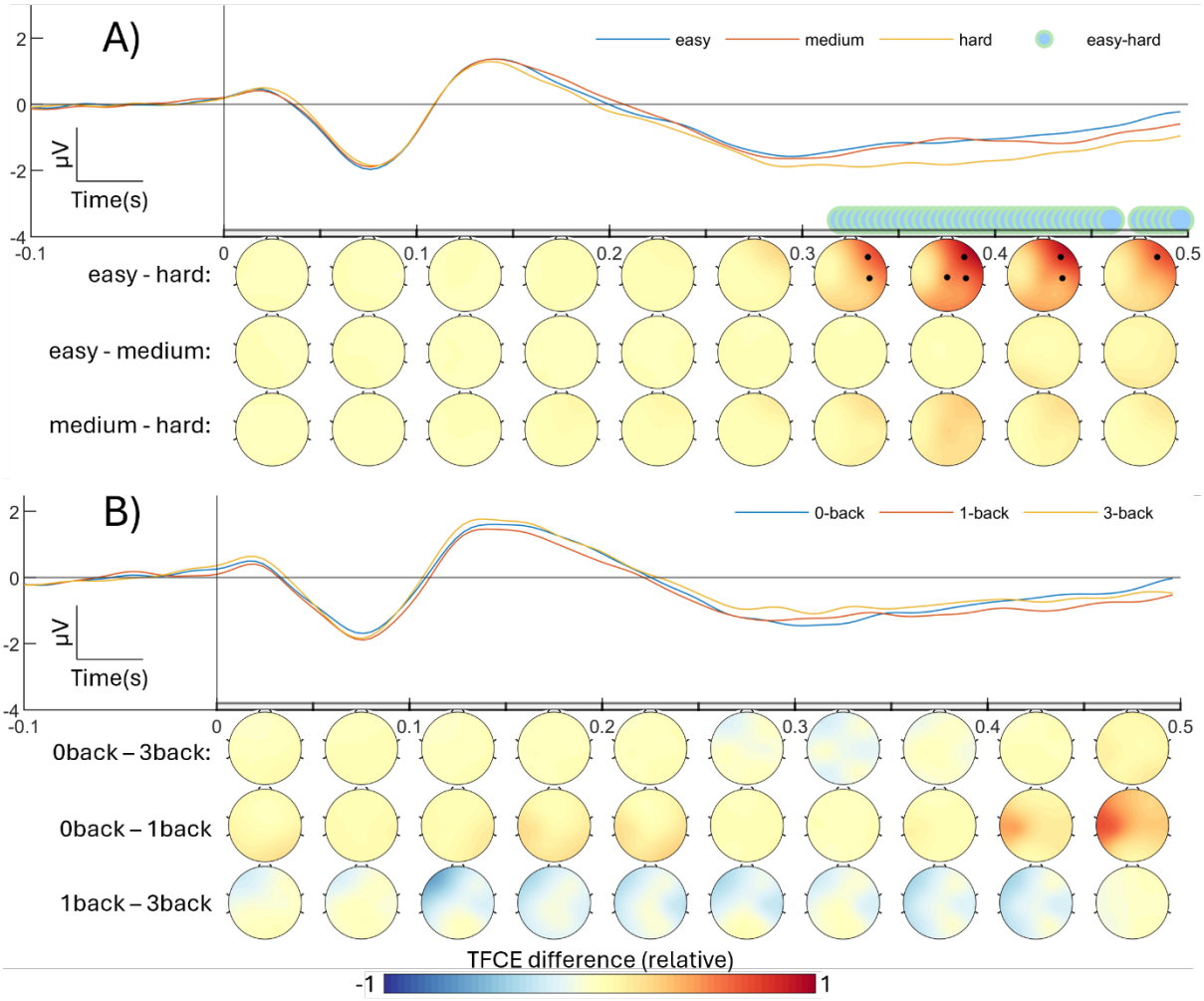
Note. Grand-average ERP at FCz per task (A. shows the MATB, B. the n-back) for each condition. Below the ERPs are topoplots illustrating the TFCE statistic of the difference score of the three condition contrasts per 50 ms time-bin. TFCE statistics were normalised by dividing all values by the largest absolute difference measured across contrasts per task. Significant channels ($p < 0.05$) are displayed as black points.

Figure 34. Lab-grade Auditory ERP Results



Note. Grand-average ERP at FCz per task (A. shows the MATB, B. the n-back) for each condition. Below the ERPs are topoplots illustrating the TFCE statistic of the difference score of the three condition contrasts per 50 ms time-bin. TFCE statistics were normalised by dividing all values by the largest absolute difference measured across contrasts per task. Significant channels ($p < 0.05$) are displayed as black points.

Figure 35. Wearable Auditory ERP Results



Note. Grand-average ERP at FCz per task (A. shows the MATB, B. the n-back) for each condition. Below the ERPs are topoplots illustrating the TFCE statistic of the difference score of the three condition contrasts per 50 ms time-bin. TFCE statistics were normalised by dividing all values by the largest absolute difference measured across contrasts per task. Significant channels ($p < 0.05$) are displayed as black points.

3.4.2 *Interim Summary*

Rapid task-irrelevant probes have previously been utilised to distinguish between fast and slow simulated driving tasks. There, the statistical tests were carried out on FCz alone, showing a reduction in the strength of the N1 and P2 components (Sugimoto et al., 2022). Here, the workload-related effects were investigated using whole-montage cluster-based permutation tests.

No workload effects were significant in the Pilot dataset. This could have been due to the initial timing issues of the probe presentation or the paradigm's lack of sensitivity, as it was previously only investigated for two widely different driving settings. As the timing of the probes was not fully accurate (markers with up to 10ms deviation from actual tone onset) in this dataset, the probes were utilised again in the following studies, this time with sub-millisecond timing accuracy (see 2.4.1).

In the Lab-grade dataset, the probes did indeed show significant effects around the P2 peak and after 400ms. The effects were, however, not present in all contrasts (low sensitivity) and were not in the expected direction (increased P2 amplitude with increasing task-load; easy-hard and medium-hard in the MATB and easy-hard in the n-back). ERP amplitudes are generally expected to reduce in amplitude with increasing workload (Deeny et al., 2014; Dyke et al., 2015, 2015; Kramer et al., 1995), putting these effects at odds with the literature.

In the Wearable dataset, only the widest MATB contrast exhibited a significant cluster, and it was located around same time as the late effects of the Lab-grade study, however, in the opposite direction (stronger negative potential in the hard MATB). These late effects were last described by Alisson and colleagues (2008), who termed them slow wave 1 (400- 650ms) and found decreasing amplitudes comparing passive viewing with on-task conditions. The difference in effect direction may suggest that either effect could represent type I errors or may arise from variations in the reference scheme, as the Lab-grade data was re-referenced to the averaged mastoids, while the Wearable data was referenced to the right earlobe.

In the case of the rapid-paradigm tested here, these effects on late potentials for both the Lab-grade and Wearable datasets have two implications. Firstly, after 400 ms, a new stimulus could appear, and in the window between 400 and 500 ms, 25% of probes occurred due to the uniformly sampled ISI range of 400-800 ms. If mental workload affected the ERP after 300-400 ms, with a lasting effect that was potentially longer than 500ms after onset, the early potentials of at least 25% of probes in that condition would be affected by the previous effect. However, the relationship could also be reversed, and its appearance here stemmed from baseline correcting the epochs. That is, mental workload affects the initial early sensory response to probes (as described by Sugimoto et al), which in turn biases the late potential of the previous probe. Overlap corrected regression ERPs could possibly elucidate this question aptly (Ehinger & Dimigen, 2019; Smith & Kutas, 2015).For now, the conceptual replications carried out in this project did not confirm the results of Sugimoto and colleagues, but rather than not finding any results, the analyses here uncovered different results that warrant further investigation.

3.4.3 Visual Probes

Less thoroughly studied than auditory probes, visual probes are often part of the task itself, like stimulus presentation epochs in an n-back task (Brouwer et al., 2012; Mühl et al., 2014) or additional items added to the trial, like the addition of a simple shape detection task in the retention period of a working memory task (Roy et al., 2016). Either have been shown to involve ERP effects in response to task-load increases.

The addition of truly task-irrelevant visual probes that do not constitute a secondary task was the topic of research in the 1980s (Papanicolaou & Johnstone, 1984) with working memory set size found to affect several ERP components elicited by brief light flashes when timed with stimulus retention or encoding intervals (Bauer et al., 1987). When such flashers were delivered irrespective of the task(i.e. trial-independent probes) these effects disappeared, however (Wilson & McCloskey, 1988). Since then, research into task-irrelevant probes has focused on auditory rather than visual probes, seemingly due to the latter being deemed too distracting (Ke et al., 2021).

In this thesis, low-amplitude steady-state visually evoked potentials (SSVEP) were considered for their potential as continuous task-irrelevant visual probes in mental workload monitoring. SSVEPs are evoked responses to periodic stimulation, which, at frequencies above 2Hz, result in induced oscillatory brain responses (Norcia et al., 2015). Typical SSVEP stimulation frequencies range from 2-20Hz and are used in clinical, cognitive, as well as BCI research (Norcia et al., 2015; Vialatte et al., 2010). In the reactive BCI realm, tagging different components on a screen with individual frequencies or code-VEPS (Castillos et al., 2023; Shirzhiyan et al., 2019) allows operators to focus on points of interest for the BCI system to trigger actions based on the dominant response measured in the EEG trace. The pronounced oscillatory peaks in the EEG's power spectrum induced by SSVEPs have a high signal-to-noise ratio (SNR), resulting in response times between .5 and 3 seconds (Maÿe et al., 2022). Recently, code-VEPs have also been used in passive BCIs to track pilot's attention to a radar monitor (Dehais et al., 2022) by using an code VEP-classifier's confidence in detecting a code-VEP as a proxy for the pilot's attention to the region of interest.

The motivation for their inclusion in the Lab-grade and Wearable datasets in this thesis was twofold. Firstly, the use of SSVEP stimuli in pBCI contexts was under-researched. This, as mentioned before, may have to do with their distracting nature as well as visual fatigue caused by the flickering stimulation (Chang et al., 2014; Ladouce et al., 2022). However, modulations to SSVEP may contain valuable information about changes in visual processing abilities with varying levels of task load or fatigue. Previously, effects of mental workload on SSVEP latency (i.e. the phase difference with the presented flickering stimulus) were deemed unreliable workload indicators (Wilson & O'Donnell, 1988). Anticipation of goal-related information on the other hand, did seem to cause a decrease in SSVEP amplitude induced by a global 13Hz flicker in a different study (Silberstein et al., 1990). Furthermore, mental workload-related degradation of SSVEP detection accuracy was observed in a dual-task combining a 4-class SSVEP detection paradigm with a verbal n-back task (Zhao et al., 2018). Taken together, these results suggest a reduction in SNR with increasing workload, which could potentially be harnessed for continuous workload monitoring. The second reason for their inclusion were recent results reporting reliable SSVEP detection using periliminal contrast settings (Ladouce &

Dehais, 2024). By drastically reducing the amplitude depths of the flickering stimulus, the user comfort could be significantly improved (Ladouce et al., 2022), which a) allows for longer experimentation using these flickers for pBCI purposes and b) reduces the distracting and fatiguing nature of visual task-irrelevant probes.

The 15Hz SSVEP, employed in the Lab-grade and Wearable datasets, was hypothesised to show reduced signal-to-noise ratios in higher task-load conditions, as previous studies suggested that SSVEP response decreased under increasing task-load (Silberstein et al., 1990; Zhao et al., 2018) or hypovigilance settings (Ladouce et al., 2025). This, just like with the task-irrelevant auditory probes, was of main interest for continuous mental workload monitoring, but here, the effects of task-load on the SSVEP SNR were also assessed at the group-level.

To this end, the data of the Lab-grade and Wearable datasets were analysed in two ways. First, sensor-space condition differences within the main (15Hz) and first harmonic (30Hz) frequencies were tested for using the same TFCE approach described for the PSD analyses. However, since we could expect pronounced peaks at the main and harmonic frequencies, the FOOOF-based tests were run on the parametrised peaks (the height of the fitted Gaussian curve), instead of the average band power after subtraction of the aperiodic fit. This was not achievable for the previous PSD analysis, as many of the literature's analysed frequency ranges do not exhibit true oscillatory peaks as found in the theta and alpha ranges. If FOOOF did not find a peak centred around the main frequency or its harmonic, the channel was instead stored as having no power in the band.

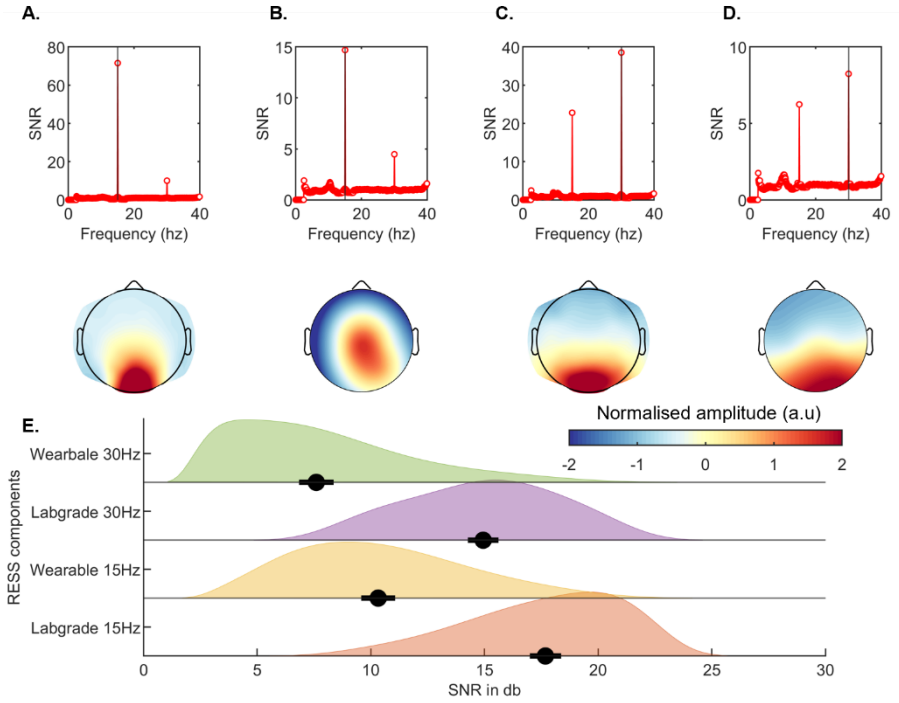
The second analysis aimed to garner more detailed insights into the dynamics of the SSVEP under different task-load settings. This analysis focused exclusively on the n-back as the MATB's uncontrolled nature likely meant that many data periods included sections where the participants were looking at the keyboard rather than the screen containing the SSVEP. Additionally, by focusing the analysis on the n-back, we can analyse the SSVEP response at the single-trial level, similar to recent work on sustained attention (Ladouce et al., 2025).

To facilitate this, the SSVEP response was amplified using a spatial filter approach termed Rhythmic Entrainment Source Separation (RESS - Cohen & Gulbinaite, 2017). RESS was computed using the minimally processed data for both the Lab-grade and Wearable datasets. Using the minimally processed data of the Lab-grade data instead of the fully processed version was done as SSVEPs tend to be robust against ocular artifacts (saccade and blink related activations tend to occupy lower frequency ranges - Perlstein et al., 2003; Vialatte et al., 2010) and there was a risk that both zapline-plus as well as the ICA pruning could have removed some of the SSVEP response from the Lab-grade data. The spatial filter in the RESS procedure was computed by first dividing the data into consecutive

10-second windows. These were filtered at the stimulation frequency or its first harmonic with a Gaussian-shaped filter using a full-width half maximum (FWHM) of 1Hz. Neighbouring frequencies +/- 1Hz were filtered in the same manner. Next, normalised covariance matrices were computed for all three filtered signals. The eigenvector of the highest eigenvalue of a generalised eigenvalue decomposition between the main and the average neighbouring frequencies covariance matrices was used as the RESS spatial filter.

Applying this spatial filter to the raw data generated a single “virtual electrode” that captured the maximised main frequency response across the full EEG montage (see Figure 36). Next, three complex Morlet wavelets, centred at the main frequency along with both neighbouring frequencies (±1 Hz), were separately convolved with the time series of the virtual electrode, maintaining a fixed temporal FWHM of 0.533 seconds (seven cycles of the 15 Hz rhythm). The instantaneous power was computed as the squared magnitude of the complex wavelet convolution. By calculating the power at 15 Hz relative to the average power of the neighbouring frequencies, a time-resolved SNR measure was derived, which was subsequently log-transformed (10*log10()) to produce a decibel-scaled measure of relative power at 15 Hz compared to its adjacent frequencies.

Figure 36. Average RESS Components



Note. Group-level averages of the first RESS component computed from the n-back data. The four columns at the top present the SNR spectrum and average topoplot after back-projection from the filter into sensor space per frequency and dataset. A) shows 15Hz in the Lab-grade data, B) shows 15Hz in the Wearable dataset, C) shows 30Hz in the Lab-grade data, and D) shows 30Hz in the Wearable dataset. E. shows the kernel density estimate (KDE) of the distributions of SNR values per filter and dataset in decibels. Black pointranges represent the mean and standard error of the mean.

Having single-trial SNR data for this analysis allowed for the fitting of a more exhaustive linear mixed-effects model. Three research questions were of main interest to this analysis

1. Does the SSVEP response to n-back stimuli differ between conditions?
2. Does the SSVEP response differ between target and non-target stimuli?
3. Does the SSVEP response differ between correct and incorrect responses?

These questions were motivated by previous research indicating that increased mental workload reduces the performance of SSVEP-based BCI systems. Stemming from these findings, it was expected that the SNR would decrease with increasing task load. On the other hand, higher n-back conditions should require more attentional resources, thereby increasing the focal attention on the stimuli and thus enhancing the SNR of the flicker (Davidson et al., 2020). It was further expected that targets would show higher SNRs than non-target trials due to the attention capture of goal-related information. Whether the SNR would change with time-on-task was uncertain.

The average SNR during stimulus presentation (0 – 500ms after stimulus onset) was analysed using linear-mixed effects models. As these tests were run on single-trial data, the linear-models were

attempted to be fitted in a maximal fashion, including subject-wise random-slopes in addition to the random-intercepts. The model complexity was reduced iteratively if the current model a) failed to converge, b) exhibited a (near) singular fit, or if the next simpler model in line did not differ significantly, tested using a likelihood ratio test (Barr et al., 2013).

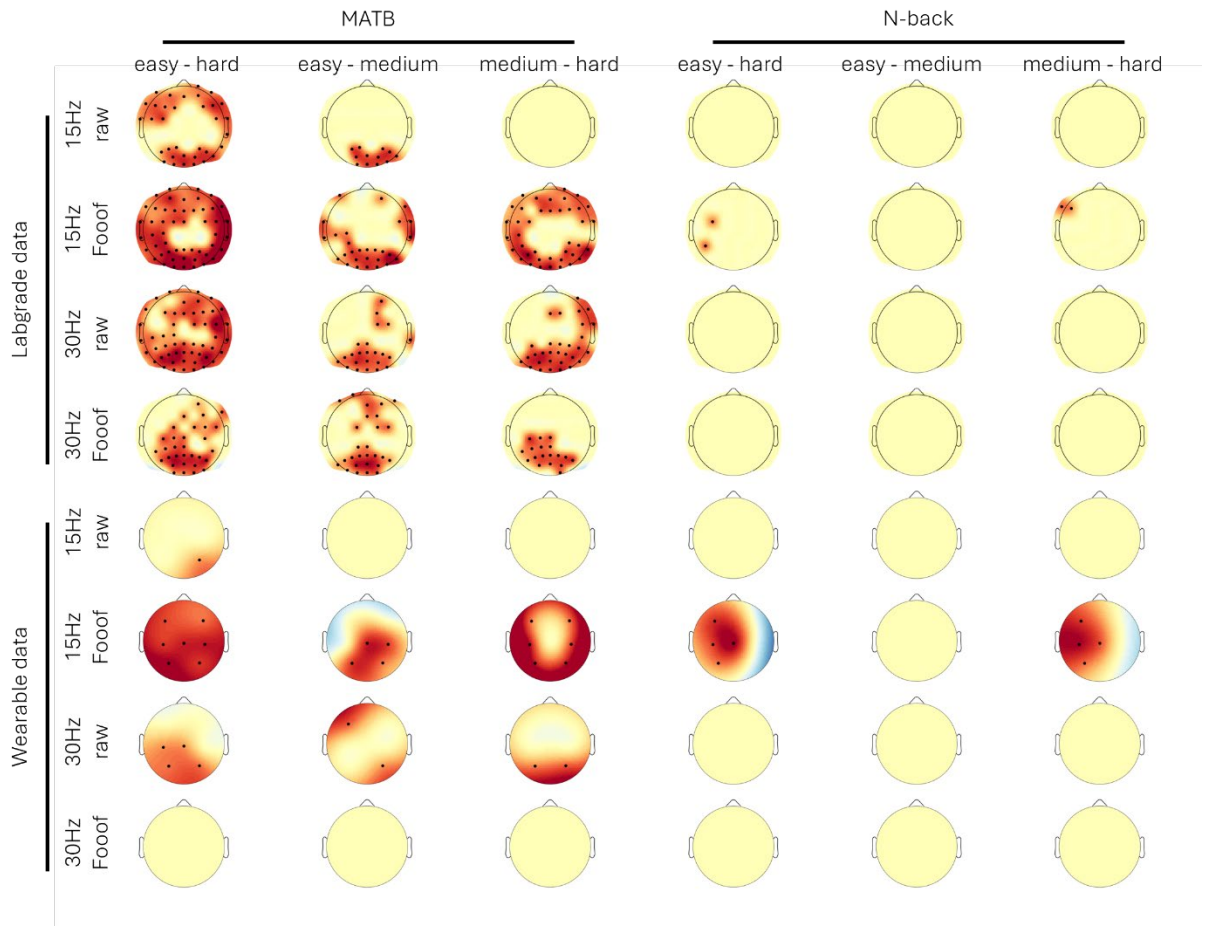
The factors *Condition* and *repetition* were treated as ordered factors, testing linear and non-linear (quadratic) effects. Factors *target* and *correct* were dummy coded (linear effects)

Models by decreasing complexity:

- `Formula A: SSVEP_SNR ~ condition * repetition + correct + target + (1 condition + repetition | subject)`
- `Formula B: SSVEP_SNR ~ condition * repetition + correct + target + (1 + condition | subject)`
- `Formula C: SSVEP_SNR ~ condition * repetition + correct + target + (1 + repetition | subject)`
- `Formula D: SSVEP_SNR ~ condition * repetition + correct + target + (1 | subject)`

Sensor-Space Results. Differences in the SSVEP response were most pronounced in the MATB task (Figure 37). Both the Lab-grade and Wearable data exhibited significant differences in all three MATB contrasts. Cluster extent appeared much wider in the Fooof estimates compared to the raw estimates when testing the main frequency range. The first harmonic's effects were more consistent in the Lab-grade data but were not present in the Fooof estimates of the Wearable dataset. Among the n-back only contrasts, only the 3-back showed significant differences. These differences were additionally only present in Fooof estimates of the main frequency.

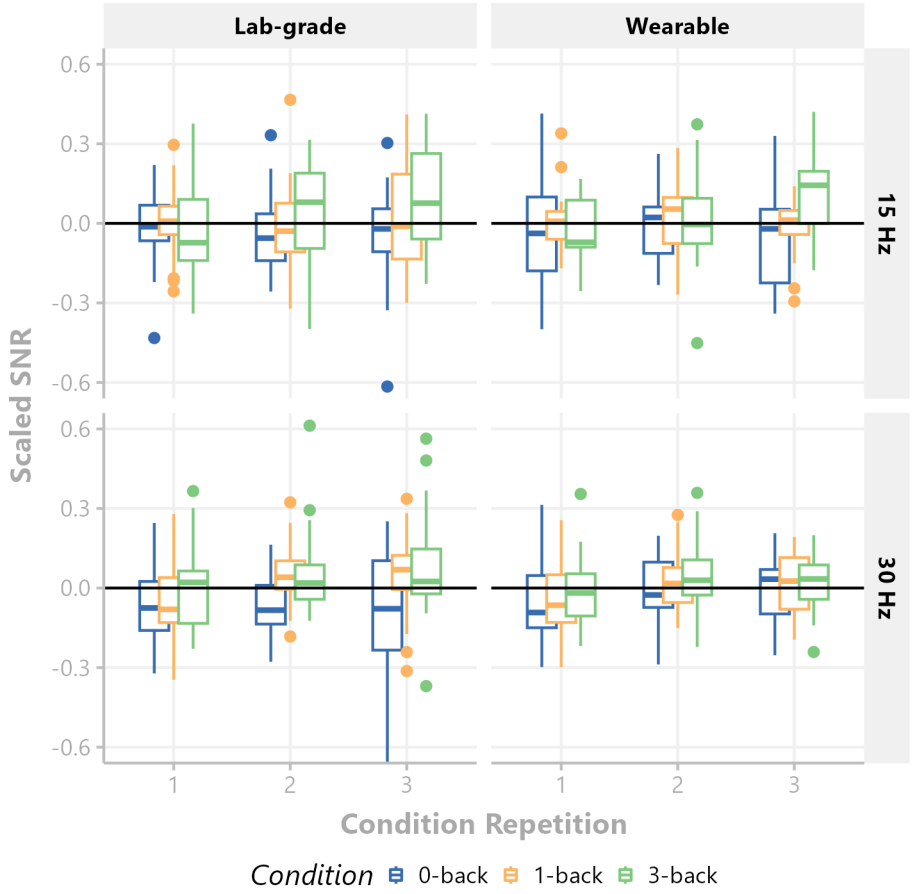
Figure 37. Sensor-Space SSVEP Results



Note. Topoplots of significant effects in the main SSVEP frequency (15 Hz) and its first harmonic (30 Hz). All tests were carried out twice per dataset and contrast: once using raw PSD estimates and once using parameterised spectra computed using the Fooof algorithm. Significant channels ($p < 0.05$) are displayed as black points.

Trial-wise RESS tests. Visually, stimulus-locked SNRs appeared to have increased with increasing n-back difficulty (Figure 38). At 15Hz the Lab-grade data was best described by Model C (including random slopes for task-load conditions). Repetition showed significant linear increases in SNR with time ($\beta = 0.21$, $SE = 0.06$, $t(11780) = 3.00$, $p = .004$). Pair-wise comparisons with tukey corrections showed that only the first and last set of condition repetitions differed significantly ($\beta = -0.292$, $SE = 0.095$, $t(\text{inf}) = -3.07$, $p = 0.006$). No other factors were significant (see Table 21). While the effect size of factor Condition was much larger than that of the other factors, its 95% interval ranged from 0 to 1. The addition of random slopes likely led to the model allowing for more between-subject variation, resulting in high variance and high uncertainty in estimating the effect.

Figure 38. N-Back RESS SNRs



Note. SNR values per stimulus representation were z-scored within subject and frequency. Boxplots display subject averages per condition and by condition repetition.

At 30Hz, the Lab-grade data was best described by Model B (including random slopes for condition repetition). Condition showed significant linear increases in SNR with increasing task-load (β = -0.278, SE = 0.075, t(11780) = 3.67, p < 0.001). Pair-wise comparisons with tukey corrections showed that SNR increased from 0-backs to 1-backs (β = -0.231, SE = 0.1902, t(inf) = -2.26, p = 0.035) as well as from 0-backs to 3-backs (β = -0.393, SE = 0.107, t(inf) = -3.75, p < 0.001). No other factors were significant (see Table 21). As before, the 95% CI of the Condition effect size is highly uncertain (0-1), suggesting the single-trial data was too variable to estimate effects.

At 15Hz the Wearable data was best described by Model D (including no random slopes). Condition showed significant linear increases in SNR with time (β = 0.24, SE = 0.09, t(inf) = 2.8, p = .005). Pair-wise comparisons with tukey corrections showed that SNR increased from 0-backs to 3-backs (β = -0.346, SE = 0.123, t(inf) = -2.8, p = 0.015) as well as from 1-backs to 3-backs (β = -0.311, SE = 0.123, t(inf) = -2.53, p = 0.018). No other factors were significant (see Table 23).

At 30Hz, the Wearable data was best described by Model D (including no random slopes. None of the tested effects were significant (see Table 24).

Table 21. 15Hz Lab-grade (Model C)

Effect	F-statistic	DoF	p-value	η^2_p
Condition	2.37	(2, 19.4)	0.12	0.20
Target	0.45	(1, 11785.3)	0.501	>0.01
Repetition	4.7	(2, 11781.1)	0.009**	>0.01
Correct	3.38	(1, 11661.6)	0.066	>0.01
Condition x Repetition	0.88	(2, 12055.1)	0.884	>0.01

Table 22. 30Hz Lab-grade (Model B)

Effect	F-statistic	DoF	p-value	η^2_p
Condition	6.89	(2, 11779.8)	0.001**	>0.01
Target	0.1	(1, 11778.4)	0.748	>0.01
Repetition	0.02	(2, 19.1)	0.164	0.17
Correct	0.1	(1, 11810.6)	0.876	>0.01
Condition x Repetition	1.42	(2, 11778.5)	0.224	>0.01

Table 23. 15Hz Wearable (Model D)

Effect	F-statistic	DoF	p-value	η^2_p
Condition	4.63	(2, 11356)	0.009**	>0.01
Target	0.85	(1, 11352)	0.355	>0.01
Repetition	0.193	(2, 11365)	0.829	>0.01
Correct	3.12	(1, 11364)	0.077	>0.01
Condition x Repetition	1.85	(2, 11355)	0.115	>0.01

Table 24. 30Hz Wearable (Model D)

Effect	F-statistic	DoF	p-value	η^2_p
Condition	0.15	(2, 11422)	0.862	>0.01
Target	0.14	(1, 11410)	0.705	>0.01
Repetition	2.52	(2, 11386)	0.08	>0.01
Correct	0	(1, 11411)	0.995	>0.01
Condition x Repetition	0.12	(4, 11420)	0.972	>0.01

3.4.4 Interim Summary

Task-load reliably decreased the SNR of the SSVEP's main frequency and its first harmonic in the MATB, with clusters extending far beyond the occipital electrodes. In the n-back condition, differences were only visible for the widest task-load contrast using parameterised estimates of the main frequency. The significant differences were lateralised on the left hemisphere for both the Lab-grade and Wearable data and not over occipital areas.

As previously mentioned in the methods, the differences in effects between both tasks were likely related to their variation in ocular activity and less to modulations of the SSVEP response throughout visual pathways. With increasing task load, the MATB required more saccades due to the rise in event rates. As a direct consequence of increased saccadic shifts, fixation periods were shortened, leading to the VEPs being interrupted more frequently, preventing them from becoming "steady" or "entrained". Furthermore, participants who were less adept at the task were more likely to look from the screen to their hands while managing communication, resource management, and system monitoring tasks. This possible shifting of the participants' gaze from the screen to the hands would have resulted in the SSVEP being completely interrupted, which could explain the reduction of average SNRs under increased workload conditions. However, the observed effects were strong and even occurred in the wearable headset, which did not possess any occipital electrodes. This finding could indicate that a low-amplitude SSVEP serves as a good, albeit noisy, proxy for ocular metrics of mental workload in a purely prediction-focused context.

To investigate whether increasing mental workload modulated the single-trial dynamics of the SSVEP, the n-back activity was analysed using RESS spatial filters to isolate the SSVEP response. Contrary to previous findings indicating that task-irrelevant visual stimulation offer unreliable task-load related information (Wilson & O'Donnell, 1988), both the Lab-grade and Wearable datasets exhibited n-back condition effects at the single-trial level. However, stimulus-specific effects akin to those reported by

Silberstein (1990) could not be observed and neither did correct and incorrect responses seem to have impacted the SNR of the SSVEP (Ladouce et al., 2025). Furthermore, the observed increases in SNR stand in contrast with previous reports reporting reduced SSVEP-based BCI accuracies under increased mental workload. Reductions in SNR were also expected due to A) the previously discussed sensor-level results, which only exhibited SNR reductions, and B) a recent study that indicated SNR reductions when participants were instructed to focus on interoception rather than visual stimuli on the screen (Kritzman et al., 2022), given that the 3-back required extensive mental rehearsal for successful target identification.

3.5 Discussion

In this chapter, the task-load settings employed in both tasks were shown to affect a range of different subjective, performance, and neurophysiological metrics. According to the RSME results, mental workload was successfully manipulated across experiments, which is why the following analyses refer to effects of mental workload, rather than task-load.

Classic band power metrics exhibited high variability across the different tests, suggesting they likely cannot offer a "one-size fits all" solution to mental workload classification. Correcting for aperiodic activation reduced effect sizes but also resulted in more interpretable effects that aligned more closely with the general literature (alpha decreases; frontal theta increases). Inconsistencies across tasks were to be expected and provided insights into why cross-task classification of mental workload is such a challenging problem. Lastly, time-on-task effects were most pronounced in the alpha band, in line with previous research.

With some of the expected effects not occurring consistently across datasets, a brief discussion regarding their absence is warranted. While the literature tends to report mostly significant effects, the file-drawer problem, indicated by Chikhi and colleague's (2022) funnel-plot analysis, may be reflected in the mixed results observed in this chapter. Three types of inconsistencies were observed: A) inconsistencies across tasks, B) inconsistencies across datasets, and C) inconsistencies across raw and aperiodic-free estimates.

Inconsistencies across tasks are of interest as they may indicate the shortcomings of the spectral metrics in representing generalisable mental workload metrics. If they were generalisable, one could argue that the subjective workload differences between tasks were the reason for the inconsistencies; for example, the n-back workload was smaller and thus the power metrics were simply not sensitive enough. Another option could be that the effects of mental workload on spectral power are task-specific, relating to Wicken's MRT of mental workload.

Inconsistencies across datasets may be due to inter-individual variation across the population. Increasing the statistical power by increasing the sample size would have likely led to more consistent effects across datasets. Especially the tests for Wearable and Multimodal datasets were of less statistical power due to the lack of spatial pooling in the TFCE tests; a higher sample size may have counteracted this. However, the need for larger sample sizes to counteract variations in effects in the population suggests power spectral metrics to likely not generalise well across subjects in applied pBCI contexts.

Lastly, the reasons for inconsistencies across raw and aperiodic-free estimates may be twofold. In theory, removing "noise" should increase effect sizes. However, aperiodic activity may fundamentally contribute to the effects reported in the literature, meaning the effects are not of an oscillatory nature alone and removing the background activity causes effects to vanish, reduce, or change direction. The other option is that the aperiodic fits carried out in this study were introducing noise themselves into the band power estimates, thereby lowering statistical power. However, residual errors of the AP fits were generally low, ranging from .01 to .06, making this latter explanation less likely.

As opposed to Ke and colleagues (2023), the aperiodic activation was not tested per band in the previous analysis. Instead, section 3.3.5 tested for condition differences using the exponent describing the 1/f slope directly. Its effects were not as widespread as to fully account for the disappearance of some of the more notable effects visible in the raw estimates. Another possibility could be that mental workload-related changes in the beta and lower gamma range (30-40Hz) may have biased the fit of the reported slope, thus subtracting too much activation from the band range in question for the higher MATB levels. This again was not visible in the goodness of fit measures, but their consistency alone cannot rule out subtle influences of the fitting procedure.

The rapid-auditory probing paradigm did not exhibit consistent workload-related effects. Rather than invalidating the results reported by Sugimoto and colleagues (2022), it qualifies the sensitivity of the method. The Wearable and Lab-grade datasets exhibiting significant effects for the widest MATB contrast, but not the n-back, suggested that the paradigm may not be useful in more nuanced monitoring scenarios. Instead, akin to ECG-based metrics, it holds potential for all-or-nothing types of monitoring tasks – i.e. "is the operator currently highly taxed or idle?"

The SSVEP results were highly significant for the MATB, in which task load likely covaried heavily with ocular activity. In the n-back, SNR differences were sparse and only became apparent using a parameterised power spectrum. After spatial filtering, the single-trial SNR of the SSVEP seemed to have increased at more difficult n-back settings rather than decreased, as indicated by the sensor-

level tests. While an SNR increase was contrary to our expectations, it may be linked to higher n-back levels requiring higher levels of focus

Importantly, most effects found in the 64-channel montages were, sometimes to a lesser degree, also usually present in the low-density wearable montage. This provides support for future workload monitoring applications that won't require the labour-intensive, high-density, gel-based systems.

4 Evaluation of Bias in Cross-Validation Methods for Passive BCIs

In the previous chapter, many of the tested EEG metrics showed sensitivity to changes in mental workload in some way or another. However, results suggesting a given metric is sensitive to changes in mental workload at the population level may not necessarily translate to high classification accuracy at the single-trial level. At the single-trial level, measurement noise and non-stationarities cannot be "averaged out", resulting in classification models having to deal with high variance and drifting data distributions that complicate the development of neurophysiological-based classifiers that can generalise across recording sessions, tasks, and people (Krusienski et al., 2011; Lotte et al., 2018; Saha & Baumert, 2020). Non-stationarities inherent to neural time-series data are highly subject- and context-dependent (Krauledat, 2008; Mayer-Kress, 1998). Of primary interest to this chapter is the fact that these non-stationarities may not just complicate the development of applied pBCIs due to shifting data distributions (Germano et al., 2023; Miladinović et al., 2021) but also because they may inflate pBCI model evaluation metrics. If continuous shifts in neural time-series are ignored, many offline model evaluation procedures can positively bias accuracy estimates by allowing models to capitalise on temporal dependencies rather than class-related differences in the data. While this problem has been pointed out before (Brouwer et al., 2015; Lemm et al., 2011), many studies either still outright ignore the problem (Li et al., 2020) or do not report their evaluation methods with enough detail (Demirezen et al., 2024) for the reader to judge whether the issue has occurred. Positively biased evaluation metrics that essentially inform future research (such as choices regarding preprocessing, feature selection, and classification techniques) counteract the desirable goal of increasing reproducibility within pBCI research.

In the field of mental state classification, a considerable portion of research focuses on the development of new methodologies and comparative analyses of existing approaches (Demirezen et al., 2024). While it would be ideal to evaluate these methods in applied settings (Aricò et al., 2018; Lotte et al., 2018), in practice, achieving such evaluations poses considerable challenges (Brouwer et al., 2015). As a result, most model evaluations are conducted offline, typically through cross-validation procedures (Lemm et al., 2011).

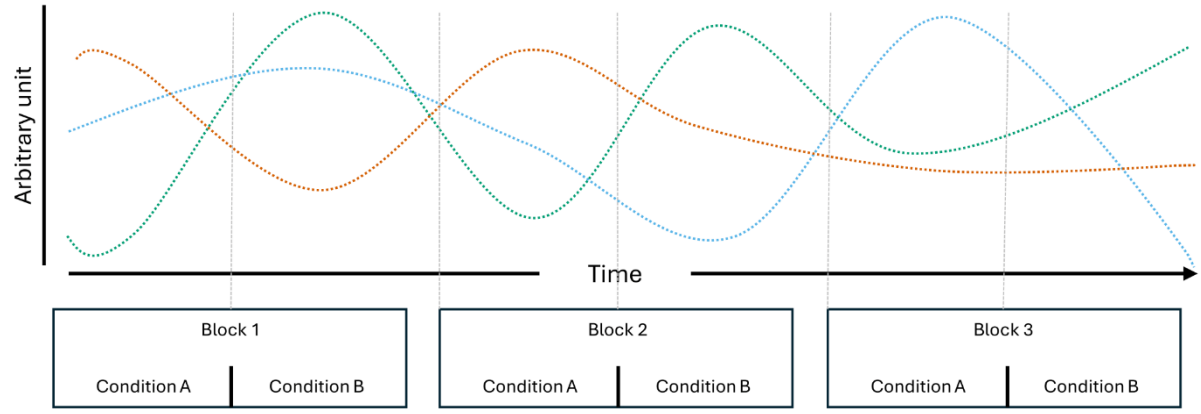
Cross-validation serves as a key technique in this context, as it maximizes the use of available data by repeatedly partitioning data into training and testing subsets to compute evaluation metrics. The

This chapter is based on the article: Schroeder, F., Fairclough, S., Dehais, F., & Richins, M. (2025). The Impact of Cross-Validation Choices on pBCI Classification Metrics: Lessons for Transparent Reporting. *Frontiers in Neuroergonomics*, *6*. https://doi.org/10.3389/fnrgo.2025.1582724

choice of cross-validation method is crucial, as it directly impacts the bias and variance of the evaluation metrics. Reducing bias typically requires larger training splits, which enhance model accuracy, whereas minimizing variance often necessitates larger testing splits, offering more robust estimates of evaluation metrics across iterations (Lemm et al., 2011).

Despite the ubiquity of cross-validation in the field, its implementation often lacks transparency (Li et al., 2020). In a review conducted by Demizeren and colleagues (2024), 93% of studies reported the cross-validation method used, while only 25% provided specific details regarding their data-splitting procedures. This lack of clarity complicates efforts to improve reproducibility - an issue highlighted as critical by researchers in the field (Gramann et al., 2024; Putze et al., 2022). Insufficient documentation of cross-validation details can hinder the assessment of bias-variance trade-offs and, in some cases, obfuscate issues regarding temporal dependencies between train and test splits (Li et al., 2020). In some cases, classification may actually be driven by temporal dependencies rather than class differences (Ivucic et al., 2024; Lemm et al., 2011; Li et al., 2020; Riascos et al., 2024; Varoquaux et al., 2017; White & Power, 2023).

Figure 39. Schematic of Multivariate Temporal Dependencies



Note. Three individual processes, which, through their combinations, suffice to uniquely identify the individual condition repetitions presented over three blocks.

Temporal dependencies in neuroimaging data are likely to arise from various sources and exist across multiple timescales. Not only are they inherent to neural time-series (Bullmore et al., 2001; Linkenkaer-Hansen et al., 2001), but they may also be introduced due to experimental design choices. The recording hardware itself may be one source of such dependencies, such as when there are minor shifts or movements in the positions of EEG sensors. Other dependencies may stem from cognitive or behavioural factors. For instance, participants who start the session feeling nervous might gradually relax as they adapt to the experimental conditions. Increasing drowsiness may be

This chapter is based on the article: Schroeder, F., Fairclough, S., Dehais, F., & Richins, M. (2025). The Impact of Cross-Validation Choices on pBCI Classification Metrics: Lessons for Transparent Reporting. *Frontiers in Neuroergonomics*, *6*. https://doi.org/10.3389/fnrgo.2025.1582724

visible in the theta and alpha (Strijkstra et al., 2003) as well as beta band of the EEG power spectrum (Aeschbach et al., 1997). Temporal dependencies may also present in more complex forms, as increasing drowsiness also affects theta- and alpha-specific connectivity metrics as well as the occurrence and prominence of microstates (Comsa et al., 2019). Initial nervousness, on the other hand, is often visible in heart rate dynamics (Lampert, 2015), which may affect the aperiodic activity (1/f slope) of the EEG power spectrum (Schmidt et al., 2024). Effects of bodily needs (i.e., hunger, thirst, dry eyes, caffeine/nicotine craving, etc.) may also exert influence after a certain point and affect EEG dynamics in unsuspecting ways. For example, when participants experience eye strain, they may start squinting their eyes, leading to power increases in higher frequency bands caused by the activations of facial muscles. In cases where data is split irrespective of the underlying block structure, the combination of a myriad of ongoing processes gives rise to multivariate temporal dependencies (Figure 39), which likely offer more information for classification than the class differences themselves (Ivucic et al., 2024; Li et al., 2020; Varoquaux et al., 2017; White and Power, 2023). The ways in which such temporal dependencies bias model evaluation metrics likely vary across cross-validation implementations, feature types, and classification algorithms, ultimately rendering conclusions drawn from underspecified cross-validation schemes unreliable sources of information.

