Improving methods of diagnosis and prognostication in neurodegenerative diseases through electroencephalography analysis and machine learning

Robin Vlieger

A thesis submitted for the degree of
Doctor of Philosophy at
The Australian National University

April 2024

Except where otherwise indicated, this thesis is my own original work.

Robin Vlieger

5 April 2024

Contents

Acknowledgements ix

COVID statement xi

Abstract xii

Publications xv

List of Abbreviations xvi

Chapter 1 Introduction 1

1.1 The necessity of earlier diagnosis of Parkinson's disease 2

1.2 The role of machine learning in Parkinson's disease diagnosis 3

1.3 Thesis objectives 4

1.4 Thesis contributions 5

1.5 Thesis outline 6

Chapter 2 Background 8

2.1 Parkinson's disease 8

2.1.1 A description of Parkinson's disease 9

2.1.2 Diagnosis and prognostication of Parkinson's disease 10

2.2 Electroencephalography 11

2.2.1 A description of electroencephalography 12

2.2.2 Signal processing of electroencephalography 16

2.2.3 Electroencephalography in Parkinson's disease 20

2.3 Machine learning 22

2.3.1 Machine learning tasks and evaluation metrics 22

2.3.2 Machine learning algorithms used in Parkinson's disease classification using electroencephalography data 24

2.3.3 Training and evaluating a machine learning model 28

2.3.4 Statistical analysis 30

2.4 Machine learning and electroencephalography in Parkinson's disease 31

2.5 Electroencephalography datasets 33

2.6 A broader perspective: Parkinson's disease and Multiple Sclerosis 35

2.7 Multiple Sclerosis 35

2.8 Chapter summary and conclusion 36

Chapter 3 Preprocessing resting-state electroencephalography data and its effects on classification metrics 38

3.1 Evaluating effects of resting-state electroencephalography data preprocessing on a machine learning classification task for Parkinson's disease 39

3.1.1 Introduction 39

3.1.2 Methods 40

3.1.3 Results 44

3.1.4 Discussion 46

3.1.5 Conclusion 48

3.2 Preprocessing without averaging epochs 48

3.2.1 Introduction 48

3.2.2 Methods 48

3.2.3 Results 49

3.2.4 Discussion 51

3.2.5 Conclusion 52

3.3 Chapter summary and conclusion 53

Chapter 4 Effects of using single or averaged epochs and electrodes from different resting-state electroencephalography datasets 54

4.1 Introduction 55

4.2 Methods 56

4.3 Results 58

4.4 Discussion 60

4.5 Conclusion 62

4.6 Chapter summary and conclusion 63

Chapter 5 Effects of sample randomisation and the moment of feature selection 65

5.1 The use of event-related potentials and machine learning to improve diagnostic testing and prognostication in Parkinson's disease 66

5.1.1 Introduction 66

5.1.2 Methods 67

5.1.3 Results 69

5.1.4 Discussion 70

5.1.5 Conclusion 70
5.2 When the science catches up 71
5.2.1 Introduction 71
5.2.2 Methods 72
5.2.3 Results 73
5.2.4 Discussion 76
5.2.5 Conclusion 77
5.3 Chapter summary and conclusion 77

Chapter 6 Effect of participant characteristics on classification outcomes 79
6.1 Introduction 80
6.2 Methods 80
6.3 Results 82
6.4 Discussion 84
6.5 Conclusion 87
6.6 Chapter summary and conclusion 88

Chapter 7 Event-related potentials in neurodegenerative diseases 89
7.1 The use of event-related potentials in the investigation of cognitive performance in people with Multiple Sclerosis: systematic review 90
7.1.1 Introduction 91
7.1.2 Methods 93
7.1.3 Results 94
7.1.4 Discussion 103
7.1.5 Conclusion 106
7.2 A comparison to event-related potentials research in Parkinson's disease 107
7.3 Chapter summary and conclusion 108

Chapter 8 Conclusion 110
8.1 Thesis summary 111
8.2 Thesis contributions 113
8.2.1 Preprocessing pipeline 113
8.2.2 Data handling and feature selection 115

8.2.3 The importance of participant characteristics 117
8.2.4 The broader context of neurodegenerative diseases 119
8.3 Limitations 120
8.4 Future work 121
8.5 Broader impact 123
8.6 Final remarks 123
References 125
Supplement 1 146
Supplement 2 152
Supplement 3 155
Supplement 4 159
Supplement 5 161
Supplement 6 164

Acknowledgements

I wish to thank the School of Computing of the College of Engineering, Computing & Cybernetics for their support, as this thesis could not have been written without them. My gratitude also goes out to the Australian government for their financial support through the PhD stipend and tuition fee waiver programmes. Additionally, I would like to thank ACT Health and The Canberra Hospital for their invaluable support. Lastly, this thesis would not have been possible without the support of and collaboration with all the wonderful researchers and staff working on the 'Our Health in Our Hands' (OHIOH) grand challenge.

I am indebted to Professor Hanna Suominen of the School of Computing, for her unwavering support and for giving me the opportunity to join the Big Data team of OHIOH as a neuroscience postgraduate with limited computer science experience. I would like to thank Dr Elena Daskalaki of the School of Computing, who helped me develop into a researcher confident in his machine learning skills. My gratitude goes out to Professor Christian J. Lueck of the School of Medicine and Psychology, who, with a simple email, first made me believe I should do a PhD and has been invaluable in helping me develop my ideas and writing. Lastly, my thanks go out to Dr Deborah Apthorp of the School of Psychology of the University of New England, whose studies into Parkinson's disease were the start of my own studies and without whose contributions and knowledge this thesis would not exist.

As my research is based on electroencephalogram recordings of people with neurological diseases and controls, I wish to thank all participants that made the time and put in the effort to make it possible for researchers to collect data. Similarly, I am indebted to the researchers at the University of New Mexico, USA, and the University of Turku, Finland, who collected the data that I have extensively used in my research. I would also like to acknowledge Dianne Walton-Sonda, who, as a research librarian at The Canberra Hospital, helped me tremendously in developing my literature research skills.

I am very fortunate to have met many wonderful people at the ANU, and my time at the ANU would not have been as enjoyable as it has been if it had not been for my wonderful fellow PhD candidates, some of whom are now Doctors of Philosophy in their own right. Thank you, Nicholas Kuo, Chirath Hettiarachchi, Sandaru Seneviratne, and Ran Cui for being wonderful colleagues and friends.

Doing a PhD can be very stressful, but fortunately I found a group of people outside of academia that have been fabulous in helping me take my mind off things. I would like to thank all the members of Chave de Ouro, the Brazilian Jiujitsu academy I have proudly been a member of since late 2018, its founder and owner Renato, and Marina for creating one of the best communities I have ever been part.

My family has been supportive throughout. Without my parents, Wim and Tilly, I would not be where I am now, supporting and encouraging me academically. Their wisdom and the lessons they have

taught me guide me in everything I do, and I could not have wished for better role models. My gratitude goes out to my mother-in-law, Jenny, who is an inspiration to me every day. The obstacles in my way pale in comparison to the ones she has overcome and the positivity with which she faces life is the stuff of legends. Lastly, there are no words in any language known to man that can express my gratitude to and the importance of my wife Sarah, the best partner I could wish for, and my support in everything I do. Without her, none of this would have been possible.

COVID Statement

It was July 2019 when I started my PhD, and, like many others, I had great expectations of what my research and thesis would look like. I had been working as a research officer at The Canberra Hospital for a few months and I was excited to use my experience there to investigate postural sway, electroencephalography, eye movements, and cognitive tests in people with Parkinson's disease (PD) and in people with Multiple Sclerosis (MS). The research into PD had been going on for a while and the prospect of comparing and contrasting the two diseases to reach a better understanding of both was exciting, not only to me but to everyone involved. So, my PhD took off: I got the laboratory at the hospital ready, prepared all the necessary documents such as information sheets and consent forms, and applied for Ethics Committee approval.

It was January 2020 when I received the Ethics approval. I had just returned from Europe and was ready to commence the experiments. At that point the news had started reporting on a novel, potentially fatal, respiratory disease which was spreading quickly but I was still optimistic and, like most of us, had no idea about what was to come. By March it had become very clear: the virus had well and truly taken a foothold in Australia, and Canberra went into its first lockdown. The chances of carrying out any in-person experiments, let alone any experiments involving medically compromised groups, were reduced to absolute zero as a result. My research plan needed a serious and complete overhaul if I were to complete this PhD.

Months of reading, planning, weighing up possibilities, and meeting with my supervisory panel followed. At this point it was unclear when restrictions would be eased, whether I would be able to do any gathering of human data at all during the course of my PhD and, if so, whether this would happen in time for me to include it in my thesis. There is often light at the end of the tunnel and, after finding additional open-access datasets, pivoting my thesis towards doing the most with the data that had already been obtained, I was once again on track.