The following section outlines previous work that explored the impact of different cross-validation schemes in neuroimaging research.

Varoquaux et al. (2017) showed that leave-one-sample-out cross-validation schemes can inflate accuracy metrics due to temporal dependencies, overestimating performance across different fMRI decoding studies by up to 43% compared to evaluations on independent test sets. Model evaluation metrics from Leave-one-sample-out schemes are not only prone to bias from temporal dependencies but also suffer from high variance, due to the test set consisting of a single sample (Lemm et al., 2011, Varoquaux et al., 2017). K-fold cross-validation reduces the variance of model evaluation metrics by splitting the available sample data into k subsets, of which k-1 are used for training and the remaining subset is used to compute the evaluation metrics. This procedure is repeated k times until each subset was once used as a test set. However, temporal dependencies may also bias the results of k-fold cross-validation schemes when the available data is split into subsets without taking account of the underlying block/trial structure of the data. Ivucic et al. (2024) demonstrated that k-fold splits independent of trial structures caused inflated accuracy estimates in three open access EEG datasets dealing with auditory attention detection. Another inquiry demonstrated that deep-

This chapter is based on the article: Schroeder, F., Fairclough, S., Dehais, F., & Richins, M. (2025). The Impact of Cross-Validation Choices on pBCI Classification Metrics: Lessons for Transparent Reporting. *Frontiers in Neuroergonomics*, *6*. https://doi.org/10.3389/fnrgo.2025.1582724

learning approaches for image and video classification based on EEG data fail entirely when stimuli are presented in a randomised rapid event-related fashion instead of a blocked design (Li et al., 2020).

With respect to pBCIs, one function of this technology is to differentiate between brain states (or cognitive/mental states), which may be considered more diffuse targets compared to the decoding of perceptual information. Manipulations aiming to induce specific emotional states or states of high mental workload may add systematic confounds that can bias model metrics (i.e. conditions differing in motor requirements; Brouwer et al., 2015). However, one major difference to domains like percept decoding or motor BCIs is that conditions tend to be presented in longer blocks of a single condition rather than short trials that allow for rapid event-based presentations with randomised condition orders. Interleaving 1 to 5-second-long trials of different motor conditions (e.g., left arm vs. right arm) or image conditions (e.g., houses vs. faces) assures that temporal dependencies are evenly spread across conditions. In the case of the Multi-Attribute Task Battery (MATB), a popular workload manipulation paradigm, a recent review has found duration of single condition presentations to range from 4 – 15 minutes (Pontiggia et al., 2024). Another popular paradigm for manipulating mental workload is the n-back, where block durations can be as short as 40 seconds (Shin et al., 2018) or as long as 10 minutes (Ke et al., 2021). Such long block durations increase the number of samples that share not just condition-specific dynamics but also the same temporal dependencies. Designing experiments with long blocks also reduces the number of repetitions of single conditions, which could be presented in a randomised order, in a standard-length recording session. Together, this range of factors complicates the evaluation of pBCIs in offline analyses.

White and Power (2023) focused on mental state classification and investigated the difference between block-independent k-fold splits and block-wise splits that assured samples from a single trial/block did not occur in both train and test subsets. Using open access EEG datasets manipulating emotional valence, they showed that k-fold accuracies were systematically higher than block-wise accuracies. They further showed that randomly reassigning class labels to half the blocks did not reduce the accuracy of the k-fold evaluations, concluding that the classifiers evaluated via k-fold made use of temporal dependencies rather than class differences. Further, they collected their own data, varying the trial durations of single condition repetitions (5, 15, and 60 seconds), thereby manipulating the number of samples sharing both class-labels and temporal dependencies. Here again, they demonstrated that a trial structure independent k-fold scheme overestimated accuracies, even using the short 15 seconds block durations. However, they also argue that their tested

This chapter is based on the article: Schroeder, F., Fairclough, S., Dehais, F., & Richins, M. (2025). The Impact of Cross-Validation Choices on pBCI Classification Metrics: Lessons for Transparent Reporting. *Frontiers in Neuroergonomics*, *6*. https://doi.org/10.3389/fnrgo.2025.1582724

classifiers seemed to overfit on block-specific temporal dependencies when evaluated using the block-wise cross-validation scheme, leading to underestimated performance metrics (compared to extracting a single sample per block in the 5-second condition). While this study offered insight into how temporal dependencies can bias pBCI cross-validation results, their results may underestimate the issue, as their maximum trial length (60 seconds) was not reflecting the long block durations usually reported in pBCI experiments. Furthermore, all their tested classifiers used canonical band power features. However, the extent to which temporal dependencies bias cross-validation results may differ between the type of feature extracted from EEG, especially when additional dimensionality reduction techniques are applied during the training phase.

In the current chapter, three separate n-back datasets were used to deepen the inquiry into biased cross-validation. By I) exploring multiple classification pipelines and II) adding additional cross-validation schemes, the following analysis expands on previous works (Ivucic et al., 2024; Varoquaux et al., 2017; White & Power, 2023). As cross-validation methods tend to ignore chronological order, they may underestimate the impact non-stationarities can have on model metrics (Riascos et al., 2024). Hence, a pseudo-online evaluation method was also included, in which only the very first occurrence of a condition was used for training. Lastly, a worst-case scenario in which k-fold splits are not carried out on sequentially ordered data but rather on shuffled samples, which likely exacerbates the bias (Brouwer et al., 2015; Riascos et al., 2024), was also included.

4.1 Methods

To assess the impact of various cross-validation schemes on pBCI evaluation metrics independently of experimental design decisions, this analysis made use of the n-back data of the minimally processed Pilot datasets as well as two open-access datasets (Hinss et al., 2023; Shin et al., 2018), which also included three repeated presentations of three n-back conditions. The Pilot and two open-access datasets were utilised to avoid issues of double-dipping in the following machine learning chapters of the thesis, in which the focus will lie with the Lab-grade, Wearable, and Multimodal datasets.

The three datasets used in the current analysis differed in several aspects, including the n-back conditions employed, their presentation order, the intervals between repeated condition presentations, and the specifics of their EEG montage configurations. Following a general overview of the individual datasets, the methods for preprocessing and data selection to improve comparability are described.

This chapter is based on the article: Schroeder, F., Fairclough, S., Dehais, F., & Richins, M. (2025). The Impact of Cross-Validation Choices on pBCI Classification Metrics: Lessons for Transparent Reporting. *Frontiers in Neuroergonomics*, *6*. https://doi.org/10.3389/fnrgo.2025.1582724

4.1.1 *Dataset Descriptions*

Because the three datasets used slightly different terminology, it is important to first establish a consistent vocabulary for the following sections. At the smallest level, a single stimulus-response sequence was called a **trial**. Consecutive trials with the same n-back condition, bounded by break periods, formed a **run**. Several runs, typically including all n-back conditions, made up a **block**. In two of the datasets, n-back blocks were interleaved with other tasks.

For the purpose of this chapter, the original block structures of the datasets were broken up and runs were regrouped into **sets**. A set contained one full instance of all n-back conditions. By reshaping all three datasets to consist of three sets, which grouped together data that were collected close in time relative to other such groupings, the cross-validation methods of interest could be applied in a consistent manner across datasets. Finally, a **session** referred to a complete EEG recording visit for a participant. Only one of the three datasets included more than one session per participant, recorded on separate days.

Shin et al., 2018. This study involved 26 participants completing 0-back, 2-back, and 3-back conditions. Numbers from 0-9 were used as stimuli, with targets making up 30% of the 20 trials that were presented per run. Each trial presented the stimulus for 0.5s, followed by a 1.5s fixation cross. Participants were instructed to respond to both targets and non-targets using a single hand (numpad 7 and 8 keys). Participants received instructions at the start of each run, but the authors provided no details about possible training periods before the data recording.

Conditions were presented in a counterbalanced order in blocks of 9 with 20s breaks in-between runs. Three of these blocks were completed one after another before the experiment continued with different tasks not relevant to the current inquiry. It is unclear if participants took breaks between blocks, and if so, how much time passed between blocks.

EEG was recorded at 1000Hz with 30 active electrodes arranged according to the 10-5 system. Electrode TP9 was used as the online reference and TP8 as the ground. The authors did not report details about the impedance of the EEG electrodes.

Hinss et al., 2023. This study involved 29 participants completing 0-back, 1-back, and 2-back conditions. Numbers from 1-9 were used as stimuli, with targets making up 33% of the 48 trials that were presented per run. Each trial presented the stimulus for 0.5s, followed by a 1.5s fixation cross. Participants were instructed to only respond to targets using a single hand (spacebar).

This chapter is based on the article: Schroeder, F., Fairclough, S., Dehais, F., & Richins, M. (2025). The Impact of Cross-Validation Choices on pBCI Classification Metrics: Lessons for Transparent Reporting. *Frontiers in Neuroergonomics*, *6*. https://doi.org/10.3389/fnrgo.2025.1582724

Conditions were presented in a blocked format, each block containing 3 runs of a single condition without any reported times regarding breaks in-between runs. The blocks were randomly spread out over a 65-80 minute-long recording period containing other tasks not relevant to the current inquiry, so the exact time between conditions is unknown. However, a special difference to the previous datasets is that participants returned twice (one and two weeks after the first session) to repeat the experiment.

EEG was recorded at 500Hz with 63 active electrodes arranged according to the 10-20 system. One electrode was sacrificed to record ECG. Electrode FCz was used as the online reference and Fpz as the ground. Electrode impedance was kept below 25kOhm during the experiment.

Pilot dataset. 20 participants completed 1-back, 3-back, and 6-back conditions. A selection of letters (B, F, G, H, K, M, P, R, S, T, X, Z) was used as stimuli, with targets making up 30% of the 70 trials that were presented per run. Each trial presented the stimulus for 0.5s followed by a 1.5s question mark. Participants were instructed to respond to both targets and non-targets using both hands (Keyboard Z and M keys). Participants received extensive training (3 runs per condition) on a day prior to the experiment and practised once more before recording data (1 run per condition).

Conditions were presented in blocks of 3, containing all 3 conditions in a randomized order. Exact data for how much time passed between runs within a block is not clear as participants were instructed to take self-paced breaks. In-between blocks, participants completed another task not relevant to the current inquiry, leading to at least 15 minutes separating the three blocks.

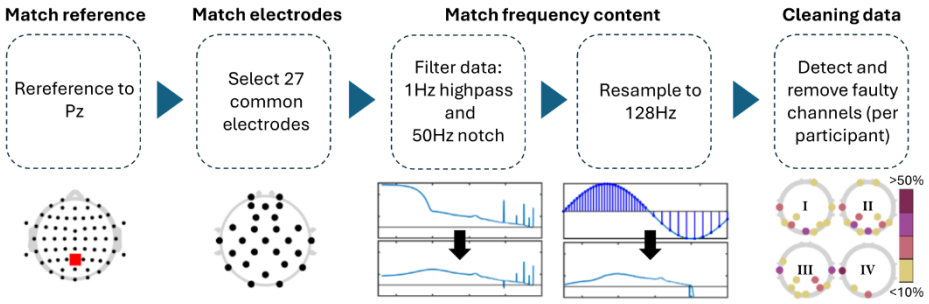
EEG was recorded at 500Hz with 64 active electrodes arranged according to the 10-20 system. The Iz electrode was sacrificed and instead used as a peri-ocular EEG channel. Electrode FCz was used as the online reference and Fpz as the ground. Electrode impedance was kept below 25kOhm during the experiment. The Following analysis uses the minimally processed data.

4.1.2 Dealing with the differences between datasets

Differences in the EEG recordings concerned the sampling frequency, online reference, number of channels used, and, to a lesser extent, the spacing of the electrodes. In order to make the EEG data more comparable across datasets, all three datasets were preprocessed to contain the same channel locations, reference electrode, and frequency content (Figure 40).

This chapter is based on the article: Schroeder, F., Fairclough, S., Dehais, F., & Richins, M. (2025). The Impact of Cross-Validation Choices on pBCI Classification Metrics: Lessons for Transparent Reporting. *Frontiers in Neuroergonomics*, *6*. https://doi.org/10.3389/fnrgo.2025.1582724

Figure 40. Combined Preprocessing for All Datasets



Note. All preprocessing steps carried out to make datasets more comparable. The topologies under the data cleaning step showcase that channel removal was not homogeneous across experiments, but mostly focused on posterior channels (I: single-day Hinss et al, II: multi-day Hinss et al, III: Shin et al, IV: Pilot dataset).

Table 25. Channel Removal Across Datasets

Dataset	Average # channels removed	Maximum # channels removed
I. *Single-day Hinss et al*	0.86	5
II. *Multi-day Hinss et al*	1.69	6
III. *Shin et al*	1.19	4
IV. *Pilot dataset*	0.79	2

*Note***.** Average number of channels removed across participants and maximum number of channels removed for a single participant per dataset.

For the comparison, all channels in the Pilot dataset and Hinss et al data not contained in Shin et al were removed. The only electrodes contained in Shin et al that were not recorded in the other two datasets were AFF4 and AFF5. These are close in space to AF4 and AF5, which were retained in their stead in the Pilot and Hinss et al data. Since Cz was missing in the first 9 participants of Hinss et al, it was also removed from all three datasets. FCz was used as the online reference in the Pilot data and was recovered after computing a common average reference. For the comparison, all data was re-referenced to Pz, as this was one of only two shared midline electrodes between the datasets. Additionally, all data was notch filtered at 50Hz (zero-phase, non-causal, with –6db cutoff frequencies at 49.25 and 50.75), highpass filtered at 1Hz (zero-phase, non-causal, with -6dB cutoff frequency at .5Hz), and resampled to 128Hz to assure comparable frequency content across datasets. Lastly, EEGLAB's clean_rawdata function was used to remove channels when their correlation with neighbouring channels was below 0.8 (see Table 6 for details). While the datasets

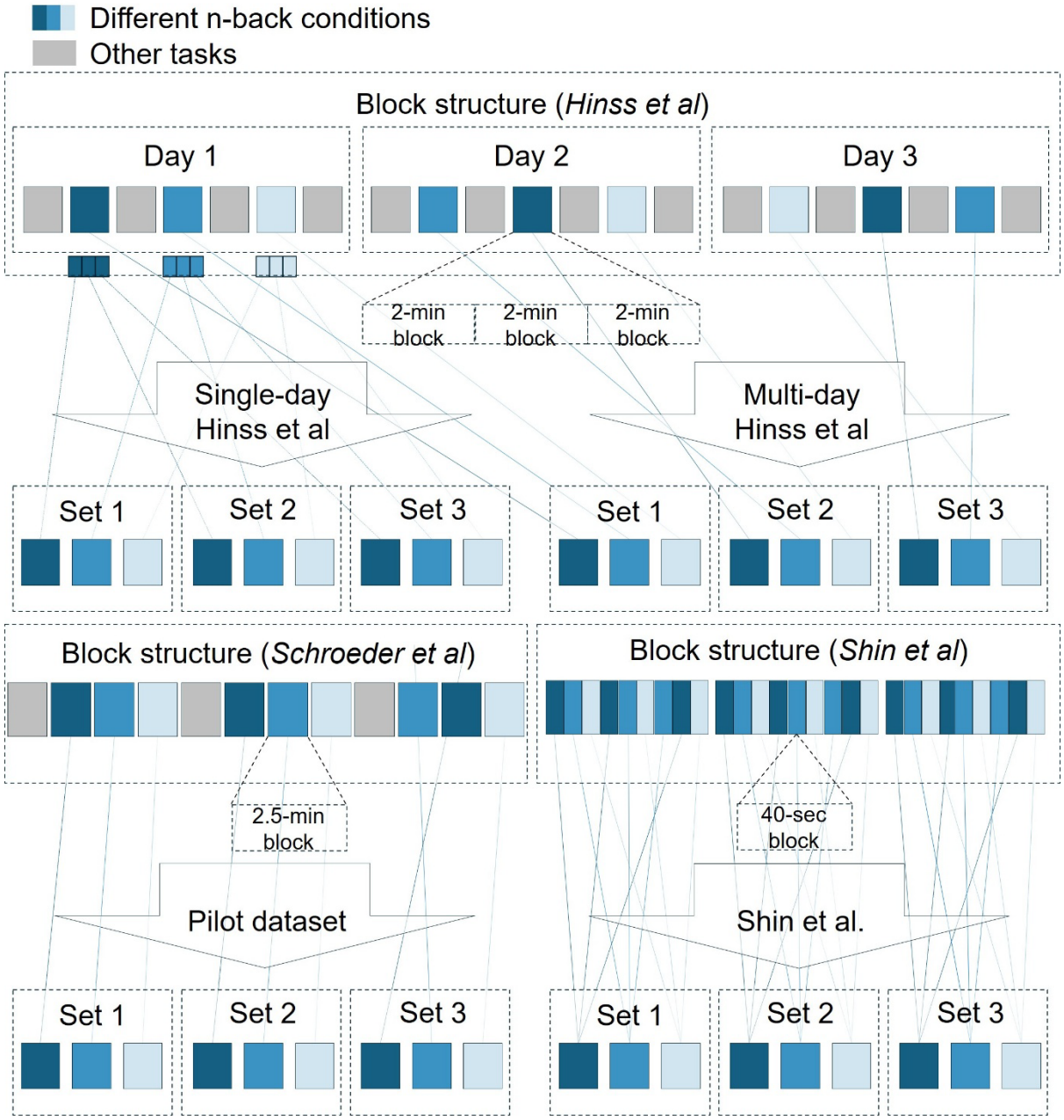
This chapter is based on the article: Schroeder, F., Fairclough, S., Dehais, F., & Richins, M. (2025). The Impact of Cross-Validation Choices on pBCI Classification Metrics: Lessons for Transparent Reporting. *Frontiers in Neuroergonomics*, *6*. https://doi.org/10.3389/fnrgo.2025.1582724

listed above all made use of the n-back task, their n-back implementations likely differed in how demanding they were. The Pilot dataset placed the greatest working memory load on participants by utilising a 6-back. However, the impossible nature of the 6-back may have led to effort withdrawal, rendering the condition less taxing than the dataset's 3-back. Both Hinss et al and Shin et al presented a 0-back, which removes the working memory aspect of the n-back and could be considered psychometrically distinct from other n-back conditions. For the comparison we decided to follow previous works described in the introduction and included two contrast per dataset – the widest difference in workload comparing the easiest and hardest conditions making for a higher class-separability contrast (Shin et al: 0 vs 3-back; Hinss et al: 0 vs 2-back; Pilot dataset: 1 vs 6-back) and the smallest difference in workload which we determined was the hardest and second hardest conditions making for a lower class-separability contrast (Shin et al: 2 vs 3-back; Hinss et al: 1 vs 2-back; Pilot dataset: 3 vs 6-back).

Also important for the current inquiry was the order and spacing with which the n-back conditions were presented. Within a single session, the Hinss et al dataset did randomise condition order, but repetitions of a single condition were grouped together in a single block. This likely results in temporal dependencies being highly informative to distinguish between conditions. However, the addition of two further recording days leaves this dataset with the most spread-out condition repetitions. For the Pilot data as well as Shin et al, each participant's data was recorded on a single day with three spaced-out blocks of n-backs containing all three conditions in a pseudo-randomised order (Figure 41). Shin et al placed these three blocks one after another, whereas condition repetitions were spaced out with the MATB blocks in between in the Pilot dataset. The blocks themselves were designed differently in Shin et al and the Pilot dataset. In the Pilot dataset, blocks contained one 140-second run per condition, whereas Shin et al presented each condition three times in smaller 40-second runs concatenated together in a pseudo-random order, assuring a single condition was not presented twice in a row. Consequently, temporal dependencies are likely somewhat less informative for classification in the Shin et al data compared to the other two datasets.

This chapter is based on the article: Schroeder, F., Fairclough, S., Dehais, F., & Richins, M. (2025). The Impact of Cross-Validation Choices on pBCI Classification Metrics: Lessons for Transparent Reporting. *Frontiers in Neuroergonomics*, *6*. https://doi.org/10.3389/fnrgo.2025.1582724

Figure 41. Block Structures and Splitting Procedures



Note. For each dataset, the available data were divided into three equal-sized sets. Temporal dependencies were anticipated to be stronger within sets than between them, as n-back blocks presented in conjunction with one another in the original experiments were grouped together within a single set. The presentation order of the n-back conditions was not necessarily as depicted in the schematic.

For the comparison, each participant's data was arranged into three sets (Figure 41). For Hinss et al, two separate procedures were adopted. The first only made use of the first recording day (96 seconds per condition per set) as an example of the additional bias introduced by the lack of a randomised and spaced-out presentation order. The second procedure made use of all three recording days. Each day was used as a separate set (288 seconds per condition per set). For Shin et

This chapter is based on the article: Schroeder, F., Fairclough, S., Dehais, F., & Richins, M. (2025). The Impact of Cross-Validation Choices on pBCI Classification Metrics: Lessons for Transparent Reporting. *Frontiers in Neuroergonomics*, *6*. https://doi.org/10.3389/fnrgo.2025.1582724

al, runs of single condition (40 seconds each) within a block were sliced out of the continuous data and merged into a single file (120 seconds per condition per set). For the Pilot dataset, the three condition repetitions were already spread out over the course of the experiment (140 seconds per condition per set). Samples within a set are expected to share more condition-unrelated information with each other, compared to samples from different sets, due to their temporal proximity in the original experiments (excluding the single-day Hinss et al sets).

4.1.3 Classification approaches

All features were computed from 2-second windows, corresponding to the length of a single n-back. Windows were extracted without overlap from the continuous data, after classifier-specific filter operations were carried out. By extracting non-overlapping windows, we ensured that any bias to the performance metrics in the k-fold cross-validation schemes stemmed from underlying temporal dependencies and not from reusing the same data in successive samples. The classification approaches throughout the thesis were implemented using the pyRiemann (v0.7; Barachant et al., 2025), sci-kit learn (v1.2.2; Pedregosa et al., 2011) and MNE (v1.6.1; Gramfort et al., 2014) python libraries. The following lists the four approaches contained in this chapter in the order in which they were expected to overfit on training data specific temporal dependencies (likely showing greater bias in block-structure independent cross-validation strategies). This order was based on two ideas. The first being that with an increasing number of free parameters, the propensity of a classifier to overfit to training-specific information increases (Domingos, 2012; Lemm et al., 2011). The second being that Riemannian classification has previously been shown to generalise well to unseen data (Congedo et al., 2017; Yger et al., 2017).

Broadband Riemann Minimum Distance to Mean (RMDM). Before windowing, the data for this classifier was bandpass filtered (1Hz – 25Hz; default MNE FIR filter with -6 dB cutoff frequency at 0.50 Hz and 28,12). Covariance matrices were computed per 2-second window with ledoit-wolf shrinkage (Ledoit and Wolf, 2003) to ensure semi-positive definite matrices. No hyperparameters were tuned, limiting the chance to overfit on trends in the training data. Classification was carried out using a Riemann minimum distance classifier (Barachant et al., 2010, 2012).

Narrowband Riemann Minimum Distance to Mean (Narrow-RMDM). Using a filter bank of Butterworth filters (zero-phase, non-causal, passband ripple = 3db, stopband attenuation = 10db, transition bandwidth = 1Hz), the data for this classifier was separately bandpass filtered with passbands in canonical delta (1 – 4Hz), theta (4 – 7Hz), alpha (8 – 12Hz), and beta (13 – 25Hz) ranges.

This chapter is based on the article: Schroeder, F., Fairclough, S., Dehais, F., & Richins, M. (2025). The Impact of Cross-Validation Choices on pBCI Classification Metrics: Lessons for Transparent Reporting. *Frontiers in Neuroergonomics*, *6*. https://doi.org/10.3389/fnrgo.2025.1582724

Ledoit-wolf shrinkage was carried out separately per frequency bin, after which all four frequency bins' covariance matrices were combined into a diagonal block matrix for classification. This procedure was inspired by the block-diagonal matrices proposed for the classification of SSVEPs (Congedo et al., 2017) and previous efforts to focus Riemannian classification on specific neurophysiological aspects (Näher et al., 2024; Yamamoto et al., 2023). To keep the computational requirements manageable, a previously proposed algorithm for efficient electrode selection on covariance matrices (Barachant and Bonnet, 2011) was used to reduce each frequency bin to its most informative combination of 8 electrodes at each training step. This pipeline was included as an example of a Riemannian classifier with an added train-set specific tuning step. Due to the fact that different channels could be selected per frequency band, all off-diagonal elements in the block-covariance matrix (theoretically containing cross-frequency coupling information) were set to 0.

Narrowband power LDA (PSD-LDA). After windowing, we computed the power spectral density within canonical delta, theta, alpha, and beta ranges per electrode using a one-dimensional discrete Fourier Transform. During training, all features were normalised using the mean and standard deviation of the training set. Additionally, every training iteration, the 18 most informative and least correlated features were selected from the 4 (frequency bands) x n_electrodes number of computed features using a minimum redundancy maximum relevancy algorithm (Peng et al., 2005). Classification was carried out using an sLDA model.

Filter Bank Common Spatial Pattern LDA (FBCSP). Using a filter bank of Butterworth filters (zero-phase, non-causal, passband ripple = 3db, stopband attenuation = 10db, transition bandwidth = 1Hz), the data for this classifier was separately bandpass filtered into 4Hz wide frequency bins ranging from 3-25Hz in steps of 2Hz. Each filtered signal was used to compute 8 spatial filters via common spatial pattern analysis (4 highest and 4 lowest eigenvalues) (Ang et al., 2012; implemented in MNE) and the log-variance of the filtered 2-second windows was extracted as the classification feature. Every training iteration, a subset of 18 features was selected from the 80 computed features (10 frequency bins x 8 spatial filters) using a minimum redundancy maximum relevancy algorithm to select the most informative and least correlated features (Peng et al., 2005). Classification was also carried out using am sLDA model.

4.1.4 Cross-validation strategies

Four different validation strategies were tested (Figure 42), ranging from likely producing conservative performance estimates to likely producing inflated performance estimates and

This chapter is based on the article: Schroeder, F., Fairclough, S., Dehais, F., & Richins, M. (2025). The Impact of Cross-Validation Choices on pBCI Classification Metrics: Lessons for Transparent Reporting. *Frontiers in Neuroergonomics*, *6*. https://doi.org/10.3389/fnrgo.2025.1582724

presented in that order below. Classification accuracy was chosen as the performance metric, as all cross-validation methods used assured no class imbalances in the train or test sets (see Table 26 for exact sample sizes).

Pseudo-online. The most conservative of the four tested cross-validation strategies. In it, only the first set of conditions was used for training, and the remaining two sets were used for testing. This evaluation scheme represents situations in which limited calibration data is available, such a "cold-starting" a new model for a user in an applied setting. The single training set mimics a calibration round while the test sets offer two separate opportunities to estimate classifier performance without making use of future data, as is the case in the following cross-validation strategies.

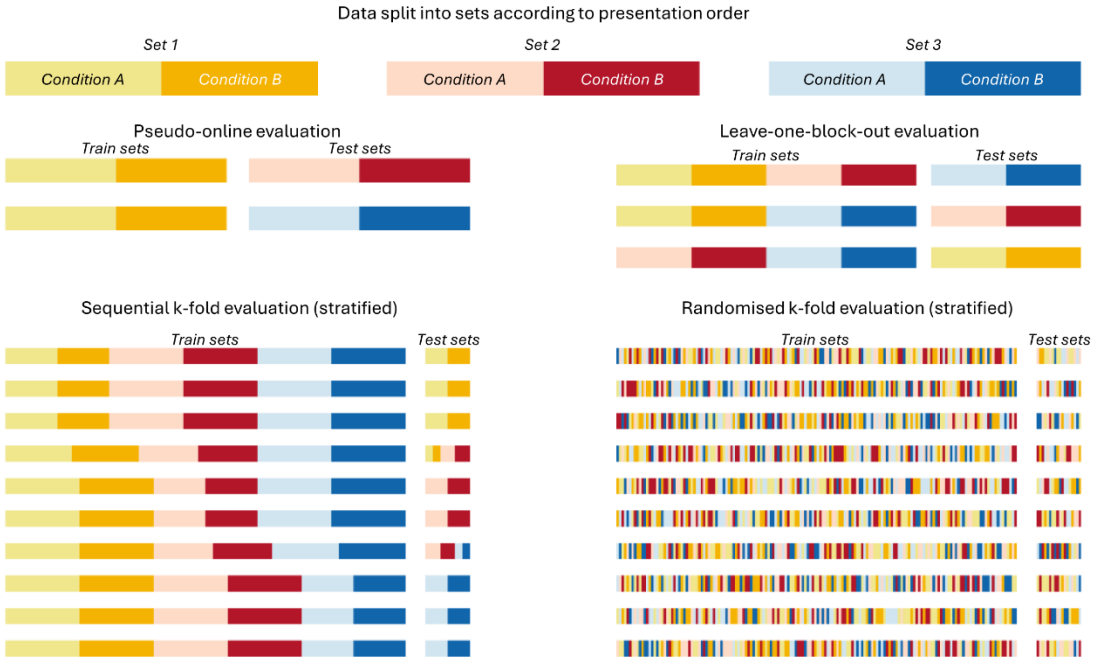
Leave-one-block-out. This strategy assured that no data of the same set occurred in both the testing and training data. However, using this strategy, some folds will use training data that occurred after the testing data. Since this is impossible for a real BCI system, its performance estimates could be considered somewhat artificial and might be overestimated (Riascos et al., 2024).

Sequential K-fold. This is a common default cross-validation strategy in which the data are split into, here, 10 equal-sized segments. 9 of the 10 segments are used for training, while the 10th is held out for testing. This process is repeated 10 times until all samples were once used for testing. The risk of splitting the data into equal sized segments of an arbitrary size is that data from a single block may occur both in the training and testing sets within a single fold. A stratified k-fold procedure was employed to avoid class imbalances.

Randomised K-fold. Here as well, the data were split into 10 equal-sized segments, using a stratified procedure to avoid class imbalances. However, in this version, all samples are first shuffled randomly before being split into 10 folds, a step that should only be performed if all observations are statistically independent from each other. It was included as an example of the worst-case scenario for overestimating classifier performance on data from block-based experiments.

This chapter is based on the article: Schroeder, F., Fairclough, S., Dehais, F., & Richins, M. (2025). The Impact of Cross-Validation Choices on pBCI Classification Metrics: Lessons for Transparent Reporting. *Frontiers in Neuroergonomics*, *6*. https://doi.org/10.3389/fnrgo.2025.1582724

Figure 42. Visualisation of Cross-Validation Strategies



Note. Visualisation of how the separate sets built in Figure 40 were split for training and testing the classifiers within the four cross-validation procedures.

4.1.5 *Statistical Analysis*

As a first step, the distributions and descriptive statistics of subject-wise classification accuracies were visualised for each dataset, cross-validation strategy, and classification approach (see Figure 42).

To analyse differences between cross-validation strategies, bootstrapped 95% confidence intervals of the differences in accuracy between the conservative pseudo-online cross-validation scheme and the other tested schemes were computed over 10.000 iterations per classifier. Each iteration randomly sampled 15 subjects with replacement across datasets and class-separability contrasts to compute the 3 difference scores per classifier (within-subject).

Table 26. Average train/test Sample Sizes

CV-strategy	Dataset	Condition A training samples	Condition B training samples	Condition A testing samples	Condition B testing samples
Pseudo-online	Single-day Hinss et al	51	51	98	98

This chapter is based on the article: Schroeder, F., Fairclough, S., Dehais, F., & Richins, M. (2025). The Impact of Cross-Validation Choices on pBCI Classification Metrics: Lessons for Transparent Reporting. *Frontiers in Neuroergonomics*, *6*. https://doi.org/10.3389/fnrgo.2025.1582724

	Multi-day Hinss et al	158	158	295	295
	Pilot data	70	70	140	140
	Shin et al	60	60	120	120
Leave-one-block-out	Single-day Hinss et al	102	102	49	49
	Multi-day Hinss et al	310	310	149	149
	Pilot data	140	140	70	70
	Shin et al	120	120	60	60
Sequential k-fold	Single-day Hinss et al	138	138	14	14
	Multi-day Hinss et al	418	418	44	44
	Pilot data	189	189	21	21
	Shin et al	162	162	18	18
Randomsied k-fold	Single-day Hinss et al	138	138	14	14
	Multi-day Hinss et al	418	418	44	44
	Pilot data	189	189	21	21
	Shin et al	162	162	18	18

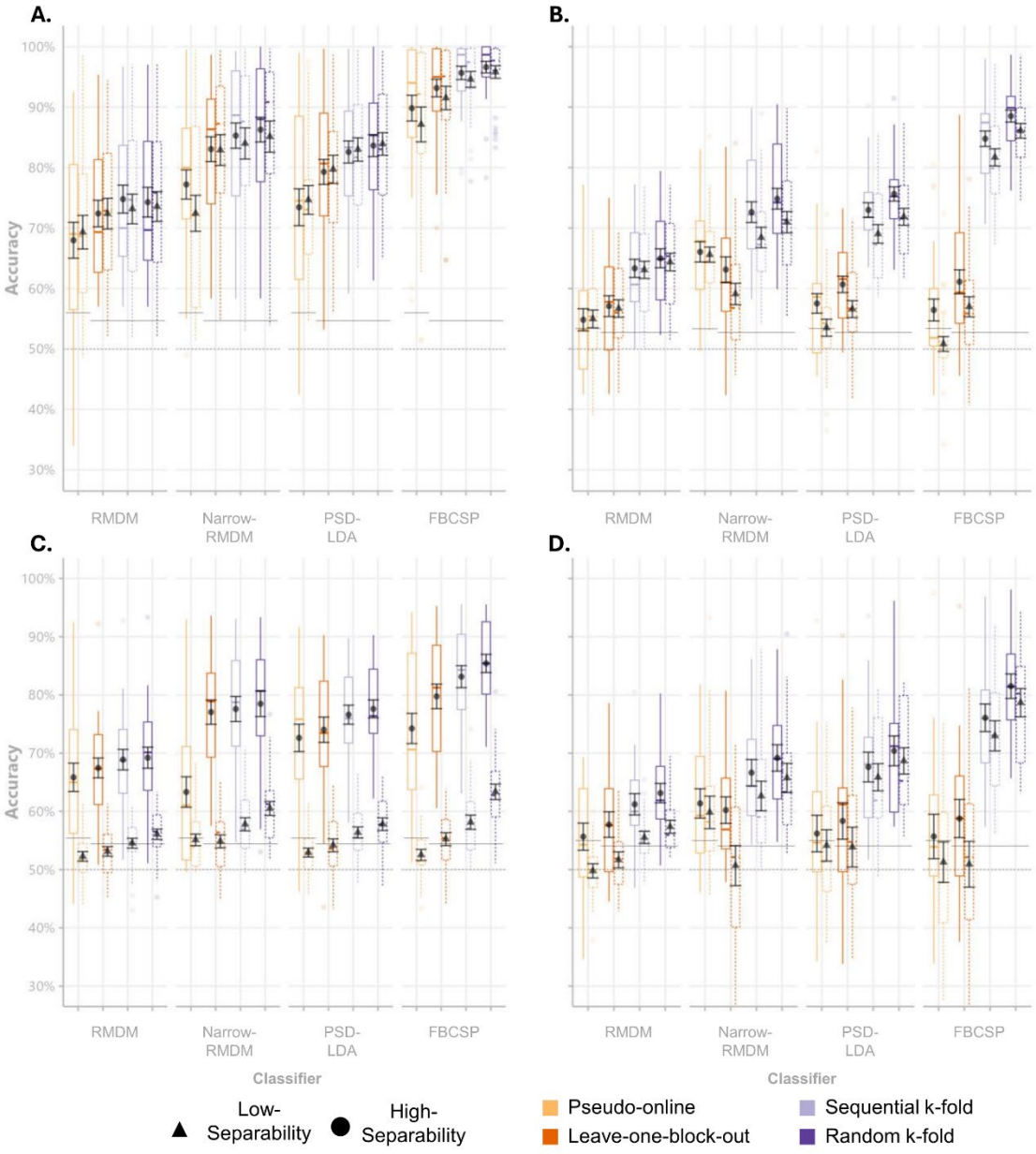
Note. The average sample size per cross-validation fold available for training and testing the four classifiers.

The next analysis dealt with the impact of cross-validation schemes on comparisons between classifiers across datasets. To investigate this, we tested for significant differences among the four classifiers within the low and high class-separability contrasts for each cross-validation scheme. Non-parametric Friedman tests were employed to account for the non-normality of the data. Additionally, post-hoc Durbin-Conover pairwise comparisons were conducted to provide more detailed insights into whether the differences between classifiers varied across evaluation schemes. P-values of the pairwise comparisons were adjusted using the Benjamini-Hochberg (BH) procedure to control the false discovery rate (FDR) at $\alpha = 0.05$ (using the PMCRplus package - Pohlert, 2024).

This chapter is based on the article: Schroeder, F., Fairclough, S., Dehais, F., & Richins, M. (2025). The Impact of Cross-Validation Choices on pBCI Classification Metrics: Lessons for Transparent Reporting. *Frontiers in Neuroergonomics*, *6*. https://doi.org/10.3389/fnrgo.2025.1582724

4.2 Results

Figure 43. Classification Accuracy Results



Note. Boxplots of mean classification accuracies for single-day Hinss et al (A.), multi-day Hinss et al (B.), Shin et al (C.), and Pilot dataset (D.). Black point ranges represent the mean accuracies across subjects and their standard errors. The dashed horizontal lines display the theoretical chance level, while the solid horizontal lines display the average sample size corrected chance levels.

Across all datasets, one could observe the expected increases going from pseudo-online to the randomised k-fold cross-validation scheme (Figure 43). The most noticeable inflation from conservative to the block-structure independent schemes was visible for the FBCSP classifier. Interestingly, differences between the high and low-separability contrasts seem to be maintained

This chapter is based on the article: Schroeder, F., Fairclough, S., Dehais, F., & Richins, M. (2025). The Impact of Cross-Validation Choices on pBCI Classification Metrics: Lessons for Transparent Reporting. *Frontiers in Neuroergonomics*, *6*. https://doi.org/10.3389/fnrgo.2025.1582724

even in the inflated accuracy estimates of the two k-fold approaches (e.g., if we see a difference between high and low separability in the conservative schemes, it is also visible in the biased schemes). A Friedman test across datasets, classifiers and cross-validation strategies ($\chi^2(1) = 25.252$, $p < .001$), followed by pairwise Wilcoxon signed-rank tests conducted per dataset, revealed that only the data from Shin et al exhibited significant differences in the classification accuracy between the low and high class-separability contrasts across cross-validation schemes and classifiers ($p < .001$).

Looking at the two panels belonging to the Hinss et al dataset, one could observe the effect of not interleaving conditions with each other. Using only the first day of their dataset, in which three repetitions of a single condition were presented in sequence, classification accuracy was on average 10.1% higher compared to the panel next to it (which used data from all three days), even for the conservative pseudo-online and leave-one-block out cross-validation schemes.

Table 27. Average Accuracies and Cross-Validation Differences

Classifier	Pseudo-online	Leave-one-block-out	Sequential K-fold	Randomized K-fold
Broad-RMDM	59.4%	61.7%	65%	66%
	(-)	(-1%, 4.4%)	(2.5%,9%)	(3.6%,10.7%)
Narrow-RMDM	65.7%	67.4%	72.7%	74.6%
	(-)	(-5.8%, 4%)	(1.67%, 10.2%)	(3.8%, 12.7%)
PSD-LDA	62.5%	65.4%	72.4%	73.3%
	(-)	(-0.8%, 4.3%)	(6.6%, 14.6%)	(8.4%, 17.1%)
FBCSP	65.9%	69.8%	81.8%	85.2%
	(-)	(-.1%, 7.9%)	(13.3%, 26%)	(17.5%, 30.4%)

Note. Average classification accuracies across class-separability contrasts and experiments (excluding *single-day Hinss et al* results) with bootstrapped 95% confidence intervals of the differences between each cross-validation scheme and the pseudo-online results.

This chapter is based on the article: Schroeder, F., Fairclough, S., Dehais, F., & Richins, M. (2025). The Impact of Cross-Validation Choices on pBCI Classification Metrics: Lessons for Transparent Reporting. *Frontiers in Neuroergonomics*, *6*. https://doi.org/10.3389/fnrgo.2025.1582724

4.2.1 *Impact of Cross-Validation Choices Across Datasets*

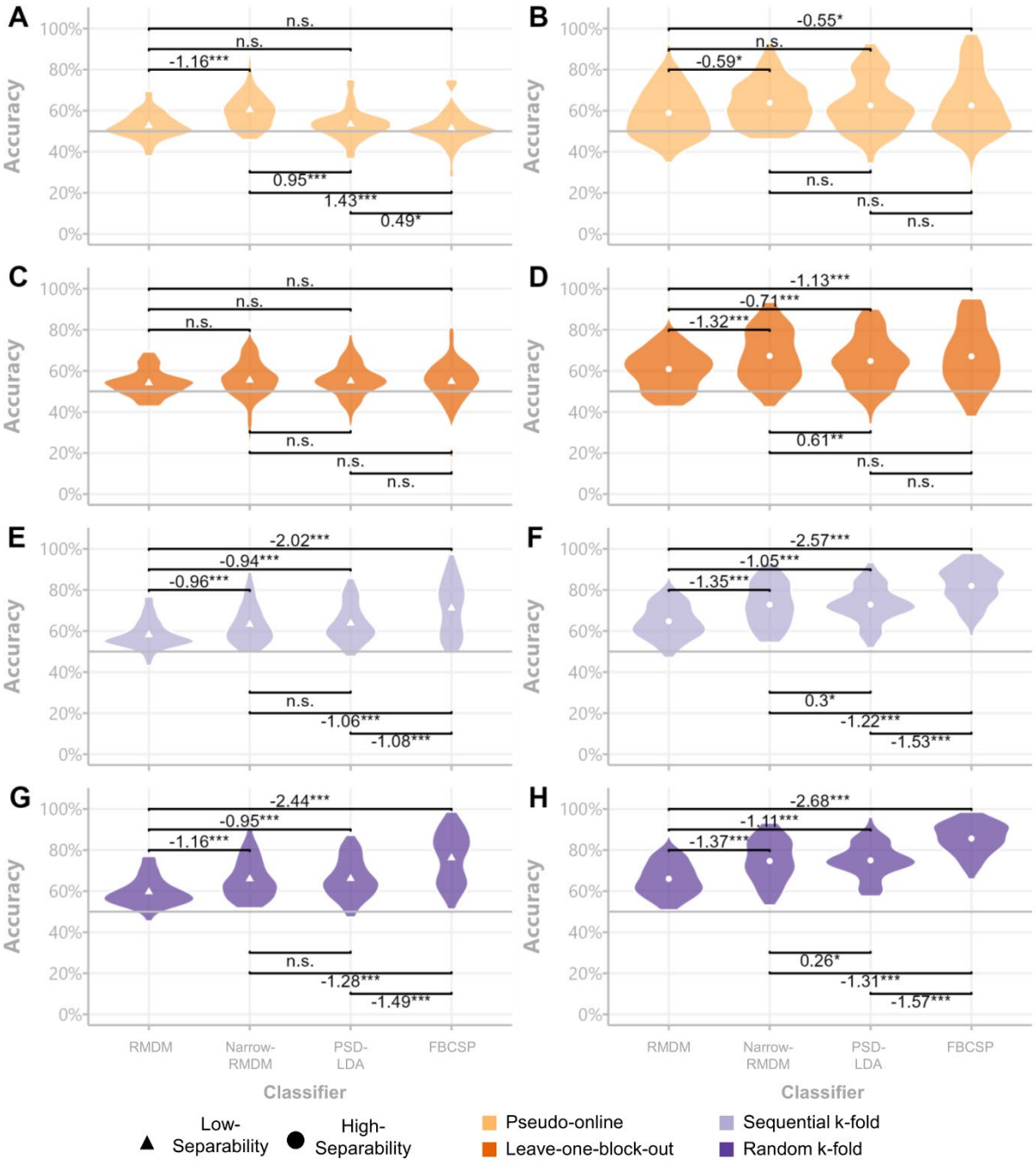
To get a general idea of how different cross-validation strategies affected estimates of model accuracy, Table 27 displays the bootstrapped mean classification accuracies across datasets with additional bootstrapped 95% confidence intervals for each cross-validation strategy's difference to the most conservative strategy (pseudo-online). The confidence intervals were computed across datasets and class-separability contrasts, excluding the single-day Hinss et al dataset because the lack of randomization for that dataset portrayed a different violation of independence as described in section 2.2.

4.2.2 *Pair-wise comparisons across datasets*

The Friedman tests conducted on the pseudo-online evaluations revealed significant effects for both the low class-separability contrast ($\chi^2(3) = 51.867$, $p < .001$) and the high class-separability contrast ($\chi^2(3) = 9.795$, $p < .001$). In the low class-separability context, pair-wise comparisons showed the Narrow-RMDM classifier outperformed all other classification approaches (see Figure 44, A). In the high class-separability context, the PSD-LDA, Narrow-RMDM, and FBCSP approaches did not differ significantly from each other (Figure 44, B). For the leave-one-block-out evaluation scheme, the low class-separability contrast did not show significant differences between classifiers ($\chi^2(3) = 5.475$, $p = .14$). However, significant differences were observed for the high class-separability contrast ($\chi^2(3) = 45.944$, $p < .001$), with pair-wise comparisons now showing the FBCSP classifier to be significantly more accurate than the broadband RMDM and PSD-LDA classifiers. The performance differences between the Narrow-RMDM and FBCSP classifiers were no longer significant (see Figure 44, D).

This chapter is based on the article: Schroeder, F., Fairclough, S., Dehais, F., & Richins, M. (2025). The Impact of Cross-Validation Choices on pBCI Classification Metrics: Lessons for Transparent Reporting. *Frontiers in Neuroergonomics*, *6*. https://doi.org/10.3389/fnrgo.2025.1582724

Figure 44. Classifier Comparisons



Note. Violin plots of subject-wise accuracy scores per class separability contrasts. Low-separability results are displayed in the left column (A, C, E, G) and high-separability results in the right column (B, D, F, H). Significance was assessed via Durbin-Conover pair-wise comparisons and average rank differences (left minus right) are displayed next to the significance signifiers. $p < .05$ = *, $p < .01$ = **, $p < .001$ = ***

This chapter is based on the article: Schroeder, F., Fairclough, S., Dehais, F., & Richins, M. (2025). The Impact of Cross-Validation Choices on pBCI Classification Metrics: Lessons for Transparent Reporting. *Frontiers in Neuroergonomics*, *6*. https://doi.org/10.3389/fnrgo.2025.1582724

In the sequential k-fold evaluation, both the low ($\chi^2(3) = 91.276$, $p < .001$) and high class-separability contrasts ($\chi^2(3) = 150.74$, $p < .001$) displayed significant differences. In this case, the FBCSP classifier demonstrated significantly better performance compared to the Narrow-RMDM classifier (see Figure 44, panels E and F). This performance difference remained highly significant in the randomised k-fold evaluation, where significant differences were again observed for both the low ($\chi^2(3) = 135.05$, $p < .001$) and high ($\chi^2(3) = 162.84$, $p < .001$) class-separability contrasts.

4.3 Discussion

The objective of the current chapter was to investigate the extent to which pBCI model evaluation metrics may be biased when temporal dependencies between train and test samples are not considered in cross-validation. To achieve this, n-back data from the Pilot and two additional open-access datasets was used. The analysis included four classifiers, ranging from models using Riemannian minimum distance metrics on minimally pre-processed broadband EEG data to models using supervised dimensionality reduction on narrowband filtered EEG data. In all four classifiers, cross-validation methods which did not observe temporal dependencies biased model evaluation metrics to a great degree, as previously reported in other areas dealing with neuroimaging-based classification (Ivucic et al., 2024; Li et al., 2020; Varoquaux et al., 2017; White and Power, 2023). Importantly, the here presented results additionally showed that this bias is not equal across classifiers. Models with greater propensity to tune to the available training data (Domingos, 2012; Lemm et al., 2011) tended to outperform alternative models in block-structure independent cross-validation schemes, while they performed at similar levels or even significantly less accurately in more conservative evaluation schemes. Consequently, model comparisons based on offline cross-validation, as will be done in the next chapter, need to ensure that cross-validation splits are not carried out irrespective of the experiment's block structure.

The most conservative validation method assessed in this chapter, pseudo-online evaluation, adhered strictly to the chronological order of the data and utilised only a single data block for model training. This reflects scenarios such as calibrating a classifier for a new user in real-time classification. The pseudo-online evaluation likely provided overly conservative performance estimates. It did indeed result in the lowest classification accuracies (Table 27), frequently failing to surpass adjusted chance levels in all datasets (Figure 43). In contrast, the leave-one-block-out cross-validation scheme, which also preserved the block structure but disregarded the temporal order of samples, achieved better-than-chance classification across a greater number of classifiers by training on two data blocks instead of one. Notably, none of the four tested classification approaches showed

This chapter is based on the article: Schroeder, F., Fairclough, S., Dehais, F., & Richins, M. (2025). The Impact of Cross-Validation Choices on pBCI Classification Metrics: Lessons for Transparent Reporting. *Frontiers in Neuroergonomics*, *6*. https://doi.org/10.3389/fnrgo.2025.1582724

significantly higher accuracy comparing the leave-one-block-out evaluation to the pseudo-online evaluation (all bootstrapped 95% CIs in Table 2 included 0), even though their training data was double in size.

When comparing the conservative evaluation methods to the two k-fold cross-validation approaches, where data splitting disregarded the experimental block structure, significant inflations in classification accuracy were observed across all classifiers we tested. For the simple RMDM approach, accuracy increased by up to 9%, while the electrode-selection variant of RMDM displayed increases of up to 12.7% accuracy. Even more pronounced increases were evident in the two LDA classifiers, which utilised canonical band power features combined with additional dimensionality reduction techniques. When only employing feature selection on band power features during the training phase, accuracy estimates rose by up to 17.1% compared to the pseudo-online evaluation. Adding another dimensionality reduction step (FBCSP) further inflated accuracy estimates, increasing them by up to 30.4% (Table 27). Against expectations, the upper bounds of the confidence intervals for the sequential and random k-fold approaches did not differ substantially across the three datasets tested (Table 27). This was surprising, as the random k-fold approach represents a more obvious violation of the assumption of independence. The observation that merely sharing non-overlapping data of a single block in train and test sets can cause similar biases in the accuracy metrics, demonstrates the issue of temporal dependencies in offline pBCI model evaluations aptly.

Although the four evaluation schemes varied in the sizes of their training sets, these variations alone do not explain the results presented. The leave-one-block-out cross-validation, despite two times the training data compared to the pseudo-online evaluation, showed negligible accuracy gains. In contrast, k-fold methods differed much more from the pseudo-online evaluation with a not quite threefold increase in training set size.

Furthermore, since the degree of inflation to model accuracy seemed to differ between feature/classifier types, model comparisons based on evaluation metrics computed on block-structure independent cross-validation schemes may lead to erroneous conclusions that would not replicate in applied settings. This was demonstrated in section 4.2.2, where the FBCSP classifier showed great advantages over the other tested classifiers in the k-fold evaluations, but was either not significantly different to its alternatives in the high-separability case and was actually outperformed by the Narrow-RMDM in the low-class separability case (Figure 44). In general, it is to be expected that the propensity to overfit on training-specific information increases with classifier complexity (Domingos, 2012; Lemm et al., 2011) - a problem underlined by the extreme differences

This chapter is based on the article: Schroeder, F., Fairclough, S., Dehais, F., & Richins, M. (2025). The Impact of Cross-Validation Choices on pBCI Classification Metrics: Lessons for Transparent Reporting. *Frontiers in Neuroergonomics*, *6*. https://doi.org/10.3389/fnrgo.2025.1582724

found between evaluation schemes used for deep-learning-based M/EEG decoding (Ivucic et al., 2024; Li et al., 2020).

Interestingly, of the tested high-separability contrasts, the Shin et al dataset showed the best performance in the pseudo-online cross-validation scheme (Figure 43). The difference to the other datasets might be linked to the manner in which conditions were interleaved with each other within a single presentation block in the Shin et al dataset. Interleaving several shorter repetitions of different conditions likely caused different classes to share the same temporal trends, leaving the trends less informative for the classification task. Another contributing factor could have been that the high separability contrast contained the 0-back condition. Reacting to a single stimulus in a stream rather than having to repeatedly encode and maintain new stimuli could be considered psychometrically distinct from the other n-back conditions, making for an easier classification problem (Gerjets et al., 2014).

In the example of only making use of a single day of the Hinss et al dataset (see Figure 43, panel A; not included in the other analyses) all blocks of a single condition were presented in sequence together, leading to samples of a single class sharing temporal dependencies regardless of whether they stemmed from a single or separate blocks. Due to the lack of randomisation, classifiers were likely utilising temporal dependencies regardless of the cross-validation scheme used. Results garnered using the data from a single day were, on average, 10.1% higher than those from the full multi-day dataset. Using the Pseudo-online or leave-one-block-out cross-validation scheme on a single day (rather than all three days) would erroneously lead to the conclusion that the FBCSP classifier performed the best of the four tested classifiers (i.e. the classifier with the highest propensity to utilise temporal dependencies instead of class differences).

The various differences between the methods of the three datasets described in 4.2.1 could be seen as an obstacle to cross-dataset analyses. However, since the reported effects persisted despite the variability caused by differences in participant training regimes, presentation order, condition contrasts, and so on, they should instead be considered a strength of this analysis. Similar results could likely be garnered from various kinds of block-based experimental designs with (pseudo-) randomised condition orders. However, as demonstrated through the example of only using a single day of the Hinss et al data, the results may differ if conditions were presented without randomisation. Furthermore, slower physiological signals, such as fNIRS or ECG, may exhibit even stronger biases in block-structure independent cross-validation due to their slowly evolving nature, which may be accompanied by longer-lasting temporal dependencies (Blanco et al., 2024). Finally, as

This chapter is based on the article: Schroeder, F., Fairclough, S., Dehais, F., & Richins, M. (2025). The Impact of Cross-Validation Choices on pBCI Classification Metrics: Lessons for Transparent Reporting. *Frontiers in Neuroergonomics*, *6*. https://doi.org/10.3389/fnrgo.2025.1582724

alluded to in the introduction, the bias stemming from splitting training and testing data irrespective of experimental structure can be avoided in event-based experimental designs, provided that the condition order is fully randomised and no more than a single sample is drawn from any given trial (White & Power, 2023).

Lastly, the effects presented here should be built upon in future research as various factors may exacerbate or potentially minimise the bias of ignoring temporal dependencies in cross-validation. Separate inquiries focusing on specific facets in pBCI processing pipelines could together form a solid base for readers and reviewers to contextualise findings in the pBCI literature. For example, the current analysis did not include data cleaning steps beyond the removal of faulty channels for the sake of parsimony. Their implementation details, as well as the length and selection of calibration data, may produce different dynamics across cross-validation methods. The tuning hyperparameters to individual subjects could also potentially exacerbate the bias, as it would allow for even more overfitting to training-data specific trends. Another dedicated inquiry could focus on the offline evaluation of adaptive machine learning approaches, where accounting for or correcting mismatches in the data distribution between train and tests sets is a key goal (Kumar et al., 2019; Lotte et al., 2018; Schlögl et al., 2010). The relationship between the calibration data, used for the adaptation, and test data, could be actively manipulated in such a study to inform possible sources of bias.

This chapter is based on the article: Schroeder, F., Fairclough, S., Dehais, F., & Richins, M. (2025). The Impact of Cross-Validation Choices on pBCI Classification Metrics: Lessons for Transparent Reporting. *Frontiers in Neuroergonomics*, *6*. https://doi.org/10.3389/fnrgo.2025.1582724

5 Continuous Workload Monitoring

The development of applied neuroadaptive systems that can monitor mental workload and adapt the human-machine interaction accordingly faces a number of "grand challenges" (Fairclough & Lotte, 2020). The challenge of reliability and robustness can be seen as the overarching theme of this thesis, as neuroadaptivity based on inaccurate inference would likely cause more frustration than relief in a human-machine interaction (Parasuraman et al., 1999). Consequently, as long as pBCI technologies cannot provide reliable inference in applied scenarios, their usage will remain limited to research laboratories. However, the required level of reliability may be context-dependent in applied settings, as errors would likely be acceptable in a video game but may cause actual harm in safety-critical contexts like aviation or the operating room.

The challenge of reliability and robustness of pBCI inference in the real world pertains to the issue of cross-subject classification as well as cross-session or cross-task classification within a single operator. Here, intra- and inter-operator variability requires adapting classifiers to shifts in the underlying data distributions to remain accurate (Jayaram et al., 2016; Zanini et al., 2018). Moving pBCIs into the real world further requires research into the reliability and robustness of cheaper and user-friendly wearable sensing devices, like the X.on used in the Wearable dataset. Methods tested on gel-based high-density montages may not translate one-to-one to signals recorded using sparse montages of sponge-based or dry electrodes with higher recording impedances.

Two examples that showcase real-world implications of unreliable BCI inference were recently published in reports on a self-paced speech BCI and a BCI based on imaginary movements used by a paralysed patient to call for their carer. In the case of self-paced speech BCIs, users may be discouraged if the BCI produces false positives, making them utter words they did not intend to (Luo et al., 2025). The alarm-sounding BCI, on the other hand, was found to produce many false positives at nighttime, disrupting the sleep of both the patient and the carer and making the use of the system less appealing (Vansteensel, 2024). In both cases, the accumulation of evidence over a longer timeframe could alleviate the frequency of false positives by trading response time for increased accuracy (Leinders et al., 2024; Luo et al., 2025).

Much of the research cited in the BCI literature tends to classify perceptual or motor-related signals, rather than global brain-state changes. Being largely informed by research around active and reactive BCIs, pBCIs tend to also operate on fixed windows for data extraction, as one would when epoching data around a certain event. However, taking advantage of the passive nature by removing the need

for stimulus-locked predictions may allow for tracking brain-state changes across various time-scales, such as short-term fluctuations over a few seconds or long-term fluctuations spanning across minutes. Passive BCI performance can likely be improved by pooling information over time, trading responsiveness for accuracy. Such strategies have previously also been termed temporal ensembling or temporal aggregation and successfully employed to boost deep learning performance in BCI (Laine & Aila, 2017) and image-based trajectory prediction (Hong et al., 2024). Past information may additionally be directly used as input for the next prediction to improve accuracy for "one-step-ahead predictions" in deep-learning based time-series forecasting (Lim & Zohren, 2021).

Fusing information is a common strategy to decrease model bias in machine learning (Bloch, 1996; Kuncheva et al., 2001). Fusing at the feature-level, providing the classification algorithms with higher-dimensional data, leads to more complex models that sometimes lead to more informative decision boundaries. Fusing information at the decision-level results in an ensemble approach akin to that of the Random Forest, where multiple independently trained classifiers reduce the likelihood of overfitting, improving the generalisation to unseen data. In the decision-level fusion, a soft or hard pooling can be carried out, using the average class probabilities as the final output in the former case or the majority vote in the latter case. Previous work already showed that BCI accuracy can be improved by increasing the window size over which features are extracted (Brouwer et al., 2012) or by aggregating over predictions of several sequential task-irrelevant probes (Ke et al., 2021), or sliding windows in a motor imagery trial (Padfield et al., 2021) .

In this chapter, decision-level fusion of a single classifier's output over time ("temporal aggregation" or "temporal ensembling") will be utilised and the classification performance in the Lab-grade and Wearable datasets systematically compared. This will be carried out both within and across subjects to ascertain whether the use of a wearable EEG headset incurred significant classification performance decrements.

Research Questions:

1. Does classification accuracy decrease when moving from Lab-grade to Wearable EEG?
   Expectation: Wearable accuracy will be lower than Lab-grade accuracy
2. Can the reliability of mental workload classification be improved by trading the pBCIs responsiveness for accuracy?
   Expectation: Accuracy will improve at larger pooling windows but may level out at larger extraction windows
3. How well do findings generalise across datasets and tasks? (Is there a single domineering approach across montages and tasks?)

Expectation: No single feature extraction method will dominate all montage-by-task combinations

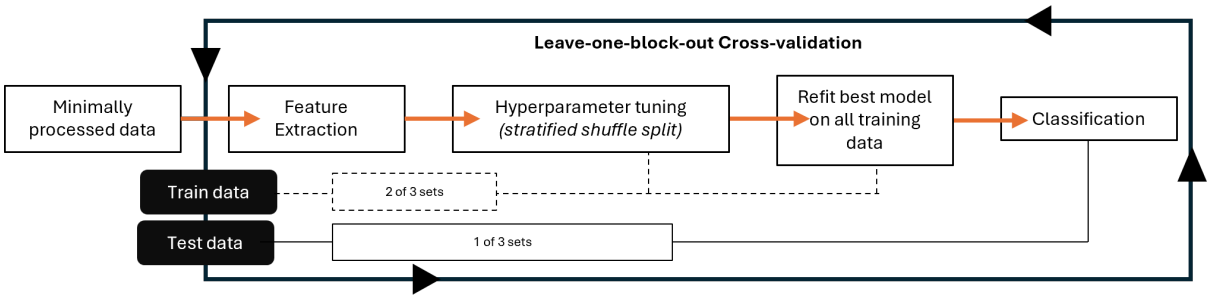
5.1 Methods

For the following analyses, only the minimally processed versions of the Lab-grade and Wearable datasets were used. To increase the comparability of the Lab-grade and Wearable datasets, the Lab-grade data was utilised not just in its 64-channel form but also in a reduced form, in which only the channels shared with the X.on headset were retained (F3, F4, C3, Cz, C4, P3, P4). The Multimodal and Pilot datasets were not included in this analysis. The Pilot dataset was excluded due to its mismatch in task-load conditions with the remaining three datasets. The Multimodal dataset was excluded as the insights gained from this chapter will be built upon in the following chapter, where the Multimodal data will be treated as a sort of hold-out test set to confirm some of the results from the current analysis.

Comparing the reduced Lab-grade data with the Wearable data aims to compare classification using the same spatial coverage but differing electrode designs. The wearable data used wet sponge-based electrodes, while the Lab-grade data used traditional Ag/AgCl gel-based electrodes. Furthermore, impedance levels were calibrated with different thresholds (<25kOhm in case of the Lab-grade data vs. <120kOhm for the Wearable data).

Following the results from Chapter 4, leave-one-block-out cross-validation was used to estimate the classification accuracy per subject (see Figure 45), using two out of the three condition repetition sets for training and the remaining set of blocks for testing. In addition to the cross-validation procedure already introduced in Chapter 4, the current analysis included a nested cross-validation in order to tune various hyperparameters.

Figure 45. Within-Subject Classification Pipeline



Note. A schematic of how the different classifiers were evaluated in the within-subject classification. Steps like data segmentation or applying filter banks were usually carried out outside the cross-validation loop to save computing resources. All classification steps that transformed data were entirely based on the training data and only carried out within the cross-validation loop. Training data consisted of

2 out of 3 total condition repetitions. The two training sets (one block per condition per set) were again split using a shuffle-split in a nested cross-validation in which a grid-search was used for hyperparameter tuning.

5.1.1 *Feature Extraction*

All but the auditory-probe related feature extraction procedures were repeated using 1-second, 2-second, 5-second, 10-second, and 20-second long extraction windows that were segmented from the continuous EEG signal with 33% overlap. The overlap percentage of 33% was chosen to increase the sample size, specifically at higher extraction window values, while minimising the collinearity of temporally adjacent features.

Raw-narrow. Similar to the feature extraction carried out for the PSD-LDA approach in chapter 4 (4.1.3). An FFT with a 1Hz resolution was computed per electrode per window, and the average power within the theta (4-7Hz), low alpha (8-10Hz), high alpha (10-13Hz), beta (17-28Hz) and a low gamma band (32-40Hz) was extracted. Frequency ranges of the beta and gamma bands were selected to avoid the fundamental and first harmonic frequency of the SSVEP. This resulted in 5 (frequency bands) by 64 (electrodes) features in the Lab-grade and 5 (frequency bands) by 7 (electrodes) in the reduced Lab-grade and Wearable data. All features were z-scored using training data within the cross-validation.

Narrow Covariance. Similar to the narrow-RMDM approach in Chapter 4 (4.1.3), a filter bank of Butterworth filters (zero-phase, noncausal, passband ripple = 3db, stopband attenuation = 10 db, transition bandwidth = 1 Hz) was used to filter the data into the same band ranges as for the Raw-narrow feature extraction, again, avoiding the fundamental and first harmonic of the SSVEP. Next, the data was segmented and covariance matrices were computed per frequency band. In the case of the Lab-grade data, each band's covariance matrices were reduced to 7 electrodes using the class-separability based approach also utilised in chapter 4 (Barachant & Bonnet, 2011). This reduced the computational load of the following steps to the same level as for the Wearable and reduced Lab-grade data. Next, the 5 individual 7 by 7 covariance matrices were combined into a block-diagonal matrix with off-diagonal elements set to 0.

FBCSP. Again, similar to the FBCSP in chapter 4 (4.1.3), a filter bank of Butterworth filters (zero-phase, noncausal, passband ripple = 3 dB, stopband attenuation = 10 dB, transition bandwidth = 1 Hz), the data was bandpass filtered into 4 Hz wide frequency bins ranging from 3–25 Hz in steps of 2 Hz. As two of the tested montages only contained 7 electrodes, only 4 filters were used for the feature extraction per band (2 highest and 2 lowest eigenvalues). The average log-variance per window was used as the final feature for classification, resulting in 5 (frequency bands) by 4 (CSP filters) number of features.

The following two feature extraction methods were specific to the auditory and visual probes used in the Lab-grade and Wearable datasets.

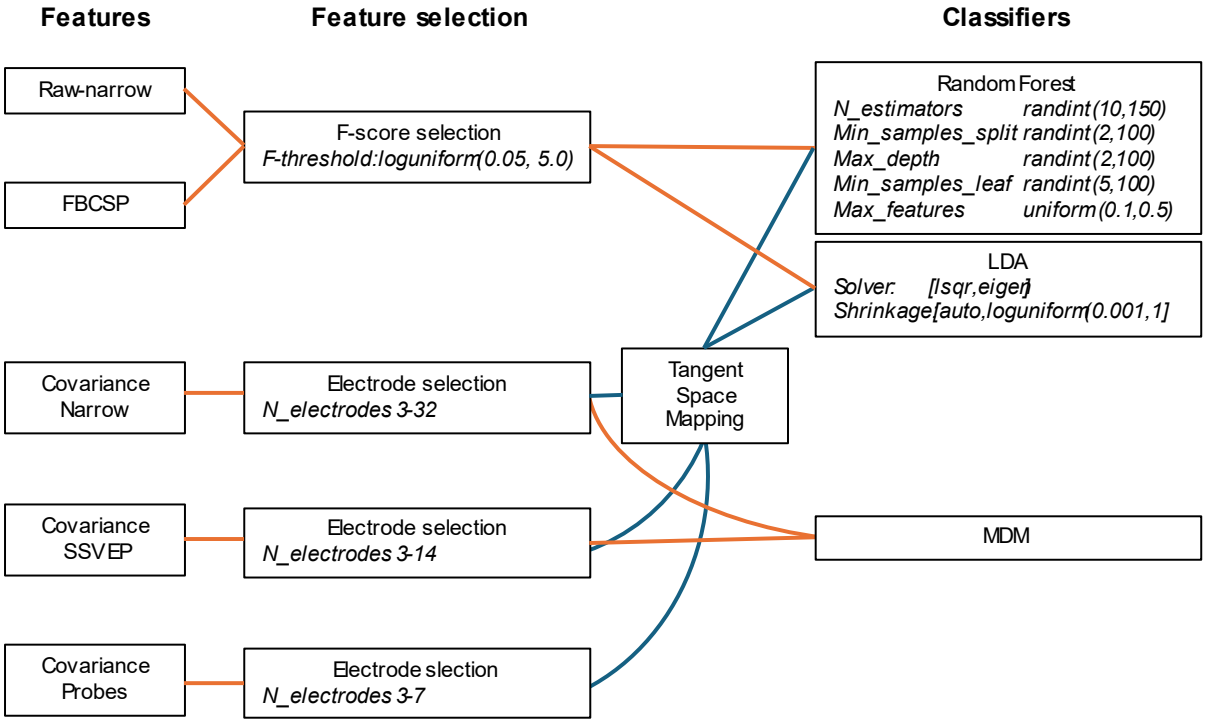
Covariance SSVEP. Similar to the Narrow Covariance approach, a block-diagonal covariance matrix was built. Before segmentation, the data was separately filtered around 15Hz (passband edges at 14 and 16Hz) and 30Hz (passband edges at 29 and 31Hz) using the same Butterworth filters as in the Narrow Covariance and FBCSP approaches. The remainder of this method was the same as for the Narrow Covariance approach, resulting in a 14 by 14 block-diagonal covariance matrix per window, containing the two 7 by 7 covariance matrices, with off-diagonal elements set to 0.

Covariance Probes. The auditory probes were also classified using covariance matrices after bandpass filtering the signal around 1 - 13Hz. The chosen frequency range still contained all the information of the relevant ERP components (Luck, 2005; Woodman, 2010) while limiting the influence of high-frequency noise. What set the probes apart from the other features extracted before was that the extraction window size was not varied, as the data was not segmented using sliding windows, but from probe onset to 400ms past the probe onset. Time-locking the covariance estimation to the onset of the auditory probes was motivated by the idea that basing each SCM on activation stemming from the same “neural generators” that give rise to the characteristic N1 and P2 components should result in less variability across samples, compared to running windows that sample spontaneous activations. More sophisticated covariance estimation incorporating phase information of the ERP (Barachant & Congedo, 2014; Ladouce et al., 2024) were forgone, as no task-load related latency differences were expected (Sugimoto et al., 2022).

5.1.2 Hyperparameter tuning

Classification was performed using Riemann Minimum Distance to Mean (MDM), Linear Discriminant Analysis (LDA), and Random Forest (RF) models for covariance matrices, or two classifiers in the case of the FBCSP and Raw-narrow features (LDA and RF). To classify the Covariance matrices using the LDA or RF classifiers, they were mapped onto the tangent space of the training data’s Riemannian mean (see section 2.7.3).

Figure 46. Hyperparamter Overview



Note. The different types of classification features were classified with two or three different classifiers. Depending on the classifier, a number of hyperparameters were tuned within each cross-validation iteration. The hyperparameters are shown in this figure.

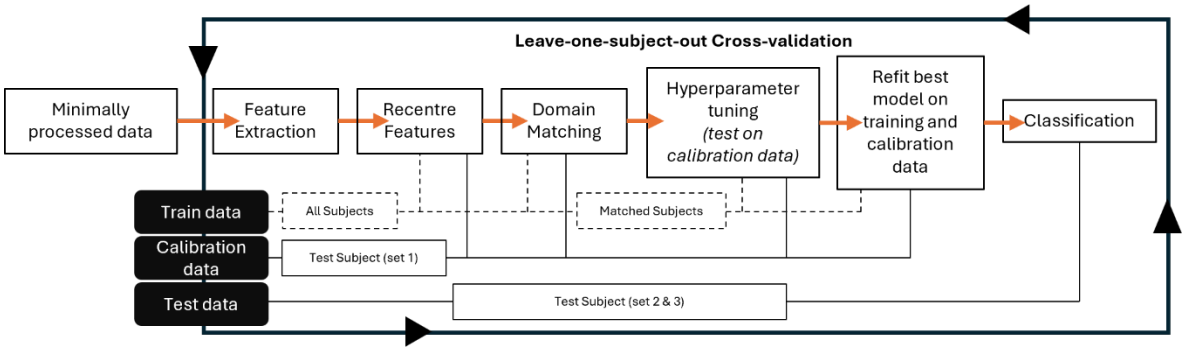
Within each cross-validation iteration, a number of hyperparameters were tuned via a Random Grid Search to improve classification accuracy. The primary hyperparameters for the LDA and MDM classification were in the feature selection step, where uninformative features were removed from the feature set, saving resources and reducing variance for the subsequent classification. For the RF classifier, a number of integral model parameters were tuned:

- N_estimatores: The number of trees in the forest (i.e. number of decision trees). The more trees, the higher the computation time, but the more complex the decision boundary.
- Min_samples_split: Required samples to split a node into further leafs of a single decision tree. Higher values prevent overfitting to minute details in the training data.
- Max_depth: Limits how deep a single decision tree can grow. Deeper trees may generate more complex decision boundaries, but risk overfitting to the training data.
- Min_samples_leaf: Similar to Min_samples_split, it limits overfitting by preventing a further split if the final node (termed leaf) ends up with too few matching training samples.
- Max_features: The fraction of features to consider when splitting a node. Lower values may increase diversity among trees.

The grid search was performed over 25 iterations with the available training samples split using a stratified shuffle split (nested cross-validation), dividing the two blocks per condition into shuffled train and test sets using a 80% / 20% split. Figure 46 displays how the different feature selection procedures and classifiers were combined with the 5 feature categories.

5.1.3 *Cross-subject classification*

Figure 47.Cross-Subject Classification Pipeline



Note. A schematic of the different classifiers' evaluation in the cross-subject classification. Steps like data segmentation or applying filter banks were usually carried out outside the cross-validation loop to save computing resources. Training data consisted of all but the current test subject. Features were either z-scored within subject (Raw-Narrow and FBCSP), or shifted to the Riemannian mean of the calibration data in the case of the SCM-based methods. Domain matching selected seven subjects whose feature covariance had the closest Riemann distance to that of the test subject.

Generalisable machine learning models require ample, high-quality training data, making within-subject approaches less suited for real-world applications. Ideally, data from previous operators could be used instead of requiring lengthy calibration sessions to train a new classifier from scratch for each operator. To this end, the here presented cross-subject classification made use of two transfer learning approaches. In machine learning, transfer learning refers to leveraging prior knowledge for new but related problems. In the BCI literature, this usually refers to reusing data from previous recording sessions or other subjects to train a classifier (Congedo et al., 2017; Jayaram et al., 2016). Reusing data in a new context faces the aforementioned hurdles of non-stationary signals and data distribution/covariance shifts. Transfer learning approaches attempt to adapt data to make it match the new context (or domain).

Re-calibrating decision boundaries or class centroids in pre-trained classifiers has been done successfully in the past, such as the lightweight PMean adjustment to LDA models (Perdikis et al., 2016; Vidaurre et al., 2011). Adjusting pre-trained models would be favourable in terms of saving computational resources; however, in the current analysis, models were trained from scratch for each new subject as an additional domain matching step, selecting only those subjects whose feature covariance structure was similar to that of the test subject, was used before training a

classifier. This was done to counteract mixing training data with opposing trends that may occur from inter-subject differences in cognitive strategies or neuroanatomical differences (such as sulci morphology).

The classification pipeline for the cross-subject analysis was extended by recentring and domain matching steps (see Figure 47). Changes to the previously described within-subject classification pertained to:

Cross-validation. Cross-validation was carried out by using one of the available subjects for testing and calibration, while the remaining subjects were used for training. The calibration data consisted of the first condition repetitions (set 1) and the test data of the 2nd and 3rd set of the test subject.

Recentring. Recentring in the Raw-Narrowband and FBCSP feature categories was achieved by z-scoring each subject's features. For the test subject, this was achieved using the mean and standard deviation of the calibration data. For the SCM-based methods, recentring was carried out on the SPD manifold (Zanini et al., 2018) using the Riemannian mean of the calibration data as the reference point (via the TLCenter function of the pyriemann package).

Domain Matching. After recentring the features, the next step was to select one-third of the available subjects with the most similar covariance structure. For the covariance-based features, this was done by sorting the distance of each subject's Riemann mean to that of the calibration data of the test subject. For the raw-narrowband and FBCSP features, this was also done using Riemann distance for consistency's sake, however, the dimensionality of the features was first reduced with PCA – only retaining the first 10 dimensions - as the higher extraction window settings would not have provided enough observations for reliably computing SPD covariance matrices spanning the full feature space. Retaining the first 10 dimensions of the PCA allowed for capturing a significant proportion of the data's variance while still providing well-conditioned SPD covariance matrices (the covariance of the data after projecting them into lower-dimensional PCA space) even in higher extraction window settings. Domain matching was carried out for the global covariance across classes.

Hyperparameter tuning. After the features were recentred and two-thirds of the available training subjects were discarded, the hyperparameter tuning was carried out using the data of the remaining ~7 subjects for training and the calibration data of the test subject for testing. The same parameter distributions (Figure 46) were tested over 25 iterations as in the within-subject classification.

For the cross-subject classification, the 64-channel Lab-grade data was omitted due to computational resource constraints. Instead, only the reduced Lab-grade and Wearable data was used.

One last difference in this cross-subject classification pertained to the FBCSP features. The CSP analysis was computed across all training subjects, and the computed spatial filters were applied to the test subject's data without further adjustment. Furthermore, previous research suggested that the small frequency bins commonly used for the FBCSP approach were not ideal for computing spatial filters across subjects. Instead, a multi-resolution approach was found to perform better, in which wider frequency bins of varied size were employed (Lotte et al., 2009). Here, instead of the 2Hz wide frequency bins, the cross-subject FBCSP used a theta (4-7Hz), alpha (8-13Hz), beta (17-28Hz), gamma (32-45Hz) and a combined theta+alpha (4-13Hz) bin.

5.1.4 *Temporal Aggregation*

Following the classification of the test data was the temporal aggregation step. As this was lightweight in computational resources, a range of decision windows from 2 to 60 seconds was tested (skipping decision windows that were smaller than the current extraction window). This temporal aggregation step used a simple sliding window approach that moved from one prediction to the next, returning the majority label of all past predictions that fit into the decision window. The same window was used across the predicted and the true labels to compute the aggregated accuracy score. Tie-breaks were handled using the most recent prediction's label.

The question of whether there was an optimal extraction/decision-window combination may be conceptually answered through modelling. However, the unstable nature of brain states is likely to complicate the answer and warrants an empirical investigation. In theory, if small extraction windows provide sufficient accuracy, aggregating over many of them should produce higher accuracy than larger extraction windows, even when accuracy also rises with extraction window size. However, there may be cases where smaller extraction windows provide too noisy predictions and feature extraction should be carried out over wider windows before aggregation to alleviate rapid fluctuations in the predictions. In those cases, cross-subject classification can be of great help as it mitigates the loss of training samples when using wider extraction windows.

5.1.5 *Statistics*

Six independent variables were included in this analysis: 1) data source (Lab-grade, Lab-grade reduced and Wearable), 2) type of task (n-back vs. MATB), 3) type of feature (Raw-narrow, Narrow-covariance, FBCSP, Covariance-SSVEP, and Covariance-probes), 4) extraction window size (0.4, 1, 2, 5, 10, 20-seconds), 5) single vs aggregated predictions, and 6) cross- or within-subject classification. Lastly, various classifiers were utilised (MDM, LDA, RF).

In the first instance, a linear mixed model was fitted to the subject-wise average accuracy scores on the smallest possible extraction windows (.4 and 1 seconds) to investigate differences between the montages, particularly in their interaction with the different feature categories. Within and cross-subject classification results were fitted separately.

```
Formula: logit(accuracy) ~ (Task + Contrast + Feature + Classifier + Dataset)^2 + (1 | Subject)
```

Fixed factors were dummy-coded with MATB (Task), high class-separability (Contrast), Covariance-probe (Feature), LDA (Classifier) and Lab-grade (Dataset) as the reference levels. The omnibus tests were computed using type III sums of squares with Satterthwaite's approximation for degrees of freedom. Planned pairwise comparisons using estimated marginal means (EMMs) assessed differences between the datasets/montages within each feature, task, and contrast combination. P-values were corrected for multiple comparisons with the Benjamini–Hochberg (BH) procedure to control the false discovery rate (FDR) at $\alpha = 0.05$ (using the PMCRplus package - Pohlert, 2024).

Next, the effect of increasing the size of the extraction window as well as aggregating over sequential predictions was analysed. The aim was to ascertain whether trading the responsiveness of a pBCI could yield significant accuracy improvements. This was done, firstly, by increasing the size of the feature extraction windows to reduce the influence of measurement noise on the classification, and secondly, by aggregating over sequential predictions using a simple majority-voting process, which may smooth over occasional false predictions, thereby potentially reducing their occurrences and preventing unwanted behaviour in neuroadaptive applications. Such techniques could further mitigate the potential accuracy decrements of the low-density montages tested here.

To ascertain whether ideal accuracy/responsiveness sweet spots existed for the different classification scenarios, the 95% confidence intervals of the average accuracy by scenario were computed per feature, extraction window, and classifier. The sweet spot was defined as the combination of the aforementioned factors that offers the fastest response time (smallest decision window) while still falling into the confidence interval of the maximal accuracy (combination of factors with the highest average accuracy).

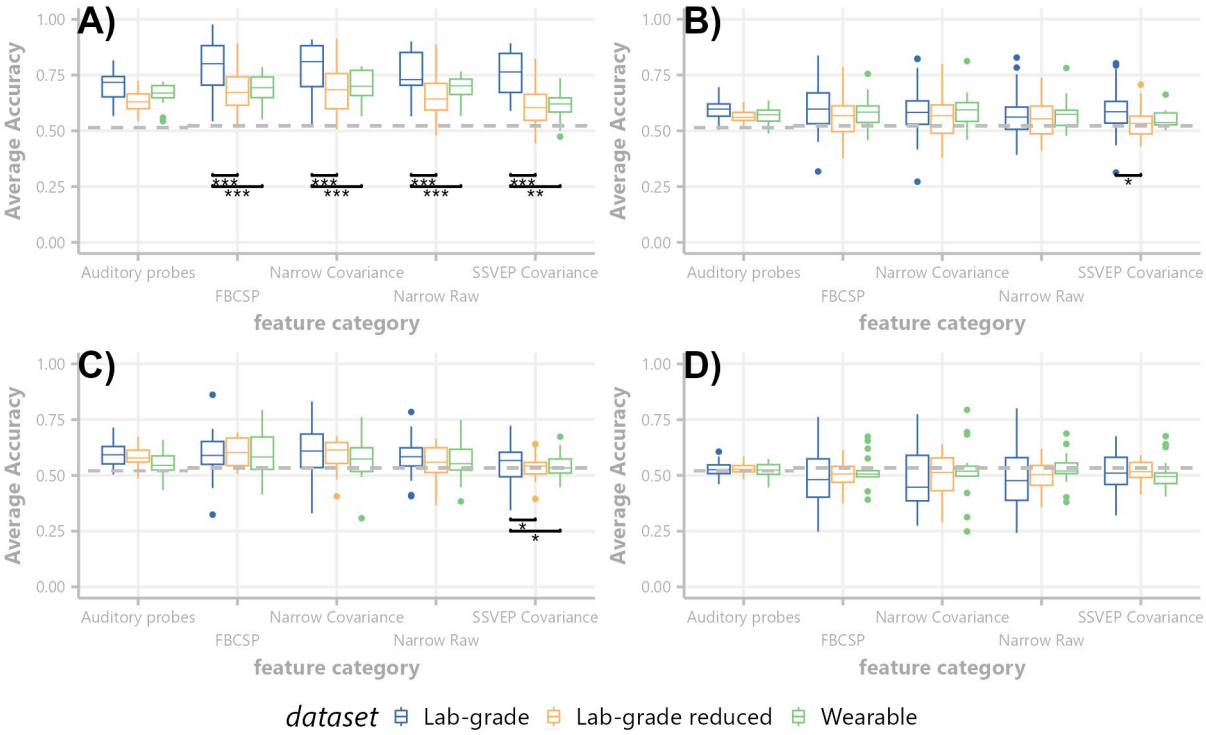
5.2 Results

Whether classification accuracy in the Wearable dataset was significantly worse than the Lab-grade dataset was tested via a linear-mixed effects model on subject-wise logit-transformed accuracy scores. In the section below, first the within, and then the cross-subject classification results will be described.

5.2.1 *Lab-grade vs Wearable*

Significant main effects for the within-subject classification results were present for different features ($F(4,2864.00) = 20.28$, $p < 0.001$), the tasks ($F(1,2864.00) = 977.39$, $p < 0.001$), and contrasts ($F(1,2864.00) = 943.15$, $p < 0.001$), but not the datasets or classifiers.