On the other side of the world things were, unfortunately, not going that well. My home country of the Netherlands had not been managing the COVID situation nearly as well as Australia. My mother works as a sector manager in the largest hospital in the country and she was overworked to an extent that had not been seen ever before. In addition, several of my family members ended up in the Intensive Care Unit, while others suffered long-term effects from COVID. In all, it was a very stressful time for me.

This thesis reports the sum of my work over the last four years during this very difficult time. While the times were troubling, the situation did help to make my work more focused. I hope you enjoy reading it as much as I enjoyed writing it.

Abstract

Parkinson's disease (PD) is a neurodegenerative disorder affecting 3% of the world's population. Its incidence increases with age, and it is a source of major disability. Currently, there is no definitive diagnostic test. Instead, clinical diagnosis is based on the detection of relevant motor features, the exclusion of other diseases, and the response to dopaminergic medication. Identification of clinically useful biomarkers of PD is the subject of widespread research because such biomarkers would be of enormous benefit to clinicians treating *people with PD* (PwPD): they could potentially enable earlier diagnosis, allow tailoring of treatment to individuals in the form of personalised treatment plans, and provide a more accurate prognosis.

One potential biomarker of interest to researchers is *electroencephalography* (EEG). EEG data are multidimensional and are collected over several minutes at high sampling rates, often over 50 Hz, which gives rise to very large datasets. Consequently, *machine learning* (ML) methods have been applied to the data, with the aim of developing diagnostic and/or prognostic models of PD. Published studies have used many different methods to clean and preprocess the EEG data they report, ranging from nothing (i.e., using raw data) to advanced filtering and artifact rejection, featurisation (including different methods of using single time-series and electrodes, or averaging them), classification (using conventional algorithms such as *Random Forest Classifiers* (RFCs) and *Support Vector Machines* (SVMs), and deep learning algorithms such as *recurrent* and *convolutional neural networks* (RNNs and CNNs). The large variety of methods used by different investigators makes it difficult to compare or combine results across studies and, in addition, no-one has yet investigated how the different methods of cleaning and preprocessing might influence the accuracy of classification tasks.

In this interdisciplinary PhD thesis in predominantly computer science, as well as clinical neuroscience, I report on five experimental studies and a systematic literature review examining the effects of varying methods of data collection, preprocessing methods, methods that use averaging of electrodes and samples (to increase the *signal-to-noise ratio,* SNR), or those that use single electrodes and samples (to increase the size of the dataset) and the use of different ML classification algorithms on performance metrics in classification tasks of people with a neurodegenerative disease and controls. The aim was to provide insight into the precise effects of using different methods and, ultimately, to offer recommendations for collecting, processing, and analysing data, with a particular focus on using EEG to develop diagnostic and prognostic models for PD in the future. Specifically, these studies investigated methods of data collection, cleaning, outlier removal, data handling and feature selection, as well as clarifying the effects of demographics such as *Movement Disorder Society-Unified Parkinson's Disease Rating Scale* (MDS-UPDRS) score, sex, and age on classification metrics.

Resting-state EEG (RSEEG) analysis and ML were addressed in my first three studies. In my first study, I looked at RSEEG and investigated the effect of using different preprocessing methods on classification accuracy. My findings indicated that, ideally, preprocessing should include filtering, electrode removal, and interpolation. In my second study, I investigated four combinations of different methods of using single electrodes and epochs, and methods of averaging them that are commonly used in ML and EEG research. From this experiment, I concluded that analysis of features extracted from regions of interest (created by grouping electrodes based on their location) performed better than analysis of features derived from single electrodes.

In my third study, I reported on ML experiments using *event-related potentials* (ERPs) instead of RSEEG to classify PwPD and controls in the context of a Go/No-Go task, a task in which participants had to press a button when a particular stimulus appeared but had to refrain from pressing it when another appeared. Focusing on the *Contingent Negative Variation* (CNV), an ERP that appears in anticipation of a motor response, the experiment showed that the difference between early and late stages of the CNV can be used to distinguish PwPD from controls.

In my fourth study, I investigated the pitfalls of the various methods of sample randomisation which have been applied to RSEEG data, the dangers of data leakage, and the effects that these can have on experimental results. I demonstrated that performing both feature selection outside of cross-validation and randomising samples when using single epochs from individual participants can positively bias classification results. In my investigation, this bias was larger when the number of features per sample was smaller.

In my fifth study, I investigated the effects of participant demographics on classification outcome. My results provided evidence that classification outcomes are strongly influenced by sex, age, and severity of PD expressed as UPDRS score, indicating a need for matching sex, age, and disease severity in the experimental design of future studies.

Finally, a literature review into the use of ERPs to investigate cognitive tasks in *people with Multiple Sclerosis* (PwMS) has been included as my sixth study to begin to extrapolate the conclusions to neurodegenerative diseases in general. As indicated in my COVID Statement above, the initial plan for this PhD thesis was to include experiments investigating both PwPD and PwMS, but the COVID-19 pandemic made data collection from PwMS impossible. This literature review is compared and contrasted with a similar review into ERPs in PD by other researchers.

The findings in this thesis have demonstrated the potential for using EEG, especially RSEEG, as a tool for classifying PwPD and controls, and hence its potential to assist clinicians in managing PwPD. The findings have also highlighted the shortcomings of the methods used in many previous published studies. By providing a detailed analysis of the factors that may prevent optimal classification, this thesis has been able to provide recommendations regarding the most appropriate way

to collect and preprocess EEG data, thereby assisting future research into its potential use as a biomarker of PD and MS.

Keywords: diagnosis, electroencephalography, evaluation study, event-related potentials, machine learning, Multiple Sclerosis, neurodegenerative diseases, Parkinson’s disease

Publications

Included in this thesis:

Vlieger, R., Austin, D., Apthorp D., Daskalaki, E., Lenskiy, A., Walton-Sonda, D., Suominen, H., Lueck, C.J. (2024). The use of event-related potentials in the investigation of cognitive performance in people with Multiple Sclerosis: systematic review, *Brain Research*, 148827.

Vlieger, R., Daskalaki, E., Apthorp, D., Lueck, C. J., & Suominen, H. (2024). Evaluating Effects of Resting-State Electroencephalography Data Pre-Processing on a Machine Learning Task for Parkinson's Disease. *Studies in Health Technology and Informatics*, 310, 1480–1481.

Vlieger, R., Suominen, H., Apthorp, D., Lueck, C.J., Daskalaki, E. (2023). Evaluating methods of oversampling and averaging resting-state electroencephalography data in classifying Parkinson's disease, *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, *2023*, 1-5

Vlieger, R., Daskalaki, E., Apthorp, D., Lueck, C.J., & Suominen, H. (2021). The use of event-related potentials and machine learning to improve diagnostic testing and prediction of disease progression in Parkinson's disease. *Studies in Health Technology and Informatics*, 284, 333-335.

Additional publications not related to the PhD:

Lensky, A., Lueck, C., Suominen, H., Jones, B., Vlieger, R., & Ahluwalia, T. (2024). Explaining predictors of discharge destination assessed along the patients' acute stroke journey. *Journal of Stroke and Cerebrovascular Diseases*, 33(2), 107514.

Rai, B.B., Van Kleef, J.P., Sabeti, F., Vlieger, R., Suominen, H., Maddess, T. (2023). Early diabetic eye damage: comparing detection methods using diagnostic power. *Survey of Ophthalmology*, 69(1), 24-33

Ahluwalia, T., Lenskiy, A., Jones, B., Vlieger, R., Suominen, H., & Lueck, C. (2022). Predicting discharge destination in acute stroke patients using machine learning. *International Journal of Stroke*, 17(2_Suppl), 18 (Published abstract).

Apthorp, D., Smith, A., Ilschner, S., Vlieger, R., Das, C., Lueck, C. J., & Looi, J. C. (2020). Postural sway correlates with cognition and quality of life in Parkinson's disease. *BMJ Neurology Open*, 2(2), 1-7.