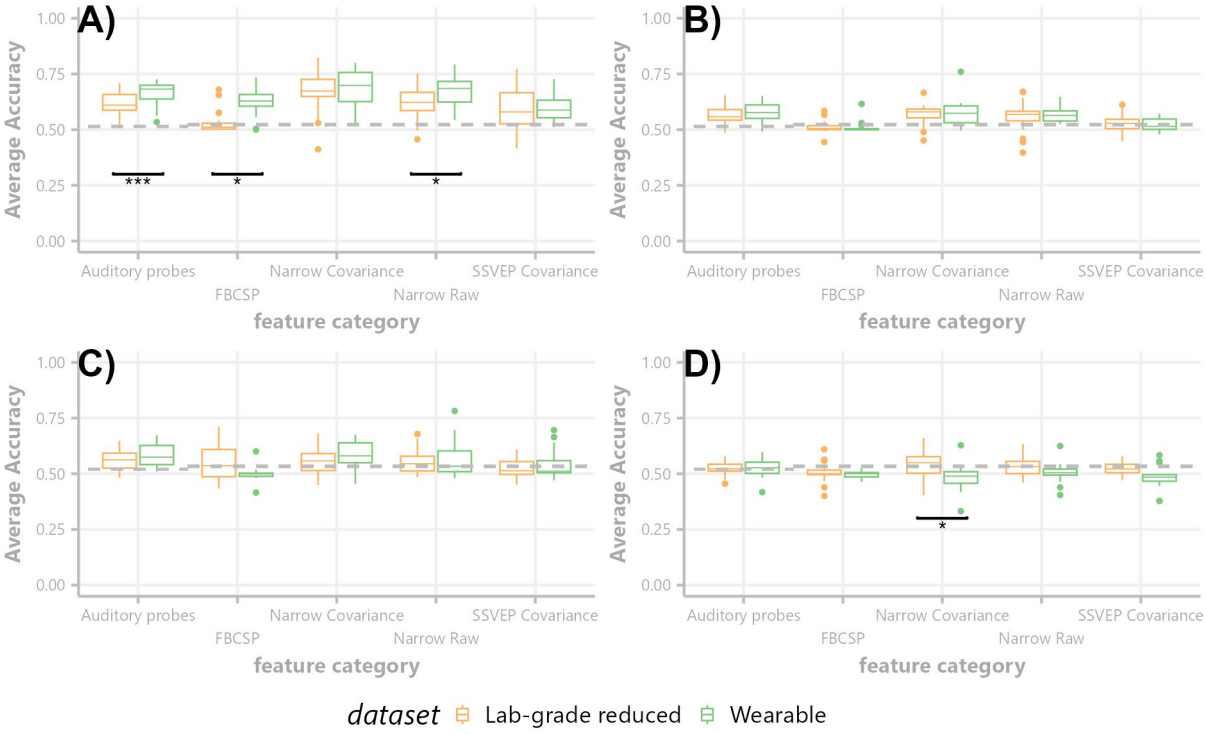
Figure 48. Within-Subject Montage Comparisons



Note. Average classification results from the within-subject classification for the high separability MATB data (A), low separability MATB data (B), high separability n-back data (C), and low separability n-back data (D). The dashed grey lines represent sample size adjusted chance levels. Significant differences between the montages were highlighted. p<.05 = *, p<.01 = **, p<.001

The interaction between dataset and feature category was not significant ($F(8,2964.00) = 1.77$, $p = 0.07$), but planned pair-wise comparisons were still carried out, as the main interest in differences between the different datasets was established a priori. Significant contrasts were mostly isolated to the MATB's high-separability scenario, where the reduced Lab-grade and Wearable data exhibited significantly lower accuracy results compared to the full Lab-grade data in all but the Covariance-probe features (see Figure 48 panel A). Only the Covariance-SSVEP feature exhibited further significant differences in the MATB's low-separability and n-back's high-separability scenario (see Figure 48 panels B and C). In all significant contrasts, the reduced Lab-grade and Wearable data exhibited lower accuracy scores than the full Lab-grade data.

Figure 49. Cross-Subject Montage Comparisons



Note. Average classification results from the cross-subject classification for the high separability MATB data (A), low separability MATB data (B), high separability n-back data (C), and low separability n-back data (D). The dashed grey lines represent sample size adjusted chance levels. Significant differences between the montages were highlighted. p<.05 = *, p<.01 = **, p<.001

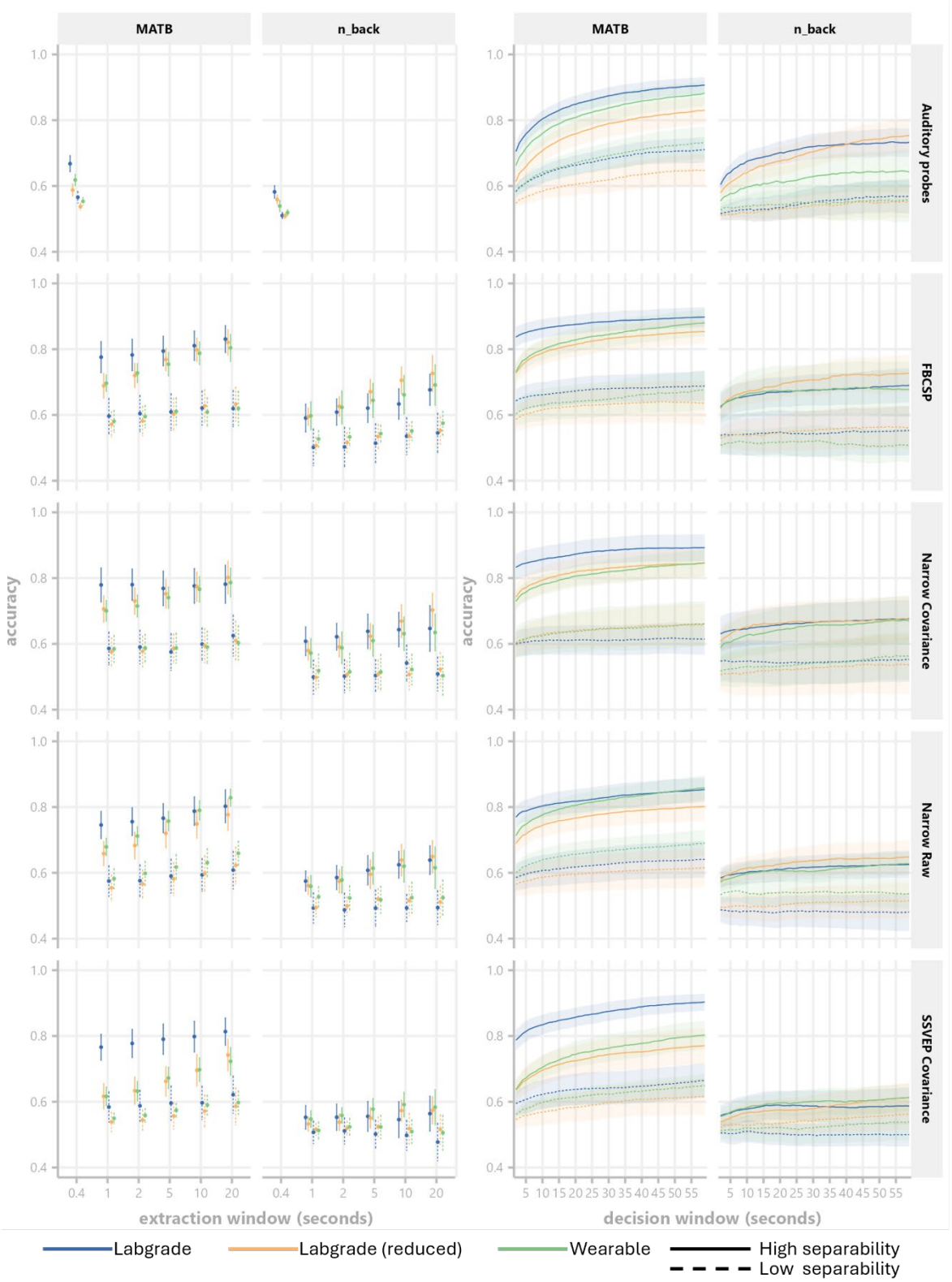
Significant main effects for the cross-subject classification results were present for different features ($F(4,1852.00) = 95.65$, $p < 0.001$), the tasks ($F(1,1852.00) = 443.03$, $p < 0.001$), and contrasts ($F(1,1852.00) = 457.94$, $p < 0.001$). The main effect for datasets and classifiers was again non-significant. However, the interaction between features and datasets was significant for the cross-subject classification ($F(4,1852.00) = 4.32$, $p = 0.002$). Significant pair-wise comparisons were again mostly limited to the MATB's high-separability scenario (see Figure 49). Here, the Wearable dataset exhibited higher accuracy scores than the reduced Lab-grade dataset. A single pair-wise comparison in the n-back's low-separability scenario exhibited significant differences in favour of the reduced Lab-grade data. This was the case for the Narrow Covariance features, however, the median classification accuracy was only marginally above the sample-size corrected chance level.

5.2.2 *Temporal Aggregation*

The next stage of the analysis was to consider how the size of the feature extraction window interacts with classification accuracy. Data for both MATB and n-back at low and high separability for the within-subjects case are illustrated in Figure 50. Differences between the full Lab-grade and the

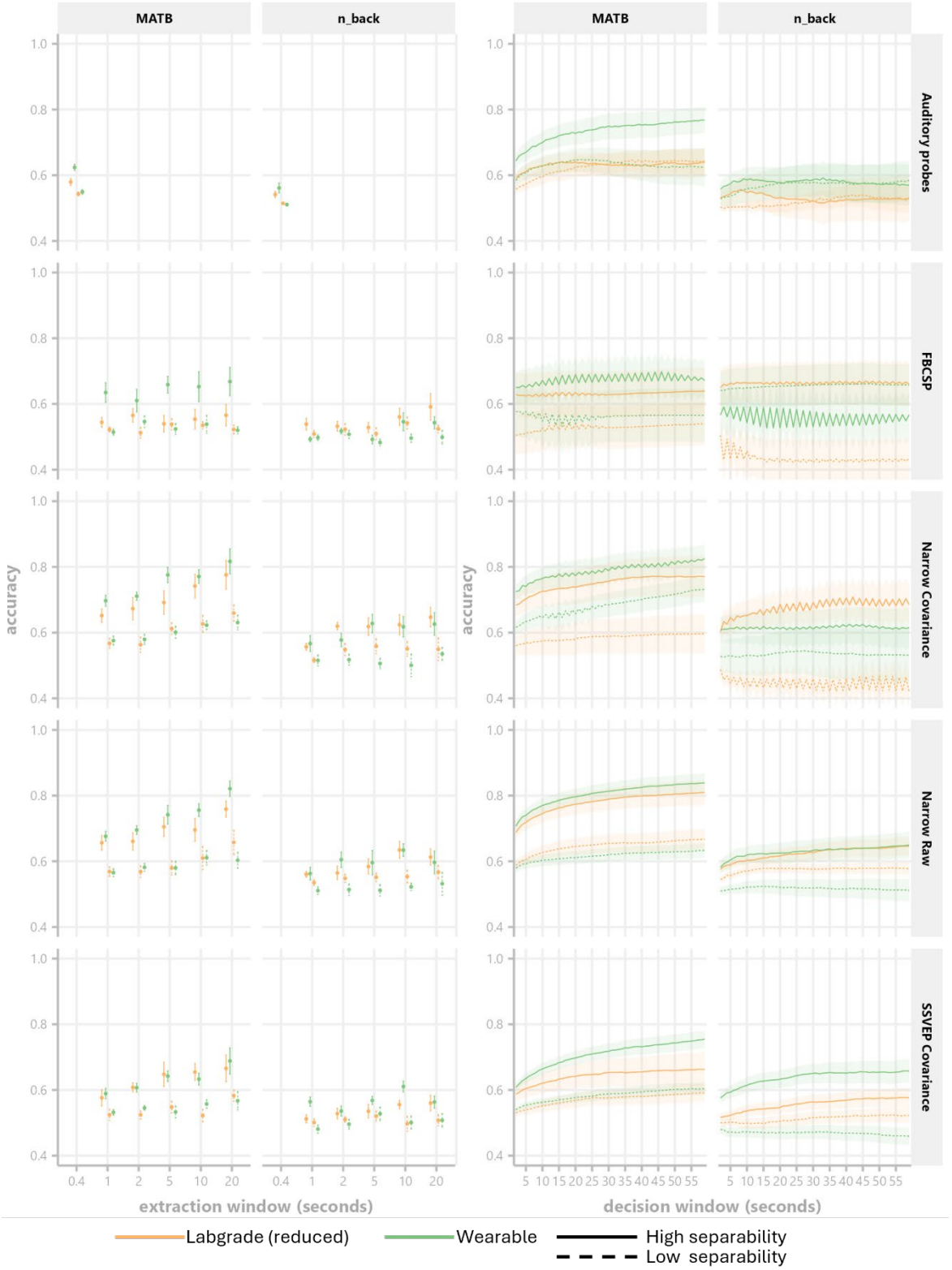
other two data sources in MATB's high separability scenario shrunk at higher extraction window values. Temporal aggregation results, which were plotted using the lowest possible extraction window (.4 or 1-second) in Figure 50, did not show as consistent reductions, as initial accuracy advantages seem to carry over to higher decision window values in many of the tested feature categories. Nonetheless, it seems that with "single-trial" observations being sufficiently informative (around 60% accuracy), the temporal aggregation of sequential samples using a simple majority vote improves the accuracy by a great margin. This effect was most pronounced for the .4 second extraction windows of the auditory probes.

Figure 50. Within-Subject Extraction Window and Aggregation Results



Note. Average accuracy over subjects per montage for single extraction windows on the left, and aggregated decision windows (over 1-second extraction windows) to the right. Error bars (or shaded area) represent the 95% confidence intervals.

Figure 51. Cross-Subject Extraction Window and Aggregation Results



Note. Average accuracy over subjects per montage for single extraction windows on the left, and aggregated decision windows (over 1-second extraction windows) to the right. Error bars (or shaded area) represent the 95% confidence intervals.

Low-separability scenarios where “single-trial” observations are much noisier did not show significant improvements after aggregation. For the n-back, increasing the extraction window showed improvements by up to 10% in the high-separability scenario, but no such improvements were visible for the n-back’s low-separability scenario.

Similar to the within-subject results, the cross-subject classification accuracies tended to increase with increasing extraction window and decision window size. Here, however, noticeable exceptions were the FBCSP features that did not improve using either technique. The low-separability scenarios also did not show improvements, similar to the within-subject classification, using either larger extraction windows or temporal aggregation.

Table 28 and Table 29 display the “sweet-spot” results, listing the most accurate and responsive combination of extraction window, decision window, feature category, and classifier per task, class separability contrast and EEG montage. The displayed pipelines were determined by finding the highest classification performance across subjects for one of the aforementioned data cases (usually at 60s decision windows), selecting all pipelines that still fell within its 95% confidence interval and storing the one with the smallest decision window, resulting in a list of pipelines that exhibited the highest accuracy in combination with the fasted responsiveness.

For the within-subject classification, Table 28 suggested an advantage of the auditory probe technique in the full Lab-grade and reduced Lab-grade data. The Wearable data exhibited varied feature category and extraction window choices. The decision window time ranged from 21 to 40 seconds. The cross-subject results in Table 29 echoed a similar heterogeneity, with the auditory probes now chosen for both n-back scenarios in the Wearable data and the high class-separability scenario of reduced Lab-grade’s MATB. The Random Forest model, however, ended up being chosen 6 out of 8 times, suggesting its non-linear decision boundaries to be advantageous for the cross-subject classification. Decision windows, again, ranged from 21 to 41 seconds, with the low class-separability n-back data being the notable exception, as classification accuracy tended not to improve using either larger extraction windows or temporal aggregation (see also Figure 51).

Table 28. Classifiers with Best Accuracy/Responsiveness Trade-Off (Within-Subject)

Task	Contrast	Dataset	Accuracy	CI	Extraction window	Decision window	Classifier	Feature Category
MATB	High separability	Lab-grade	90.5%	88.2%-92.8%	.4	40	RF	Auditory Probes
		Lab-grade (red.)	88.1%	85.6%-90.7%	.4	40	RF	Auditory Probes
		Wearable	89.3%	86.5%-92.2%	10	41	RF	FBCSP
	Low separability	Lab-grade	74.7%	72.3%-77.1%	.4	43	MDM	Auditory Probes
		Lab-grade (red.)	69.3%	66.4%-72.2%	.4	31	RF	Auditory Probes
		Wearable	68.7%	65.1%-72.3%	5	27	LDA	Narrow raw
n-back	High separability	Lab-grade	75.7%	72%-79.3%	.4	31	LDA	Auditory Probes
		Lab-grade (red.)	76.4%	73.1%-79.7%	.4	32	LDA	Auditory Probes
		Wearable	68.7%	63.8%-73.7%	5	21	LDA	FBCSP
	Low separability	Lab-grade	58.4%	55.1%-61.8%	.4	21	RF	Auditory Probes
		Lab-grade (red.)	62.8%	60.3%-65.3%	.4	25	RF	Auditory Probes
		Wearable	58.2%	54.3%-62.1%	20	27	MDM	SSVEP Covariance

Table 29. Classifiers with Best Accuracy/Responsiveness Trade-Off (Cross-Subject)

Task	Contrast	Dataset	Accuracy	CI	Extraction window	Decision window	Classifier	Feature Category
MATB	High separability	Lab-grade	-	-	-	-	-	-
		Lab-grade (red.)	84.7%	81.7%-87.7%	.4	37	RF	Auditory Probes
		Wearable	89.3%	86.7%-91.9%	20	41	RF	Narrow Raw
	Low separability	Lab-grade	-	-	-	-	-	-
		Lab-grade (red.)	70.9%	67.7-74.1%	20	27	RF	Narrow Covariance
		Wearable	68.8%	65.5%-72.1%	20	27	RF	Narrow Raw
n-back	High separability	Lab-grade	-	-	-	-	-	-
		Lab-grade (red.)	69.7%	65.2%-74.1%	10	21	LDA	Narrow Covariance
		Wearable	68.2%	64.1%-72.3%	.4	18	LDA	Auditory Probes
	Low separability	Lab-grade	-	-	-	-	-	-
		Lab-grade (red.)	59.1%	55.8%-62.4%	1	5	RF	Narrow Covariance
		Wearable	55.7%	53.5%-57.9%	.4	6	RF	Auditory Probes

5.3 Discussion

In this chapter, data from the Lab-grade and Wearable datasets were classified using various feature extraction and classification strategies. This was primarily done to determine whether lab-based results would translate to more convenient wearable EEG headsets. The expectation was that, due to its reduced spatial coverage and higher recording impedance, the accuracy in the Wearable dataset would be significantly worse than in the Lab-grade dataset.

While the best-case scenario for mental workload prediction (high separability; MATB) did suggest the full 64-channel lab-grade montage's superior performance, none of the other tested classification scenarios favoured one over the other montages consistently. Therefore, there was no evidence suggesting that the sparse montage with high-impedance wet electrodes could not serve as a viable option for monitoring mental workload in more applied contexts.

The second aim of the here presented analyses was to determine whether accuracy could be significantly improved by reducing the responsiveness of the system by either increasing the size of the feature extraction windows or aggregating over sequential predictions. Expanding feature extraction windows did seem to produce sizable accuracy increases across all tested feature extraction methods. Similar effects could also be confirmed for aggregating over past predictions. No optimal combination of extraction window and decision window parameters could be determined to work across all tested feature categories. However, in the high separability MATB scenario, increasing the feature extraction window as well as aggregating over smaller extraction windows allowed to the sparse montages tested to achieve similar accuracy as the 64-channel Lab-grade data. Consequently, increasing feature extraction window sizes or aggregating over sequential predictions may be a crucial strategy when operating pBCIs in applied settings, in which predictions from small data segments (<2 seconds) may be too noisy for accurate mental state inference.

Harking back to the no free lunch theorem (Wolpert, 1996b, 1996a) introduced in section 1.1.5, the third aim was to determine whether any of the five tested feature extraction methods would consistently produce the best results across tasks and montages. Table 28 and 29 presented the pipelines with the best accuracy/responsiveness trade-offs. Neither for the within nor for the cross-subject classification could this identify a consistent winner across tasks or montages, giving further credence to the no free lunch theorem. However, looking at figures 48 and 49, band power-based metrics tended to perform at similar levels, and which of them ended up as the “best” pipeline in Tables 28 and 29 could likely have been due to chance. Of the tested band power features, the FBCSP method performed markedly worse in the cross-subject classification. While the extraction of

multiresolution CSP filters was undertaken to counter inter-subject variability issues, these results suggest that sensor-level features or the classification of covariance-based features would be preferable for more generalisable pBCIs.

The analyses presented in this chapter echoed the results previously gleaned by the group-level analysis. Differences in the EEG data between task-load conditions were most pronounced when comparing the easy and hard task-load conditions, while the differences in the n-back tended to be less consistent across measures and datasets. This puts the sensitivity of the here tested EEG metrics into question. A pBCI that is only capable of distinguishing between extreme task-load differences would certainly have fewer use cases for neuroadaptive interfaces than one that could detect gradual changes with the same accuracy.

However, the analyses presented here, while exhaustive, by no means covered the whole spectrum of possible EEG metrics that could be used for mental state classification. The band-power metrics used in the Raw Narrow approach could, for example, be improved by removing the aperiodic activity (Ke et al., 2023). This, however, would likely only work if the power spectra were relatively noise-free, requiring longer extraction windows (10 seconds could be enough already, according to some preliminary tests not reported here). Information theory-inspired or Connectivity-based features may also be computed if longer extraction windows were to be used.

Lastly, the Decision-level fusion employed here may be improved in several ways. Ensemble learning has a long history in machine learning, and many strategies have been developed to reduce variance through the smart combination of disparate information. For example, by employing different weightings of the available information (Ho et al., 1994). In an online system where workload is expected to fluctuate, a weighted aggregation with weights that reduce the further the predictions lie in the past may further enhance performance.

6 Multimodal Decision Fusion

In the previous chapter, classification of mental workload using the Lab-grade and Wearable datasets showed high accuracies in the widest class-separability contrast of the MATB. In the other three classification scenarios, accuracy fell to below 75% in most cases, even after temporal aggregation over 60-second-long decision windows (see Figure 50 and 51). The low class separability contrast (0-back vs 1-back) produced the most uncertain classification scenario (on average 57% in the cross-subject classification and 60% in the within-subject classification). This finding highlighted the limited sensitivity of commonly reported pBCI approaches to mental workload classification, where wide differences in mental workload are easily distinguishable, but more subtle differences are not. These subtle differences in mental workload, when experienced by an operator over prolonged periods, cause strain (Jalali et al., 2023; Mahdavi et al., 2024) that an adaptive system would ideally be able to mitigate and should thus be able to detect.

One possible way to increase sensitivity may lie in combining information not only across time but also across feature categories. Separate measurement modalities, such as ECG, eye-tracking, or fNIRS can complement each other in such ensemble approaches (Vortmann et al., 2022). Combining the information from multiple classifiers works best with a diverse set of classifiers that tend to err on different samples (Polikar, 2006). The process of combining their individually noisy predictions has previously been described as a sort of low-pass filter (Polikar, 2006), smoothing over the noise (similar to the temporal aggregation from Chapter 5). The practice has been described under various names, such as mixture of experts (Jacobs et al., 1991), stacked generalisation (Wolpert, 1992), consensus aggregation (Benediktsson & Swain, 1992), classifier fusion (Bloch, 1996; Kuncheva et al., 2001), and many more.

Another potential advantage of combining multiple neurophysiological indices for mental workload classification arises in the context of cross-subject classification. Not all features may perform equally well across different subjects; some subjects may have their individual best-performing feature category or modality. A multi-feature ensemble with a subject-specific meta-classifier that combines information from various feature categories based on their informational value for that specific individual could provide a generalisable architecture that adapts to the operator and scenario through calibration sessions or adaptive refitting procedures. Viewing different subjects as different datasets relates this to the no free lunch theorem, from which follows, as observable in Chapter 5, that no single classifier will be ideal for every dataset. Ensemble approaches have previously been shown to not produce the highest average accuracy scores across datasets. Instead, their benefit stemmed from being ranked the highest on average, meaning that while a specialised classifier may

reach higher accuracy on suitable datasets, an ensemble trades some accuracy for higher robustness across datasets (Gómez & Rojas, 2016). In this chapter, the multimodal dataset will be utilised for such a (stacked) ensemble approach, fusing two of the EEG-based classifiers tested previously with two types of fNIRS classifiers.

FNIRS-based classification of mental workload is commonly carried out by extracting time-series statistics like the mean, standard deviation, skewness, and kurtosis (Angsuwatanakul et al., 2020; Naseer et al., 2016; Sun et al., 2015). The slope of the HbO and HbR signals has also been shown to contain information about different n-back levels and is sometimes computed using a least-squares regression (Herff et al., 2014) or by subtracting the average HbO concentration from the first half of a trial from the second half of the trial (Coyle et al., 2007). However, due to fNIRS' high spatial specificity, the montage may be a deciding factor in whether such findings replicate in new data.

Alternatively, a recent preprint proposed a method for classifying fNIRS data by utilising the Riemannian geometry of spatial covariance and other kernel matrices (akin to similarity matrices) of HbO and HbR data (Näher et al., 2024). They presented their method's superior classification performance in distinguishing between a battery of cognitive tasks compared to the aforementioned traditional fNIRS feature categories. In their Covariance-based approach, the covariance of the HbO and HbR data is computed in isolation to allow for different shrinkage levels and other data-driven data augmentation procedures. This is akin to the Riemann-Narrow classifier that was already presented in Chapters 4 and 5, where dimensionality reduction and shrinkage were computed for each frequency band before combining them into a block-diagonal matrix.

Multimodal fusion of fNIRS and EEG has previously been explored for the detection of affective states (Sun et al., 2015), drowsiness (Nguyen et al., 2017), stress (Al-Shargie et al., 2016), as well as mental workload in n-back tasks (Liu et al., 2017; Mandal et al., 2020) and a mental arithmetic task (Qiu et al., 2022). Data-level fusion, feature-level fusion, and decision-level fusion were all demonstrated to improve performance over individual, modality-specific classifiers. However, as with the EEG literature, it is not uncommon to either find no detailed information on cross-validation data splitting procedures or to find block-structure independent splitting schemes in the fNIRS and hybrid mental state classification literature. While Chapter 4 did not investigate how the bias stemming from such block-independent splits compares between EEG and fNIRS, it is likely that it is at least equivalent and may even be more pronounced, given the slowly evolving nature of the haemodynamic response. Consequently, interesting results suggesting either superior feature-level or decision-level accuracies over single-modality approaches become difficult to interpret/compare. Unclear cross-validation descriptions like "we adopted leave-one-out approach for cross validation among the 22

subjects" (Al-Shargie et al., 2016) leave the reader in the dark as to whether the authors may have presented impressive cross-subject results (93% accuracy distinguishing between self-paced or timed mental arithmetic blocks) or an overfitted support vector machine within a leave-one-sample-out cross-validation. Block-structure independent procedures tend to also be reported (Liu et al., 2017; Mandal et al., 2020; Nguyen et al., 2017) as well as non-descriptive cross-validation methods such as "8-fold cross-validation", not detailing whether samples were stratified to assure class-balances, or perhaps even randomised before data splitting (Qiu et al., 2022). As previously demonstrated in Chapter 4, model comparison of models with increasing complexity using block-structure independent cross-validation may result in significant differences in favour of the more complex classifiers, as they tend to exploit the block/trial-dependent multivariate patterns more effectively.

However, ample examples outside the mental-state classification literature also report classification performance improvements testing EEG-fNIRS hybrid classifiers, such as the classification of ALS patients (Deligani et al., 2021), motor-imagery (Fazli et al., 2012), or real-time quad-copter control (Khan & Hong, 2017). Consequently, the fusion of these two modalities surely warrants further research in the context of mental state classification.

In the current chapter, a multimodal decision-level fusion approach was tested against modality-specific classifiers. The main research question that this analysis aimed to answer was whether the fusion of fNIRS and EEG data can improve classification beyond that achieved by EEG classifiers alone. Potential improvements are of particular interest in the low-class separability cases, where the results of Chapter 5 suggested subpar classification accuracies.

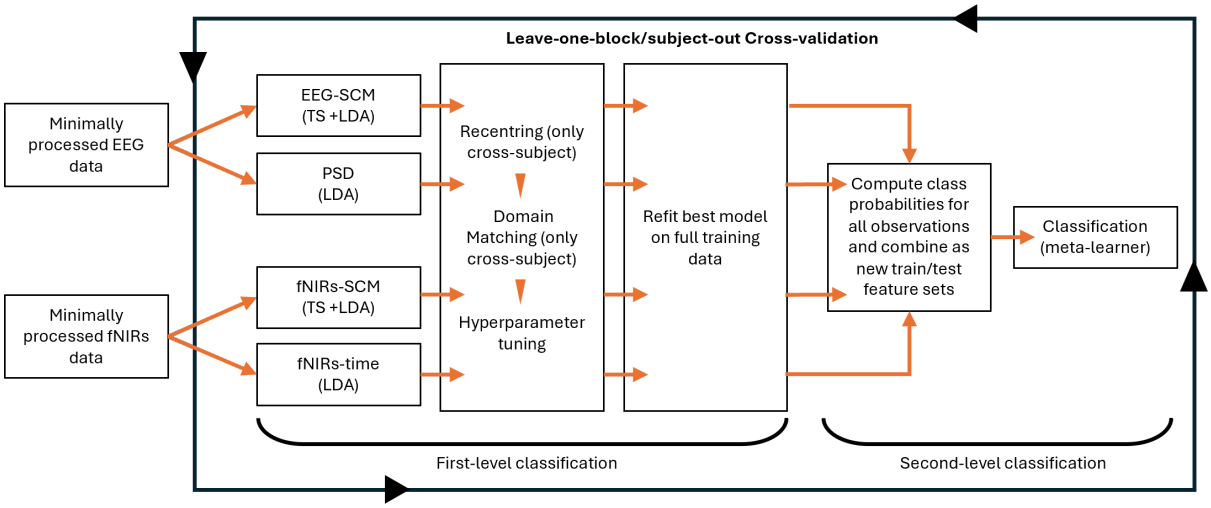
6.1 Methods

The Multimodal dataset consisted of 19 participants, of whom only 14 completed all blocks of the experiment. These 14 were included in this analysis. Preprocessing of the fNIRS data, consisting of 6 parietal and 11 frontal channels, was detailed in Section 2.3.2. The minimally processed EEG data, consisting of 32 channels arranged according to the 10-20 layout, was used for the EEG-based classifiers. Additionally, as in chapter 5, the same tests were also repeated using a reduced montage, which only included the 7 channel locations also present in the X.on headset (F3,F4,C3,Cz,C4,P3,P4). Modality-specific classification (first-level) was carried out exactly as previously described in Chapter 5 with a minor tweak. The cross-validation splits, domain matching, and recentring approaches were the same. However, instead of testing three types of classifiers, the current analysis no longer included MDM or Random Forest classifiers due to the lack of a classifier main effect in section 5.2.1. The main change compared to Chapter 5 was the distinction between first-level and second-level classifiers. The first-level classifiers can be compared to those tested in Chapters 4 and 5. Their class

probability estimates were treated as feature vectors and stacked to form a new feature space, which was used to train a second-level classifier - a meta-learner. The meta-learner ensures that the information from multiple first-level classifiers is optimally combined. Together, they form a so-called stacked ensemble (see Figure 52)

6.1.1 *Stacked Ensemble*

Figure 52. Stacked Ensemble Schematic



Note. Schematic of the stacked ensemble classification pipeline. Steps in the first-level classification were carried out per classifier and followed the same logic as in the more detailed Figures 46 and 47.

In the stacked ensemble, decision-level fusion was handled not by heuristics but by a second-level classifier, a so-called meta-learner, that aimed to capitalise on highly predictive first-order classifiers while subduing the influence of less-informative classifiers. In the current analysis, logistic regression was used as a meta-learner model. An L2 (Ridge) penalty was applied to ensure that no single classifier would dominate the meta-learner. An L1(LASSO) penalty could have been chosen alternatively to zero out non-informative features. Both approaches, as well as many other types of classifier selection and meta-learner model choice, may have their own advantages. However, the no free lunch theorem seems to also hold in the selection of decision functions for ensemble-based predictions, as no single method performs optimally across contexts (Polikar, 2006). Polikar (2006) suggested that using a meta-learner, such as a weighted decision function, should be preferred over simple means or multiplication rules of posterior probabilities, if the accuracy of individual classifiers can be estimated. Here, the meta-learner was trained on the first-level classifiers' class-probability estimates from the training/calibration data, resulting in a subject-specific meta-learner.

6.1.2 FNIRs Classifiers

FNIRs-SCM. As described in Näher et al. (2024), covariance matrices of HbO and HbR data were separately regularised using Ledoit-Wolf shrinkage. The same dimensionality reduction technique used for the EEG SCMs was further applied separately, reducing each SCM to 10 dimensions or channels. After shrinkage and dimensionality reduction, the SCMs were concatenated into a block-diagonal matrix with the off-diagonal blocks set to 0. Finally, the block-diagonal matrices were projected onto their tangent space. An sLDA classifier was used for the first-level classification.

FNIRS-time. The classic fNIRS classification literature tends to focus on first- and second-order statistics for classification. Here, the mean, variance, kurtosis, skewness, peak value, and slope were extracted from each channel's HbO and HbR time-series, resulting in a 17 (number of channels) by 12 (number of statistics) feature matrix per window. The peak value was computed as the maximum value within a given window. The slope was calculated using a first-order polynomial fit via numpy's polyfit function. Again, an sLDA classifer was used for the first-level classification.

6.1.3 EEG Classifiers

Rather than incorporating the full array of previously validated EEG classifiers, which would have significantly increased the complexity of the fusion model and the subsequent analysis, the selection was restricted to two representative classifiers.

EEG-SCM. A simple Sample Covariance-based (SCM) Riemannian classifier akin to the one used for the auditory probe's was used as the first EEG-based classifier. Before segmentation, the EEG data was bandpass filtered between 1-13Hz, just like for the Covariance-Probe features in section 5.1.1. The efficient electrode selection already utilised in the previous EEG classification was used to reduce each SCM to 10 dimensions/channels. Finally, the SCMs were projected onto their tangent space. An LDA classifier was used for the first-level classification.

EEG-power. The same Narrow-power classifier already described in section 5.1.1 was reused here. Band-power features were z-scored based on the training data in the within-subject classification and the calibration data in the cross-subject classification. As with the EEG-SCM features, an LDA classifier was used for the first-level classification.

While other classifiers from Chapter 5 (such as FBCSP or Narrow-band Covariance) offered variations on these themes, restricting the ensemble inputs to these two fundamental representations (spectral and spatial) serves to reduce potential redundancy and collinearity. Furthermore, by limiting the EEG inputs to two classifiers a balanced ratio with the fNIRS modality (2:2) was

maintained, keeping the dimensionality of the final ensemble within a range that allowed for a clear and interpretable analysis of the fusion results.

6.1.4 Statistics

As in Chapter 5, mental workload classification was conducted over a range of feature extraction windows (2,5,10,20 seconds, with 33% overlap) for both high-separability and low-separability scenarios in the n-back and MATB tasks. Within- and cross-subject classification was tested. Due to fNIRS' slow nature, the 1-second extraction window was excluded from the following analysis.

Whether the Meta-learner performed better than any of the first-level classifiers was tested using linear-mixed effects models on logit-transformed subject-wise averaged accuracy estimates. Pairwise comparisons were FDR corrected using the Benjamini-Hochberg method with the FDR controlled at 5%.

```
Formular: logit(Accuracy) ~ Classifier * Cross-subject* Montage +
Extraction window + (1 | subject)
```

Subject ID was used to compute random intercepts. The categorical factors “Montage”, “Cross-subject”, and “Classifier” were dummy-coded with the full montage, within-subject classification, and the meta classifier as the respective reference levels. The factor “Extraction window” was treated as a continuous variable.

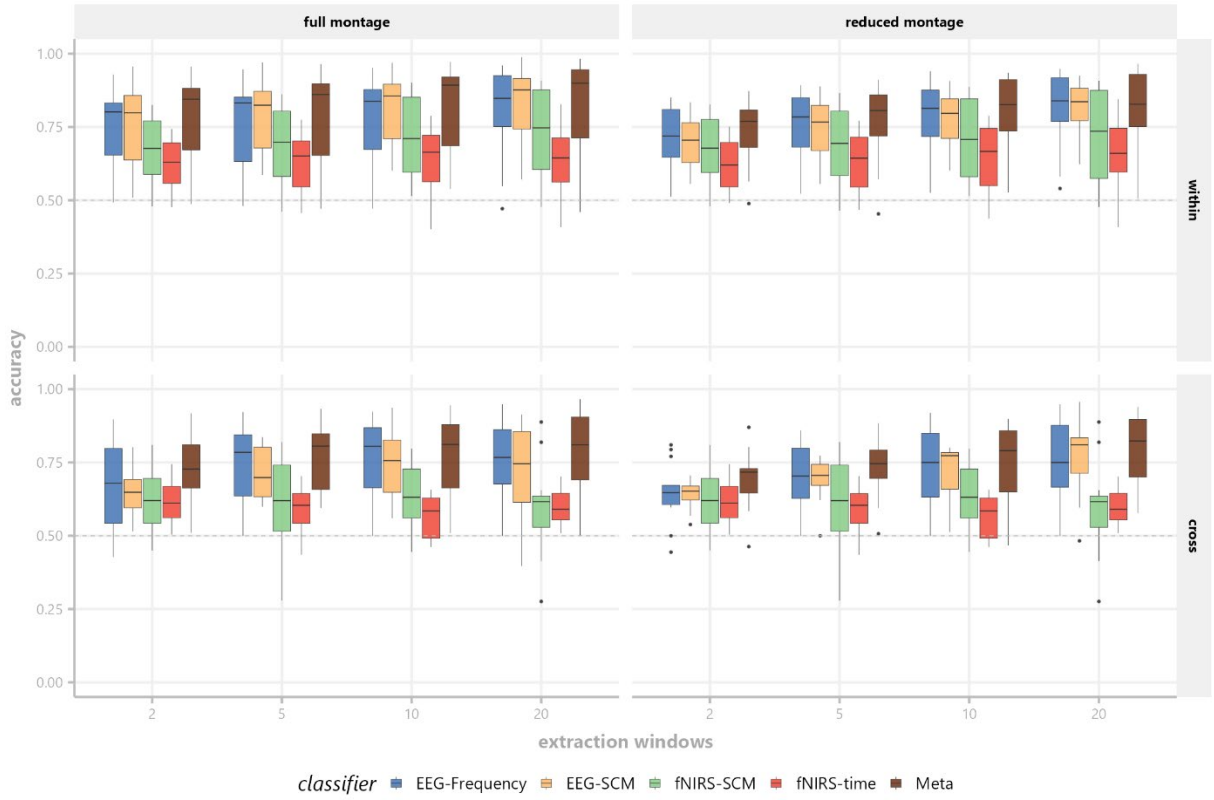
The temporal aggregation from Chapter 5 was also performed, but only for the meta-learner’s predictions, to provide a final estimate of the proposed pipeline's accuracy in mental workload monitoring. Tables containing the optimal combination of extraction and decision window were computed as was done previously in Chapter 5.

6.2 Results

Of main interest to this Chapter’s analysis was whether stacking different pBCI classifiers could improve the performance above and beyond any single classification approach (first-level classifier). Of further interest was whether fNIRS as an additional neurophysiological source of information, could improve the lacking sensitivity of the EEG-based mental workload classifiers when faced with the low class-separability contrasts.

6.2.1 *Meta-learner vs Single Classifiers*

Figure 53. Multimodal MATB Classification Results (High Class-Separability)



Note. Boxplot of subject-wise average accuracies

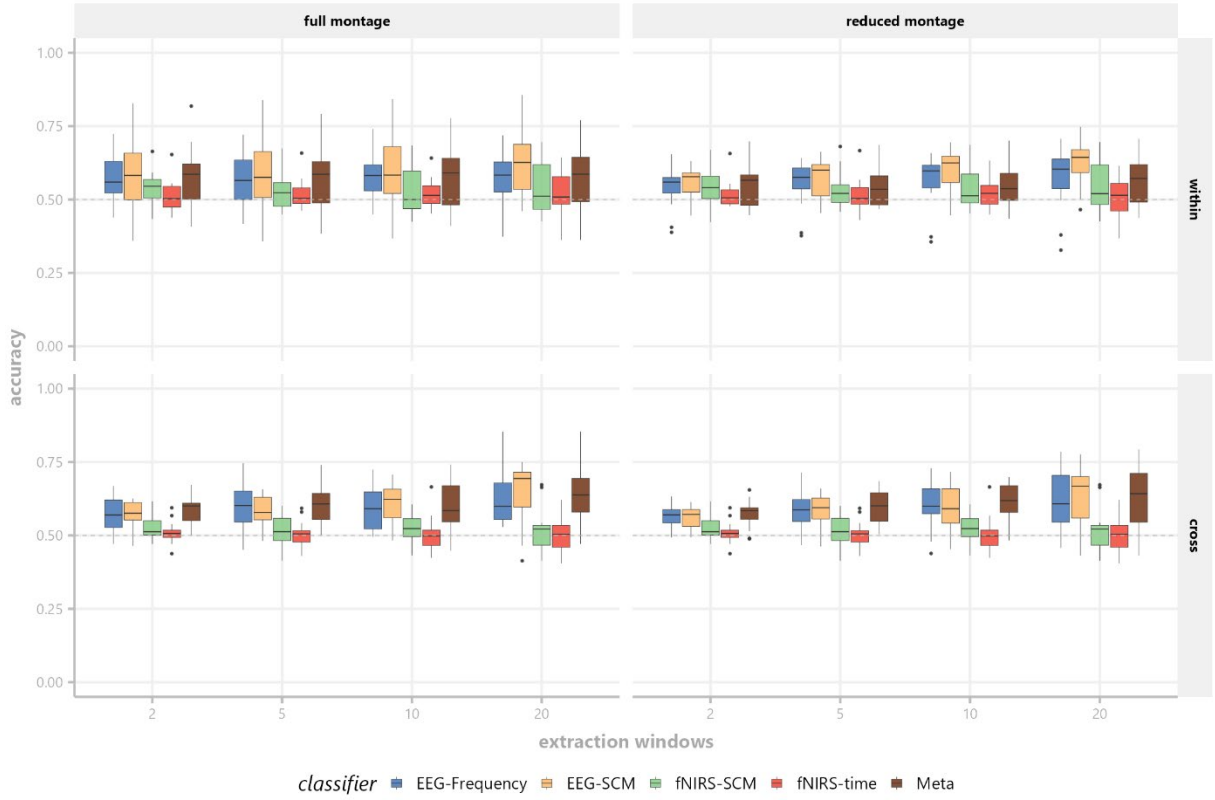
MATB. In the high-separability scenario of the MATB (Figure 53), the factors classifier ($F(4, 1086) = 153.18$, $p < 0.001$), montage-size ($F(1, 1086) = 13.35$, $p < 0.001$), cross-subject/within-subject comparison ($F(1, 1086) = 146.32$, $p < 0.001$) and extraction window size ($F(1, 1086) = 103.85$, $p < 0.001$) were significant and so were the interaction of classifier by cross-subject classification ($F(4, 1086) = 3.25$, $p = 0.005$), and classifier by montage-size ($F(4, 1086) = 2.35$, $p = 0.03$).

Table 30. MATB High Class-Separability First- and Second-level Comparisons

	Full-montage		Reduced-montage	
	Within	Cross	Within	Cross
EEG-SCM	-0.47%	-5.49%***	-3.95%*	-1.97%**
EEG-Frequency	-2.93%**	-2.93%*	-2.68%	-1.7%
fNIRs-SCM	-9.73%***	-15%***	-7.89%***	-12.7%***
fNIRs-time	-17%***	-17.2%***	-14.5%***	-14.9%***
Meta (accuracy)	80%	76.4%	77.9%	74.1%

Table 30 presents the BH corrected post-hoc comparisons of the first-level classifiers with the meta-learner across extraction windows and montages. Individual cells of the first-level classifiers display the average difference to the meta-learner. Significant negative differences indicate that the meta learner performed better than the respective first-level classifier, which was the case for all but the EEG-Frequency LDA model using the reduced 7-channel montage.

Figure 54. Multimodal MATB Classification Results (Low Class-Separability)



Note. Boxplot of subject-wise average accuracies

In the low-separability scenario of the MATB (Figure 54), the factors classifier ($F(4, 1086) = 74.81$, $p < 0.001$) cross-subject/within-subject comparison ($F(1, 1086) = 7.4$, $p = 0.006$) and extraction window size ($F(1, 1086) = 25.64$, $p < 0.001$) as well as interaction of classifier and cross-subject/within-subject comparison ($F(4, 1086) = 8.34$, $p < 0.001$) were significant. The montage size or other interaction terms were not significant.

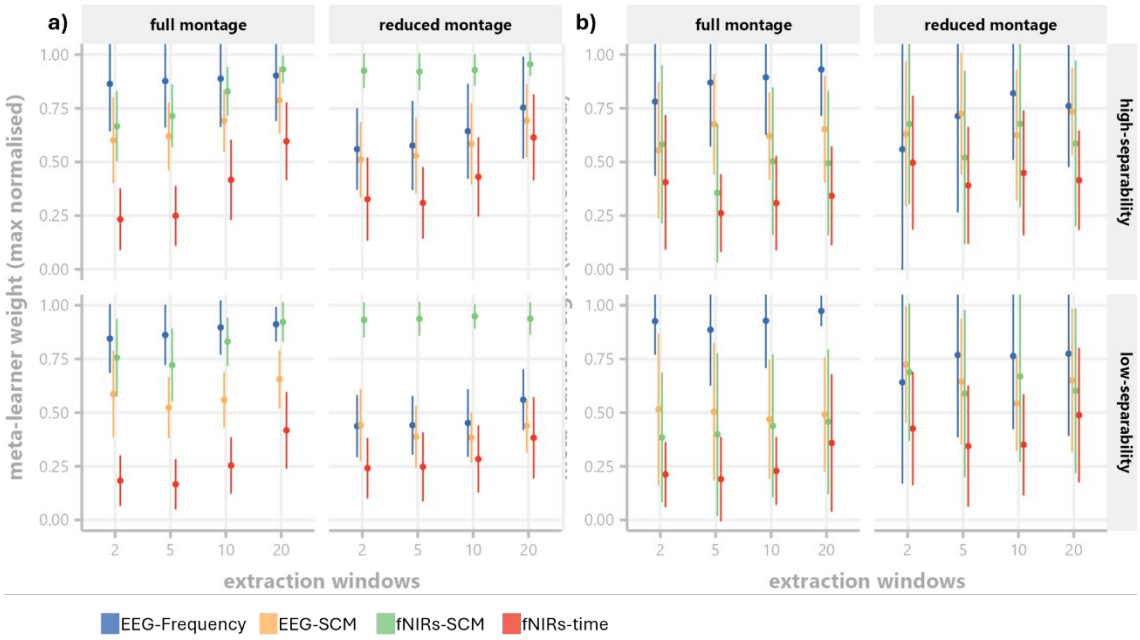
Table 31 presents the BH corrected post-hoc comparisons of the first-level classifiers with the meta-learner. While the meta-learner still performed significantly better than the first-level fNIRS classifiers, the EEG classifiers no longer differed significantly, and in the case of the covariance-based EEG-SCM approach, the reduced montage even indicated superior performance of the first-level classifier in within-subject classification.

Table 31. MATB Low Class-Separability First- and Second-level Comparisons

	Full-montage		Reduced-montage	
	Within	Cross	Within	Cross
EEG-SCM	1.41%	-0.73%	3.72%*	-0.89%
EEG-Frequency	-0.37%	-1.06%	0.67%	-0.99%
fNIRs-SCM	-4.35%***	-8.52%***	-1.1%	-8%***
fNIRs-time	-5.93%***	-10.3%***	-3.39%*	-9.74%***
Meta (accuracy)	57.7%	60.7%	55%	60.2%

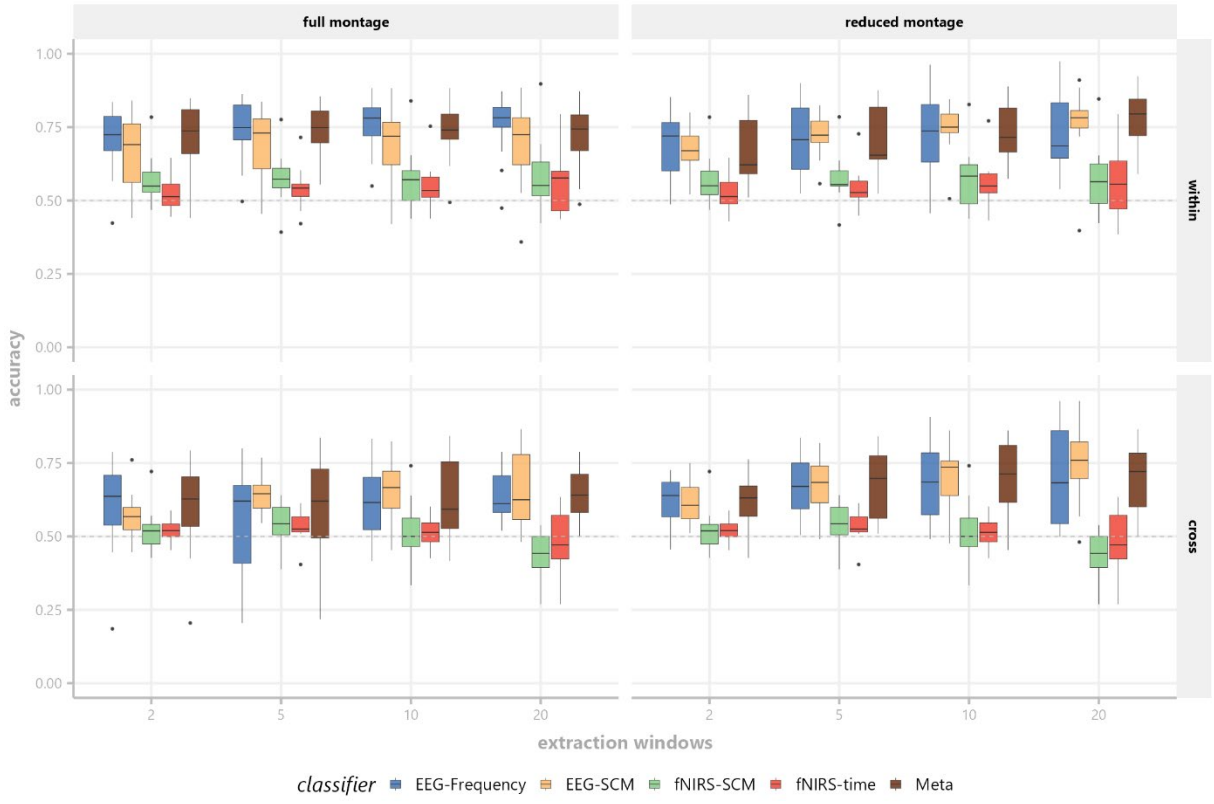
The weights assigned to each of the first-level classifiers by the meta-learner can be seen in Figure 55. The meta-learner tended to assign the lowest weight to the fNIRS-time classifier across all MATB classification scenarios. Interestingly, the fNIRS-SCM classifier was weighted similarly to the more accurate EEG-Frequency classifier and was even given more importance in the reduced montage cases for the within-subject classification. In the cross-subject classification, the first noticeable difference from the within-subject cases was the much larger standard deviations across first-level classifiers. Additionally, the weighting of the fNIRS-SCM classifier was reduced in the cross-subject classification, suggesting that the fNIRS-SCM approach did not generalise as well as the EEG ones across participants. The larger standard deviations could have also been due to high collinearity, however, in that case, it would be expected that the mean weights in the figure would be closer together without any single extreme superior weighted first-level classifier.

Figure 55. Multimodal MATB Meta-Learner Weights



Note. Mean and standard deviation of subject-wise averaged and max-normalised meta-learner weights for a) within-subject and b) cross-subject classification.

Figure 56. Multimodal N-back Classification Results (High Class-Separability)



Note. Boxplot of subject-wise average accuracies

N-back. In the high-separability scenario of the n-back (Figure 56), the factors classifier ($F(4, 1086) = 164.21$, $p < 0.001$), montage-size ($F(1, 1086) = 13.2$, $p < 0.001$), cross-subject/within-subject comparison ($F(1, 1086) = 16.64$, $p < 0.001$) and extraction window size ($F(1, 1086) = 21.92$, $p < 0.001$) were significant and so were the interaction of classifier by cross-subject classification ($F(4, 1086) = 4.22$, $p = 0.002$), classifier by montage-size ($F(4, 1086) = 2.59$, $p = 0.034$), cross-subject classification by montage-size ($F(1, 1086) = 9.72$, $p = 0.002$) and their three-way interaction ($F(4, 1086) = 2.47$, $p = 0.042$).

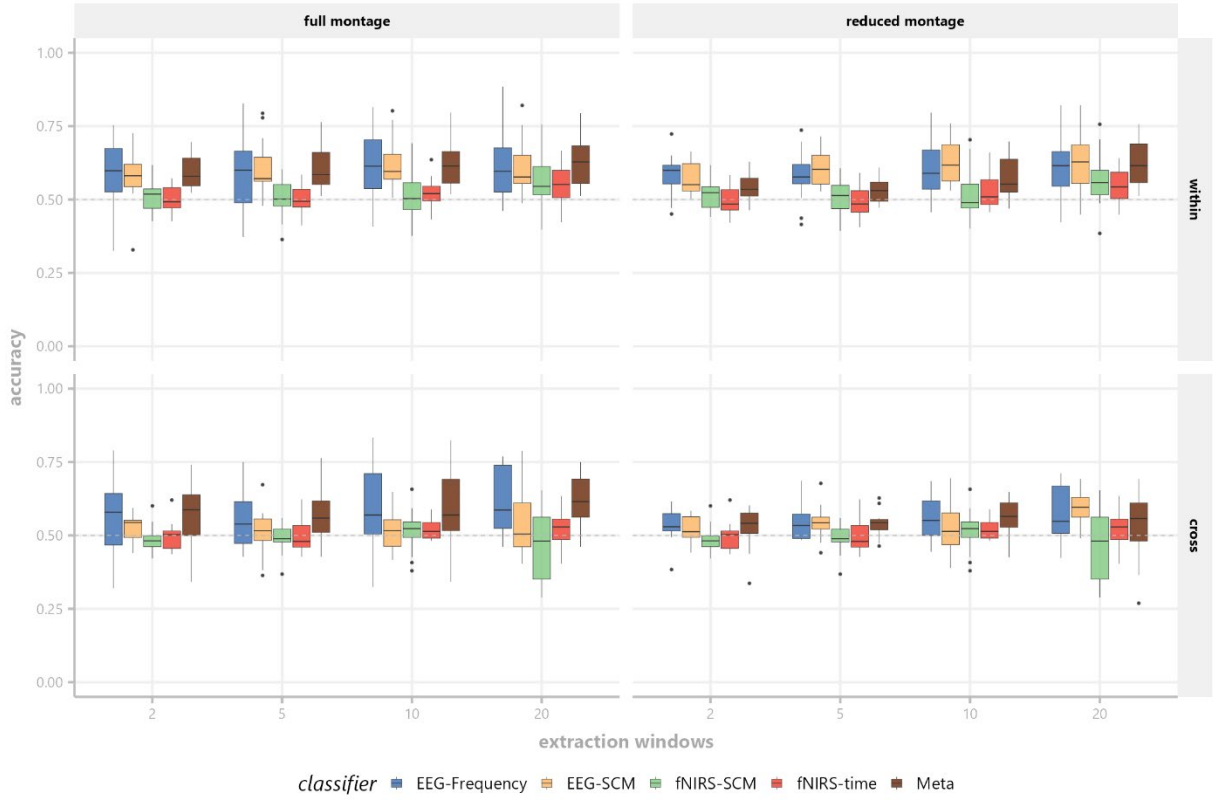
Table 32. N-back High Class-Separability First- and Second-level Comparisons

	Full-montage		Reduced-montage	
	Within	Cross	Within	Cross
EEG-SCM	-0.38%	1.5%	0.94%	0.96%
EEG-Frequency	1.31%	-0.83%	0.43%	-0.59%
fNIRs-SCM	-15.2%***	-10.9%***	-14.7%***	-17%***
fNIRs-time	-18.6%***	-10.3%***	-17.3%***	-16.3%***
Meta (accuracy)	72.7%	61.4%	71.8%	67.5%

Table 32 presents the post-hoc comparisons of the first-level classifiers with the meta-learner, which mimicked the results of the low class-separation scenario of the MATB, with the first-level classifiers not differing significantly from the meta-learner.

In the low-separability scenario of the n-back (Figure 57), the factors classifier ($F(4, 1086) = 48.49$, $p < 0.001$), montage size ($F(1, 1086) = 6.14$, $p < 0.001$), cross-subject/within-subject comparison ($F(1, 1086) = 55.7$, $p < 0.001$) and extraction window size ($F(1, 1086) = 41.17$, $p < 0.001$) were significant and so were the interaction of classifier by cross-subject classification ($F(4, 1086) = 5.44$, $p = 0.002$) and classifier by montage-size ($F(4, 1086) = 4.85$, $p = 0.034$). Cross-subject classification by montage size and the three-way interaction term were not significant. Table 33 presents the post-hoc comparisons of the first-level classifiers with the meta-learner.

Figure 57. Multimodal N-back Classification Results (Low Class-Separability)

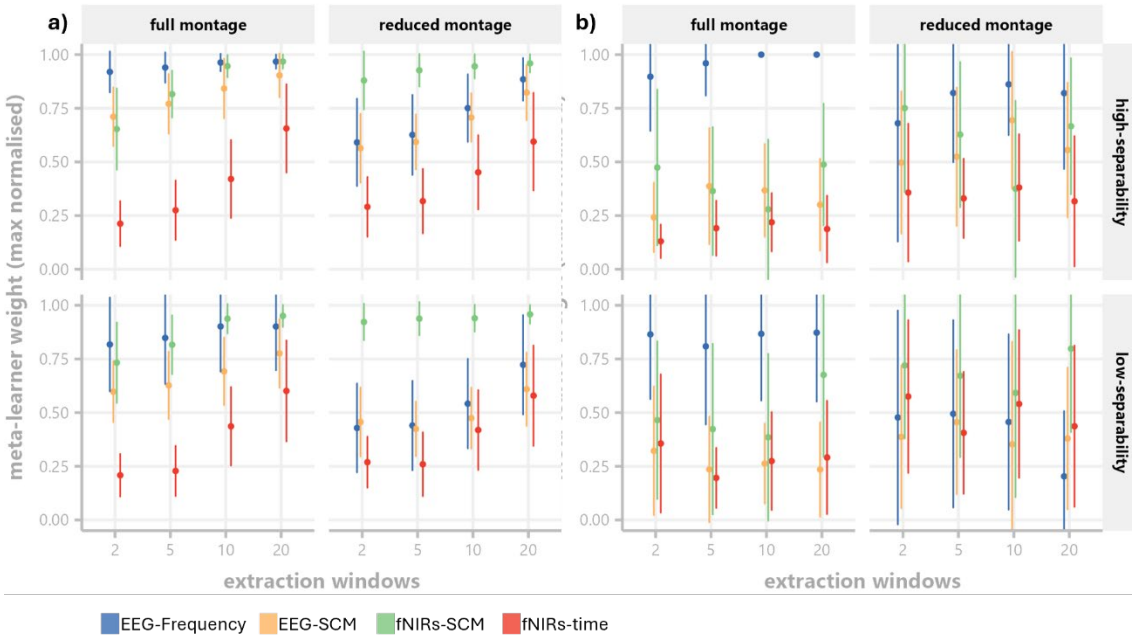


Note. Boxplot of subject-wise average accuracies

Table 33. N-back Low Class-Separability First- and Second-level Comparisons

	Full-montage		Reduced-montage	
	Within	Cross	Within	Cross
EEG-SCM	-0.53%	6.57%*	3.76%*	1.49%
EEG-Frequency	1.14%	-0.83%	2.71%	-0.84%
fNIRs-SCM	-8.55%***	-9.87%***	-4.14%*	-4.8%**
fNIRs-time	-9.38%***	-7.69%***	-5.36%**	-2.71%
Meta (accuracy)	61.2%	58.8%	56.9%	53.8%

Figure 58. Multimodal N-back Meta-Learner Weights



Note. Mean and standard deviation of subject-wise averaged and max-normalised meta-learner weights for a) within-subject and b) cross-subject classification.

The meta-learner weights for the n-back task (Figure 58) mimicked those of the MATB classification. Standard deviations were increased in the cross-subject classification, suggesting either that different subjects require different combinations of first-level classifiers or that the lack of truly robust results caused somewhat random weight assignments. The fNIRS-time classifier was once again weighted with the least importance in most scenarios and, consistent with the MATB results, the fNIRS-SCM classifier rose in importance when the EEG montage was reduced to seven electrodes.

6.2.2 *Aggregation results*

The best compromise between classification responsiveness and accuracy presented in Table 34 and Table 35 suggested that the addition of fNIRS information did not improve the classification beyond the results presented Chapter 5's Table 28 and 29 and, in some cases, the accuracy in the low-separability scenarios was reduced in the meta-learner results (Figures 59 and 60). However, the decision windows tended to be smaller than those presented in Chapter 5. Both results may be influenced by the absence of a Narrow-SCM classifier in this chapter, which was intentionally omitted here due to its slow training times.

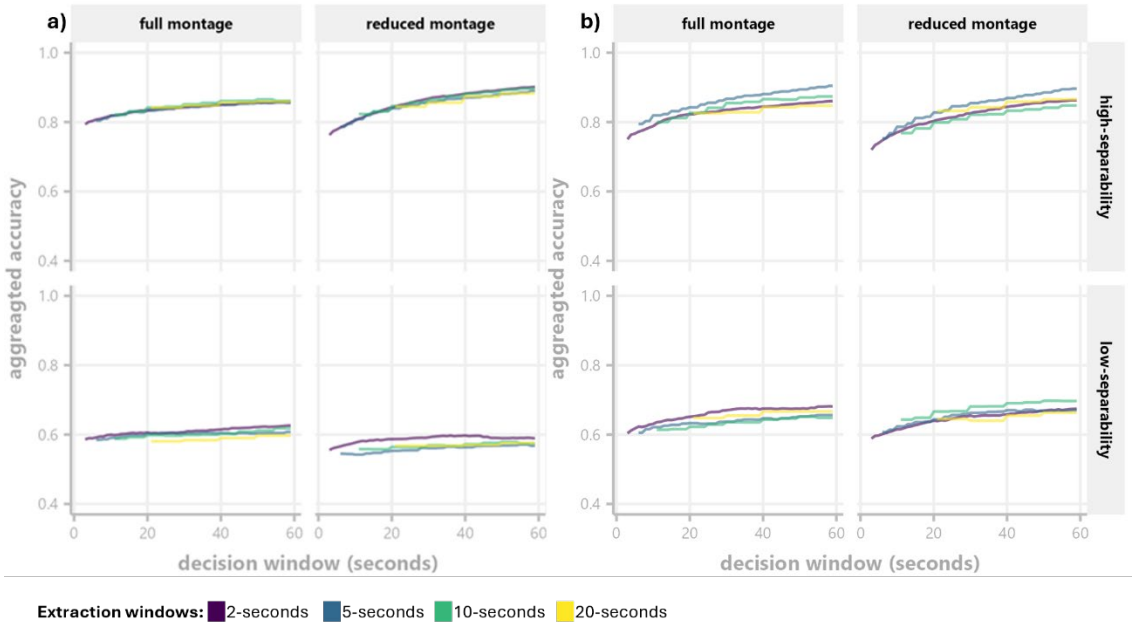
Table 34. Meta-Learner with Best Accuracy/Responsiveness Trade-Off (Within-Subject)

Task	Contrast	Montage	Accuracy	95%CI	Extraction Window	Decision Window
MATB	High-separability	Full	80.7%	75.1%-86.4%	2	6
		Reduced	85.7%	80.2% - 90.4%	2	24
	Low-separability	Full	58.6%	54.4% - 62.8%	2	3
		Reduced	55.6%	52.6% - 58.5%	2	3
N-back	High-separability	Full	75.7%	72.1% - 79.4%	5	8
		Reduced	81.2%	77.8% - 84.7%	20	21
	Low-separability	Full	65.9%	62.3% - 69.5%	10	20
		Montage	63.9%	60.5% - 67.5%	20	21

Table 35. Meta-Learner with Best Accuracy/Responsiveness Trade-Off (Cross-Subject)

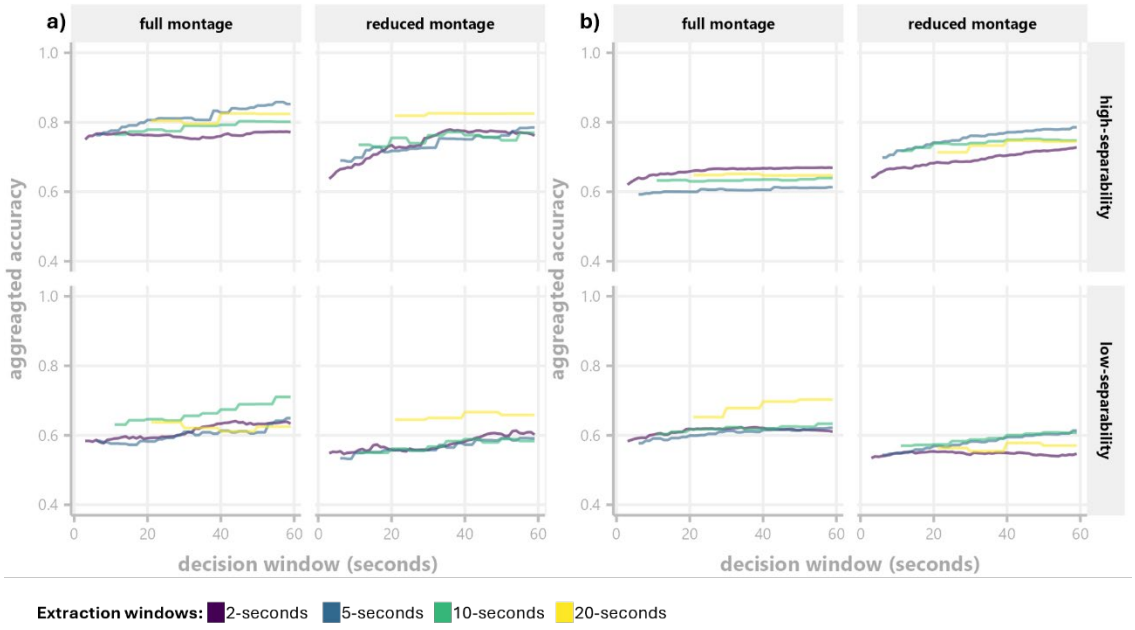
Task	Contrast	Dataset	Accuracy	95%CI	Extraction Window	Decision Window
MATB	High-separability	Full	86.7%	82.2%-91.2%	5	30
		Reduced	85.5%	81.1% - 89.8%	5	30
	Low-separability	Full	65.4%	62.5% - 68.3%	2	22
		Reduced	66.7%	63% - 70.3%	10	20
N-back	High-separability	Full	62%	55.8%-68.2%	2	3
		Reduced	72.3%	67%-77.6%	5	13
	Low-separability	Full	65.2%	61.5%-69%	20	21
		Reduced	57%	54.7%- 59.3%	10	11

Figure 59. Meta-Learner Aggregation Results (MATB)



Note. Average classification accuracies after temporal aggregation for a) within-subject and b) cross-subject classification.

Figure 60. Meta-Leaner Aggregation Results (N-back)



Note. Average classification accuracies after temporal aggregation for a) within-subject and b) cross-subject classification.

6.3 Discussion

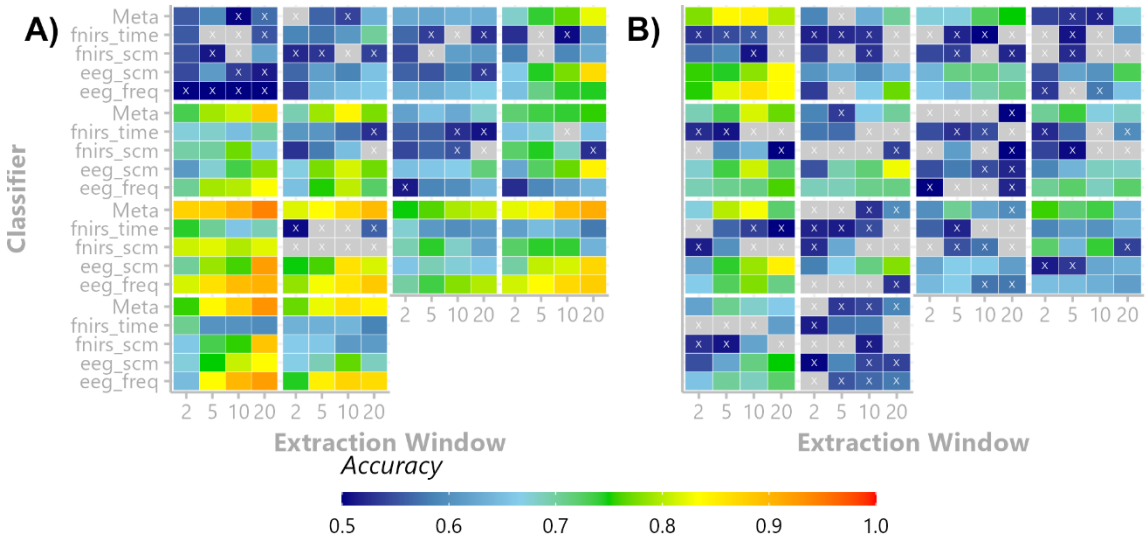
In this chapter, the multimodal dataset was used to test the potential of fusing fNIRS and EEG-based classifiers using a subject-specific meta-learner.

FNIRS-based classifiers consistently underperformed compared to the EEG-based classifiers. Contrasting the first-level classifiers with the results of the meta-learner showed decreased accuracy in the first-level fNIRS comparisons, and usually non-significant, as well as a few significant accuracy benefits for the EEG-based classifiers. In all but the high class-separability MATB contrasts, the first-level EEG-based classifiers exhibited higher classification accuracy than the meta-learner, suggesting that the meta-learner may have been impeded by the inclusion of the chance-level fNIRS classifiers. However, this likely pertained specifically to the fNIRS-time classifier, as the fNIRS-SCM classifier tended to be weighted highly in the meta-classifier, suggesting that its class-probability outputs provide some useful and independent information for mental workload detection. (see Figure 55 and Figure 58).

In the n-back, the results were most clearly in favour of the EEG-based classifiers. With the fNIRS-based first-level classifiers generally having been outperformed by the meta-learner by a wide margin in both high and low-class separability cases.

For the tested scenarios (two tasks; two class-separability contrasts; within- & cross-subject; full & sparse EEG montage), it can be said that fNIRS classification of mental workload benefited greatly from additional EEG information, even with a sparse 7-electrode montage. Conversely, the conclusion is less definitive, but at large, fNIRs did not seem to markedly improve the classification performance beyond that already achieved with EEG-based classifiers alone. Only in the high-separability MATB case did the average meta-learner accuracies rise significantly above those of the first-level EEG classifiers.

Figure 61. Per-Subject Variability

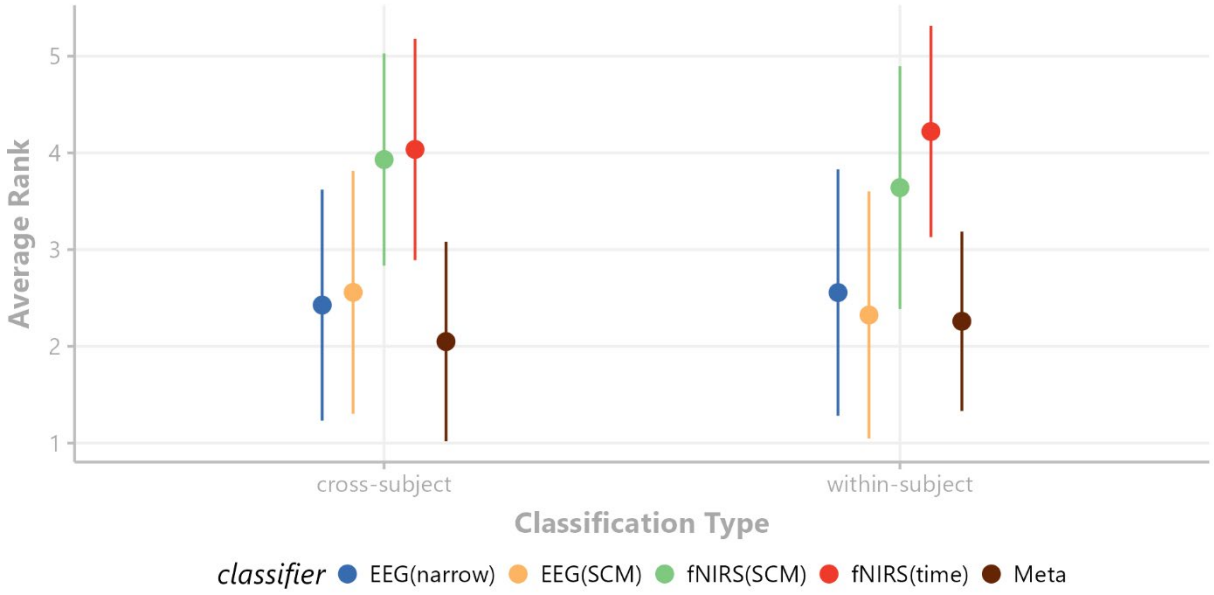


Note. Individual results per classifier per subjects (N=14) for the cross-subject classification of the high class-separability scenario for A) the MATB and B) the n-back. Accuracies that were below sample-size corrected chance levels are marked with a white X.

However, the probably most useful property of a stacked ensemble for generalisable pBCI architectures may not stem from increasing accuracy across all subjects, but rather from providing robust classification for multiple subjects using a single classification strategy. When inspecting the individual classification results by subject (Figure 61), it becomes clear that no single modality assured good performance across subjects. Inter-subject variability may result in some features performing well for one individual but not another. This can be countered by stacking the outputs of multiple classifiers and letting another classifier, tuned to the specific subject (the meta learner), optimally weigh the individual classifiers. Optimal here refers to as best as possible given the available data – limited training data or non-stationarities causing covariate shifts between train and test data may still be an issue in cases where the meta-learner and its respective first-level classifiers are not retrained or recalibrated over time.

The meta-learner tested here likely did not fully capitalise on the flexible nature of such a stacked ensemble approach, as a combination of four classifiers may have not provided enough diversity in the class-probability estimates. However, per classification scenario, the meta-learner still tended to be ranked the highest on average (see Figure 62), with a Friedman test ($Chi(4) = 44.758, p < .001$), followed by Durbin-Conover pair-wise tests, showing significant advantages in ranks (Meta vs. EEG-SCM: $p = .007$, EEG-narrow: $p = .008$, fNIRS-SCM: $p < .001$, EEG-SCM: $p < .001$). This matches the previous conclusion that, while accuracy may on average be reduced, stacking classifiers offers more robust classification across datasets (e.g. subjects) and contexts (e.g. tasks, montages). The approach tested here could likely benefit from increasing classifier diversity by adding further feature categories, for example, further fNIRS-based kernel matrices (Näher et al., 2024), more specific narrow-band EEG-SCM approaches like those used in Chapters 4 and 5, as well as information-theoretical features, such as different entropy and complexity metrics (Angsuwatanakul et al., 2020; Cao et al., 2017; Liu et al., 2019), which were not utilised in this thesis.

Figure 62. Average Classifier Rankings



Note. Average ranks per classification type of the five different classifiers. Classifiers were ranked within-subject per task, montage, and class-separability contrast. Error bars show the standard deviation of the ranks.

Increasing the diversity of the ensemble may offer additional benefits in average accuracy and would further bolster the generalisability of the approach to unseen subjects. A more exhaustive ensemble, based on additional feature categories and modalities, could also allow for estimating the confidence of the ensemble estimate. Cases in which the individual classifiers agree would be deemed highly confident predictions, whereas cases with heterogeneous decisions across individual classifiers would be deemed unreliable. This could be used as an additional control signal in neuroadaptive interfaces.

In conclusion, as implemented here, fNIRS could not provide the hoped-for accuracy improvements in the MATB and n-back low class-separability cases. If this is an inherent shortcoming of the modality, or if it was montage specific or even caused by the design choices in our experiment, cannot be said with certainty. The concept of stacking multiple pBCIs may, however, be an effective way to design robust pBCIs that, combined with further domain adaptation methods, could improve generalisability across aspects like the type of task than any single classifier would.

7 General Discussion

The final chapter of the thesis aims to summarise its main findings, assess how well the project's aims have been achieved, and relate its contents to the existing literature as well as future developments.

7.1 Main Findings

This thesis set out to contribute to the development of neuroadaptive interfaces by A) comparing a range of metrics and classification algorithms across lab-grade and wearable EEG headsets, B) investigating the value of adding continuous task-irrelevant probes to common mental workload paradigms, and C) test whether different ensemble approaches can improve mental workload classification in within and cross-subject scenarios. Lastly, (D) various elements of the thesis aimed to support reproducibility in the field of pBCI research by exploring the consistency of common metrics across independent samples, measuring possible biases in pBCI model evaluation and comparison methods, and curating a set of datasets with varying sensor montages that helps to fill the gap of open access data for research on mental workload and will provide the community with more ways to validate novel methods.

7.1.1 Wearable vs Lab-grade EEG

The starting point of this thesis was the need for a more thorough validation of commonly reported pBCI approaches using wearable EEG sensors. Wearable EEG devices have become cheaper and more numerous on the market (Niso et al., 2023) and could potentially offer more user-friendly (i.e. less intrusive, thereby increasing operator acceptability) solutions for applied pBCI systems. Chapter 5 contained a thorough comparison of the wet-electrode 7-channel system used in this project with a 64-channel gel-based system. Overall, accuracy estimates across the two datasets were mostly comparable. Only in the case of the MATB did the 64-channel system show clear accuracy advantages over the wearable system.

Two approaches were tested to address this performance gap. Feature extraction windows were increased to reduce the influence of noisy data periods, and a temporal ensemble strategy was additionally utilised to further combat unstable predictions over sequential extraction windows. By increasing the length of the feature extraction windows and aggregating predictions, differences between the lab-grade and wearable systems could be alleviated for most of the tested feature extraction methods, in both the within- and cross-subject classification. The resulting “sweet-spots” between the system’s responsiveness and accuracy tended to range from 20-40 seconds for different

feature categories. These results provided important insights into the operator acceptability of neuroadaptive systems. The analysis demonstrated that aggregating features over these larger windows could decrease misclassifications, which could mitigate the risk of pBCI informaed system adaptations degrading performance—a likely unacceptable risk in safety-critical contexts.

Of further interest for applied settings were the limited sensitivity to detect subtle differences in mental workload for either montage. Lower confidence interval bounds of neither the lab-grade nor the wearable headset surpassed 60% for the low-class separability contrast of the n-back. As both performance and subjective metrics indicated task-load related differences, applied systems based on these and comparable methods may struggle to detect when operators experience more nuanced changes in mental workload.

7.1.2 Task-irrelevant probes

A secondary aim of this thesis was to test the potential of using task-irrelevant probes for continuous workload monitoring. Task-irrelevant probes had previously been shown to evoke ERPs containing mental workload-related information in various tasks and probe designs (Allison & Polich, 2008; Dyke et al., 2015; Ladouce et al., 2025; Roy et al., 2016). The use of task-irrelevant probes has appeal for applied workload monitoring, as it may offer neurophysiological metrics independent of task and environmental contexts (Papanicolaou & Johnstone, 1984). Here, two sensory probing paradigms that did not require additional input or attention from the operator were selected: a visual probe utilising a low-contrast 15Hz flicker and an auditory probe, presented at very fast inter-stimulus intervals to allow for nearly continuous workload assessment.

The auditory probes had previously been tested in a simulated driving experiment, where several ERP components exhibited promising effect sizes to differentiate between fast and slow driving conditions at the population level (Sugimoto et al., 2022). The results gleaned from the population-level tests presented in Chapter 3 were unexpected, as they A) did not exhibit consistent effects across tasks and datasets, and B) did not exhibit the same effect directions that the original authors reported. P2 amplitude, as well as a late potential around 400-500ms, exhibited increased amplitudes with increasing workload in the Lab-grade dataset. For the Wearable dataset, only the late potential exhibited significant differences in MATB alone, and these differences were in the opposite direction to those observed in the Lab-grade dataset. Originally expected, however, were N1 and P2 amplitude reductions with increasing workload. The conceptual replication in Chapter 3 further suggested limited sensitivity as the medium task-load levels did not exhibit significant differences from the easy levels (with a single exception). The auditory stress operators may experience with this constant stimulation may not be justified, given these results.

For the visual probe, population-level results indicated widespread SNR reductions in the MATB across all task-load contrasts. A likely explanation for these strong effects could lie in changes to ocular dynamics with increasing MATB levels. Nonetheless, due to the ease with which SSVEP information can be extracted from EEG, these results may offer a convenient alternative to more processing-heavy eye-tracking approaches to mental workload monitoring. In the n-back the effects were not as strong. While sensor-space tests suggested SNR decreases on the left hemisphere, single-trial analysis of spatially filtered SSVEP dynamics suggested opposing SNR increases with increasing n-back loads. Although task-load effects were present, the linear-mixed effects models fitted to trial-wise data required different random effect structures and struggled with very high standard errors, suggesting that the SSVEP did not change homogeneously with task-load increases. Rather, trial-wise dynamics were likely more complex, and the task-related parameters entered into the model did not suffice to capture this variance. A more sophisticated signal processing or modelling strategy may, however, reveal response-related effects, as the current models showed relatively high F-statistics for the factor “Correct”, which distinguished between correct and incorrect responses.

For either probe, the current thesis could not offer conclusive evidence in favour of their usefulness for mental workload monitoring. The task-irrelevant rapid-probe paradigm would likely benefit from introducing ideas of classic oddballs back into the design, as the SNR of reorienting related ERPs would likely be advantageous for the single-trial classification and offer higher power for population-level (Dyke et al., 2015; Ke et al., 2021). The SSVEP paradigm yielded interesting results that require further study. Contrast settings, as well as stimulation frequencies, may produce different dynamics that could potentially offer highly convenient mental workload monitoring features. It may be that lower contrast settings, approaching periliminal levels, offer better insights into the waxing and waning of task focus, as a recent study reported robust SNR decreases preceding missed events in a sustained attention task (Ladouce et al., 2025). The contrast chosen for the current thesis was selected to balance user-comfort and SNR, which may have caused the SNR to spike too quickly, whereas lower contrast settings would likely require prolonged fixation periods to reach higher SNR values (Ladouce et al., 2022).

7.1.3 Multimodal Workload Monitoring

In Chapter 6, the focus was on the Multimodal dataset. While EEG was the primary modality throughout this thesis, fNIRS as an additional independent source of information was theorised to improve upon the EEG classification by using a stacked ensemble with decision-level fusion. To this end, the EEG montage was reduced to a lower 32-channel count to accommodate an additional

frontoparietal fNIRS montage. The analysis of the ensemble weights did suggest a sizable influence of the fNIRS data on the classification results. However, the accuracy in the low-separability cases did not rise substantially over those previously reported with EEG classifiers alone.

While the EEG classifiers often did not differ significantly from the meta-learner's classification accuracies, an analysis of ranks within subjects suggested that the meta-learner tended to perform more consistently at reasonably high accuracy levels. In contrast, the EEG-based classifiers would exhibit higher variability in performance across participants, leading to significantly lower average rankings in both within- and cross-subject classification. This underlined the benefits of drawing from multiple classifiers rather than relying on a single approach when developing generalisable mental workload monitoring systems. First-level classifier diversity could likely be improved upon by utilising a filter bank approach akin to the Narrow-Covariance approach presented in Chapters 4 and 5. Instead of combining the different frequency bands into a single covariance matrix, lightweight RMDM classifiers would be used as first-level classifiers for an ensemble approach such as the one presented in Chapter 6. Additionally, information-theoretical as well as connectivity-based features may offer a further source of diversity for such ensemble approaches in future research.

7.1.4 Reproducible pBCI Research

Chapter 3 aimed to not just replicate the results of the rapid task-irrelevant auditory probe paradigm, but also to interrogate the robustness of commonly cited band power effects in the mental workload literature. This was additionally supplemented by repeating said band power tests on aperiodic-free PSD estimates to validate that the effects of interest are of actual oscillatory (or burst) nature and not due to shifts in the aperiodic "background" activity of the EEG.

The results showed alpha related changes to be most consistent across datasets and tasks. However, alpha power also exhibited the strongest time-on-task effects, and these duration-related effects on alpha differed in their spatial distributions between the two tasks, offering insights into why pBCIs without adaptive components struggle to maintain high accuracy across time and across tasks. Furthermore, accounting for changes to the aperiodic activation changed effect directions in the alpha band, uncovering possible reasons for mixed results present in the mental workload literature (Borghini et al., 2014; Chikhi et al., 2022). Together, the inconsistencies, especially across tasks, render band power estimates somewhat unreliable for mental workload detection outside controlled laboratory conditions. However, the cross-subject classification results in Chapters 5 and 6 suggested that they nonetheless possessed predictive power for within-task mental workload monitoring.