List of Abbreviations

ACT	Australian Capital Territory
AD	Alzheimer's disease
ANN	artificial neural network
ANOVA	Analysis of Variance
ANT	Attention Network Test
ANU	Australian National University
ASR	Artifact Subspace Reconstruction
AUC	Area Under the Curve
BDI	Beck's Depression Inventory
CH	Canberra Hospital
CI	confidence interval
CINAHL	Cumulative Index of Nursing and Allied Health Literature
CI	confidence interval
CIS	Clinically Isolated Syndrome
CNN	convolutional neural network
CNV	contingent negative variation
CT	computed tomography
CV	cross-validation
DFA	discriminant function analysis
DL	deep learning
EC	eyes closed
EDSS	Expanded Disability Status Scale
EMBC	Annual International Conference of the IEEE Engineering in Medicine and Biology Society
EEG	electroencephalography
EMG	electromyography
EO	eyes open
ERN	Error-Related Negativity
ERP	event-related potential
ES	effect size
FDA	Food and Drug Administration
fMRI	functional magnetic resonance imaging
FPR	false positive rate
GBC	gradient boosting classifier
HD	Huntington's disease
HOS	higher-order spectra
IC	independent component
ICA	Independent Component Analysis
IEEE	Institute of Electrical and Electronics Engineers
kNN	k-Nearest Neighbours
LOOCV	leave-one-out cross-validation
MAE	Mean Absolute Error
MARA	Multiple Artifact Rejection Algorithm

MDS-UPDRS	Movement Disorder Society – Unified Parkinson's Disease Rating Scale
ML	machine learning
MMN	mismatch negativity
MMSE	Mini Mental State Exam
MRI	magnetic resonance imaging
MS	Multiple Sclerosis
MSE	Mean Squared Error
NIH	National Institutes of Health
OHIOH	Our Health in Our Hands
PD	Parkinson's disease
PET	positron emission tomography
PPMS	primary progressive Multiple Sclerosis
PPV	positive predictive value
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROSPERO	International prospective register of systematic reviews
PwMS	people with Multiple Sclerosis
PwPD	people with Parkinson's disease
RANSAC	random sample consensus
RBD	rapid eye movement sleep behaviour disorder
RBF	radial basis function kernel
RFC	Random Forest Classifier
RMSE	Root Mean Squared Error
RoBANS	Risk of Bias Assessment Tool for Non-randomized Studies
ROC	Receiver Operating Characteristic
ROI	region-of-interest
RRMS	relapsing remitting Multiple Sclerosis
RSEEG	resting-state electroencephalography
SNpc	substantia nigra pars compacta
SNR	signal-to-noise ratio
SPMS	secondary progressive Multiple Sclerosis
SVM	Support Vector Machine
TPR	true positive rate
UK	United Kingdom
UNM	University of New Mexico
UPDRS	Unified Parkinson's Disease Rating Scale
US	United States
UTU	University of Turku
WHO	World Health Organization

Chapter 1: Introduction

Parkinson's disease (PD) is the second most prevalent neurodegenerative disorder after *Alzheimer's disease* (AD; Nutt & Wooten, 2005), affecting 3% of the world's population. Its incidence increases with age, and it is a source of major disability. Currently, there is no definitive diagnostic test (Jankovic, 2008; Kalia & Lang, 2015). Instead, clinical diagnosis is based on the detection of relevant motor features, the exclusion of other diseases, and the response to dopaminergic medication. Identification of clinically useful biomarkers of PD is the subject of widespread research (Wu, Le, & Jankovic, 2011) as such biomarkers would be of enormous benefit to clinicians treating *people with PD* (PwPD). They could potentially enable earlier diagnosis, allow tailoring of treatment to individuals in the form of personalised treatment plans, and provide more accurate prognosis.

One potential biomarker of interest to researchers is *electroencephalography* (EEG; Biasiucci, Franceschiello, & Murray, 2019). EEG data are multidimensional and are collected over several minutes at high sampling rates, often over 50 Hz, and this gives rise to datasets with large numbers of measurements. In response, *machine learning* (ML) methods have been applied to the data, aiming to develop diagnostic and/or prognostic models of PD (Maitín, García-Tejedor, & Romero Muñoz, 2020; Maitín, Romero Muñoz, & García-Tejedor, 2022). The published studies have used many different methods to clean and preprocess the EEG data, ranging from no preprocessing (i.e., using raw data) to advanced filtering and artifact rejection, methods to increase dataset size and improve the *signal-to-noise ratio* (SNR), and classification methods (using conventional algorithms such as *Random Forest Classifiers* (RFCs; Ho, 1995; Breiman, 2001), *Support Vector Machines* (SVMs; Noble, 2006), and *deep learning* (DL) algorithms (Jain, Mao, & Mohiuddin, 1996; Krogh, 2008)). The large variety of methods used by different investigators makes it difficult to compare or combine results across studies and, in addition, how the different methods of preprocessing might influence the accuracy of classification tasks has not been investigated in detail.

The work presented here is interdisciplinary in nature, being in computer science, as well as clinical neuroscience. Specifically, the work in this thesis can be best described as integrative applied research, defined by Gabriele Bammer in her book "Disciplining Interdisciplinarity" (2013) as "a research style that deals with complex real-world problems by bringing together disciplinary and stakeholder knowledge and explicitly dealing with remaining unknowns, in order to use that integrated research to support policy and practice change." For this thesis, this means the analysis and integration of methods commonly used in the fields of computer science and clinical neuroscience, the expert knowledge of clinicians, and the lived experiences of people with neurodegenerative diseases.

In this interdisciplinary PhD thesis in predominantly computer science and clinical neuroscience, five studies will be discussed that examine the effects of varying preprocessing methods,

methods to increase dataset size and improve SNR, and ML classification algorithms on performance metrics in classification tasks of PwPD and controls. The aim has been to provide insight into the precise effects of using different preprocessing methods, to investigate how differences in datasets, such as those obtained during recording with *eyes open* (EO) or *eyes closed* (EC) lead to different outcomes, and to clarify how participant characteristics influence classification outcomes.

1.1 The necessity of earlier diagnosis of Parkinson's disease

In an ideal world, it would be possible to state definitively whether an individual did, or did not, have a particular disease. This information could potentially be used to inform diagnosis, prognosis, choice of therapy, and to assess any response to therapy. It is possible to define some diseases in this way but, in practice, many conditions cannot be defined so precisely. For example, definitive diagnosis may require histological examination of brain tissue which cannot usually be obtained while a patient is alive. In this situation, it becomes necessary to work on the basis of the best available standard for diagnosis. In the field of medicine, this is generally referred to as the 'gold standard' (Versi, 1992). Gold standards such as this are not absolute: as a result of medical and scientific progress, gold standards can be, and often are, superseded by newer versions that are considered to be more accurate. Therefore, medical gold standards should be understood in the same way the gold standard is understood in the field of finance: just as a currency can be compared to the value of gold at any given time point, a gold standard in medicine is the benchmark to which other methods of diagnosis and prognostication can be compared at a particular time (Claassen, 2005).

At present, there is no clinically useful definitive diagnostic test for PD. Definite disease can only be confirmed by the finding of Lewy body pathology and degeneration of the *substantia nigra pars compacta* (SNpc) in a pathological examination after death (Kalia & Lang, 2015). Clinical diagnosis is therefore based on the presence of features such as rest tremor, increased tone, posture and gait impairment, and bradykinesia, the exclusion of other diseases, and the response to medicinal dopaminergic treatment (Postuma et al., 2015). Unfortunately, by the time motor features become apparent, the SNpc has already lost at least 50% of its dopaminergic neurons (Fearnley & Lees, 1991), which means the disease has already been progressing for many years.

In terms of a gold standard, the *Unified Parkinson's Disease Rating Scale* (UPDRS) is the most commonly used tool for assessing disease state in people with PD (Ramaker et al., 2002). The UPDRS was introduced in 1987 and revised in 2008 by the *Movement Disorder Society* as the MDS-UPDRS (Goetz et al., 2008) in response to criticism highlighting several weaknesses in the original version, such as an overemphasis on motor symptoms and ambiguously phrased questions. The MDS-UPDRS was assessed for internal validity at the time of its initial release, and the Spanish version has subsequently been independently validated, confirming its reliability, internal validity, and precision (Martinez-Martin et al., 2013).

However, as Evers et al. (2019) point out, the aforementioned studies have only shown that the MDS-UPDRS can be used to distinguish between different people with PD at a specific timepoint. Applying a linear Gaussian state space model to a dataset of repeated MDS-UPDRS measurement from early-stage PwPD, they demonstrated substantial intra-subject variability when predicting follow-up scores based on baseline scores. Therefore, the MDS-UPDRS cannot be used as a reliable tool to predict disease progression in an individual. Accordingly, the MDS-UPDRS is not a good 'gold standard' for the purpose of prognostication and, as with many diseases, clinicians and scientists are searching for an alternative tool.

In summary, PD lacks a reliable clinical gold standard for diagnosis and prognostication: post-mortem validation is not clinically useful and the MDS-UPDRS cannot be used to predict disease progression with any certainty. Unfortunately, this also means that the MDS-UPDRS cannot be used to benchmark any potential candidate gold standards: if the MDS-UPDRS does not reliably measure disease progression, comparing new measures to it becomes a futile endeavour. What is required instead is a method that can track disease progression more accurately and objectively over time, preferably starting before the cardinal symptoms of the disease appear, for example by measuring the physiological changes in the nervous system that occur in PwPD. One such method under consideration is the use of EEG (Biasiucci, Franceschiello, & Murray, 2019), the measurement of the electrophysiological activity of the brain through electrodes placed on the scalp. Differences in EEG signals have been observed between PwPD and controls as well as people at different stages in their disease (Livia Fantini et al., 2003; Morita et al., 2009). Abnormalities have also been found in people with *rapid eye movement sleep behaviour disorder* (RBD), a condition which often presages PD (Schenck, Bundlie, & Mahowald, 1996; Livia Fantini et al., 2003), lending credence to the idea that EEG could be used to track disease development.