Another effort to foster reproducible research in the pBCI community was put forward in Chapter 4, where the effects of underreported or inappropriately selected cross-validation methods were not just pointed out, but empirically demonstrated. The results could paint a clear picture that the block-structures common to mental workload experiments can induce serious bias into model comparisons based on offline cross-validation procedures and that models with more degrees of freedom are likely to capitalise on temporal trends to a higher degree than their simpler counterparts. Consequently, more complex models would outshine their simpler counterparts in such model comparisons, which, if such practices are not curbed by reviewers and stakeholders, could hinder progress by falsely guiding the field to ever more complex solutions.

The final contribution of this thesis to making pBCI research more reproducible concerns the data sharing upon project completion. All four datasets presented were collected with making them openly accessible in mind. Open-access datasets dealing with mental workload are scarce to non-existent, with only a few exceptions at this point in time (Hinss et al., 2021). Allowing researchers to validate their approaches on unseen data should bolster the validity of novel approaches. Furthermore, by varying the montage designs across the datasets, they may offer a diverse test bed for future research.

7.2 Neuroadaptive interfaces – a wider scope

Already in the 1980s, O'Donnell and Eggemeier defined six criteria to measure the utility of a mental workload metric (O'Donnell & Eggemeier, 1986). This thesis was mainly concerned with the criteria of sensitivity, intrusiveness, implementation requirements, and operator acceptability. Hence, focus lay with comparing the sensitivity between wearable and lab-grade EEG headsets, as reducing the intrusive nature of sensing devices represents an important step towards improving operator acceptability. Across tasks and mental workload contrasts, the wearable system appeared to match the sensitivity of the lab-grade system. Furthermore, for applied scenarios, both ensemble methods offered promising solutions to reducing misclassifications. A caveat of the here within presented results concerned the sensitivity to subtle differences in mental workload (related to the performance plateau from de Waard's (1996) figure) Even though the metrics that were tested within this thesis were only assessed within a single task - a likely more achievable goal for a pBCI than the creation of a task-agnostic monitoring system - accuracy dropped sharply when more subtle differences in task-load needed to be distinguished. These sharp drop-offs were visible for the traditional PSD-based classification but also for the state-of-the-art Riemannian classifiers. Such classification errors could not just harbour potential for creating safety risks - the opposite of what neuroadaptive systems set out to do - but also severely lower the operator acceptability, regardless

of the form factor and other convenience aspects the physical design of a wearable EEG system could offer.

7.2.1 Consequences for the Workplace

Importantly, even a hypothetically infallible system would likely face hurdles for acceptance by the broader population. So far, only briefly hinted at in the beginning of Chapter 1, the societal implications of neuroadaptive systems should not be ignored when researching a technology that ultimately should find application in our everyday lives. While the initial motivations of reducing risks or improving work satisfaction by reducing time spent in over- or underloaded states may be worthwhile, the same technology could fuel an ever-growing trend of algorithmic control, or "algorithmic management", in the workplace.

Algorithmic control in its current form allows employers to exert novel forms of control over their workforce. The emerging literature revolving around algorithmic control initially focused on gig-workers at food and grocery delivery services such as Uber and Instacart, where braking and steering behaviour was, for example, automatically processed to recommend drivers to take breaks in case of erratic driving behaviour (Rosenblat & Stark, 2016). Today, a more exhaustive taxonomy of algorithmic control systems showcases their wide usage across types of workplaces, from gig-workers to office-workers (Alizadeh et al., 2023). According to Kellog et al. (2020), employers possess six main mechanisms (six R's) to direct workers using such algorithms. Workers may be directed by *restricting* and *recommending* actions, they may be evaluated by *recording* and *rating*, and they may be disciplined by *replacing* and *rewarding*. Recent reviews all call for further research and exploration of the various interlinked phenomena related to algorithmic control in order to mitigate negative outcomes for workers (Alizadeh et al., 2023; Dutta et al., 2024; Kellogg et al., 2020). However, as of now, neuroadaptive technologies seem not to be included in any of these discussions. At the same time, issues related to algorithmic control tend not to be addressed in the neuroadaptive literature. With an increasingly well-funded neuro-tech industry (NeuroX [Uk], Hird [Denmark], Optohive [Switzerland], Bitbrain [Spain], Kernel [USA], Neurable [USA], Zanderlabs [Germany], and many more), an integration of neuroadaptive technology into the discussion of algorithmic control may be warranted

Already somewhat more established are emotion recognition technologies, which operate on text, voice recordings, facial photography, or video. The European Union's 2024 AI Act aimed to protect users from adverse effects of AI systems (Council of the European Union, 2023) such as biased and flawed emotion recognition technologies. Current emotion recognition technologies were deemed

high risk, and serious concerns about the reliability, lack of specificity, and limited generalizability of the current state of the art solutions were cited for its regulation in the workplace (Prégent, 2025).

Judging by the results of this thesis, as well as the 2022 pBCI competition (Roy et al., 2022), mental workload monitoring may face similar concerns, as the accuracy of mental state classifiers generally declines with increasing ecological validity. Furthermore, premature or overly controlling implementation of open or closed-loop pBCIs could potentially even damage future operator acceptance. Misaligned support systems that were initially supposed to improve safety and well-being, like tachometers in lorries, have previously been reported to be tampered with as their directions may be difficult to comply with, given delivery pressures. Furthermore, their information may be used to inform disciplinary actions (Kave, 2017), rendering them an adversary rather than a support system to the employees.

7.3 Limitations

Several limitations need to be qualified to contextualise the results of this thesis. While sample sizes around or below 20 participants are commonplace in the pBCI literature (Demirezen et al., 2024), the individual datasets of this thesis only consisting of 20 participants may complicate a conclusive comparison between them, as sample-dependent dynamics may obfuscate smaller effects at such small participant numbers. This issue became especially apparent in Chapter 3, where many effects were visible in the data but failed to meet significance due to high between-subject variance.

Secondly, the machine learning results presented in chapters 4 – 6 were all task-specific and would likely not generalise across tasks. The two tasks chosen for the four datasets were quite different in the nature of their task-load manipulations, and while a common neurophysiological marker of mental workload may theoretically exist, generalising multivariate patterns between these two tasks likely poses a considerable challenge, as the differing patterns in Chapter 3 alluded to. However, some previous work did successfully test classifiers within-subject across MATB and n-back data (Ke et al., 2021) using complex sounds with slow presentation rates as their task-irrelevant probing technique. The downside of testing mental workload classification within a task is that motor confounds may be highly predictive of the task-load condition (Brouwer et al., 2015; Lemm et al., 2011), such as increased eye or joystick movements in the hard MATB tasks or adapted blink rates in the 3-back task.

Thirdly, as explored in chapter 4, the block length of 5 minutes for the MATB and 2.5 minutes for the n-back was probably unnecessarily long. Long block durations could increase the likelihood of effort withdrawal if the task becomes too difficult. Furthermore, particularly in the MATB, the randomised

generation of scenarios may have led to uneven event distributions and caused some scenarios to consist of high and low-load periods. Either factor could have led to a less stable “ground truth” of mental workload than the static labels suggested to the classification algorithms, resulting in unnecessary variance during training and testing. Shorter blocks would have also enabled more condition repetitions, allowing for additional cross-validation splits and potentially reducing the risk of overfitting to block-specific temporal trends in the within-subject classification analyses (White & Power, 2023).

Lastly, the inclusion of the visual and auditory probes throughout all of the Lab-grade and Wearable datasets required slightly atypical frequency ranges for the feature extraction techniques informed by canonical band power metrics (e.g., starting the beta band at 17Hz to avoid the 15Hz flicker). Their inclusion also means that the datasets will need to be used with great care to avoid accidentally capitalising on the SSVEP’s highly informative nature in the MATB task, which will have to be clearly emphasised when sharing the data publicly.

7.4 Future Research

The datasets collected for this thesis still hold considerable potential for further research. For example, mental workload classification across tasks and, less commonly studied, across EEG montages will be possible using the 80-participant strong dataset collection. Totalling around 60 hours of MATB and 28 hours of n-back recordings, the data may also be useful for evaluating deep-learning methods for mental workload detection. However, the ultimate goal of composing this collection of MATB and n-back data was to offer validation data other researcher can use to report machine learning results above and beyond their own data collection efforts. Reproducibility in pBCI research could be improved if researchers could compare their most recent discoveries using convenient and transparent, publicly available datasets. The MATB and n-back data of the 2022 BCI competition already offered a valuable testbed for cross-session classification. This thesis’s data will add to a now hopefully growing corpus of open-access mental workload data and could motivate novel cross-subject, and more interestingly, cross-montage classification research.

To extend the topics already covered in this thesis, future research should deepen the inquiry into cross-subject mental workload classification. Sources of inter-individual differences require further research, as robust classification methods which can handle such variations will be needed for the successful integration of pBCIs into applied settings. An interesting candidate that was not explored in greater depth in this thesis was the domain matching step in chapters 5 and 6. Using a test-subject’s calibration data to select only those previous subjects’ data whose feature distributions best match those of the test-subject sounds sensible in theory. Still, it requires further research to

demonstrate its merits. Additionally, due to covariance shifts, it may be required to continuously check for domain mismatches and update the classifier accordingly. The data provided in this thesis could potentially offer a sort of library to draw from in such an adaptive cross-subject approach. Subjects with highly matching covariance patterns could be grouped together to form a battery of classifiers, which could then be adaptively weighted to best match the target subject's patterns.

Extending on the meta-learner results, future research should also attempt to study how to maximise the diversity of the first-level classifiers to boost performance. Results presented in this thesis suggested that a very simple ensemble architecture can yield more consistent cross-subject classification accuracies. Classifier diversity may be increased by exploring more varied feature extraction methods, such as information-theoretically informed complexity measures, connectivity metrics, or methods that boost the SNR of the task-irrelevant probing technique. The addition of further modalities, such as eye-tracking or galvanic skin responses, may also help increase the first-level classifier diversity and, with that, increase the accuracy of a meta-learner approach. The subject-specific meta-learner presented in Chapter 6 could already show that cross-subject classification does not necessarily incur great accuracy decrements and cross-subject results could likely be included in analyses that may initially only plan on reporting within-subject results.

Ultimately, the real test for pBCIs lies in their integration into neuroadaptive technologies. True performance estimates will lie in their potential to improve task performance and safety, not in their raw prediction accuracies. With the results demonstrated in this thesis, it is likely that such applied systems will have to rely on temporal evidence accumulation to avoid large numbers of false positives/negatives. Further logic, such as user models, heuristic decision rules based on class probabilities of multiple classifiers and the integration of contextual information about the system in question, could likely already offer valuable information for system adaptation in office and training-related contexts. However, rather than aiming solely for higher classification accuracies, future research will need to test their systems in applied contexts where operators/users can be studied directly, to ascertain the concrete benefits of implementing pBCIs in real-world settings.

7.5 Conclusion

Passive Brain-Computer Interfaces represent an emerging technology that may allow for novel human-computer interaction paradigms. Convenient wearable systems that users can likely apply themselves, such as the x.On, paired with robust meta-learner approaches, already offer high prediction accuracies when it comes to differentiating between reasonable and very high subjective states of mental workload. For more subtle mental workload distinctions, the here-tested classification approaches did not yield comparable accuracy, suggesting that currently state-of-the-

art pBCI solutions may not offer sufficient sensitivity for monitoring mental workload at a high resolution. Their use may instead be limited to warning systems, in which particularly high (e.g., 'overloaded') states can be detected.

In order to further their development, it will be paramount to normalise the public sharing of data so that more thorough meta-analysis can be carried out. These meta-analyses will be able to offer essential ground-truths for mental state detection, which were so far difficult to attain due to the idiosyncratic nature of a field in which research is focused on novelty rather than robustness. Similar to the benchmarking efforts for active and reactive BCI paradigms, pBCI research could greatly benefit from such large-scale analyses, as they may not only provide insights into which models offer the highest prediction accuracy but also inform how inter-individual differences in the expression of mental states can be dealt with for the development of generalisable classifiers.

To conclude, the hardware and classification methods for applied pBCIs already exist. Without extensive preprocessing, wide differences in mental workload can already be detected with reasonable accuracy using minimal operator-specific calibration data. Future research on pBCI should thus focus on how to apply these existing methods to yield tangible benefits for employees, rather than solely developing new machine learning pipelines that result in minor accuracy increases in lab-based experiments. The development of such methods is certainly also important, but should hopefully be validated on public datasets rather than small and privately held samples. This would foster reproducible research standards and allow practitioners to easily compare and select methods for their specific applications.

It is also these applications that require much more research at this point in time. Support systems in aviation and heavy machinery are frequently cited in the mental workload literature, where pBCIs could adaptively modify warning indicators or implement automation solutions based on the operator's mental workload. The safety-critical nature of these applications explains why the pBCI literature concentrates on them, but it can also act as a barrier to more applied research, since deploying research systems in these contexts is highly complex. A less safety-critical and thus possibly risk-free application for mental workload monitoring may also lie in supporting skill acquisition in varied contexts. A functional mental workload monitoring system could record data during skill acquisition or fact learning. Rather than analysing heaps of sensor-level data, the continuous outputs of the classifier could subsequently be analysed to design adaptive learning strategies for knowledge retention, as already done using behavioural data alone (Sense & van Rijn, 2022). Regardless of the application context, what will be important will be the evaluation by the operators themselves, as their acceptance rates will ultimately drive the technology's adoption in the

real world. Exhaustive surveys, performance evaluations, and possibly even qualitative interviews will be required to truly assess the benefits of pBCIs for their users. Hence, increasing the focus on the operator rather than the neurophysiological data will likely mark an essential next step that will bring neuroadaptive out of the laboratory and into the workplace.

8 References

Aasted, C. M., Yücel, M. A., Cooper, R. J., Dubb, J., Tsuzuki, D., Becerra, L., Petkov, M. P., Borsook, D., Dan, I., & Boas, D. A. (2015). Anatomical guidance for functional near-infrared spectroscopy: AtlasViewer tutorial. *Neurophotonics*, *2*(2), 020801. https://doi.org/10.1117/1.NPh.2.2.020801

Aeschbach, D., Matthews, J. R., Postolache, T. T., Jackson, M. A., Giesen, H. A., & Wehr, T. A. (1997). Dynamics of the human EEG during prolonged wakefulness: Evidence for frequency-specific circadian and homeostatic influences. *Neuroscience Letters*, *239*(2), 121–124. https://doi.org/10.1016/S0304-3940(97)00904-X

Afzal, U., Prouzeau, A., Lawrence, L., Dwyer, T., Bichinepally, S., Liebman, A., & Goodwin, S. (2022). Investigating Cognitive Load in Energy Network Control Rooms: Recommendations for Future Designs. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.812677

Ahlstrom, U., & Friedman-Berg, F. J. (2006). Using eye movement activity as a correlate of cognitive workload. *International Journal of Industrial Ergonomics*, *36*(7), 623–636. https://doi.org/10.1016/j.ergon.2006.04.002

alexandre Barachant, Quentin Barthélemy, Gabriel Wagner vom Berg, Alexandre Gramfort, Jean-Rémi KING, Pedro L. C. Rodrigues, Bruna Junqueira Lopes, Dave, Emanuele Olivetti, Vladislav Goncharenko, gcattan, maxdolle, Maria Sayu Yamamoto, GhilesReguig, Apolline Mellot, toncho11, stonebig, Steven Mortier, mhurte, … Ben Beasley. (2025). *pyRiemann/pyRiemann: V0.8* (Version v0.8) [Computer software]. Zenodo. https://doi.org/10.5281/ZENODO.593816

Alizadeh, A., Hirsch, F., Jiang, J., Wiener, M., & Benlian, A. (2023). A Taxonomy of Algorithmic Control Systems. *ICIS 2023 Proceedings*. https://aisel.aisnet.org/icis2023/techandfow/techandfow/8

Allison, B. Z., & Polich, J. (2008). Workload assessment of computer gaming using a single-stimulus event-related potential paradigm. *Biological Psychology*, *77*(3), 277–283. https://doi.org/10.1016/j.biopsycho.2007.10.014

Al-Shargie, F., Kiguchi, M., Badruddin, N., Dass, S. C., Hani, A. F. M., & Tang, T. B. (2016). Mental stress assessment using simultaneous measurement of EEG and fNIRS. *Biomedical Optics Express*, *7*(10), 3882–3898. https://doi.org/10.1364/BOE.7.003882

Ang, K. K., Chin, Z. Y., Wang, C., Guan, C., & Zhang, H. (2012). Filter Bank Common Spatial Pattern Algorithm on BCI Competition IV Datasets 2a and 2b. *Frontiers in Neuroscience*, *6*. https://doi.org/10.3389/fnins.2012.00039

Angsuwatanakul, T., O'Reilly, J., Ounjai, K., Kaewkamnerdpong, B., & Iramina, K. (2020). Multiscale Entropy as a New Feature for EEG and fNIRS Analysis. *Entropy*, *22*(2), Article 2. https://doi.org/10.3390/e22020189

Aricò, P., Borghini, G., Di Flumeri, G., Sciaraffa, N., & Babiloni, F. (2018). Passive BCI beyond the lab: Current trends and future directions. *Physiological Measurement*, *39*(8), 08TR02. https://doi.org/10.1088/1361-6579/aad57e

Ayaz, H., Shewokis, P. A., Bunce, S., Izzetoglu, K., Willems, B., & Onaral, B. (2012). Optical brain monitoring for operator training and mental workload assessment. *NeuroImage*, *59*(1), 36–47. https://doi.org/10.1016/j.neuroimage.2011.06.023

Bai, O., Lin, P., Vorbach, S., Floeter, M. K., Hattori, N., & Hallett, M. (2007). A high performance sensorimotor beta rhythm-based brain–computer interface associated with human natural motor behavior. *Journal of Neural Engineering*, *5*(1), 24. https://doi.org/10.1088/1741-2560/5/1/003

Baldwin, C. L., & Penaranda, B. N. (2012). Adaptive training using an artificial neural network and EEG metrics for within- and cross-task workload classification. *NeuroImage*, *59*(1), 48–56. https://doi.org/10.1016/j.neuroimage.2011.07.047

Balfe, N., Crowley, K., Smith, B., & Longo, L. (2017). Estimation of Train Driver Workload: Extracting Taskload Measures from On-Train-Data-Recorders. In L. Longo & M. C. Leva (Eds.), *Human Mental Workload: Models and Applications* (pp. 106–119). Springer International Publishing. https://doi.org/10.1007/978-3-319-61061-0_7

Barachant, A., & Bonnet, S. (2011). Channel Selection Procedure using Riemannian distance for BCI applications. *International IEEE EMBS Conference on Neural Engineering*, TBA. https://hal.science/hal-00602707

Barachant, A., Bonnet, S., Congedo, M., & Jutten, C. (2010). Riemannian Geometry Applied to BCI Classification. In V. Vigneron, V. Zarzoso, E. Moreau, R. Gribonval, & E. Vincent (Eds.), *Latent Variable Analysis and Signal Separation* (pp. 629–636). Springer. https://doi.org/10.1007/978-3-642-15995-4_78

Barachant, A., & Congedo, M. (2014). *A Plug&Play P300 BCI Using Information Geometry* (No. arXiv:1409.0107). arXiv. https://doi.org/10.48550/arXiv.1409.0107

Barnova, K., Mikolasova, M., Kahankova, R. V., Jaros, R., Kawala-Sterniuk, A., Snasel, V., Mirjalili, S., Pelc, M., & Martinek, R. (2023). Implementation of artificial intelligence and machine learning-based methods in brain–computer interaction. *Computers in Biology and Medicine*, *163*, 107135. https://doi.org/10.1016/j.compbiomed.2023.107135

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 10.1016/j.jml.2012.11.001. https://doi.org/10.1016/j.jml.2012.11.001

Barthelemy, Q., Mayaud, L., Ojeda, D., & Congedo, M. (2019). The Riemannian Potato Field: A Tool for Online Signal Quality Index of EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering: A Publication of the IEEE Engineering in Medicine and Biology Society*, *27*(2), 244–255. https://doi.org/10.1109/TNSRE.2019.2893113

Bassett, D. S., & Gazzaniga, M. S. (2011). Understanding complexity in the human brain. *Trends in Cognitive Sciences*, *15*(5), 200–209. https://doi.org/10.1016/j.tics.2011.03.006

Bauer, L. O., Goldstein, R., & Stern, J. A. (1987). Effects of Information-Processing Demands on Physiological Response Patterns. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *29*(2), 213–234. https://doi.org/10.1177/001872088702900208

Benedetto, S., Pedrotti, M., Minin, L., Baccino, T., Re, A., & Montanari, R. (2011). Driver workload and eye blink duration. *Transportation Research Part F: Traffic Psychology and Behaviour*, *14*(3), 199–208. https://doi.org/10.1016/j.trf.2010.12.001

Benediktsson, J. A., & Swain, P. H. (1992). Consensus theoretic classification methods. *IEEE Transactions on Systems, Man, and Cybernetics*, *22*(4), 688–704. https://doi.org/10.1109/21.156582

Beniczky, S., & Schomer, D. L. (2020). Electroencephalography: Basic biophysical and technological aspects important for clinical applications. *Epileptic Disorders*, *22*(6), 697–715. https://doi.org/10.1684/epd.2020.1217

Blankertz, B., Tomioka, R., Lemm, S., Kawanabe, M., & Muller, K. (2008). Optimizing Spatial filters for Robust EEG Single-Trial Analysis. *IEEE Signal Processing Magazine*, *25*(1), 41–56. https://doi.org/10.1109/MSP.2008.4408441

Bliss, J. P. (2003). Investigation of Alarm-Related Accidents and Incidents in Aviation. *The International Journal of Aviation Psychology*, *13*(3), 249–268. https://doi.org/10.1207/S15327108IJAP1303_04

Bloch, I. (1996). Information combination operators for data fusion: A comparative review with classification. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, *26*(1), 52–67. https://doi.org/10.1109/3468.477860

Blum, S., Jacobsen, N. S. J., Bleichner, M. G., & Debener, S. (2019). A Riemannian Modification of Artifact Subspace Reconstruction for EEG Artifact Handling. *Frontiers in Human Neuroscience*, *13*. https://www.frontiersin.org/articles/10.3389/fnhum.2019.00141

Boere, K., Hecker, K., & Krigolson, O. E. (2024). Validation of a mobile fNIRS device for measuring working memory load in the prefrontal cortex. *International Journal of Psychophysiology*, *195*, 112275. https://doi.org/10.1016/j.ijpsycho.2023.112275

Bollmann, S., & Barth, M. (2021). New acquisition techniques and their prospects for the achievable resolution of fMRI. *Progress in Neurobiology*, *207*, 101936. https://doi.org/10.1016/j.pneurobio.2020.101936

Borghini, G., Astolfi, L., Vecchiato, G., Mattia, D., & Babiloni, F. (2014). Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. *Neuroscience & Biobehavioral Reviews*, *44*, 58–75. https://doi.org/10.1016/j.neubiorev.2012.10.003

Boustani, S. E., Marre, O., Béhuret, S., Baudot, P., Yger, P., Bal, T., Destexhe, A., & Frégnac, Y. (2009). Network-State Modulation of Power-Law Frequency-Scaling in Visual Cortical Neurons. *PLOS Computational Biology*, *5*(9), e1000519. https://doi.org/10.1371/journal.pcbi.1000519

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brickwedde, M., Anders, P., Kühn, A. A., Lofredi, R., Holtkamp, M., Kaindl, A. M., Grent-'t-Jong, T., Krüger, P., Sander, T., & Uhlhaas, P. J. (2024). Applications of OPM-MEG for translational neuroscience: A perspective. *Translational Psychiatry*, *14*(1), 341. https://doi.org/10.1038/s41398-024-03047-y

Brigadoi, S., & Cooper, R. J. (2015). How short is short? Optimum source–detector distance for short-separation channels in functional near-infrared spectroscopy. *Neurophotonics*, *2*(2), 025005. https://doi.org/10.1117/1.NPh.2.2.025005

Brookhuis, K. A., De Waard, D., & Fairclough, S. H. (2003). Criteria for driver impairment. *Ergonomics*, *46*(5), 433–445. https://doi.org/10.1080/001401302/1000039556

Brouwer, A.-M., Hogervorst, M. A., Erp, J. B. F. van, Heffelaar, T., Zimmerman, P. H., & Oostenveld, R. (2012). Estimating workload using EEG spectral power and ERPs in the n-back task. *Journal of Neural Engineering*, *9*(4), 045008. https://doi.org/10.1088/1741-2560/9/4/045008

Brouwer, A.-M., Zander, T. O., van Erp, J. B. F., Korteling, J. E., & Bronkhorst, A. W. (2015). Using neurophysiological signals that reflect cognitive or affective state: Six recommendations to

avoid common pitfalls. *Frontiers in Neuroscience*, *9*. https://www.frontiersin.org/articles/10.3389/fnins.2015.00136

Buckner, R. L. (1998). Event-related fMRI and the hemodynamic response. *Human Brain Mapping*, *6*(5–6), 373–377. https://doi.org/10.1002/(SICI)1097-0193(1998)6:5/6&%2523x0003c;373::AID-HBM8&%2523x0003e;3.0.CO;2-P

Budd, T. W., Barry, R. J., Gordon, E., Rennie, C., & Michie, P. T. (1998). Decrement of the N1 auditory event-related potential with stimulus repetition: Habituation vs. refractoriness. *International Journal of Psychophysiology: Official Journal of the International Organization of Psychophysiology*, *31*(1), 51–68. https://doi.org/10.1016/s0167-8760(98)00040-3

Bullmore, E., Long, C., Suckling, J., Fadili, J., Calvert, G., Zelaya, F., Carpenter, T. A., & Brammer, M. (2001). Colored noise and computational inference in neurophysiological (fMRI) time series analysis: Resampling methods in time and wavelet domains. *Human Brain Mapping*, *12*(2), 61–78. https://doi.org/10.1002/1097-0193(200102)12:2%253C61::AID-HBM1004%253E3.0.CO;2-W

Buxton, R. B. (2012). Dynamic models of BOLD contrast. *NeuroImage*, *62*(2), 953–961. https://doi.org/10.1016/j.neuroimage.2012.01.012

Buzsáki, G., Anastassiou, C. A., & Koch, C. (2012). The origin of extracellular fields and currents—EEG, ECoG, LFP and spikes. *Nature Reviews Neuroscience*, *13*(6), 407–420. https://doi.org/10.1038/nrn3241

Cao, Z., Prasad, M., & Lin, C.-T. (2017). Estimation of SSVEP-based EEG complexity using inherent fuzzy entropy. *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 1–5. https://doi.org/10.1109/FUZZ-IEEE.2017.8015730

Castillos, K., Cabrera, Ladouce, S., Darmet, L., & Dehais, F. (2023). Burst c-VEP Based BCI: Optimizing stimulus design for enhanced classification with minimal calibration data and improved user experience. *NeuroImage*, *284*, 120446. https://doi.org/10.1016/j.neuroimage.2023.120446

Cavanagh, J. F., & Frank, M. J. (2014). Frontal theta as a mechanism for cognitive control. *Trends in Cognitive Sciences*, *18*(8), 414–421. https://doi.org/10.1016/j.tics.2014.04.012

Cavanagh, J. F., Zambrano-Vazquez, L., & Allen, J. J. B. (2012). Theta lingua franca: A common mid-frontal substrate for action monitoring processes. *Psychophysiology*, *49*(2), 220–238. https://doi.org/10.1111/j.1469-8986.2011.01293.x

Chang, M. H., Baek, H. J., Lee, S. M., & Park, K. S. (2014). An amplitude-modulated visual stimulation for reducing eye fatigue in SSVEP-based brain-computer interfaces. *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology*, *125*(7), 1380–1391. https://doi.org/10.1016/j.clinph.2013.11.016

Charles, R. L., & Nixon, J. (2019). Measuring mental workload using physiological measures: A systematic review. *Applied Ergonomics*, *74*, 221–232. https://doi.org/10.1016/j.apergo.2018.08.028

Chen, W.-L., Wagner, J., Heugel, N., Sugar, J., Lee, Y.-W., Conant, L., Malloy, M., Heffernan, J., Quirk, B., Zinos, A., Beardsley, S. A., Prost, R., & Whelan, H. T. (2020). Functional Near-Infrared Spectroscopy and Its Clinical Application in the Field of Neuroscience: Advances and Future Directions. *Frontiers in Neuroscience*, *14*. https://doi.org/10.3389/fnins.2020.00724

Chen, Y., & Huang, X. (2016). Modulation of Alpha and Beta Oscillations during an n-back Task with Varying Temporal Memory Load. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.02031

Chenarboo, F. J., Hekmatshoar, R., & Fallahi, M. (2022). The influence of physical and mental workload on the safe behavior of employees in the automobile industry. *Heliyon*, *8*(10). https://doi.org/10.1016/j.heliyon.2022.e11034

Chevallier, S., Carrara, I., Aristimunha, B., Guetschel, P., Sedlar, S., Lopes, B., Velut, S., Khazem, S., & Moreau, T. (2024a). *The largest EEG-based BCI reproducibility study for open science: The MOABB benchmark* (No. arXiv:2404.15319). arXiv. https://doi.org/10.48550/arXiv.2404.15319

Chevallier, S., Carrara, I., Aristimunha, B., Guetschel, P., Sedlar, S., Lopes, B., Velut, S., Khazem, S., & Moreau, T. (2024b). *The largest EEG-based BCI reproducibility study for open science: The MOABB benchmark* (No. arXiv:2404.15319). arXiv. https://doi.org/10.48550/arXiv.2404.15319

Chikhi, S., Matton, N., & Blanchet, S. (2022). EEG power spectral measures of cognitive workload: A meta-analysis. *Psychophysiology*, *59*(6), e14009. https://doi.org/10.1111/psyp.14009

Chitnis, D., Cooper, R. J., Dempsey, L., Powell, S., Quaggia, S., Highton, D., Elwell, C., Hebden, J. C., & Everdell, N. L. (2016). Functional imaging of the human brain using a modular, fibre-less, high-density diffuse optical tomography system. *Biomedical Optics Express*, *7*(10), 4275–4288. https://doi.org/10.1364/BOE.7.004275

Cohen, M. X., & Gulbinaite, R. (2017). Rhythmic entrainment source separation: Optimizing analyses of neural responses to rhythmic sensory stimulation. *NeuroImage*, *147*, 43–56. https://doi.org/10.1016/j.neuroimage.2016.11.036

Colle, H. A., & Reid, G. B. (1998). Context Effects in Subjective Mental Workload Ratings. *Human Factors*, *40*(4), 591–600. https://doi.org/10.1518/001872098779649283

Comsa, I. M., Bekinschtein, T. A., & Chennu, S. (2019). Transient Topographical Dynamics of the Electroencephalogram Predict Brain Connectivity and Behavioural Responsiveness During Drowsiness. *Brain Topography*, *32*(2), 315–331. https://doi.org/10.1007/s10548-018-0689-9

Congedo, M., Barachant, A., & Andreev, A. (2013). *A New Generation of Brain-Computer Interface Based on Riemannian Geometry* [Research Report]. GIPSA-lab. https://hal.science/hal-00879050

Congedo, M., Barachant, A., & Bhatia, R. (2017). Riemannian geometry for EEG-based brain-computer interfaces; a primer and a review. *Brain-Computer Interfaces*, *4*(3), 155–174. https://doi.org/10.1080/2326263X.2017.1297192

Council of the European Union. (2023). *Artificial intelligence act: Council and Parliament strike a deal on the first rules for AI in the world*. Council of the European Union.

https://www.consilium.europa.eu/en/press/press-releases/2023/12/09/artificial-intelligence-act-council-and-parliament-strike-a-deal-on-the-first-worldwide-rules-for-ai/

Coyle, S. M., Ward, T. E., & Markham, C. M. (2007). Brain–computer interface using a simplified functional near-infrared spectroscopy system. *Journal of Neural Engineering*, *4*(3), 219. https://doi.org/10.1088/1741-2560/4/3/007

Critchley, H. D., & Garfinkel, S. N. (2018). The influence of physiological signals on cognition. *Current Opinion in Behavioral Sciences*, *19*, 13–18. https://doi.org/10.1016/j.cobeha.2017.08.014

Darvishi-Bayazi, M.-J., Law, A., Romero, S. M., Jennings, S., Rish, I., & Faubert, J. (2023). Beyond performance: The role of task demand, effort, and individual differences in ab initio pilots. *Scientific Reports*, *13*(1), 14035. https://doi.org/10.1038/s41598-023-41427-4

Davidson, M. J., Mithen, W., Hogendoorn, H., van Boxtel, J. J., & Tsuchiya, N. (2020). The SSVEP tracks attention, not consciousness, during perceptual filling-in. *eLife*, *9*, e60031. https://doi.org/10.7554/eLife.60031

De Cheveigné, A. (2020). ZapLine: A simple and effective method to remove power line artifacts. *NeuroImage*, *207*, 116356. https://doi.org/10.1016/j.neuroimage.2019.116356

De Rivecourt, M., Kuperus, M. N., Post, W. J., & Mulder, L. J. M. (2008). Cardiovascular and eye activity measures as indices for momentary changes in mental effort during simulated flight. *Ergonomics*, *51*(9), 1295–1319. https://doi.org/10.1080/00140130802120267

de Waard, D. (1996). *The measurement of drivers' mental workload* [PhD Thesis]. University of Groningen.

Deeny, S., Chicoine, C., Hargrove, L., Parrish, T., & Jayaraman, A. (2014). A Simple ERP Method for Quantitative Analysis of Cognitive Workload in Myoelectric Prosthesis Control and Human-Machine Interaction. *PLOS ONE*, *9*(11), e112091. https://doi.org/10.1371/journal.pone.0112091

Dehais, F., Cabrera Castillos, K., Ladouce, S., & Clisson, P. (2024). Leveraging textured flickers: A leap toward practical, visually comfortable, and high-performance dry EEG code-VEP BCI. *Journal of Neural Engineering*, *21*(6), 066023. https://doi.org/10.1088/1741-2552/ad8ef7

Dehais, F., Duprès, A., Blum, S., Drougard, N., Scannella, S., Roy, R. N., & Lotte, F. (2019). Monitoring Pilot's Mental Workload Using ERPs and Spectral Power with a Six-Dry-Electrode EEG System in Real Flight Conditions. *Sensors*, *19*(6), Article 6. https://doi.org/10.3390/s19061324

Dehais, F., Hodgetts, H. M., Causse, M., Behrend, J., Durantin, G., & Tremblay, S. (2019). Momentary lapse of control: A cognitive continuum approach to understanding and mitigating perseveration in human error. *Neuroscience & Biobehavioral Reviews*, *100*, 252–262. https://doi.org/10.1016/j.neubiorev.2019.03.006

Dehais, F., Ladouce, S., Darmet, L., Nong, T.-V., Ferraro, G., Torre Tresols, J., Velut, S., & Labedan, P. (2022). Dual Passive Reactive Brain-Computer Interface: A Novel Approach to Human-Machine Symbiosis. *Frontiers in Neuroergonomics*, *3*. https://www.frontiersin.org/articles/10.3389/fnrgo.2022.824780

Dehais, F., Lafont, A., Roy, R., & Fairclough, S. (2020). A Neuroergonomics Approach to Mental Workload, Engagement and Human Performance. *Frontiers in Neuroscience*, *14*. https://doi.org/10.3389/fnins.2020.00268

Dehais, F., Roy, R. N., Durantin, G., Gateau, T., & Callan, D. (2018). EEG-Engagement Index and Auditory Alarm Misperception: An Inattentional Deafness Study in Actual Flight Condition. In C. Baldwin (Ed.), *Advances in Neuroergonomics and Cognitive Engineering* (pp. 227–234). Springer International Publishing. https://doi.org/10.1007/978-3-319-60642-2_21

Deiber, M.-P., Missonnier, P., Bertrand, O., Gold, G., Fazio-Costa, L., Ibañez, V., & Giannakopoulos, P. (2007). Distinction between Perceptual and Attentional Processing in Working Memory Tasks: A Study of Phase-locked and Induced Oscillatory Brain Dynamics. *Journal of Cognitive Neuroscience*, *19*(1), 158–172. https://doi.org/10.1162/jocn.2007.19.1.158

Deligani, R. J., Borgheai, S. B., McLinden, J., & Shahriari, Y. (2021). Multimodal fusion of EEG-fNIRS: A mutual information-based hybrid classification framework. *Biomedical Optics Express*, *12*(3), 1635–1650. https://doi.org/10.1364/BOE.413666

Delorme, A., & Makeig, S. (2004). *EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics Journal of Neuroscience Methods 134:9-21 Includes details of EEGLAB ICA and time/frequency methods. Please cite this paper to reference EEGLAB in publications.* [Computer software].

Demirezen, G., Taşkaya Temizel, T., & Brouwer, A.-M. (2024). Reproducible machine learning research in mental workload classification using EEG. *Frontiers in Neuroergonomics*, *5*. https://doi.org/10.3389/fnrgo.2024.1346794

Dewiputri, W. I., & Auer, T. (2013). Functional magnetic resonance imaging (FMRI) neurofeedback: Implementations and applications. *The Malaysian Journal of Medical Sciences: MJMS*, *20*(5), 5–15.

Di Stasi, L. L., Antolí, A., Gea, M., & Cañas, J. J. (2011). A neuroergonomic approach to evaluating mental workload in hypermedia interactions. *International Journal of Industrial Ergonomics*, *41*(3), 298–304. https://doi.org/10.1016/j.ergon.2011.02.008

Di Stasi, L. L., Marchitto, M., Antolí, A., Baccino, T., & Cañas, J. J. (2010). Approximation of on-line mental workload index in ATC simulated multitasks. *Journal of Air Transport Management*, *16*(6), 330–333. https://doi.org/10.1016/j.jairtraman.2010.02.004

Dimigen, O. (2020). Optimizing the ICA-based removal of ocular EEG artifacts from free viewing experiments. *NeuroImage*, *207*, 116117. https://doi.org/10.1016/j.neuroimage.2019.116117

Domingos, P. (2012). A few useful things to know about machine learning. *Commun. ACM*, *55*(10), 78–87. https://doi.org/10.1145/2347736.2347755

Donoghue, T., Schaworonkow, N., & Voytek, B. (2022). Methodological considerations for studying neural oscillations. *European Journal of Neuroscience*, *55*(11–12), 3502–3527. https://doi.org/10.1111/ejn.15361

Dumoulin, S. O., Fracasso, A., Van Der Zwaag, W., Siero, J. C. W., & Petridou, N. (2018). Ultra-high field MRI: Advancing systems neuroscience towards mesoscopic human brain function. *NeuroImage*, *168*, 345–357. https://doi.org/10.1016/j.neuroimage.2017.01.028

Durantin, G., Dehais, F., Gonthier, N., Terzibas, C., & Callan, D. E. (2017). Neural signature of inattentional deafness. *Human Brain Mapping*, *38*(11), 5440–5455. https://doi.org/10.1002/hbm.23735

Durantin, G., Gagnon, J.-F., Tremblay, S., & Dehais, F. (2014). Using near infrared spectroscopy and heart rate variability to detect mental overload. *Behavioural Brain Research*, *259*, 16–23. https://doi.org/10.1016/j.bbr.2013.10.042

Dussault, C., Jouanin, J.-C., Philippe, M., & Guezennec, C.-Y. (2005). EEG and ECG changes during simulator operation reflect mental workload and vigilance. *Aviation, Space, and Environmental Medicine*, *76*(4), 344–351.

Dutta, S., Pramanik, H. S., Rajan, S. G., Rajan, R. G., & Satapathy, S. (2024). Proliferations in Algorithmic Control: Review of the Phenomenon and Its Implications. In S. K. Sharma, Y. K. Dwivedi, B. Metri, B. Lal, & A. Elbanna (Eds.), *Transfer, Diffusion and Adoption of Next-Generation Digital Technologies* (pp. 44–54). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-50188-3_5

Dyke, F. B., Leiker, A. M., Grand, K. F., Godwin, M. M., Thompson, A. G., Rietschel, J. C., McDonald, C. G., & Miller, M. W. (2015). The efficacy of auditory probes in indexing cognitive workload is dependent on stimulus complexity. *International Journal of Psychophysiology*, *95*(1), 56–62. https://doi.org/10.1016/j.ijpsycho.2014.12.008

Edell, D. J., Toi, V. V., McNeil, V. M., & Clark, L. D. (1992). Factors influencing the biocompatibility of insertable silicon microshafts in cerebral cortex. *IEEE Transactions on Biomedical Engineering*, *39*(6), 635–643. IEEE Transactions on Biomedical Engineering. https://doi.org/10.1109/10.141202

Eggemeier, F. T. (1988). Properties of Workload Assessment Techniques. In P. A. Hancock & N. Meshkati (Eds.), *Advances in Psychology* (Vol. 52, pp. 41–62). North-Holland. https://doi.org/10.1016/S0166-4115(08)62382-1

Ehinger, B. V., & Dimigen, O. (2019). Unfold: An integrated toolbox for overlap correction, non-linear modeling, and regression-based EEG analysis. *PeerJ*, *7*, e7838. https://doi.org/10.7717/peerj.7838

Elgendi, M., Jonkman, M., & De Boer, F. (2010). Frequency bands effects on QRS detection: 3rd International Conference on Bio-inspired Systems and Signal Processing, BIOSIGNALS 2010. *BIOSIGNALS 2010 - Proceedings of the 3rd International Conference on Bio-Inpsired Systems and Signal Processing*, *1*, 428–431.