1.2 The role of machine learning in Parkinson's disease diagnosis

In an effort to develop diagnostic models of PD, researchers have used ML methods to classify participants based on their EEG data (Maitín, García-Tejedor, & Romero Muñoz, 2020; Maitín, Romero Muñoz, & García-Tejedor, 2022). The methods used in the ML studies vary widely, beginning with data collection: some studies have used *resting-state EEG* (RSEEG) data collected while the participants were either sitting or lying down in a relaxed state, while others have used data collected during specific tasks. Because EEG data are collected from the scalp, other sources of electrical activity are also recorded, which need to be removed when the aim is to study brain activity. Methods of preprocessing data with the aim to remove these non-neural signals vary widely between studies, and the choice of which features to extract from the data has also shown heterogeneity. A wide variety of algorithms have been used, some studies employing conventional algorithms such as SVMs (Noble, 2006) and RFCs (Ho, 1995; Breiman, 2001), while others have used DL methods such as certain

implementations of multi-layer *artificial neural networks* (ANNs; Jain, Mao, & Mohiuddin, 1996; Krogh, 2008).

As EEG data are time-consuming to collect, the datasets used are often rather small. For example, the three datasets used in this thesis (which will be discussed in Section 2.5) only contain 38, 56, and 40 participants, respectively, with other studies reporting participant numbers as few as 18 (Guo et al., 2021). Because it is difficult to recruit large numbers of subjects, reported data generally consist of PwPD at many different stages of the disease (as measured by the MDS-UPDRS), and people of both sexes. It is well-known that the EEG signals of PwPD change as the disease progresses (Soikkeli et al., 1991; Morita et al., 2009; Benz et al., 2014), and significant differences between male and female PwPD have also been observed (Cerri, Mus, & Blandini, 2019), mirroring the general differences in EEG data between females and males that have been reported in the literature (Greischar et al., 2004; Li et al., 2007). These observations mean that it may not be best practice to mix data from both sexes and from people at different stages of PD when investigating EEG as a method of classification in PD.

This considerable heterogeneity makes it difficult to compare results across studies and draw reliable conclusions regarding best practice in the investigation of EEG as a biomarker of PD. It is known from the EEG literature that different preprocessing approaches may have influenced experimental results: different algorithms have individual advantages and disadvantages that potentially make them appropriate for one situation but not for another. For EEG to become the gold standard, or part of the gold standard, for PD diagnosis and/or prognosis, it would be necessary to develop more uniform methods and clarify the effects of experimental choices. Moreover, the small dataset sizes make it difficult to draw robust conclusions as the precise composition of participants in terms of different characteristics, such as sex and disease status, may influence results. As yet, the effects of preprocessing steps to clean EEG data, the differences between feature selection methods, the varying algorithms, and the effects of participant characteristics on classification tasks in PD have not been formally investigated.

1.3 Thesis objectives

The main objective of the work presented in this thesis is to investigate methods employed when applying ML to EEG data in the classification of PwPD and controls, resulting in the proposal of an end-to-end processing pipeline. This would aid in the comparison of EEG methods to the current gold standards of PD diagnosis and prognosis. To accomplish this, four sub-objectives need to be accomplished, involving an exploration of the various methods used in PD classification tasks in the literature and then applying these to different datasets, either separately or combined, to investigate the effects of each step systematically.

First, it is necessary to clarify the effects of different preprocessing methods on classification outcomes, as previous studies have used a large variety of preprocessing methods. Second, the difference in data obtained while participants had their eyes open or closed will be investigated, as well as the effects of using data for which epochs, created by dividing the time series into shorter parts, have been averaged for each participant or for which multiple epochs taken from the same sample have been treated as independent samples. Third, the effects of feature selection at different time points in the pipeline and differing methods of sample randomisation will be investigated, paying special attention to methods that might introduce data leakage. Fourth, the effects of combining and splitting datasets based on participant characteristics, such as age, sex and MDS-UPDRS score, will be investigated. Lastly, the findings will be considered in the broader perspective of other neurodegenerative diseases, by comparing the experimental results in this thesis to results obtained from a systematic review into the use of *event-related potentials* (ERPs), which are EEG signals time-locked to a stimulus (Luck & Kappenman, 2011; Luck, 2014), in the investigation of cognition in *Multiple Sclerosis* (MS).

The aim of making recommendations and proposing an end-to-end pipeline is an interdisciplinary undertaking. The methods that were used in the experiments described in this thesis are the result of decades of research in the fields of, predominantly, computer science and clinical neuroscience. Some of these methods, while common knowledge in one field, might be unfamiliar to researchers in the other field. For example, in Chapter 3 the effects of different EEG preprocessing steps are discussed, a topic that has been published on widely in clinical neuroscience, but here these effects are investigated in the context of an ML task and the differences in evaluation metrics they lead to, which is a topic that has received less attention. Therefore, the aim of making recommendations and proposing an end-to-end pipeline required building an understanding of ideas and methods in the context of both fields and integrating these ideas and methods into one interdisciplinary paradigm.

1.4 Thesis contributions

Heterogeneity of methods and participant characteristics, and uncertainty of their effects on ML classification of PwPD and controls, mean that comparison of experimental results is difficult, and raise questions about which methods might be best suited to distinguish people with neurodegenerative diseases from those unaffected by them. This also makes it difficult to compare experimental outcomes to current methods of diagnosis and prognosis. By systematically investigating the effects of the different methods used in the literature, the work presented here offers a basis for comparison of ML studies of PD classification using EEG and places the results in a clearer context. The findings in this thesis show the potential of EEG, and especially RSEEG, as a method of classifying PwPD and controls and thus potentially aiding clinicians in their decision making. Additionally, by comparing methods found in the literature and investigating their effects step-by-step, the research presented here provides

a further analysis of the factors that might influence successful classification. In doing so, this thesis provides recommendations for future research into EEG as a useful biomarker of PD.

1.5 Thesis outline

This thesis consists of a further seven chapters. Each chapter builds on the chapters that precede it.

Chapter 2 contains the background and methods sections, as well as a discussion of previous studies investigating PD classification using ML methods and EEG data. I will provide details of PD as a disease and its progression; discuss electroencephalography, its methods, and the various ways in which it can be used; describe ML, and what is entailed in training and evaluating an ML model; and address the use of ML in medicine in general and in PD classification specifically.

In Chapter 3, I will look at the effects that different levels of preprocessing have on classification outcomes of PwPD and controls. To do so, I used the recommendations of the creators of EEGLAB (Delorme & Makeig, 2004), specifically those of Makoto Miyakoshi, and divided the full preprocessing pipeline into separate stages that correspond to the different levels of preprocessing that other ML studies have used. All experiments were performed on three separate datasets obtained from different universities.

In Chapter 4, I compare different methods of averaging and methods to increase sample numbers, using data that were recorded while the participants had their eyes open and while they had their eyes closed. To do this, the three independent datasets used in Chapter 3 were merged to produce one large dataset. Researchers have used varying methods of oversampling and averaging, but it is unclear which method is best. To complicate matters, EEG recordings differ depending on whether a person has their eyes open or closed: while most studies have only used eyes closed data, it is unclear whether one or the other would produce better results.

In Chapter 5, I present the results of experiments performed to find out whether the results reported on by Demircioğlu (2021) in the study on feature selection in radiomics experiments also hold true for EEG data by performing experiments in which feature selection either forms part of the *cross-validation* (CV) process or, alternatively, takes place before the CV process. Additionally, I discuss the effects of different methods of randomising samples, some of which introduce data leakage, on classification outcomes, and combine both the feature selection methods and randomisation methods to show the effects that their combination has on the outcomes.

In Chapter 6, the combined dataset is used to create separate smaller datasets with comparable characteristics in terms of age, sex, and MDS-UPDRS score to see whether it is necessary to create separate models for individual groups with different characteristics. Research has shown differences in RSEEG signals exist between men and women, as well as between people at different stages of PD as

measured using the MDS-UPDRS. The experiments presented here are used to clarify the aforementioned differences in the context of an ML classification task and to investigate which features are responsible for the improved evaluation outcomes.

In Chapter 7, I put the research into PD into a broader perspective of neurodegenerative diseases in general by reporting on a literature review into the use of ERPs in the investigation of cognition in MS. I compared and contrasted the results of this literature review with research into PD (Seer et al., 2016) to make suggestions and recommendations regarding the direction of future research, and what might be necessary to improve the role of EEG in the diagnosis of neurodegenerative diseases.

Lastly, in Chapter 8 I bring together the findings discussed in this thesis, and how these relate to the current body of knowledge that exists regarding the use of ML to enhance EEG as a tool for the classification of PwPD and controls. Recommendations for future research are provided, and the findings put into the broader context of research into biomarkers of neurodegenerative diseases. By doing this, the work presented here contributes to laying the groundwork for EEG to be the, or be included in, the gold standard for diagnosis and prognosis.

Chapter 2: Background

To understand the research into the use of *machine learning* (ML) for *Parkinson's disease* (PD) classification using *electroencephalography* (EEG) data, it is necessary to understand PD as a disease, to understand EEG and its applications, and to explore ML methods. Accordingly, the disease PD will be discussed, its history and disease features, methods of diagnosis and prognostication, as well as its treatment. Then EEG as a method to measure brain activity will be introduced, highlighting the different ways in which EEG data can be acquired, and the ways in which EEG data are often preprocessed will be discussed. Lastly, ML and the tasks it can be applied to are introduced. How to train and evaluate an ML model is discussed, as well as different ML algorithms that can be used to classify *people with PD* (PwPD) and controls.

Generally, an ML experiment looking at the use of EEG for PD classification follows a similar outline. First, the datasets that are to be used need to be defined. Then, EEG data needs to be preprocessed to remove noise and unwanted signals. This step can vary widely between studies but can include downsampling of the data, filtering, electrode removal and interpolation of noisy electrodes, and removal of signals of non-neural origin, known as artifacts. Unless the choice is to use the whole time series, as is the case in many neural network approaches, individual features are then extracted. The ML algorithm is then trained on a selected subset of 'training' data before it is applied to the 'test' data. This process is repeated several times with new train-and-test splits to make sure the results are as unbiased as possible. Finally, the overall performance of the ML algorithm is assessed using several different evaluation metrics, such as accuracy, sensitivity, and/or specificity.

In this chapter, the pipeline described in the previous paragraph will be followed, explaining each step and its relevance. As there are multiple methods that have been developed to analyse EEG data and there are just as many ML methods that have been (or could be) applied to the matters discussed in this study, detailed explanations of core concepts that apply across all experiments are provided. For situations where this would be impractical, such as, for example, covering all possible ML algorithms, explanations have been limited to the methods used in the studies in this thesis.

2.1 Parkinson's disease

Throughout history, observers have written about people who, often in old age, start to drool, develop "the shakes", and have difficulty moving. A papyrus scroll written between 1350 and 1200 BCE describes a king whose jaw slackened in old age and who started to drool (Lees, 2007). Clinical features such as tremor are described in Sanskrit writings on Ayurveda, Indian traditional medicine, dating back to the 2nd century BCE, and these are, in turn, based on older texts (Stern, 1989). In the 2nd century AD, Galen, the court physician of Marcus Aurelius Antoninus who was considered to have been one of the

greatest doctors of his time, described both tremor and gait disorders (Currier, 1996). In the 15th century AD, another intellectual powerhouse, Leonardo da Vinci, wrote about "how nerves sometimes operate by themselves without any command from other functioning parts or the soul. This is clearly apparent for you will see paralytics and those who are shivering and benumbed by cold move their trembling parts, such as their head or hands without permission of the soul; which soul with all its forces cannot prevent these parts from trembling" (Lees, 2007).

Descriptions similar to those provided above appeared in many time periods and in different parts of the world, but it was not until 1817 that a full description of the disease was provided in an essay titled "An Essay on the Shaking Palsy" (Parkinson, 2002). In that essay, James Parkinson, an English surgeon and apothecary, described six cases of a slowly progressing disease, referencing physicians before him, such as Galen and Sylvius de la Boë, who had described similar features. He described a disease that initially caused slight weakness, a tremor in one body part that gradually spread to other body parts, which went on to cause problems with posture that were most noticeable while walking, and fatigue. As the disease progressed, the patients started to shuffle when walking, developed trouble writing, and experienced disturbed sleep, constipation, and difficulty with speech. As the title of the essay suggests, the disease was named *shaking palsy* or *paralysis agitans*, until, after a campaign by Jean-Martin Charcot, considered the father of French neurology, it was renamed after the author of "An Essay on the Shaking Palsy", James Parkinson, and given the name we know it by today: Parkinson's disease (Lees, 2007).

2.1.1 A description of Parkinson's disease

PD is a neurodegenerative disease in which dopaminergic cells, predominantly in the *substantia nigra pars compacta* (SNpc) of the midbrain, degenerate, resulting in a variety of motor and non-motor symptoms (Jankovic, 2008; Kalia & Lang, 2015). The cause of the disease is currently unknown, though the previous hypothesis that PD was caused by environmental factors has now been replaced by a view that it results from a complex interplay between genetics and the environment (Kalia & Lang, 2015). While degeneration of the SNpc is a cardinal feature of the disease, it is not the only brain area affected by the disease, with neuronal loss occurring in other areas such as the locus coeruleus, amygdala, and hypothalamus (Kalia & Lang, 2015). Apart from the degeneration of the dopaminergic cells themselves, the presence of Lewy bodies is another cardinal feature of PD, since the misfolding of proteins occurs in the disease, as it does in many other neurodegenerative diseases such as *Alzheimer's disease* (AD). In PD, evidence shows a relationship between the misfolding of the protein α-synuclein and cognitive impairment. Based on the observed correlation between the location of Lewy body pathology and clinical disease progression, Braak et al. (2003) have proposed a model of six stages that categorise disease progression from its prodromal stage to late-stage PD.

While PD is generally known for its motor features, the non-motor features of the disease often present before the motor features (Hawkes, Del Tredici, & Braak, 2010). For example, research shows that approximately 38% of people with *rapid eye movement sleep behaviour disorder* (RBD) go on to develop PD, with an average time interval between diagnoses of about 12 years (Schenck, Bundlie, & Mahowald, 1996). Other non-motor features reported by PwPD in the prodromal stage of their disease are constipation, problems with smell, and depression (Hawkes, Del Tredici, & Braak, 2010). This is not to say that non-motor features do not continue to develop after diagnosis of PD: pain and fatigue are commonly reported in the later stages, and 83% of people with a disease duration of 20 years or more develop dementia (Hely et al., 2008). Using the Braak staging (Braak et al., 2003), gastrointestinal problems such as constipation often occur in Stage 1, about 20 years before the clinical diagnosis of PD. It is also at this point that a loss of smell is reported, with accumulation of misfolded α-synuclein found in the olfactory system. In Stage 2, about 10 years before clinical diagnosis, sleep problems, such as RBD and depression, become prevalent.

The cardinal motor features of PD are bradykinesia (slowness of initiation of movement), rigidity of muscles, resting tremor, and impairment of posture and gait (Kalia & Lang, 2015). Motor features generally start unilaterally, and research shows that by this point the SNpc has lost at least 50% of its dopaminergic neurons (Fearnley & Lees, 1991). This is referred to as Stage 3 by Braak et al. and, apart from the SNpc, other brain structures like the amygdala are now progressively impacted by the disease (Braak et al., 2003; Hawkes, Del Tredici, & Braak, 2010). Once the patient enters Stage 4, the disease has become bilateral and Lewy body pathology has become so widespread that clinical correlation with individual pathological features becomes difficult. Around 10 years after clinical diagnosis of PD, in Stage 5, balance has deteriorated, leading to an increased risk of falls, cognitive decline has set in, and the person becomes more and more dependent on others. In Stage 6, the last stage of the disease, the person is often bed-bound with dementia (Hawkes, Del Tredici, & Braak, 2010).

2.1.2 Diagnosis and prognostication of Parkinson's disease

As mentioned in the introduction, PD is currently diagnosed on the basis of its cardinal motor features, the exclusion of other diseases, and the response to dopaminergic medication. The medication that is generally first prescribed when someone is diagnosed with, or is suspected to have, PD is levodopa. Levodopa is a naturally occurring amino acid and a precursor to dopamine. When administered on its own, only about 5% to 10% is ultimately converted to dopamine in the brain after crossing the blood-brain barrier because the remainder is metabolised to dopamine elsewhere in the body. This increase in peripheral dopamine levels can lead to side effects such as nausea and low blood pressure, so levodopa is often administered with an additional medication like carbidopa, which inhibits the peripheral metabolism of levodopa.