Fairclough, S. H. (2009). Fundamentals of physiological computing. *Interacting with Computers*, *21*(1–2), 133–145. https://doi.org/10.1016/j.intcom.2008.10.011

Fairclough, S. H. (2022). Designing human-computer interaction with neuroadaptive technology. In *Current Research in Neuroadaptive Technology* (pp. 1–15). Elsevier. https://doi.org/10.1016/B978-0-12-821413-8.00006-3

Fairclough, S. H., & Lotte, F. (2020). Grand Challenges in Neurotechnology and System Neuroergonomics. *Frontiers in Neuroergonomics*, *1*. https://doi.org/10.3389/fnrgo.2020.602504

Fairclough, S. H., Stamp, K., & Dobbins, C. (2023). Functional connectivity across dorsal and ventral attention networks in response to task difficulty and experimental pain. *Neuroscience Letters*, *793*, 136967. https://doi.org/10.1016/j.neulet.2022.136967

Fairclough, S. H., Venables, L., & Tattersall, A. (2005). The influence of task demand and learning on the psychophysiological response. *International Journal of Psychophysiology*, *56*(2), 171–184. Scopus. https://doi.org/10.1016/j.ijpsycho.2004.11.003

Farwell, L. A., & Donchin, E. (1988). Talking off the top of your head: Toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and Clinical Neurophysiology*, *70*(6), 510–523. https://doi.org/10.1016/0013-4694(88)90149-6

Fawagreh, K., Gaber ,Mohamed Medhat, & and Elyan, E. (2014). Random forests: From early developments to recent advancements. *Systems Science & Control Engineering*, *2*(1), 602–609. https://doi.org/10.1080/21642583.2014.956265

Fazli, S., Mehnert, J., Steinbrink, J., Curio, G., Villringer, A., Müller, K.-R., & Blankertz, B. (2012). Enhanced performance by a hybrid NIRS–EEG brain computer interface. *NeuroImage*, *59*(1), 519–529. https://doi.org/10.1016/j.neuroimage.2011.07.084

Fernández-Palleiro, P., Rivera-Baltanás, T., Rodrigues-Amorim, D., Fernández-Gil, S., del Carmen Vallejo-Curto, M., Álvarez-Ariza, M., López, M., Rodriguez-Jamardo, C., Luis Benavente, J., de las Heras, E., Manuel Olivares, J., & Spuch, C. (2020). Brainwaves Oscillations as a Potential Biomarker for Major Depression Disorder Risk. *Clinical EEG and Neuroscience*, *51*(1), 3–9. https://doi.org/10.1177/1550059419876807

Ferrari, M., Mottola, L., & Quaresima, V. (2004). Principles, Techniques, and Limitations of Near Infrared Spectroscopy. *Canadian Journal of Applied Physiology*, *29*(4), 463–487. https://doi.org/10.1139/h04-031

Ferrari, M., & Quaresima, V. (2012). A brief review on the history of human functional near-infrared spectroscopy (fNIRS) development and fields of application. *NeuroImage*, *63*(2), 921–935. https://doi.org/10.1016/j.neuroimage.2012.03.049

Fife, D. A., & D'Onofrio, J. (2023). Common, uncommon, and novel applications of random forest in psychological research. *Behavior Research Methods*, *55*(5), 2447–2466. https://doi.org/10.3758/s13428-022-01901-9

Fishburn, F. A., Norr, M. E., Medvedev, A. V., & Vaidya, C. J. (2014). Sensitivity of fNIRS to cognitive state and load. *Frontiers in Human Neuroscience*, *8*. https://doi.org/10.3389/fnhum.2014.00076

Fisher, A. J., Medaglia, J. D., & Jeronimus, B. F. (2018). Lack of group-to-individual generalizability is a threat to human subjects research. *Proceedings of the National Academy of Sciences*, *115*(27), E6106–E6115. https://doi.org/10.1073/pnas.1711978115

Flemisch, F. O., & Onken, R. (2002). Open a Window to the Cognitive Work Process! Pointillist Analysis of Man–Machine Interaction. *Cognition, Technology & Work*, *4*(3), 160–170. https://doi.org/10.1007/s101110200015

Fournier, L. R., Wilson, G. F., & Swain, C. R. (1999). Electrophysiological, behavioral, and subjective indexes of workload when performing multiple tasks: Manipulations of task difficulty and training. *International Journal of Psychophysiology*, *31*(2), 129–145. https://doi.org/10.1016/S0167-8760(98)00049-X

Foxe, J. J., & Snyder, A. C. (2011). The Role of Alpha-Band Brain Oscillations as a Sensory Suppression Mechanism during Selective Attention. *Frontiers in Psychology*, *2*. https://doi.org/10.3389/fpsyg.2011.00154

Friesen, C. L., Lawrence, M., Ingram, T. G. J., Smith, M. M., Hamilton, E. A., Holland, C. W., Neyedli, H. F., & Boe, S. G. (2022). Portable wireless and fibreless fNIRS headband compares favorably to a stationary headcap-based system. *PLOS ONE*, *17*(7), e0269654. https://doi.org/10.1371/journal.pone.0269654

Gacek, A., & Pedrycz, W. (2011). *ECG Signal Processing, Classification and Interpretation: A Comprehensive Framework of Computational Intelligence*. Springer Science & Business Media.

Gao, Q., Wang, Y., Song, F., Li, Z., & Dong, X. (2013). Mental workload measurement for emergency operating procedures in digital nuclear power plants. *Ergonomics*, *56*(7), 1070–1085. https://doi.org/10.1080/00140139.2013.790483

Gao, R. (2016). Interpreting the electrophysiological power spectrum. *Journal of Neurophysiology*, *115*(2), 628–630. https://doi.org/10.1152/jn.00722.2015

Gao, R., Peterson, E. J., & Voytek, B. (2017). Inferring synaptic excitation/inhibition balance from field potentials. *NeuroImage*, *158*, 70–78. https://doi.org/10.1016/j.neuroimage.2017.06.078

Gerjets, P., Walter, C., Rosenstiel, W., Bogdan, M., & Zander, T. O. (2014). Cognitive state monitoring and the design of adaptive instruction in digital environments: Lessons learned from cognitive workload assessment using a passive brain-computer interface approach. *Frontiers in Neuroscience*, *8*. https://doi.org/10.3389/fnins.2014.00385

Germano, D., Sciaraffa, N., Ronca, V., Giorgi, A., Trulli, G., Borghini, G., Di Flumeri, G., Babiloni, F., & Aricò, P. (2023). Unsupervised Detection of Covariate Shift Due to Changes in EEG Headset Position: Towards an Effective Out-of-Lab Use of Passive Brain–Computer Interface. *Applied Sciences*, *13*(23), Article 23. https://doi.org/10.3390/app132312800

Gerster, M., Waterstraat, G., Litvak, V., Lehnertz, K., Schnitzler, A., Florin, E., Curio, G., & Nikulin, V. (2022). Separating Neural Oscillations from Aperiodic 1/f Activity: Challenges and Recommendations. *Neuroinformatics*, *20*(4), 991–1012. https://doi.org/10.1007/s12021-022-09581-8

Gevins, A., Smith, M. E., McEvoy, L., & Yu, D. (1997). High-resolution EEG mapping of cortical activation related to working memory: Effects of task difficulty, type of processing, and practice. *Cerebral Cortex*, *7*(4), 374–385. https://doi.org/10.1093/cercor/7.4.374

Ghanbary Sartang, A., School of Health, Isfahan University of Medical Sciences, Isfahan, Iran., Ashnagar, M., Bandar Abbas, Iran., Habibi, E., Professor, Dept. of Occupational Health Engineering, School of Health, Isfahan University of Medical Sciences, Isfahan, Iran., Sadeghi, S., & MSc of Industrial Engineering, Dept. of Industrial Engineering, Ilam Branch, Islamic Azad University, Ilam, Iran. (2016). Evaluation of Rating Scale Mental Effort (RSME) effectiveness for mental workload assessment in nurses. *Journal of Occupational Health and Epidemiology*, *5*(4), 211–217. https://doi.org/10.18869/acadpub.johe.5.4.211

Ghani, U., Signal, N., Niazi, I. K., & Taylor, D. (2020). ERP based measures of cognitive workload: A review. *Neuroscience & Biobehavioral Reviews*, *118*, 18–26. https://doi.org/10.1016/j.neubiorev.2020.07.020

Gómez, D., & Rojas, A. (2016). An Empirical Overview of the No Free Lunch Theorem and Its Effect on Real-World Machine Learning Classification. *Neural Computation*, *28*(1), 216–228. https://doi.org/10.1162/NECO_a_00793

Gong, D., Li, Y., Yan, Y., Yao, Y., Gao, Y., Liu, T., Ma, W., & Yao, D. (2019). The high-working load states induced by action real-time strategy gaming: An EEG power spectrum and network study. *Neuropsychologia*, *131*, 42–52. https://doi.org/10.1016/j.neuropsychologia.2019.05.002

Gramann, K., Fairclough, S. H., Zander, T. O., & Ayaz, H. (2017). Editorial: Trends in Neuroergonomics. *Frontiers in Human Neuroscience*, *11*. https://doi.org/10.3389/fnhum.2017.00165

Gramann, K., Jung, T.-P., Ferris, D. P., Lin, C.-T., & Makeig, S. (2014). Toward a new cognitive neuroscience: Modeling natural brain dynamics. *Frontiers in Human Neuroscience*, *8*. https://doi.org/10.3389/fnhum.2014.00444

Gramann, K., Lotte, F., Dehais, F., Ayaz, H., Vukelić, M., Karwowski, W., Fairclough, S., Brouwer, A.-M., & Roy, R. N. (2024). Editorial: Open science to support replicability in neuroergonomic research. *Frontiers in Neuroergonomics*, *5*, 1459204. https://doi.org/10.3389/fnrgo.2024.1459204

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Parkkonen, L., & Hämäläinen, M. S. (2014). MNE software for processing MEG and EEG data. *NeuroImage*, *86*, 446–460. https://doi.org/10.1016/j.neuroimage.2013.10.027

Gratton, G., Brumback, C. R., Gordon, B. A., Pearson, M. A., Low, K. A., & Fabiani, M. (2006). Effects of measurement method, wavelength, and source-detector distance on the fast optical signal. *NeuroImage*, *32*(4), 1576–1590. https://doi.org/10.1016/j.neuroimage.2006.05.030

Greene, A. S., Horien, C., Barson, D., Scheinost, D., & Constable, R. T. (2023). Why is everyone talking about brain state? *Trends in Neurosciences*, *0*(0). https://doi.org/10.1016/j.tins.2023.04.001

Grubov, V. V., Khramova, M. V., Goman, S., Badarin, A. A., Kurkin, S. A., Andrikov, D. A., Pitsik, E., Antipov, V., Petushok, E., Brusinskii, N., Bukina, T., Fedorov, A. A., & Hramov, A. E. (2024). Open-Loop Neuroadaptive System for Enhancing Student's Cognitive Abilities in Learning. *IEEE Access*, *12*, 49034–49049. https://doi.org/10.1109/ACCESS.2024.3383847

Gruenwald, J., Kapeller, C., Guger, C., Ogawa, H., Kamada, K., & Scharinger, J. (2017). Comparison of Alpha/Beta and high-gamma band for motor-imagery based BCI control: A qualitative study. *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2308–2311. https://doi.org/10.1109/SMC.2017.8122965

Gu, X., Cao, Z., Jolfaei, A., Xu, P., Wu, D., Jung, T.-P., & Lin, C.-T. (2021). EEG-Based Brain-Computer Interfaces (BCIs): A Survey of Recent Studies on Signal Sensing Technologies and Computational Intelligence Approaches and Their Applications. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *18*(5), 1645–1666. IEEE/ACM Transactions on Computational Biology and Bioinformatics. https://doi.org/10.1109/TCBB.2021.3052811

Günther, M., Schuster, L., Boßelmann, C., Lerche, H., Ziemann, U., Feil, K., & Marquetand, J. (2023). Sponge EEG is equivalent regarding signal quality, but faster than routine EEG. *Clinical Neurophysiology Practice*, *8*, 58–64. https://doi.org/10.1016/j.cnp.2023.03.002

Haatveit, B. C., Sundet ,Kjetil, Hugdahl ,Kenneth, Ueland ,Torill, Melle ,Ingrid, & and Andreassen, O. A. (2010). The validity of d prime as a working memory index: Results from the "Bergen n-back task." *Journal of Clinical and Experimental Neuropsychology*, *32*(8), 871–880. https://doi.org/10.1080/13803391003596421

Hämäläinen, M. S. (1992). Magnetoencephalography: A tool for functional brain imaging. *Brain Topography*, *5*(2), 95–102. https://doi.org/10.1007/BF01129036

Hancock, P. A., & Matthews, G. (2019). Workload and Performance: Associations, Insensitivities, and Dissociations. *Human Factors*, *61*(3), 374–392. https://doi.org/10.1177/0018720818809590

Hanslmayr, S., Matuschek, J., & Fellner, M.-C. (2014). Entrainment of Prefrontal Beta Oscillations Induces an Endogenous Echo and Impairs Memory Formation. *Current Biology*, *24*(8), 904–909. https://doi.org/10.1016/j.cub.2014.03.007

Hanslmayr, S., Staresina, B. P., & Bowman, H. (2016). Oscillations and Episodic Memory: Addressing the Synchronization/Desynchronization Conundrum. *Trends in Neurosciences*, *39*(1), 16–25. https://doi.org/10.1016/j.tins.2015.11.004

Hanzu-Pazara, R., Barsan, E., Arsenie, P., Chiotoroiu, L., & Raicu, G. (2008). Reducing of maritime accidents caused by human factors using simulators in training process. *Journal of Maritime Research*, *5*(1), 3–18.

Harris, L. R., Carnevale, M. J., D'Amour, S., Fraser, L. E., Harrar, V., Hoover, A. E. N., Mander, C., & Pritchett, L. M. (2015). How our body influences our perception of the world. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00819

Hart, S. G. (2006). Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*(9), 904–908. https://doi.org/10.1177/154193120605000909

Herff, C., Heger, D., Fortmann, O., Hennrich, J., Putze, F., & Schultz, T. (2014). Mental workload during n-back task—Quantified in the prefrontal cortex using fNIRS. *Frontiers in Human Neuroscience*, *7*. https://doi.org/10.3389/fnhum.2013.00935

Herrmann, C. S., Fründ, I., & Lenz, D. (2010). Human gamma-band activity: A review on cognitive and behavioral correlates and network models. *Neuroscience & Biobehavioral Reviews*, *34*(7), 981–992. https://doi.org/10.1016/j.neubiorev.2009.09.001

Hettinger, L. J., Branco, P., Encarnacao, L. M., & Bonato, P. (2003). Neuroadaptive technologies: Applying neuroergonomics to the design of advanced interfaces. *Theoretical Issues in Ergonomics Science*, *4*(1–2), 220–237. https://doi.org/10.1080/1463922021000020918

Hinrichs, H., Scholz, M., Baum, A. K., Kam, J. W. Y., Knight, R. T., & Heinze, H.-J. (2020). Comparison between a wireless dry electrode EEG system with a conventional wired wet electrode EEG

system for clinical applications. *Scientific Reports*, *10*(1), 5218. https://doi.org/10.1038/s41598-020-62154-0

Hinss, M. F., Jahanpour, E. S., Somon, B., Pluchon, L., Dehais, F., & Roy, R. N. (2023). Open multi-session and multi-task EEG cognitive Dataset for passive brain-computer Interface Applications. *Scientific Data*, *10*(1), Article 1. https://doi.org/10.1038/s41597-022-01898-y

Hinss, M. F., Somon, B., Dehais, F., & Roy, R. N. (2021). Open EEG Datasets for Passive Brain-Computer Interface Applications: Lacks and Perspectives. *2021 10th International IEEE/EMBS Conference on Neural Engineering (NER)*, 686–689. https://doi.org/10.1109/NER49283.2021.9441214

Hirshfield, L. M., Wickens, C., Doherty, E., Spencer, C., Williams, T., & Hayne, L. (2024). Toward Workload-Based Adaptive Automation: The Utility of fNIRS for Measuring Load in Multiple Resources in the Brain. *International Journal of Human–Computer Interaction*, *40*(22), 7404–7430. https://doi.org/10.1080/10447318.2023.2266242

Ho, T. K., Hull, J. J., & Srihari, S. N. (1994). Decision combination in multiple classifier systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *16*(1), 66–75. https://doi.org/10.1109/34.273716

Höller, Y., Thomschewski, A., Bergmann, J., Kronbichler, M., Crone, J. S., Schmid, E. V., Butz, K., Höller, P., & Trinka, E. (2013). EEG-Response Consistency across Subjects in an Active Oddball Task. *PLOS ONE*, *8*(9), e74572. https://doi.org/10.1371/journal.pone.0074572

Höller, Y., Trinka, E., Kalss, G., Schiepek, G., & Michaelis, R. (2019). Correlation of EEG spectra, connectivity, and information theoretical biomarkers with psychological states in the epilepsy monitoring unit—A pilot study. *Epilepsy & Behavior*, *99*. https://doi.org/10.1016/j.yebeh.2019.106485

Hong, K.-Y., Wang, C.-C., & Lin, W.-C. (2024). *Multi-modal Motion Prediction using Temporal Ensembling with Learning-based Aggregation* (No. arXiv:2410.19606). arXiv. https://doi.org/10.48550/arXiv.2410.19606

Hughes, A. M., Hancock, G. M., Marlow, S. L., Stowers, K., & Salas, E. (2019). Cardiac Measures of Cognitive Workload: A Meta-Analysis. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *61*(3), 393–414. https://doi.org/10.1177/0018720819830553

İşcan, Z., & Nikulin, V. V. (2018). Steady state visual evoked potential (SSVEP) based brain-computer interface (BCI) performance under different perturbations. *PLoS ONE*, *13*(1), e0191673. https://doi.org/10.1371/journal.pone.0191673

Ismail, L. E., & Karwowski, W. (2020). Applications of EEG indices for the quantification of human cognitive performance: A systematic review and bibliometric analysis. *PLOS ONE*, *15*(12), e0242857. https://doi.org/10.1371/journal.pone.0242857

Ivucic, G., Pahuja, S., Putze, F., Cai, S., Li, H., & Schultz, T. (2024). The Impact of Cross-Validation Schemes for EEG-Based Auditory Attention Detection with Deep Neural Networks. *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 1–4. https://doi.org/10.1109/EMBC53108.2024.10782636

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive Mixtures of Local Experts. *Neural Computation*, *3*(1), 79–87. https://doi.org/10.1162/neco.1991.3.1.79

Jalali, M., Esmaeili, R., Habibi, E., Alizadeh, M., & Karimi, A. (2023). Mental workload profile and its relationship with presenteeism, absenteeism and job performance among surgeons: The mediating role of occupational fatigue. *Heliyon*, *9*(9), e19258. https://doi.org/10.1016/j.heliyon.2023.e19258

Jayaram, V., Alamgir, M., Altun, Y., Scholkopf, B., & Grosse-Wentrup, M. (2016). Transfer Learning in Brain-Computer Interfaces. *IEEE Computational Intelligence Magazine*, *11*(1), 20–31. IEEE Computational Intelligence Magazine. https://doi.org/10.1109/MCI.2015.2501545

Jayaram, V., & Barachant, A. (2018). MOABB: Trustworthy algorithm benchmarking for BCIs. *Journal of Neural Engineering*, *15*(6), 066011. https://doi.org/10.1088/1741-2552/aadea0

Jia, H., Feng, F., Caiafa, C. F., Duan, F., Zhang, Y., Sun, Z., & Solé-Casals, J. (2023). Multi-Class Classification of Upper Limb Movements With Filter Bank Task-Related Component Analysis.

IEEE Journal of Biomedical and Health Informatics, *27*(8), 3867–3877. IEEE Journal of Biomedical and Health Informatics. https://doi.org/10.1109/JBHI.2023.3278747

Jokisch, D., & Jensen, O. (2007). Modulation of Gamma and Alpha Activity during a Working Memory Task Engaging the Dorsal or Ventral Stream. *Journal of Neuroscience*, *27*(12), 3244–3251. https://doi.org/10.1523/JNEUROSCI.5399-06.2007

Jonas, E., & Kording, K. P. (2017). Could a Neuroscientist Understand a Microprocessor? *PLOS Computational Biology*, *13*(1), e1005268. https://doi.org/10.1371/journal.pcbi.1005268

Jorna, P. G. A. M. (1991). Operator Workload as a Limiting Factor in Complex Systems. In J. A. Wise, V. D. Hopkin, & M. L. Smith (Eds.), *Automation and Systems Issues in Air Traffic Control* (pp. 281–292). Springer. https://doi.org/10.1007/978-3-642-76556-8_28

Kahneman, D. (1973). *Attention and effort*. Prentice-Hall.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In *Heuristics and biases: The psychology of intuitive judgment* (pp. 49–81). Cambridge University Press. https://doi.org/10.1017/CBO9780511808098.004

Kałamała, P., Gyurkovics, M., Bowie, D. C., Clements, G. M., Low, K. A., Dolcos, F., Fabiani, M., & Gratton, G. (2024). Event-induced modulation of aperiodic background EEG: Attention-dependent and age-related shifts in E:I balance, and their consequences for behavior. *Imaging Neuroscience*, *2*, imag–2–00054. https://doi.org/10.1162/imag_a_00054

Kalunga, E. K., Chevallier, S., Barthélemy, Q., Djouani, K., Monacelli, E., & Hamam, Y. (2016). Online SSVEP-based BCI using Riemannian geometry. *Neurocomputing*, *191*, 55–68. https://doi.org/10.1016/j.neucom.2016.01.007

Karran, A. J., Demazure, T., Leger, P.-M., Labonte-LeMoyne, E., Senecal, S., Fredette, M., & Babin, G. (2019). Toward a Hybrid Passive BCI for the Modulation of Sustained Attention Using EEG and fNIRS. *Frontiers in Human Neuroscience*, *13*. https://doi.org/10.3389/fnhum.2019.00393

Kave, R. (2017, September 24). *Rise in lorry tachograph tampering on UK roads*. https://www.bbc.com/news/uk-41361351

Ke, Y., Jiang, T., Liu, S., Cao, Y., Jiao, X., Jiang, J., & Ming, D. (2021). Cross-Task Consistency of Electroencephalography-Based Mental Workload Indicators: Comparisons Between Power Spectral Density and Task-Irrelevant Auditory Event-Related Potentials. *Frontiers in Neuroscience*, *15*. https://www.frontiersin.org/articles/10.3389/fnins.2021.703139

Ke, Y., Wang, T., He, F., Liu, S., & Ming, D. (2023). Enhancing EEG-based cross-day mental workload classification using periodic component of power spectrum. *Journal of Neural Engineering*, *20*(6), 066028. https://doi.org/10.1088/1741-2552/ad0f3d

Kellogg, K. C., Valentine, M. A., & Christin, A. (2020). Algorithms at Work: The New Contested Terrain of Control. *Academy of Management Annals*, *14*(1), 366–410. https://doi.org/10.5465/annals.2018.0174

Khan, M. J., & Hong, K.-S. (2017). Hybrid EEG–fNIRS-Based Eight-Command Decoding for BCI: Application to Quadcopter Control. *Frontiers in Neurorobotics*, *11*. https://doi.org/10.3389/fnbot.2017.00006

Kim, H., Luo, J., Chu, S., Cannard, C., Hoffmann, S., & Miyakoshi, M. (2023). ICA's bug: How ghost ICs emerge from effective rank deficiency caused by EEG electrode interpolation and incorrect re-referencing. *Frontiers in Signal Processing*, *3*. https://doi.org/10.3389/frsip.2023.1064138

Kim, M.-J., Kim, Y.-J., Yum, M.-S., & Kim, W. Y. (2022). Alpha-power in electroencephalography as good outcome predictor for out-of-hospital cardiac arrest survivors. *Scientific Reports*, *12*(1), 10907. https://doi.org/10.1038/s41598-022-15144-3

Kirchner, W. K. (1958). Age differences in short-term retention of rapidly changing information. *Journal of Experimental Psychology*, *55*(4), 352–358. https://doi.org/10.1037/h0043688

Klein, F., Kohl, S. H., Lührs, M., Mehler, D. M. A., & Sorger, B. (2024). From lab to life: Challenges and perspectives of fNIRS for haemodynamic-based neurofeedback in real-world environments. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *379*(1915), 20230087. https://doi.org/10.1098/rstb.2023.0087

Klug, M., & Kloosterman, N. A. (2022). Zapline-plus: A Zapline extension for automatic and adaptive removal of frequency-specific noise artifacts in M/EEG. *Human Brain Mapping*, *43*(9), 2743–2758. https://doi.org/10.1002/hbm.25832

Kosti, M. V., Georgiadis, K., Adamos, D. A., Laskaris, N., Spinellis, D., & Angelis, L. (2018). Towards an affordable brain computer interface for the assessment of programmers' mental workload. *International Journal of Human-Computer Studies*, *115*, 52–66. https://doi.org/10.1016/j.ijhcs.2018.03.002

Kramer, A. F. (1991). Physiological metrics of mental workload: A review of recent progress. In *Multiple Task Performance*. CRC Press.

Kramer, A. F., Trejo, L. J., & Humphrey, D. (1995). Assessment of mental workload with task-irrelevant auditory probes. *Biological Psychology*, *40*(1–2), 83–100. https://doi.org/10.1016/0301-0511(95)05108-2

Kramer, M. A., & Chu, C. J. (2023). The 1/f-like behavior of neural field spectra are a natural consequence of noise driven brain dynamics. *bioRxiv*, 2023.03.10.532077. https://doi.org/10.1101/2023.03.10.532077

Krauledat, J. M. (2008). *Analysis of Nonstationarities in EEG Signals for Improving Brain-Computer Interface Performance*. https://depositonce.tu-berlin.de/items/urn:nbn:de:kobv:83-opus-18155

Kritzman, L., Eidelman-Rothman, M., Keil, A., Freche, D., Sheppes, G., & Levit-Binnun, N. (2022). Steady-state visual evoked potentials differentiate between internally and externally directed attention. *NeuroImage*, *254*, 119133. https://doi.org/10.1016/j.neuroimage.2022.119133

Krol, L. R., & Zander, T. (2017). *Passive Bci-Based Neuroadaptive Systems*. 7th Graz Brain-Computer Interface Conference, Graz, Austria. https://doi.org/10.3217/978-3-85125-533-1-46

Krusienski, D. J., Grosse-Wentrup, M., Galán, F., Coyle, D., Miller, K. J., Forney, E., & Anderson, C. W. (2011). Critical issues in state-of-the-art brain–computer interface signal processing. *Journal of Neural Engineering*, *8*(2), 025002. https://doi.org/10.1088/1741-2560/8/2/025002

Kumar, S., Yger, F., & Lotte, F. (2019). Towards Adaptive Classification using Riemannian Geometry approaches in Brain-Computer Interfaces. *2019 7th International Winter Conference on Brain-Computer Interface (BCI)*, 1–6. https://doi.org/10.1109/IWW-BCI.2019.8737349

Kumaravel, V. P., Kartsch, V., Benatti, S., Vallortigara, G., Farella, E., & Buiatti, M. (2021). Efficient Artifact Removal from Low-Density Wearable EEG using Artifacts Subspace Reconstruction. *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 333–336. https://doi.org/10.1109/EMBC46164.2021.9629771

Kuncheva, L. I., Bezdek, J. C., & Duin, R. P. W. (2001). Decision templates for multiple classifier fusion: An experimental comparison. *Pattern Recognition*, *34*(2), 299–314. https://doi.org/10.1016/S0031-3203(99)00223-X

Laborde, S., Mosley, E., & Thayer, J. F. (2017). Heart Rate Variability and Cardiac Vagal Tone in Psychophysiological Research – Recommendations for Experiment Planning, Data Analysis, and Data Reporting. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.00213

Ladouce, S., Darmet, L., Torre Tresols, J. J., Velut, S., Ferraro, G., & Dehais, F. (2022). Improving user experience of SSVEP BCI through low amplitude depth and high frequency stimuli design. *Scientific Reports*, *12*(1), 8865. https://doi.org/10.1038/s41598-022-12733-0

Ladouce, S., & Dehais, F. (2024). Frequency tagging of spatial attention using periliminal flickers. *Imaging Neuroscience*, *2*, 1–17. https://doi.org/10.1162/imag_a_00223

Ladouce, S., Pietzker, M., Manzey, D., & Dehais, F. (2024). Evaluation of a headphones-fitted EEG system for the recording of auditory evoked potentials and mental workload assessment. *Behavioural Brain Research*, *460*, 114827. https://doi.org/10.1016/j.bbr.2023.114827

Ladouce, S., Torre Tresols, J. J., Goff, K. L., & Dehais, F. (2025). EEG-based assessment of long-term vigilance and lapses of attention using a user-centered frequency-tagging approach. *Journal of Neural Engineering*, *22*(3), 036018. https://doi.org/10.1088/1741-2552/add771

Laine, S., & Aila, T. (2017). *Temporal Ensembling for Semi-Supervised Learning* (No. arXiv:1610.02242). arXiv. https://doi.org/10.48550/arXiv.1610.02242

Lal, S. K. L., & Craig, A. (2002). Driver fatigue: Electroencephalography and psychological assessment. *Psychophysiology*, *39*(3), 313–321. https://doi.org/10.1017/s0048577201393095

Lampert, R. (2015). ECG signatures of psychological stress. *Journal of Electrocardiology*, *48*(6), 1000–1005. https://doi.org/10.1016/j.jelectrocard.2015.08.005

Lebedev, M. A., & Nicolelis, M. A. L. (2006). Brain–machine interfaces: Past, present and future. *Trends in Neurosciences*, *29*(9), 536–546. https://doi.org/10.1016/j.tins.2006.07.004

Ledoit, O., & Wolf, M. N. (2003). Honey, I Shrunk the Sample Covariance Matrix. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.433840

Lee, K.-S., Lee, J., & Hwang, J. (2024). Research trends in ergonomics, industrial safety and health: Semantic network and topic analyses. *International Journal of Occupational Safety and Ergonomics*, *30*(1), 20–32. https://doi.org/10.1080/10803548.2022.2157544

Leinders, S., Aarnoutse, E. J., Branco, M. P., Freudenburg, Z. V., Geukes, S. H., Schippers, A., Verberne, M. S., Boom, M. van den, Vijgh, B. van der, Crone, N. E., Denison, T., Ramsey, N. F., & Vansteensel, M. J. (2024). DO NOT LOSE SLEEP OVER IT: IMPLANTED BRAIN-COMPUTER INTERFACE FUNCTIONALITY DURING NIGHTTIME IN LATE-STAGE AMYOTROPHIC LATERAL SCLEROSIS. *medRxiv*, 2024.10.11.24315027. https://doi.org/10.1101/2024.10.11.24315027

Lemm, S., Blankertz, B., Dickhaus, T., & Müller, K.-R. (2011). Introduction to machine learning for brain imaging. *NeuroImage*, *56*(2), 387–399. https://doi.org/10.1016/j.neuroimage.2010.11.004

Li, G., Huang, S., Xu, W., Jiao, W., Jiang, Y., Gao, Z., & Zhang, J. (2020). The impact of mental fatigue on brain activity: A comparative study both in resting state and task state using EEG. *BMC Neuroscience*, *21*(1), 20. https://doi.org/10.1186/s12868-020-00569-1

Li, G.-L., Wu, J.-T., Xia, Y.-H., He, Q.-G., & Jin, H.-G. (2020). Review of semi-dry electrodes for EEG recording. *Journal of Neural Engineering*, *17*(5), 051004. https://doi.org/10.1088/1741-2552/abbd50

Li, R., Johansen, J. S., Ahmed, H., Ilyevsky, T. V., Wilbur, R. B., Bharadwaj, H. M., & Siskind, J. M. (2020). The Perils and Pitfalls of Block Design for EEG Classification Experiments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–1. https://doi.org/10.1109/TPAMI.2020.2973153

Lim, B., & Zohren, S. (2021). Time-series forecasting with deep learning: A survey. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *379*(2194), 20200209. https://doi.org/10.1098/rsta.2020.0209

Linkenkaer-Hansen, K., Nikouline, V. V., Palva, J. M., & Ilmoniemi, R. J. (2001). Long-Range Temporal Correlations and Scaling Behavior in Human Brain Oscillations. *Journal of Neuroscience*, *21*(4), 1370–1377. https://doi.org/10.1523/JNEUROSCI.21-04-01370.2001

Liu, P.-K., Beh, W.-K., Shih, C.-Y., Chen, Y.-T., & Wu, A.-Y. A. (2019). Entropy and Complexity Assisted EEG-based Mental Workload Assessment System. *2019 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, 1–4. https://doi.org/10.1109/BIOCAS.2019.8919019

Liu, Y., Ayaz ,Hasan, & and Shewokis, P. A. (2017). Mental workload classification with concurrent electroencephalography and functional near-infrared spectroscopy. *Brain-Computer Interfaces*, *4*(3), 175–185. https://doi.org/10.1080/2326263X.2017.1304020

Lobier, M., Palva, J. M., & Palva, S. (2018). High-alpha band synchronization across frontal, parietal and visual cortex mediates behavioral and neuronal effects of visuospatial attention. *NeuroImage*, *165*, 222–237. https://doi.org/10.1016/j.neuroimage.2017.10.044

Loef, B., Wong, A., Janssen, N. A. H., Strak, M., Hoekstra, J., Picavet, H. S. J., Boshuizen, H. C. H., Verschuren, W. M. M., & Herber, G.-C. M. (2022). Using random forest to identify longitudinal predictors of health in a 30-year cohort study. *Scientific Reports*, *12*(1), 10372. https://doi.org/10.1038/s41598-022-14632-w

Logothetis, N. K. (2008). What we can do and what we cannot do with fMRI. *Nature*, *453*(7197), 869–878. https://doi.org/10.1038/nature06976

Lotte, F., Bougrain, L., Cichocki, A., Clerc, M., Congedo, M., Rakotomamonjy, A., & Yger, F. (2018). A review of classification algorithms for EEG-based brain–computer interfaces: A 10 year update. *Journal of Neural Engineering*, *15*(3), 031005. https://doi.org/10.1088/1741-2552/aab2f2

Lotte, F., Guan, C., & Ang, K. K. (2009). Comparison of designs towards a subject-independent brain-computer interface based on motor imagery. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, *2009*, 4543–4546. https://doi.org/10.1109/IEMBS.2009.5334126

Lotte, F., & Jeunet, C. (2018). Defining and Quantifying Users' Mental Imagery-based BCI skills: A first step. *Journal of Neural Engineering*, *15*(4), 1–37. https://doi.org/10.1088/1741-2552/aac577

Louis, L.-E. L., Moussaoui, S., Van Langhenhove, A., Ravoux, S., Le Jan, T., Roualdes, V., & Milleville-Pennel, I. (2023). Cognitive tasks and combined statistical methods to evaluate, model, and predict mental workload. *Frontiers in Psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1122793

Luck, S. J. (2005). *An introduction to the event-related potential technique*. MIT Press.

Luo, S., Angrick, M., Coogan, C., Candrea, D. N., Wyse-Sookoo, K., Schippers, A., Ganji, R., Milsap, G. W., Anderson, W. S., Gordon, C. R., Tippett, D. C., Maragakis, N. J., Clawson, L. L., Vansteensel, M. J., Tenore, F. V., Hermansky, H., Fifer, M. S., Ramsey, N. F., & Crone, N. E.

(2025). *Self-paced silent speech brain-computer interface for device control* (p. 2025.04.09.25325542). medRxiv. https://doi.org/10.1101/2025.04.09.25325542

Luzzani, G., Buraioli, I., Demarchi, D., & Guglieri, G. (2024). A review of physiological measures for mental workload assessment in aviation: A state-of-the-art review of mental workload physiological assessment methods in human-machine interaction analysis. *The Aeronautical Journal*, *128*(1323), 928–949. https://doi.org/10.1017/aer.2023.101

Macmillan, N. A., & Creelman, C. D. (1990). Response bias: Characteristics of detection theory, threshold theory, and "nonparametric" indexes. *Psychological Bulletin*, *107*(3), 401–413. https://doi.org/10.1037/0033-2909.107.3.401

Mahdavi, N., Tapak, L., Darvishi, E., Doosti-Irani, A., & Shafiee Motlagh, M. (2024). Unraveling the interplay between mental workload, occupational fatigue, physiological responses and cognitive performance in office workers. *Scientific Reports*, *14*(1), 17866. https://doi.org/10.1038/s41598-024-68889-4

Mahini, R., Zhang, G., Parviainen, T., Düsing, R., Nandi, A. K., Cong, F., & Hämäläinen, T. (2024). Brain Evoked Response Qualification Using Multi-Set Consensus Clustering: Toward Single-Trial EEG Analysis. *Brain Topography*, *37*(6), 1010–1032. https://doi.org/10.1007/s10548-024-01074-y

Mainzer, K. (2007). The emergence of mind and brain: An evolutionary, computational, and philosophical approach. In R. Banerjee & B. K. Chakrabarti (Eds.), *Progress in Brain Research* (Vol. 168, pp. 115–132). Elsevier. https://doi.org/10.1016/S0079-6123(07)68010-8

Makeig, S., Gramann, K., Jung, T.-P., Sejnowski, T. J., & Poizner, H. (2009). Linking brain, mind and behavior. *International Journal of Psychophysiology*, *73*(2), 95–100. https://doi.org/10.1016/j.ijpsycho.2008.11.008

Mandal, S., Singh, B. K., & Thakur, K. (2020). Classification of working memory loads using hybrid EEG and fNIRS in machine learning paradigm. *Electronics Letters*, *56*(25), 1386–1389. https://doi.org/10.1049/el.2020.2710

Marco-Pallarés, J., Münte, T. F., & Rodríguez-Fornells, A. (2015). The role of high-frequency oscillatory activity in reward processing and learning. *Neuroscience & Biobehavioral Reviews*, *49*, 1–7. https://doi.org/10.1016/j.neubiorev.2014.11.014

Markow, Z. E., Trobaugh, J. W., Richter, E. J., Tripathy, K., Rafferty, S. M., Svoboda, A. M., Schroeder, M. L., Burns-Yocum, T. M., Bergonzi, K. M., Chevillet, M. A., Mugler, E. M., Eggebrecht, A. T., & Culver, J. P. (2025). Ultra high density imaging arrays in diffuse optical tomography for human brain mapping improve image quality and decoding performance. *Scientific Reports*, *15*(1), 3175. https://doi.org/10.1038/s41598-025-85858-7

Marquart, G., Cabrall, C., & De Winter, J. (2015). Review of Eye-related Measures of Drivers' Mental Workload. *Procedia Manufacturing*, *3*, 2854–2861. https://doi.org/10.1016/j.promfg.2015.07.783

Martinez, W., Benerradi, J., Midha, S., Maior, H. A., & Wilson, M. L. (2022). Understanding the Ethical Concerns for Neurotechnology in the Future of Work. *Proceedings of the 1st Annual Meeting of the Symposium on Human-Computer Interaction for Work*, 1–19. https://doi.org/10.1145/3533406.3533423

Marzetti, L., Makkinayeri, S., Pieramico, G., Guidotti, R., D'Andrea, A., Roine, T., Mutanen, T. P., Souza, V. H., Kičić, D., Baldassarre, A., Ermolova, M., Pankka, H., Ilmoniemi, R. J., Ziemann, U., Luca Romani, G., & Pizzella, V. (2024). Towards real-time identification of large-scale brain states for improved brain state-dependent stimulation. *Clinical Neurophysiology*, *158*, 196–203. https://doi.org/10.1016/j.clinph.2023.09.005

Mathewson, K. E., Harrison, T. J. L., & Kizuk, S. A. D. (2017). High and dry? Comparing active dry EEG electrodes to active and passive wet electrodes. *Psychophysiology*, *54*(1), 74–82. https://doi.org/10.1111/psyp.12536

Matthews, G., Reinerman-Jones, L. E., Barber, D. J., & Abich, J. (2015). The Psychometrics of Mental Workload: Multiple Measures Are Sensitive but Divergent. *Human Factors*, *57*(1), 125–143. https://doi.org/10.1177/0018720814539505

Maÿe, A., Rauterberg, R., & Engel, A. K. (2022). Instant classification for the spatially-coded BCI. *PLOS ONE*, *17*(4), e0267548. https://doi.org/10.1371/journal.pone.0267548

Mayer-Kress, G. (1998). Non-Linear Mechanisms in the Brain. *Zeitschrift Für Naturforschung C*, *53*(7–8), 677–685. https://doi.org/10.1515/znc-1998-7-820

McCormick, D. A., Nestvogel, D. B., & He, B. J. (2020). Neuromodulation of Brain State and Behavior. *Annual Review of Neuroscience*, *43*(Volume 43, 2020), 391–415. https://doi.org/10.1146/annurev-neuro-100219-105424

Medel, V., Irani, M., Crossley, N., Ossandón, T., & Boncompte, G. (2023). Complexity and 1/f slope jointly reflect brain states. *Scientific Reports*, *13*(1), 21700. https://doi.org/10.1038/s41598-023-47316-0

Mehler, D. M. A., & Kording, K. P. (2018). *The lure of misleading causal statements in functional connectivity research* (Version 3). arXiv. https://doi.org/10.48550/ARXIV.1812.03363

Mehta, R. K., & Parasuraman, R. (2013). Neuroergonomics: A review of applications to physical and cognitive work. *Frontiers in Human Neuroscience*, *7*. https://doi.org/10.3389/fnhum.2013.00889

Menon, R. S., Ogawa, S., Strupp, J. P., & Uğurbil, K. (1997). Ocular Dominance in Human V1 Demonstrated by Functional Magnetic Resonance Imaging. *Journal of Neurophysiology*, *77*(5), 2780–2787. https://doi.org/10.1152/jn.1997.77.5.2780

Mensen, A., & Khatami, R. (2013). Advanced EEG analysis using threshold-free cluster-enhancement and non-parametric statistics. *NeuroImage*, *67*, 111–118. https://doi.org/10.1016/j.neuroimage.2012.10.027

Meule, A. (2017). Reporting and Interpreting Working Memory Performance in n-back Tasks. *Frontiers in Psychology*, *8*, 352. https://doi.org/10.3389/fpsyg.2017.00352

Michel, C. M., & Brunet, D. (2019). EEG Source Imaging: A Practical Review of the Analysis Steps. *Frontiers in Neurology*, *10*. https://doi.org/10.3389/fneur.2019.00325

Middendorf, M., McMillan, G., Calhoun, G., & Jones, K. S. (2000). Brain-computer interfaces based on the steady-state visual-evoked response. *IEEE Transactions on Rehabilitation Engineering*, *8*(2), 211–214. https://doi.org/10.1109/86.847819

Miladinović, A., Ajčević, M., Jarmolowska, J., Marusic, U., Colussi, M., Silveri, G., Battaglini, P. P., & Accardo, A. (2021). Effect of power feature covariance shift on BCI spatial-filtering techniques: A comparative study. *Computer Methods and Programs in Biomedicine*, *198*, 105808. https://doi.org/10.1016/j.cmpb.2020.105808

Miller, K. J., Hermes, D., & Staff, N. P. (2020). *The current state of electrocorticography-based brain–computer interfaces*. https://doi.org/10.3171/2020.4.FOCUS20185

Moore, T. M., & Picou, E. M. (2018). A Potential Bias in Subjective Ratings of Mental Effort. *Journal of Speech, Language, and Hearing Research : JSLHR*, *61*(9), 2405. https://doi.org/10.1044/2018_JSLHR-H-17-0451

Morales, J. M., Ruiz-Rabelo, J. F., Diaz-Piedra, C., & Di Stasi, L. L. (2019). Detecting Mental Workload in Surgical Teams Using a Wearable Single-Channel Electroencephalographic Device. *Journal of Surgical Education*, *76*(4), 1107–1115. https://doi.org/10.1016/j.jsurg.2019.01.005

Moray, N. (Ed.). (1979). *Mental Workload*. Springer US. https://doi.org/10.1007/978-1-4757-0884-4

Morigi, J. J., Kovaleva, N., & Phan, S. (2022). Spotlight on: "Dynamic PET/CT imaging." *Clinical and Translational Imaging*, *10*(3), 239–241. https://doi.org/10.1007/s40336-022-00500-0

Mühl, C., Jeunet, C., & Lotte, F. (2014). EEG-based workload estimation across affective contexts. *Frontiers in Neuroscience*, *8*. https://doi.org/10.3389/fnins.2014.00114

Muhl, E. (2024). The challenge of wearable neurodevices for workplace monitoring: An EU legal perspective. *Frontiers in Human Dynamics*, *6*. https://doi.org/10.3389/fhumd.2024.1473893

Mullen, T., Kothe, C., Chi, Y. M., Ojeda, A., Kerth, T., Makeig, S., Cauwenberghs, G., & Jung, T.-P. (2013). Real-time modeling and 3D visualization of source dynamics and connectivity using wearable EEG. *2013 35th Annual International Conference of the IEEE Engineering in*

Medicine and Biology Society (EMBC), 2184–2187. https://doi.org/10.1109/EMBC.2013.6609968

Mushtaq, F., Welke, D., Gallagher, A., Pavlov, Y. G., Kouara, L., Bosch-Bayard, J., Van Den Bosch, J. J. F., Arvaneh, M., Bland, A. R., Chaumon, M., Borck, C., He, X., Luck, S. J., Machizawa, M. G., Pernet, C., Puce, A., Segalowitz, S. J., Rogers, C., Awais, M., … Valdes-Sosa, P. (2024). One hundred years of EEG for brain and behaviour research. *Nature Human Behaviour*, *8*(8), 1437–1443. https://doi.org/10.1038/s41562-024-01941-5

Näätänen, R., & Picton, T. (1987). The N1 wave of the human electric and magnetic response to sound: A review and an analysis of the component structure. *Psychophysiology*, *24*(4), 375–425. https://doi.org/10.1111/j.1469-8986.1987.tb00311.x

Nabian, M., Yin, Y., Wormwood, J., Quigley, K. S., Barrett, L. F., & Ostadabbas, S. (2018). An Open-Source Feature Extraction Tool for the Analysis of Peripheral Physiological Data. *IEEE Journal of Translational Engineering in Health and Medicine*, *6*, 1–11. https://doi.org/10.1109/JTEHM.2018.2878000

Näher, T., Bastian, L., Vorreuther, A., Fries, P., Goebel, R., & Sorger, B. (2024). *Riemannian Geometry for the classification of brain states with fNIRS* (p. 2024.09.06.611347). bioRxiv. https://doi.org/10.1101/2024.09.06.611347

Nakamura, T., Usui, S., Shinohara, K., & Kanda, K. (2004). The Psychological Factors concerning Human Errors as the Cause of Labour Accidents in Japan. In C. Spitzer, U. Schmocker, & V. N. Dang (Eds.), *Probabilistic Safety Assessment and Management* (pp. 1–6). Springer. https://doi.org/10.1007/978-0-85729-410-4_1

Naseer, N., & Hong, K.-S. (2015). fNIRS-based brain-computer interfaces: A review. *Frontiers in Human Neuroscience*, *9*. https://doi.org/10.3389/fnhum.2015.00003

Naseer, N., Noori, F. M., Qureshi, N. K., & Hong, K.-S. (2016). Determining Optimal Feature-Combination for LDA Classification of Functional Near-Infrared Spectroscopy Signals in Brain-

Computer Interface Application. *Frontiers in Human Neuroscience*, *10*. https://doi.org/10.3389/fnhum.2016.00237

Nguyen, T., Ahn, S., Jang, H., Jun, S. C., & Kim, J. G. (2017). Utilization of a combined EEG/NIRS system to predict driver drowsiness. *Scientific Reports*, *7*(1), 43933. https://doi.org/10.1038/srep43933

Nickel, P., & Nachreiner, F. (2003). Sensitivity and Diagnosticity of the 0.1-Hz Component of Heart Rate Variability as an Indicator of Mental Workload. *Human Factors*, *45*(4), 575–590. https://doi.org/10.1518/hfes.45.4.575.27094

Niso, G., Romero, E., Moreau, J. T., Araujo, A., & Krol, L. R. (2023). Wireless EEG: A survey of systems and studies. *NeuroImage*, *269*, 119774. https://doi.org/10.1016/j.neuroimage.2022.119774

Norcia, A. M., Appelbaum, L. G., Ales, J. M., Cottereau, B. R., & Rossion, B. (2015). The steady-state visual evoked potential in vision research: A review. *Journal of Vision*, *15*(6), 4. https://doi.org/10.1167/15.6.4

Norris, D. G. (2006). Principles of magnetic resonance assessment of brain function. *Journal of Magnetic Resonance Imaging*, *23*(6), 794–807. https://doi.org/10.1002/jmri.20587

Novi, S. L., Abdalmalak, A., Kazazian, K., Norton, L., Debicki, D. B., Mesquita, R. C., & Owen, A. M. (2023). *Improved short-channel regression for mapping resting-state functional connectivity networks using functional near-infrared spectroscopy* (p. 2023.06.12.543244). bioRxiv. https://doi.org/10.1101/2023.06.12.543244

Nunez, P. L., & Srinivasan, R. (2006). *Electric Fields of the Brain*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195050387.001.0001

O'Brien, W. J., Carlton, L., Muhvich, J., Kura, S., Ortega-Martinez, A., Dubb, J., Duwadi, S., Hazen, E., Yücel, M. A., Lühmann, A. von, Boas, D. A., & Zimmermann, B. B. (2024). ninjaNIRS: An open hardware solution for wearable whole-head high-density functional near-infrared spectroscopy. *Biomedical Optics Express*, *15*(10), 5625–5644. https://doi.org/10.1364/BOE.531501

O'Donnell, R. D., & Eggemeier, F. T. (1986). Workload assessment methodology. In *Handbook of perception and human performance, Vol. 2: Cognitive processes and performance* (pp. 1–49). John Wiley & Sons.