Nowadays, disease status in PwPD is most commonly assessed using the Movement Disorder Society's modification of the *Unified Parkinson's Disease Rating scale* (MDS-UPDRS; Goetz et al., 2008). This tool consists of four parts: Part I 'non-motor experiences of daily living', Part II 'motor experiences of daily living', Part III 'motor examination', and Part IV 'motor complications'. In addition to a few general descriptive questions, each part consists of a series of questions that are rated on a scale of 0 to 4 in 1-point increments, where 0 means 'normal' and 4 means 'severe'. For example, Question 1.1 asks *"Over the past week have you had problems remembering things, following conversations, paying attention, thinking clearly, or finding your way around the house or in town?"* A score of 0 on this question means *"no cognitive impairment"*, while a score of 4 means *"Cognitive dysfunction precludes the patient's ability to carry out normal activities and social interactions."* After scoring all questions, points are tallied for each individual part and the overall total is then calculated. In all, there are 65 questions (13 in Part I, 13 in Part II, 33 in Part III and 6 in Part IV), so the total score ranges from 0 to 260.

Whether someone is taking dopaminergic medication or not is of importance when assessing their disease status and when researching potential biomarkers of PD, as clinical features are often affected by the medication itself. Indeed, there is evidence that EEG data differ between PwPD when on medication and when off medication. Because of this, researchers often indicate whether their participants were 'ON' or 'OFF' medication by analogy with a light switch: when 'ON', the person is less affected by their disease and can move around, as opposed to the untreated situation which is referred to as 'OFF'. In the case of an individual treated with dopaminergic medication, 'OFF' generally occurs after a certain amount of time following the last dose and this needs to be taken into consideration when the participant is tested. Note that the state of having taken, or having refrained from taking, the medication itself should not be confused with the 'ON' and 'OFF' states which describe the clinical state of the participant. When being assessed with the MDS-UPDRS, the assessor will record whether the participant is prescribed medication or not and, if they are taking medication, whether they are in an ON or an OFF clinical state. The ON state is defined as *"the typical functional state when patients are receiving medication and have a good response"*, while the OFF state is defined as *"the typical functional state when patients have a poor response in spite of taking medications"* (Goetz et al., 2008). "A typical functional state" when taking levodopa and in the ON state, for example, is a reduction in rigidity and bradykinesia, while these features would increase in severity in an OFF state.

2.2 Electroencephalography

The ability to diagnose PD before the onset of its cardinal clinical features would require the discovery of relevant biomarkers. The word 'biomarker' is a shortening of 'biological marker' and has been defined in various ways by organisations and work groups that include the *World Health Organization* (WHO), the *United States* (US) *Food and Drug Administration* (FDA), and the *US National Institutes*

of Health (NIH; Strimbu & Tavel, 2010). Fortunately, the definitions largely overlap, so for the purposes of this thesis the definition used by the FDA-NIH Biomarker Working Group (2016) is used: "a defined characteristic that is measured as an indicator of normal biological processes, pathogenic processes or responses to an exposure or intervention." Biomarkers are quantifiable, objective, and can be independent from the experience of the participant they are being measured from. Examples include blood values, blood pressure, heart rate, and genetic tests, as well as findings from imaging methods such *as magnetic resonance imaging* (MRI) and c*omputed tomography* (CT).

In the search for biomarkers of PD, researchers have investigated a method known as EEG (Biasiucci, Franceschiello, & Murray, 2019). During an EEG recording, a participant is fitted with electrodes on the scalp, either embedded in an electrode cap or fixed to the scalp with an adhesive, and the electrical activity of the underlying brain is measured, either during a cognitive task or at rest. EEG has been used clinically in the diagnosis and treatment of epilepsy, sleep disorders, traumatic brain injury, and brain tumours, among others (Niedermeyer & da Silva, 2005).

2.2.1 A description of electroencephalography

EEGs can be recorded from as few as 1 electrode to several hundred electrodes (see Figure 2.1a for an example of a 64-electrode setup and Figure 2.1b for the electrode layout). Compared to other methods of measuring brain activity, such as *functional MRI* (fMRI), and *positron emission tomography* (PET), EEG has a higher temporal resolution, being able to detect changes on a scale of milliseconds. It also measures the electrical activity of the brain directly, while other methods measure it indirectly, either through metabolic changes or a haemodynamic response. Another major benefit of EEG is that the necessary equipment is much cheaper than that of other methods like, for example, MRI, which can cost several millions of dollars compared to a few thousand for EEG equipment. Additionally, methods like MRI and *magnetoencephalography* (MEG), require purpose-built facilities, while EEG recordings only require a quiet room. Additionally, some methods such as (f)MRI recordings and the drawing of blood, for the purposes of identifying blood-based markers, can be experienced as being unpleasant (Tazegul et al., 2015). MRI recordings require the participant to enter the narrow bore of the MRI machine, which can be experienced as claustrophobic (Munn et al., 2015), and the machine itself can be very loud, with noise levels sometimes reaching 117 dB (Counter et al., 1997). In the case of blood markers, the drawing of blood requires puncturing of the skin with a needle, which invariably causes a small amount of physical trauma and can cause anxiety in participants (Trost et al., 2017). Analysis of the blood sample also requires specialised equipment. EEG recordings, in contrast, are non-invasive, can be obtained with the participant comfortably sitting in a chair or lying down, and can be analysed on any modern computer.

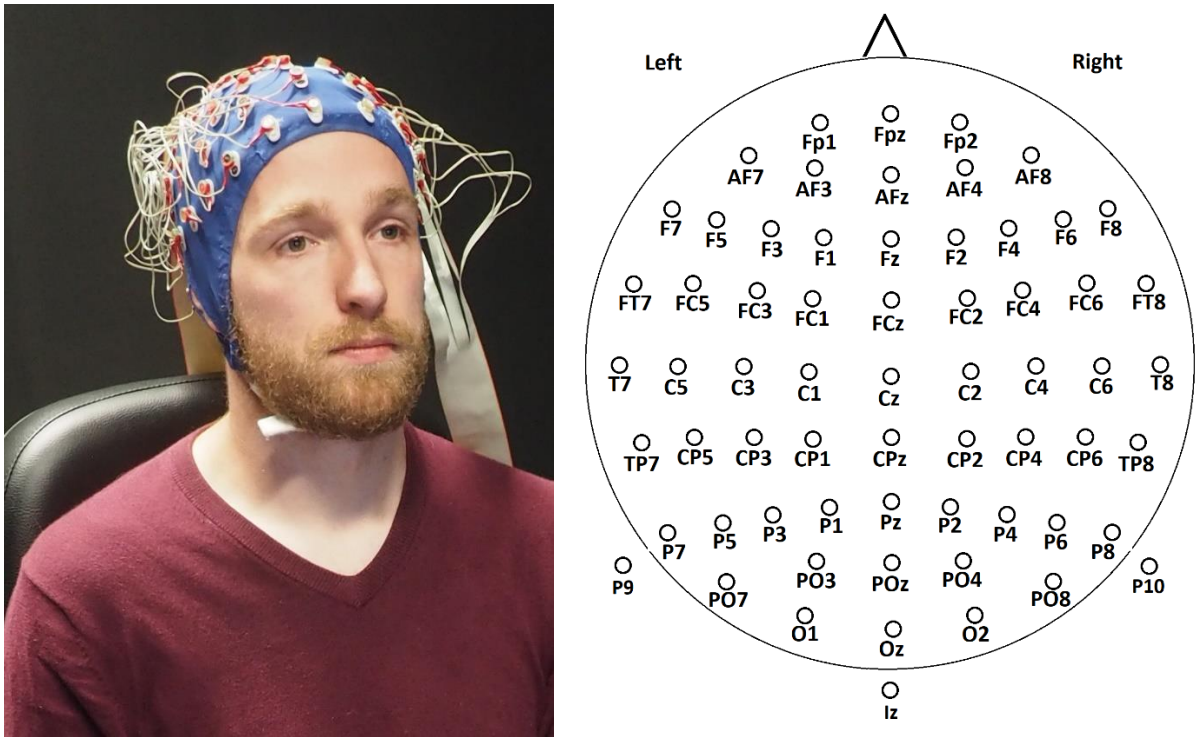


Figure 2.1a and b. A 64-electrode EEG setup and the electrode layout of the cap worn in the picture. Uneven numbers refer to electrodes on the left side of the participant's head, and even numbers to the right side. Midline electrodes have the letter z. Fp is 'prefrontal', A is 'anterior', F is 'frontal', C is 'central', P is 'parietal', O is 'occipital', and I is 'inion'.

While there is evidence that it is possible to measure voltages from deeper brain structures using high-density setups and mathematical models (Seeber et al., 2019), EEG electrodes generally detect and measure the post-synaptic potentials of groups of millions of pyramidal neurons firing together in the cortical layers of the brain (Niedermeyer & da Silva, 2005; Biasiucci, Franceschiello, & Murray, 2019). Both the orientation of the neurons and synchronicity of activation are important for the electrodes to detect a signal. If the neurons do not fire in unison, their individual potentials cancel each other out and the summated potential becomes too small to be recorded. Similarly, if the neurons are not spatially aligned, the potentials of individual neurons do not summate to generate a potential that is large enough to be measured.