O'Hanlon, J. F. (1972). *Heart Rate Variability: A New Index of Driver Alertness/Fatigue*. 720141. https://doi.org/10.4271/720141

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2011). FieldTrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data. *Computational Intelligence and Neuroscience*, *2011*(1), 156869. https://doi.org/10.1155/2011/156869

Padfield, N., Ren, J., Qing, C., Murray, P., Zhao, H., & Zheng, J. (2021). Multi-segment Majority Voting Decision Fusion for MI EEG Brain-Computer Interfacing. *Cognitive Computation*, *13*(6), 1484–1495. https://doi.org/10.1007/s12559-021-09953-3

Pan, H., Song, H., Zhang, Q., & Mi, W. (2022). Review of Closed-Loop Brain–Machine Interface Systems From a Control Perspective. *IEEE Transactions on Human-Machine Systems*, *52*(5), 877–893. https://doi.org/10.1109/THMS.2021.3138677

Pan, J., & Tompkins, W. J. (1985). A Real-Time QRS Detection Algorithm. *IEEE Transactions on Biomedical Engineering*, *BME-32*(3), 230–236. https://doi.org/10.1109/TBME.1985.325532

Papanicolaou, A. C., & Johnstone, J. (1984). Probe Evoked Potentials: Theory, Method and Applications. *International Journal of Neuroscience*, *24*(2), 107–131. https://doi.org/10.3109/00207458409089800

Pape, A. M., Wiegmann, D. A., & Shappell, S. (2001, March). *Air Traffic Control (ATC) Related Accidents and Incidents: A Human Factors Analysis: 11th International Symposium on Aviation Psychology*.

Parasuraman, R. (2003). Neuroergonomics: Research and practice. *Theoretical Issues in Ergonomics Science*, *4*(1–2), 5–20. https://doi.org/10.1080/14639220210199753

Parasuraman, R., Mouloua, M., & Hilburn, B. (1999). Adaptive aiding and adaptive task allocation enhance human-machine interaction. In *Automated Technology and Human Performance: Current Research and Trends* (pp. 119–123). Erlbaum.

Paxion, J., Galy, E., & Berthelon, C. (2014). Mental workload and driving. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.01344

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830.

Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(8), 1226–1238. https://doi.org/10.1109/TPAMI.2005.159

Perdikis, S., Leeb, R., & Millán, J. d R. (2016). Context-aware adaptive spelling in motor imagery BCI. *Journal of Neural Engineering*, *13*(3), 036018. https://doi.org/10.1088/1741-2560/13/3/036018

Pereira, D. R., Cardoso, S., Ferreira-Santos, F., Fernandes, C., Cunha-Reis, C., Paiva, T. O., Almeida, P. R., Silveira, C., Barbosa, F., & Marques-Teixeira, J. (2014). Effects of inter-stimulus interval (ISI) duration on the N1 and P2 components of the auditory event-related potential. *International Journal of Psychophysiology*, *94*(3), 311–318. https://doi.org/10.1016/j.ijpsycho.2014.09.012

Perlstein, W. M., Cole, M. A., Larson, M., Kelly, K., Seignourel, P., & Keil, A. (2003). Steady-state visual evoked potentials reveal frontally-mediated working memory activity in humans. *Neuroscience Letters*, *342*(3), 191–195. https://doi.org/10.1016/S0304-3940(03)00226-X

Piastra, M. C., Nüßing, A., Vorwerk, J., Clerc, M., Engwer, C., & Wolters, C. H. (2020). A comprehensive study on electroencephalography and magnetoencephalography sensitivity

to cortical and subcortical sources. *Human Brain Mapping*, *42*(4), 978. https://doi.org/10.1002/hbm.25272

Pieper, K., Spang, R. P., Prietz, P., Möller, S., Paajanen, E., Vaalgamaa, M., & Voigt-Antons, J.-N. (2021). Working With Environmental Noise and Noise-Cancelation: A Workload Assessment With EEG and Subjective Measures. *Frontiers in Neuroscience*, *15*. https://doi.org/10.3389/fnins.2021.771533

Pindi, P., Houenou, J., Piguet, C., & Favre, P. (2022). Real-time fMRI neurofeedback as a new treatment for psychiatric disorders: A meta-analysis. *Progress in Neuro-Psychopharmacology & Biological Psychiatry*, *119*, 110605. https://doi.org/10.1016/j.pnpbp.2022.110605

Pion-Tonachini, L., Kreutz-Delgado, K., & Makeig, S. (2019). ICLabel: An automated electroencephalographic independent component classifier, dataset, and website. *NeuroImage*, *198*, 181–197. https://doi.org/10.1016/j.neuroimage.2019.05.026

Podvalny, E., Noy, N., Harel, M., Bickel, S., Chechik, G., Schroeder, C. E., Mehta, A. D., Tsodyks, M., & Malach, R. (2015). A unifying principle underlying the extracellular field potential spectral responses in the human cortex. *Journal of Neurophysiology*, *114*(1), 505–519. https://doi.org/10.1152/jn.00943.2014

Pohlert, T. (2024). *PMCMRplus: Calculate Pairwise Multiple Comparisons of Mean Rank Sums Extended* (Version 1.9.12) [Computer software]. https://cran.r-project.org/web/packages/PMCMRplus/index.html

Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, *6*(3), 21–45. https://doi.org/10.1109/MCAS.2006.1688199

Pontiggia, A., Gomez-Merino, D., Quiquempoix, M., Beauchamps, V., Boffet, A., Fabries, P., Chennaoui, M., & Sauvet, F. (2024). MATB for assessing different mental workload levels. *Frontiers in Physiology*, *15*, 1408242. https://doi.org/10.3389/fphys.2024.1408242

Pope, A. T., Bogart, E. H., & Bartolome, D. S. (1995). Biocybernetic system evaluates indices of operator engagement in automated task. *Biological Psychology*, *40*(1), 187–195. https://doi.org/10.1016/0301-0511(95)05116-3

Pramme, L., Larra, M. F., Schächinger, H., & Frings, C. (2016). Cardiac cycle time effects on selection efficiency in vision. *Psychophysiology*, *53*(11), 1702–1711. https://doi.org/10.1111/psyp.12728

Prasetyo, R. A. B. (2024). The use of multi-attribute task battery in mental workload studies: A scoping review. *SHS Web of Conferences*, *189*, 01043. https://doi.org/10.1051/shsconf/202418901043

Prégent, A. (2025). Is there not an obvious loophole in the AI act's ban on emotion recognition technologies? *AI & SOCIETY*. https://doi.org/10.1007/s00146-025-02289-8

Puma, S., Matton, N., Paubel, P.-V., Raufaste, É., & El-Yagoubi, R. (2018). Using theta and alpha band power to assess cognitive workload in multitasking environments. *International Journal of Psychophysiology*, *123*, 111–120. https://doi.org/10.1016/j.ijpsycho.2017.10.004

Pütz, S., Mertens, A., Chuang, L., & Nitsch, V. (2024). Physiological measures of operators' mental state in supervisory process control tasks: A scoping review. *Ergonomics*, *67*(6), 801–830. https://doi.org/10.1080/00140139.2023.2289858

Putze, F., Putze, S., Sagehorn, M., Micek, C., & Solovey, E. T. (2022). Understanding HCI Practices and Challenges of Experiment Reporting with Brain Signals: Towards Reproducibility and Reuse. *ACM Trans. Comput.-Hum. Interact.*, *29*(4), 31:1-31:43. https://doi.org/10.1145/3490554

Qiu, L., Zhong, Y., He, Z., & Pan, J. (2022). Improved classification performance of EEG-fNIRS multimodal brain-computer interface based on multi-domain features and multi-level progressive learning. *Frontiers in Human Neuroscience*, *16*. https://doi.org/10.3389/fnhum.2022.973959

Quaresima, V., & Ferrari, M. (2019). Functional Near-Infrared Spectroscopy (fNIRS) for Assessing Cerebral Cortex Function During Human Behavior in Natural/Social Situations: A Concise

Review. *Organizational Research Methods*, *22*(1), 46–68. https://doi.org/10.1177/1094428116658959

Raufi, B., & Longo, L. (2022). An Evaluation of the EEG Alpha-to-Theta and Theta-to-Alpha Band Ratios as Indexes of Mental Workload. *Frontiers in Neuroinformatics*, *16*. https://doi.org/10.3389/fninf.2022.861967

Reason, J. (1997). The contribution of latent human failures to the breakdown of complex systems. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, *327*(1241), 475–484. https://doi.org/10.1098/rstb.1990.0090

Reid, G. B., & Nygren, T. E. (1988). The Subjective Workload Assessment Technique: A Scaling Procedure for Measuring Mental Workload. In P. A. Hancock & N. Meshkati (Eds.), *Advances in Psychology* (Vol. 52, pp. 185–218). North-Holland. https://doi.org/10.1016/S0166-4115(08)62387-0

Ren, M., Xu, J., Li, Y., Wang, M., Georgiev, G., Shen, L., Zhao, J., Cao, Z., Zhang, S., Wang, W., Xu, S., Zhou, Z., Chen, S., Chen, X., Shi, X., Tang, X., & Shan, C. (2023). Neural signatures for the n-back task with different loads: An event-related potential study. *Biological Psychology*, *177*, 108485. https://doi.org/10.1016/j.biopsycho.2023.108485

Riascos, J. A., Molinas, M., & Lotte, F. (2024). Machine Learning Methods for BCI: Challenges, pitfalls and promises. *ESANN 2024 Proceesdings*, 555–564. https://doi.org/10.14428/esann/2024.ES2024-4

Rivet, B., Cecotti, H., Souloumiac, A., Maby, E., & Mattout, J. (2011). Theoretical analysis of xDAWN algorithm: Application to an efficient sensor selection in a p300 BCI. *2011 19th European Signal Processing Conference*, 1382–1386. https://ieeexplore.ieee.org/document/7073970

Roscoe, A. H., & Ellis, G. A. (1990). *A Subjective Rating Scale for Assessing Pilot Workload in Flight: A Decade of Practical Use*. Royal Aerospace Establishment Farnborough.

Rosenblat, A., & Stark, L. (2016). *Algorithmic Labor and Information Asymmetries: A Case Study of Uber's Drivers* (SSRN Scholarly Paper No. 2686227). Social Science Research Network. https://doi.org/10.2139/ssrn.2686227

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Roy, R. N., Bonnet, S., Charbonnier, S., & Campagne, A. (2013). Mental fatigue and working memory load estimation: Interaction and implications for EEG-based passive BCI. *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 6607–6610. https://doi.org/10.1109/EMBC.2013.6611070

Roy, R. N., Charbonnier, S., & Bonnet, S. (2014). Detection of mental fatigue using an active BCI inspired signal processing chain. *IFAC Proceedings Volumes*, *47*(3), 2963–2968. https://doi.org/10.3182/20140824-6-ZA-1003.00897

Roy, R. N., Charbonnier, S., Campagne, A., & Bonnet, S. (2016). Efficient mental workload estimation using task-independent EEG features. *Journal of Neural Engineering*, *13*(2), 026019. https://doi.org/10.1088/1741-2560/13/2/026019

Roy, R. N., Hinss, M. F., Darmet, L., Ladouce, S., Jahanpour, E. S., Somon, B., Xu, X., Drougard, N., Dehais, F., & Lotte, F. (2022). Retrospective on the First Passive Brain-Computer Interface Competition on Cross-Session Workload Estimation. *Frontiers in Neuroergonomics*, *3*. https://doi.org/10.3389/fnrgo.2022.838342

Roy, Y., Banville, H., Albuquerque, I., Gramfort, A., Falk, T. H., & Faubert, J. (2019). Deep learning-based electroencephalography analysis: A systematic review. *Journal of Neural Engineering*, *16*(5), 051001. https://doi.org/10.1088/1741-2552/ab260c

Ryu, K., & Myung, R. (2005). Evaluation of mental workload with a combined measure based on physiological indices during a dual task of tracking and mental arithmetic. *International Journal of Industrial Ergonomics*, *35*(11), 991–1009. https://doi.org/10.1016/j.ergon.2005.04.005

Saha, S., & Baumert, M. (2020). Intra- and Inter-subject Variability in EEG-Based Sensorimotor Brain Computer Interface: A Review. *Frontiers in Computational Neuroscience*, *13*. https://doi.org/10.3389/fncom.2019.00087

Samek, W., & Müller, K.-R. (2019). Towards Explainable Artificial Intelligence. In W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, & K.-R. Müller (Eds.), *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (Vol. 11700, pp. 5–22). Springer International Publishing. https://doi.org/10.1007/978-3-030-28954-6_1

Santiago-Espada, Y. (with Langley Research Center, & United States). (2011). *The Multi-Attribute Task Battery II (MATB-II) software for human performance and workload research: A user's guide*. National Aeronautics and Space Administration, Langley Research Center.

Scharfen, H.-E., & Memmert, D. (2024). The model of the brain as a complex system: Interactions of physical, neural and mental states with neurocognitive functions. *Consciousness and Cognition*, *122*, 103700. https://doi.org/10.1016/j.concog.2024.103700

Schlögl, A., Vidaurre, C., & Müller, K.-R. (2010). Adaptive Methods in BCI Research—An Introductory Tutorial. In B. Graimann, G. Pfurtscheller, & B. Allison (Eds.), *Brain-Computer Interfaces: Revolutionizing Human-Computer Interaction* (pp. 331–355). Springer. https://doi.org/10.1007/978-3-642-02091-9_18

Schmidt, F., Danböck, S. K., Trinka, E., Klein, D. P., Demarchi, G., & Weisz, N. (2024). *Age-related changes in "cortical" 1/f dynamics are linked to cardiac activity*. https://doi.org/10.7554/eLife.100605.1

Scholkmann, F., Kleiser, S., Metz, A. J., Zimmermann, R., Mata Pavia, J., Wolf, U., & Wolf, M. (2014). A review on continuous wave functional near-infrared spectroscopy and imaging instrumentation and methodology. *NeuroImage*, *85*, 6–27. https://doi.org/10.1016/j.neuroimage.2013.05.004

Seeber, M., Cantonas, L.-M., Hoevels, M., Sesia, T., Visser-Vandewalle, V., & Michel, C. M. (2019). Subcortical electrophysiological activity is detectable with high-density EEG source imaging. *Nature Communications*, *10*(1), 753. https://doi.org/10.1038/s41467-019-08725-w

Seidel, P., Levine, S. M., Tahedl, M., & Schwarzbach, J. V. (2020). Temporal Signal-to-Noise Changes in Combined Multislice- and In-Plane-Accelerated Echo-Planar Imaging with a 20- and 64-Channel Coil. *Scientific Reports*, *10*(1), Article 1. https://doi.org/10.1038/s41598-020-62590-y

Sense, F., & van Rijn, H. (2022). *Optimizing Fact-Learning with a Response-Latency-Based Adaptive System*. OSF. https://doi.org/10.31234/osf.io/chpgv

Shalchy, M. A., Pergher, V., Pahor, A., Van Hulle, M. M., & Seitz, A. R. (2020). N-Back Related ERPs Depend on Stimulus Type, Task Structure, Pre-processing, and Lab Factors. *Frontiers in Human Neuroscience*, *14*. https://doi.org/10.3389/fnhum.2020.549966

Shaw, E. P., Rietschel, J. C., Hendershot, B. D., Pruziner, A. L., Miller, M. W., Hatfield, B. D., & Gentili, R. J. (2018). Measurement of attentional reserve and mental effort for cognitive workload assessment under various task demands during dual-task walking. *Biological Psychology*, *134*, 39–51. https://doi.org/10.1016/j.biopsycho.2018.01.009

Shi, Y., Zhu, Y., Mehta, R. K., & Du, J. (2020). A neurophysiological approach to assess training outcome under stress: A virtual reality experiment of industrial shutdown maintenance using Functional Near-Infrared Spectroscopy (fNIRS). *Advanced Engineering Informatics*, *46*, 101153. https://doi.org/10.1016/j.aei.2020.101153

Shin, J., von Lühmann, A., Kim, D.-W., Mehnert, J., Hwang, H.-J., & Müller, K.-R. (2018). Simultaneous acquisition of EEG and NIRS during cognitive tasks for an open access dataset. *Scientific Data*, *5*(1), 180003. https://doi.org/10.1038/sdata.2018.3

Shirzhiyan, Z., Keihani, A., Farahi, M., Shamsi, E., GolMohammadi, M., Mahnam, A., Haidari, M. R., & Jafari, A. H. (2019). Introducing chaotic codes for the modulation of code modulated visual

evoked potentials (c-VEP) in normal adults for visual fatigue reduction. *PLOS ONE*, *14*(3), e0213197. https://doi.org/10.1371/journal.pone.0213197

Shukla, A. K., & Kumar, U. (2006). Positron emission tomography: An overview. *Journal of Medical Physics*, *31*(1), 13. https://doi.org/10.4103/0971-6203.25665

Silberstein, R. B., Schier, M. A., Pipingas, A., Ciorciari, J., Wood, S. R., & Simpson, D. G. (1990). Steady-State Visually Evoked Potential topography associated with a visual vigilance task. *Brain Topography*, *3*(2), 337–347. https://doi.org/10.1007/BF01135443

Simões, M., Borra, D., Santamaría-Vázquez, E., GBT-UPM, Bittencourt-Villalpando, M., Krzemiński, D., Miladinović, A., Neural_Engineering_Group, Schmid, T., Zhao, H., Amaral, C., Direito, B., Henriques, J., Carvalho, P., & Castelo-Branco, M. (2020). BCIAUT-P300: A Multi-Session and Multi-Subject Benchmark Dataset on Autism for P300-Based Brain-Computer-Interfaces. *Frontiers in Neuroscience*, *14*, 568104. https://doi.org/10.3389/fnins.2020.568104

Smit, A. S., Eling, P. A. T. M., Hopman, M. T., & Coenen, A. M. L. (2005). Mental and physical effort affect vigilance differently. *International Journal of Psychophysiology*, *57*(3), 211–217. https://doi.org/10.1016/j.ijpsycho.2005.02.001

Smith, M. E., & Gevins, A. (2005). Neurophysiologic monitoring of mental workload and fatigue during operation of a flight simulator. *Biomonitoring for Physiological and Cognitive Performance during Military Operations*, *5797*, 116–126. https://doi.org/10.1117/12.602181

Smith, M. E., Gevins, A., Brown, H., Karnik, A., & Du, R. (2001). Monitoring Task Loading with Multivariate EEG Measures during Complex Forms of Human-Computer Interaction. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *43*(3), 366–380. https://doi.org/10.1518/001872001775898287

Smith, N. J., & Kutas, M. (2015). Regression-based estimation of ERP waveforms: II. Nonlinear effects, overlap correction, and practical considerations. *Psychophysiology*, *52*(2), 169–181. https://doi.org/10.1111/psyp.12320

Smith, S. M., & Nichols, T. E. (2009). Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. *NeuroImage*, *44*(1), 83–98. https://doi.org/10.1016/j.neuroimage.2008.03.061

Splawn, J. M., & Miller, M. E. (2013). Prediction of Perceived Workload From Task Performance and Heart Rate Measures. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *57*(1), 778–782. https://doi.org/10.1177/1541931213571170

Sponheim, C., Papadourakis, V., Collinger, J. L., Downey, J., Weiss, J., Pentousi, L., Elliott, K., & Hatsopoulos, N. (2021). Longevity and Reliability of Chronic Unit Recordings using the Utah, Intracortical Multi-electrode Arrays. *Journal of Neural Engineering*, *18*(6), 10.1088/1741. https://doi.org/10.1088/1741-2552/ac3eaf

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *31*(1), 137–149. https://doi.org/10.3758/BF03207704

Stenner, T., Boulay, C., Grivich, M., Medine, D., Kothe, C., tobiasherzke, Grimm, G., xloem, Biancarelli, A., Mansencal, B., chausner, Frey, J., kyucrane, Powell, S., Clisson, P., & phfix. (2022). *sccn/liblsl: V1.16.0* (Version v1.16.0) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.6387090

Strijkstra, A. M., Beersma, D. G. M., Drayer, B., Halbesma, N., & Daan, S. (2003). Subjective sleepiness correlates negatively with global alpha (8–12 Hz) and positively with central frontal theta (4–8 Hz) frequencies in the human resting awake electroencephalogram. *Neuroscience Letters*, *340*(1), 17–20. https://doi.org/10.1016/S0304-3940(03)00033-8

Sugimoto, F., Kimura, M., & Takeda, Y. (2022). Investigation of the optimal time interval between task-irrelevant auditory probes for evaluating mental workload in the shortest possible time. *International Journal of Psychophysiology: Official Journal of the International Organization of Psychophysiology*, *177*, 103–110. https://doi.org/10.1016/j.ijpsycho.2022.04.013

Sun, Y., Ayaz, H., & Akansu, A. N. (2015). Neural correlates of affective context in facial expression analysis: A simultaneous EEG-fNIRS study. *2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 820–824. https://doi.org/10.1109/GlobalSIP.2015.7418311

Tachtsidis, I., & Scholkmann, F. (2016). False positives and false negatives in functional near-infrared spectroscopy: Issues, challenges, and the way forward. *Neurophotonics*, *3*(3), 031405. https://doi.org/10.1117/1.NPh.3.3.031405

Tao, D., Tan, H., Wang, H., Zhang, X., Qu, X., & Zhang, T. (2019). A Systematic Review of Physiological Measures of Mental Workload. *International Journal of Environmental Research and Public Health*, *16*(15), 2716. https://doi.org/10.3390/ijerph16152716

Taylor, S. F., & Martz, M. E. (2023). Real-time fMRI neurofeedback: The promising potential of brain-training technology to advance clinical neuroscience. *Neuropsychopharmacology*, *48*(1), 238–239. https://doi.org/10.1038/s41386-022-01397-z

Thio, B. J., & Grill, W. M. (2023). Relative contributions of different neural sources to the EEG. *NeuroImage*, *275*, 120179. https://doi.org/10.1016/j.neuroimage.2023.120179

Thomas, E., Dyson, M., & Clerc, M. (2013). An analysis of performance evaluation for motor-imagery based BCI. *Journal of Neural Engineering*, *10*(3), 031001. https://doi.org/10.1088/1741-2560/10/3/031001

Troller-Renfree, S. V., Morales, S., Leach, S. C., Bowers, M. E., Debnath, R., Fifer, W. P., Fox, N. A., & Noble, K. G. (2021). Feasibility of assessing brain activity using mobile, in-home collection of electroencephalography: Methods and analysis. *Developmental Psychobiology*, *63*(6), e22128. https://doi.org/10.1002/dev.22128

Tsang, P. S., & Velazquez, V. L. (1996). Diagnosticity and multidimensional subjective workload ratings. *Ergonomics*, *39*(3), 358–381. https://doi.org/10.1080/00140139608964470

Tuladhar, A. M., Huurne, N. T., Schoffelen, J., Maris, E., Oostenveld, R., & Jensen, O. (2007). Parieto-occipital sources account for the increase in alpha activity with working memory load. *Human Brain Mapping*, *28*(8), 785–792. https://doi.org/10.1002/hbm.20306

Tzannatos, E. (2010). Human Element and Accidents in Greek Shipping. *The Journal of Navigation*, *63*(1), 119–127. https://doi.org/10.1017/S0373463309990312

Urigüen, J. A., & Garcia-Zapirain, B. (2015). EEG artifact removal—State-of-the-art and guidelines. *Journal of Neural Engineering*, *12*(3), 031001. https://doi.org/10.1088/1741-2560/12/3/031001

Vaadia, E., & Birbaumer, N. (2009). Grand challenges of brain computer interfaces in the years to come. *Frontiers in Neuroscience*, *3*. https://doi.org/10.3389/neuro.01.015.2009

Valeriani, D., Santoro, F., & Ienca, M. (2022). The present and future of neural interfaces. *Frontiers in Neurorobotics*, *16*. https://doi.org/10.3389/fnbot.2022.953968

Van Acker, B. B., Parmentier, D. D., Vlerick, P., & Saldien, J. (2018). Understanding mental workload: From a clarifying concept analysis toward an implementable framework. *Cognition, Technology & Work*, *20*(3), 351–365. https://doi.org/10.1007/s10111-018-0481-3

Vansteensel, M. J. (2024, October 7). *Opportunities and Challenges of Implanted ECoG-based BCIs for Communication* [Keynote]. Neuroergonomics Conference, Bordeaux, France.

Varoquaux, G., Raamana, P. R., Engemann, D. A., Hoyos-Idrobo, A., Schwartz, Y., & Thirion, B. (2017). Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. *NeuroImage*, *145*, 166–179. https://doi.org/10.1016/j.neuroimage.2016.10.038

Veltman, J. A., & Gaillard, A. W. K. (1996). Physiological indices of workload in a simulated flight task. *Biological Psychology*, *42*(3), 323–342. https://doi.org/10.1016/0301-0511(95)05165-1

Verdière, K. J., Dehais, F., & Roy, R. N. (2019). Spectral EEG-based classification for operator dyads' workload and cooperation level estimation. *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, 3919–3924. https://doi.org/10.1109/SMC.2019.8913848

Vialatte, F.-B., Maurice, M., Dauwels, J., & Cichocki, A. (2010). Steady-state visually evoked potentials: Focus on essential paradigms and future perspectives. *Progress in Neurobiology*, *90*(4), 418–438. https://doi.org/10.1016/j.pneurobio.2009.11.005

Vidal-Rosas, E. E., von Lühmann, A., Pinti, P., & Cooper, R. J. (2023). Wearable, high-density fNIRS and diffuse optical tomography technologies: A perspective. *Neurophotonics*, *10*(2), 023513. https://doi.org/10.1117/1.NPh.10.2.023513

Vidaurre, C., Kawanabe, M., von Bünau, P., Blankertz, B., & Müller, K. R. (2011). Toward Unsupervised Adaptation of LDA for Brain–Computer Interfaces. *IEEE Transactions on Biomedical Engineering*, *58*(3), 587–597. https://doi.org/10.1109/TBME.2010.2093133

von Lühmann, A., Zheng, Y., Ortega-Martinez, A., Kiran, S., Somers, D. C., Cronin-Golomb, A., Awad, L. N., Ellis, T. D., Boas, D. A., & Yücel, M. A. (2021). Toward Neuroscience of the Everyday World (NEW) using functional near-infrared spectroscopy. *Current Opinion in Biomedical Engineering*, *18*, 100272. https://doi.org/10.1016/j.cobme.2021.100272

Vortmann, L.-M., Ceh, S., & Putze, F. (2022). Multimodal EEG and Eye Tracking Feature Fusion Approaches for Attention Classification in Hybrid BCIs. *Frontiers in Computer Science*, *4*. https://doi.org/10.3389/fcomp.2022.780580

Wallston, K. A., Slagle, J. M., Speroff, T., Nwosu, S., Crimin, K., Feurer, I. D., Boettcher, B., & Weinger, M. B. (2014). Operating Room Clinicians' Ratings of Workload: A Vignette Simulation Study. *Journal of Patient Safety*, *10*(2), 95. https://doi.org/10.1097/PTS.0000000000000046

Wang, G. (2019). High Temporal-Resolution Dynamic PET Image Reconstruction Using a New Spatiotemporal Kernel Method. *IEEE Transactions on Medical Imaging*, *38*(3), 664–674. https://doi.org/10.1109/TMI.2018.2869868

Wang, Y., Yang, X., Zhang, X., Wang, Y., & Pei, W. (2023). Implantable intracortical microelectrodes: Reviewing the present with a focus on the future. *Microsystems & Nanoengineering*, *9*(1), 1–17. https://doi.org/10.1038/s41378-022-00451-6

Waschke, L., Donoghue, T., Fiedler, L., Smith, S., Garrett, D. D., Voytek, B., & Obleser, J. (2021). Modality-specific tracking of attention and sensory statistics in the human electrophysiological spectral exponent. *eLife*, *10*, e70068. https://doi.org/10.7554/eLife.70068

Weigl, M., Stefan, P., Abhari, K., Wucherer, P., Fallavollita, P., Lazarovici, M., Weidert, S., Euler, E., & Catchpole, K. (2016). Intra-operative disruptions, surgeon's mental workload, and technical performance in a full-scale simulated procedure. *Surgical Endoscopy*, *30*(2), 559–566. https://doi.org/10.1007/s00464-015-4239-1

Weiskopf, N., Sitaram, R., Josephs, O., Veit, R., Scharnowski, F., Goebel, R., Birbaumer, N., Deichmann, R., & Mathiak, K. (2007). Real-time functional magnetic resonance imaging: Methods and applications. *Magnetic Resonance Imaging*, *25*(6), 989–1003. https://doi.org/10.1016/j.mri.2007.02.007

Weiss, S., & Mueller, H. M. (2012). "Too Many betas do not Spoil the Broth": The Role of Beta Brain Oscillations in Language Processing. *Frontiers in Psychology*, *3*. https://doi.org/10.3389/fpsyg.2012.00201

White, J., & Power, S. D. (2023). k-Fold Cross-Validation Can Significantly Over-Estimate True Classification Accuracy in Common EEG-Based Passive BCI Experimental Designs: An Empirical Investigation. *Sensors*, *23*(13), Article 13. https://doi.org/10.3390/s23136077

Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science*, *3*(2), 159–177. https://doi.org/10.1080/14639220210123806

Wickens, C. D. (2008). Multiple resources and mental workload. *Human Factors*, *50*(3), 449–455. https://doi.org/10.1518/001872008X288394

Wierwille, W. W., & Eggemeier, F. T. (1993). Recommendations for Mental Workload Measurement in a Test and Evaluation Environment. *Human Factors*, *35*(2), 263–281. https://doi.org/10.1177/001872089303500205

Wiggins, C., Poser, B., Mauconduit, F., Boulant, N., & Gras, V. (2019). Universal Pulses for MRI at 9.4 Tesla—A Feasibility Study. *2019 International Conference on Electromagnetics in Advanced Applications (ICEAA)*, 1185–1188. https://doi.org/10.1109/ICEAA.2019.8879180

Wijk, G. van der, Enkhbold, Y., Cnudde, K., Szostakiwskyj, M. W., Blier, P., Knott, V., Jaworska, N., & Protzner, A. B. (2021). *One size does not fit all: Single-subject analyses reveal substantial individual variation in electroencephalography (EEG) characteristics of antidepressant treatment response* (p. 2020.11.09.20227280). medRxiv. https://doi.org/10.1101/2020.11.09.20227280

Williams, N. S., McArthur, G. M., Wit, B. de, Ibrahim, G., & Badcock, N. A. (2020). A validation of Emotiv EPOC Flex saline for EEG and ERP research. *PeerJ*, *8*, e9713. https://doi.org/10.7717/peerj.9713

Wilson, G. F., & McCloskey, K. (1988). Using Probe Evoked Potentials to Determine Information Processing Demands. *Proceedings of the Human Factors Society Annual Meeting*, *32*(19), 1400–1403. https://doi.org/10.1177/154193128803201920

Wilson, G. F., & O'Donnell, R. D. (1988). Measurement of Operator Workload with the Neuropsychological Workload Test Battery. In *Advances in Psychology* (Vol. 52, pp. 63–100). Elsevier. https://doi.org/10.1016/S0166-4115(08)62383-3

Wilson, M. R., Poolton, J. M., Malhotra, N., Ngo, K., Bright, E., & Masters, R. S. W. (2011). Development and Validation of a Surgical Workload Measure: The Surgery Task Load Index (SURG-TLX). *World Journal of Surgery*, *35*(9), 1961–1969. https://doi.org/10.1007/s00268-011-1141-4

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, *5*(2), 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1

Wolpert, D. H. (1996a). The Existence of A Priori Distinctions Between Learning Algorithms. *Neural Computation*, *8*(7), 1391–1420. https://doi.org/10.1162/neco.1996.8.7.1391

Wolpert, D. H. (1996b). The Lack of A Priori Distinctions Between Learning Algorithms. *Neural Computation*, *8*(7), 1341–1390. https://doi.org/10.1162/neco.1996.8.7.1341

Woodman, G. F. (2010). A brief introduction to the use of event-related potentials in studies of perception and attention. *Attention, Perception & Psychophysics*, *72*(8), 2031–2046. https://doi.org/10.3758/APP.72.8.2031

Wu, J., Zhou, Q., Li, J., Chen, Y., Shao, S., & Xiao, Y. (2021). Decreased resting-state alpha-band activation and functional connectivity after sleep deprivation. *Scientific Reports*, *11*(1), 484. https://doi.org/10.1038/s41598-020-79816-8

Xing, X., Pei, W., Wang, Y., Guo, X., Zhang, H., Xie, Y., Gui, Q., Wang, F., & Chen, H. (2018). Assessing a novel micro-seepage electrode with flexible and elastic tips for wearable EEG acquisition. *Sensors and Actuators A: Physical*, *270*, 262–270. https://doi.org/10.1016/j.sna.2017.12.048

Yamamoto, M. S., Lotte, F., Yger, F., & Chevallier, S. (2022, July). Class-distinctiveness-based frequency band selection on the Riemannian manifold for oscillatory activity-based BCIs: Preliminary results. *EMBC 2022- 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society*. https://doi.org/10.1109/EMBC48229.2022.9871820

Yamamoto, M. S., Mellot, A., Chevallier, S., & Lotte, F. (2023). Novel SPD Matrix Representations Considering Cross-Frequency Coupling for EEG Classification Using Riemannian Geometry. *2023 31st European Signal Processing Conference (EUSIPCO)*, 960–964. https://doi.org/10.23919/EUSIPCO58844.2023.10290043

Yang, S.-Y., & Lin, Y.-P. (2023). Movement Artifact Suppression in Wearable Low-Density and Dry EEG Recordings Using Active Electrodes and Artifact Subspace Reconstruction. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *31*, 3844–3853. https://doi.org/10.1109/TNSRE.2023.3319355

Yeh, Y.-Y., & Wickens, C. D. (1988). Dissociation of Performance and Subjective Measures of Workload. *Human Factors*, *30*(1), 111–120. https://doi.org/10.1177/001872088803000110

Yger, F., Berar, M., & Lotte, F. (2017). Riemannian Approaches in Brain-Computer Interfaces: A Review. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *25*(10), 1753–1762. IEEE Transactions on Neural Systems and Rehabilitation Engineering. https://doi.org/10.1109/TNSRE.2016.2627016

Young, M. S., Brookhuis, K. A., Wickens, C. D., & Hancock, P. A. (2015). State of science: Mental workload in ergonomics. *Ergonomics*, *58*(1), 1–17. https://doi.org/10.1080/00140139.2014.956151

Zakrzewska, M. Z., & Brzezicka, A. (2014). Working memory capacity as a moderator of load-related frontal midline theta variability in Sternberg task. *Frontiers in Human Neuroscience*, *8*. https://doi.org/10.3389/fnhum.2014.00399

Zander, T. O., & Kothe, C. (2011). Towards passive brain-computer interfaces: Applying brain-computer interface technology to human-machine systems in general. *Journal of Neural Engineering*, *8*(2), 025005. https://doi.org/10.1088/1741-2560/8/2/025005

Zander, T. O., Krol, L. R., Birbaumer, N. P., & Gramann, K. (2016). Neuroadaptive technology enables implicit cursor control based on medial prefrontal cortex activity. *Proceedings of the National Academy of Sciences*, *113*(52), 14898–14903. https://doi.org/10.1073/pnas.1605155114

Zanini, P., Congedo, M., Jutten, C., Said, S., & Berthoumieu, Y. (2018). Transfer Learning: A Riemannian Geometry Framework With Applications to Brain–Computer Interfaces. *IEEE Transactions on Biomedical Engineering*, *65*(5), 1107–1116. https://doi.org/10.1109/TBME.2017.2742541

Zhang, H., Geng, X., Wang, Y., Guo, Y., Gao, Y., Zhang, S., Du, W., Liu, L., Sun, M., Jiao, F., Yi, F., Li, X., & Wang, L. (2021). The Significance of EEG Alpha Oscillation Spectral Power and Beta Oscillation Phase Synchronization for Diagnosing Probable Alzheimer Disease. *Frontiers in Aging Neuroscience*, *13*. https://doi.org/10.3389/fnagi.2021.631587

Zhao, Y., Tang, J., Cao, Y., Jiao, X., Xu, M., Zhou, P., Ming, D., & Qi, H. (2018). Effects of Distracting Task with Different Mental Workload on Steady-State Visual Evoked Potential Based Brain Computer Interfaces—An Offline Study. *Frontiers in Neuroscience*, *12*. https://www.frontiersin.org/article/10.3389/fnins.2018.00079

Zijlstra, F., & Doorn, L. (1985). The Construction of a Scale to Measure Perceived Effort. *Department of Philosophy and Social Sciences*.

Zimeo Morais, G. A., Balardin, J. B., & Sato, J. R. (2018). fNIRS Optodes' Location Decider (fOLD): A toolbox for probe arrangement guided by brain regions-of-interest. *Scientific Reports*, *8*(1), 3341. https://doi.org/10.1038/s41598-018-21716-z