EEG can be measured under a variety of conditions. *Resting-state EEG* (RSEEG) is recorded while participants are relaxed, either sitting or lying down, without any stimuli being presented. Evoked potentials are also recorded while the participant is sitting or lying down but, in this case, the participant is presented with a stimulus. *Event-related potentials* (ERPs; Luck, 2014) are recorded while a participant engages in a task that is more cognitively demanding than required for EPs. In this thesis, only RSEEG and ERPs will be discussed.

RSEEG is most often analysed in the time-frequency domain (Niedermeyer & da Silva, 2005; Biasiucci, Franceschiello, & Murray, 2019). The participant is seated or lying down and asked either to close their eyes or to keep them open. Then, a recording with a duration from one to several minutes

will be made. The resulting EEG signals can be divided into different frequency bands: *delta* (1-4 Hz), *theta* (4-8 Hz), *alpha* (8-13 Hz), *beta* (13-30 Hz), and *gamma* (>30 Hz). The four frequency bands from *delta* to *beta*, which will be the focus of the studies here, are depicted in Figure 2.2.

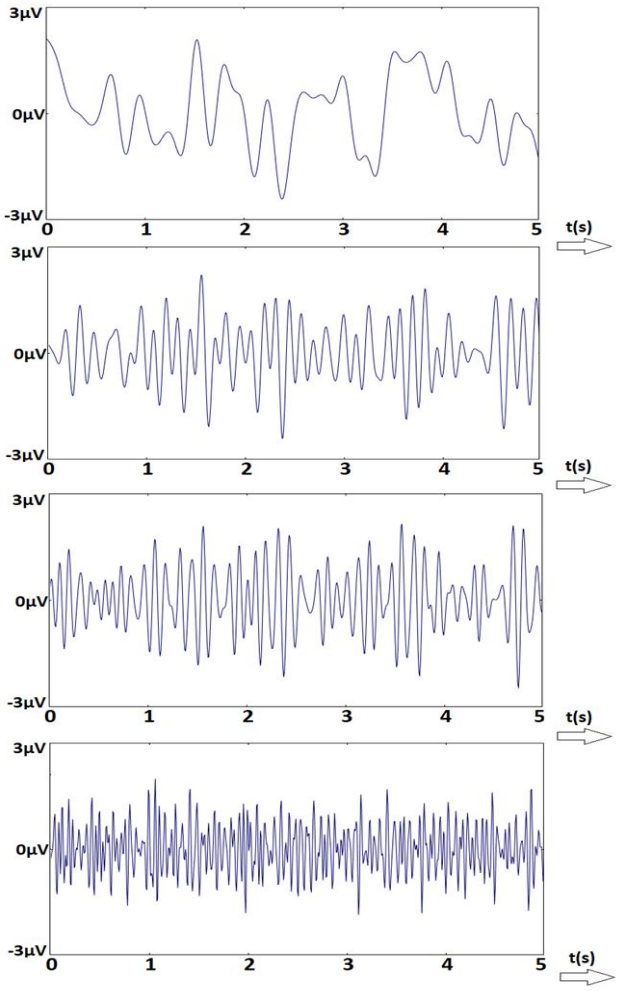


Figure 2.2. Examples of, from top to bottom, 5 seconds of data in the *delta*, *theta*, *alpha*, and *beta* frequency bands.

The individual frequency bands are associated with different processes in the brain (Kaufman & Milstein, 2012), and methods such as Fourier Transforms and Wavelet Transforms can be used to decompose the composite EEG signal into those bands (Cohen, 2014). The slowest frequencies, under 4 Hz, known as *delta* waves, are associated with deep slow wave sleep in adults and have also been associated with the inhibition of sensory activity that interferes with concentration during cognitive tasks. Frequencies between 4 and 8 Hz are called *theta* waves and are associated with drowsiness as well as response inhibition in cognitive tasks. The *alpha* band, so named because it was the first band observed by Hans Berger, the inventor of EEG (Tudor, Tudor, & Tudor, 2005), contains frequencies between 8 and 13 Hz. It becomes stronger when a participant closes their eyes and attenuates when they open them. Similarly, it is stronger when a person relaxes and weakens during cognitive tasks. *Beta* waves, between 13 and 30 Hz, are associated with wakefulness. They are observed when a person is

concentrating deeply, is thinking actively, or is anxious. Finally, the *gamma* band, over 30 Hz, is associated with sensory processing across modalities. After decomposing the raw signal, it is possible to calculate characteristics such as power and entropy in each of the separate frequency bands and to calculate power ratios between frequency bands and the whole signal, called relative power.

In contrast to RSEEG, ERPs are most often analysed in the time domain (Luck & Kappenman, 2011; Luck, 2014). The participant is presented with a particular task and the brain's response is averaged across many trials, sometimes hundreds, to obtain a stereotypical response. For example, an often-used paradigm is the oddball paradigm (Squires, Squires, & Hillyard, 1975). In this paradigm, the participant is presented with a train of stimuli that are usually either auditory or visual in nature, one of which differs from the others: this might be a 500 Hz tone for standard stimuli and a 1000 Hz tone for oddball stimuli. The participant is required either to respond when an oddball occurs or to report how many times oddballs occurred at the end of the experiment. This is done for several hundred trials, after which the EEG recordings relating to each of the oddball presentations are averaged to generate a stereotypical brain response. This can be done for many different paradigms, using different stimulus modalities, and for varying numbers of trials. The obtained signal can then be divided into several different components, each being named depending on whether it is a positive- or negative-going component, and at roughly what time after the stimulus it occurs. For example, a stereotypical ERP component in an oddball experiment is P300, which is a positive-going potential that occurs approximately 300 *milliseconds* (ms) post-stimulus. Important to note is that the first components that appear, such as P100 and N100, cannot represent cognitive activity as they occur before the participant has had a chance to notice and orient their attention towards the stimulus. Components that appear from approximately 200 ms onwards are generally considered to be cognitive components.

Differences in ERP components between groups are generally expressed in terms of either latency or amplitude (Luck & Kappenman, 2011; Luck, 2014). Amplitudes are measured in microvolts and the values are quoted in relation to the potential at another 'reference' point, the choice of which can differ based on the individual study. For example, a researcher can choose to define amplitude as the difference between the potential measured at the peak of a component within a certain time window and the potential at the moment the stimulus was presented or, alternatively, as the peak of a component within a certain time window and the potential at a 'baseline' measurement taken before the stimulus was presented. Alternatively, the amplitude of a component can be measured in relation to that of another component, as is the case in experiments that measure P300. P300 is immediately preceded by N200, so researchers can choose to define the amplitude of P300 in relation to that of N200 or to a baseline instead. The latency of a component is simply the point in time at which the peak of that component occurs following stimulus presentation.

A major disadvantage of EEG is its poor spatial resolution (Cohen, 2014). Often, when performing an experiment, the variables are defined and then data collected. This is known as a 'forward' problem, which is a well-posed problem, meaning it has a solution that is unique and changes continuously with changes in variables (Pohjolainen & Heiliö, 2016). The problem with EEG is that we have no knowledge of all the possible variables involved: we know the voltage that was measured by an electrode on the scalp, but this voltage represents the summation of signals from many possible origins, including different neuronal populations, muscle activity and eye movements, and there are many possible combinations of these sources that could have led to the exact voltage that was recorded at the scalp. This is known as an 'inverse' problem.

Additionally, EEG recordings can be affected by a wide variety of individual participant characteristics. For example, the nature of a participant's hair can have an influence on the EGG recordings. Studies have indicated that recordings from people of African descent, who commonly have hair that is coarse and curly, are often of poorer quality due to most EEG electrodes not being designed for use with this type of hair (Choy, Baker, & Stavropoulos, 2022). This problem is not isolated to specific ethnic groups. Studies investigating differences between male and female participants have found differences in data quality between the sexes attributable to the difference in hair length and styles (Greischar et al., 2004). For example, being bald or shaving one's head can lead to a thicker scalp than that of people who have hair which, due to different tissues having different conductivity, leads to differences in measured voltages. Differences in conductivity of other tissues are also important, as thickness of the skull itself also leads to different measurements and women tend to have thicker skulls than men (Li et al., 2007).

These limitations should not be taken as a reason not to investigate EEG as a clinical method for the diagnosis and prognostication of neurodegenerative disease such as PD. Rather, they simply require more avenues of research, such as the development of source localisation methods to deal with the inverse problem, the development of better hardware such as special electrodes to accommodate differences between participants, and the development of protocols to ensure data quality and quantity. EEG is a relatively old method, going back to its discovery by Hans Berger in 1924 (Haas, 2003), and it has been used with great success to investigate diseases such as PD, as will be shown in Section 2.2.3. But, before discussing the value of EEG research in the diagnosis and prognostication of PD, it is important to outline the methods of signal processing and analysis used in this research.

2.2.2 Signal processing of electroencephalography

Electrodes used in EEG experiments cannot distinguish between what is a signal of neural origin and what is not. An electrode will detect any electrical activity, whether it arises from the brain or sources other than the brain. In practice this means that electrodes record not only the activity of the brain, but also signals of non-neural origin. These signals of non-neural origin, called artifacts, consist of activity

arising from both physiological and non-physiological sources (Kaya, 2019). Examples of physiological sources include the muscles of the scalp, eye movements and blinks, heartbeat, respiration, and perspiration. Examples of non-physiological artifacts include movements of the head, limbs, and body, movement of the electrodes, and interference from other electrical equipment. Artifacts can also be divided into stationary artifacts, ones that occur in a regular pattern such as eye blinks, and non-stationary artifacts, such as when a participant adjusts in their chair. Figure 2.3a provides an example of raw data before any preprocessing has taken place.

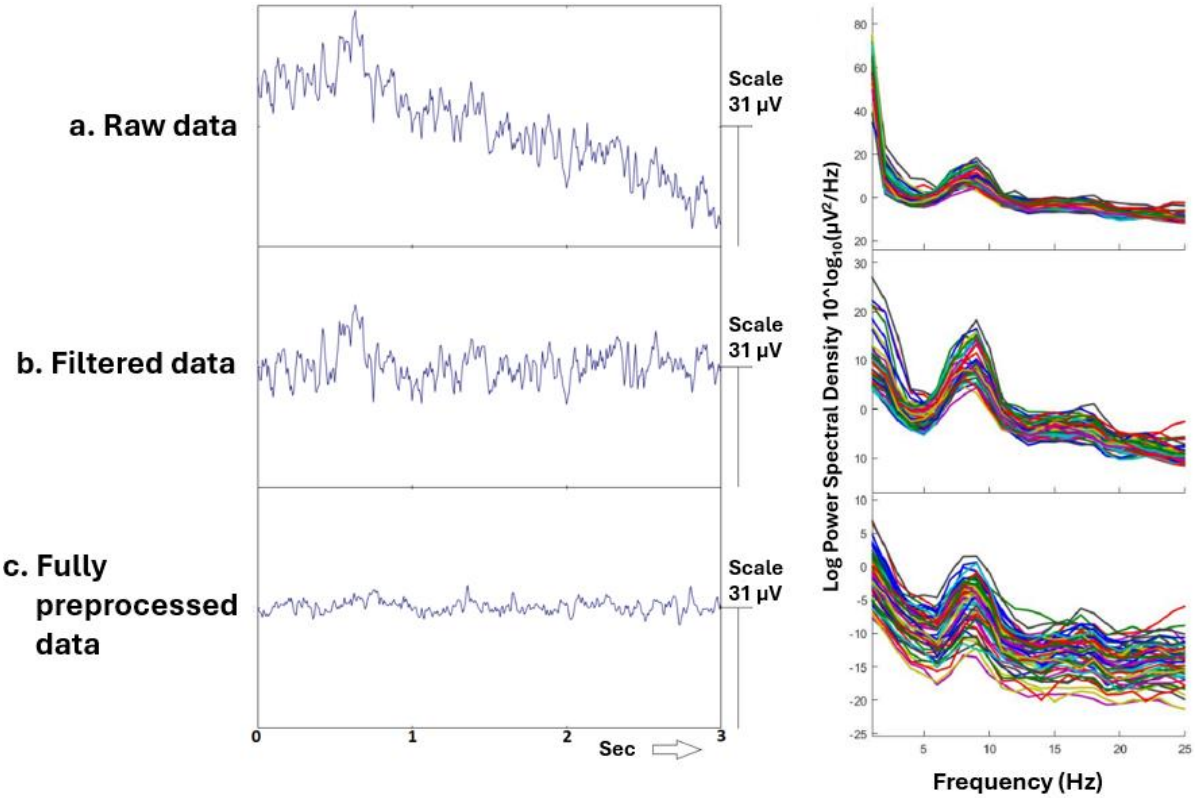


Figure 2.3. Examples of a) raw data, b) filtered data, and c) data that has undergone artifact rejection. On the left, we see a 3-second time series and voltages in µV, and on the right, we see power spectral density plots for all three preprocessing steps.

To be able to make claims which specifically relate to neural activity alone, it is necessary to remove artifacts from the EEG data through a set of preprocessing steps. Preprocessing of EEG data is a research field in itself with improvements being made every year. Researchers have developed tools to enable easier and more consistent preprocessing, such as FieldTrip (Oostenveld et al., 2011) and EEGLAB (Delorme & Makeig, 2004). Additionally, researchers have developed plugins for these toolboxes for use in specific steps of the preprocessing pipeline, such as *Artifact Subspace Reconstruction* (ASR; Mullen et al., 2015; Chang et al., 2019) and ICLabel (Pion-Tonachini, Kreutz-Delgado, & Makeig, 2019), both of which will be discussed in more depth later.

There is currently no agreed-upon protocol for the preprocessing of EEG data because any preprocessing steps, and the manner in which they are implemented, will depend on the specifics of

what the researchers are investigating. For example, the choice of filter settings in ERP research depends on whether a study focuses on slow-building potentials like the *Contingent Negative Variation* (CNV), or on components like P300 that build up much faster. Additionally, in RSEEG research, experimenters may choose to apply specialised filters because they are only interested in activity within a particular frequency range. While specifics might vary depending on the research being conducted, and the EEG hardware being used, most EEG studies will include down-sampling, filtering, electrode rejection and interpolation, as well as artifact rejection in their preprocessing pipeline (Miyakoshi, n.d). These terms are defined and discussed below:

Sampling rate: EEG data are often collected at sampling rates of 250 Hz or higher, but the frequencies most researchers are interested in are often much lower than this. To speed up analysis, researchers may choose to down-sample the data they are using (Miyakoshi, n.d). The lowest frequency to which they can down-sample is based on the Shannon-Nyquist theorem (Luck, 2014), which states that the sampling rate of a signal needs to be at least twice that of the highest frequency of interest (otherwise known as the Nyquist frequency). For example, if a researcher is interested in activity in the *gamma* band between 30 and 50 Hz, the sampling rate of the signal needs to be at least 100 Hz. In practice, most researchers will not down-sample to the lowest acceptable frequency but, rather, to a frequency that remains higher than the Nyquist frequency but does not noticeably slow down analysis.

Filtering: Filtering is often the next step in the preprocessing procedure. Generally, researchers apply a high-pass filter to remove baseline drift. In EEG recordings, this drift arises from a range of different factors, such as changes in conductivity resulting from the participant sweating or the conductive gel drying out. A low-pass filter is sometimes used to remove so-called line noise, caused by adjacent electrical equipment. The frequency cut-off depends on where the recording was made due to differences in the electrical net. The cut-off frequency in Europe and Australia is 50 Hz, while the cut-off frequency in North America is 60 Hz. It is also possible to apply either a notch-filter, which filters around a pre-defined frequency, or to use a plugin such as CleanLine (Mullen, 2012), which has been specifically developed to remove interference from the electrical net. Figure 2.3b provides an example of filtered data.

Electrode removal and interpolation: In almost every experiment there are electrodes that show a pattern that deviates from what would be considered typical. This can be a flat recording, meaning that something has gone wrong with the conductivity between the scalp and the electrode and so the electrode did not record at all, or an electrode might show activity with fluctuations that differ from what is normally seen, sometimes caused by, once again, poor conductivity or by the particular location of the electrode (Kaya, 2019). Removal of a particular electrode recording before subsequent analysis can be done manually, with one or more researchers visually inspecting the recording and removing any electrodes that are deemed deviant, or by using automated procedures. EEG preprocessing

toolboxes like EEGLAB (Delorme & Makeig, 2004), MNE (Gramfort et al., 2014), and FieldTrip (Oostenveld et al., 2011) all have built-in functions to remove and interpolate deviant channels. An example of this is the above mentioned EEGLAB plugin ASR (Miyakoshi, n.d; Mullen et al., 2015; Chang et al., 2019) which will reject an electrode if it correlates by less than a pre-specified value with the electrodes around it during a pre-specified time window. Using the signals from locations around the removed electrode, researchers can then decide whether or not to reconstruct the output that theoretically would have been obtained from the missing electrode or to manage without it.

Artifact and noise removal: As with the deviant electrodes, artifacts can be dealt with by researchers inspecting the signal and removing any part of the signal that shows artifactual activity, either manually or by using automated procedures. Once again, ASR can be used here. ASR (Miyakoshi, n.d; Mullen et al., 2015; Chang et al., 2019) uses a sliding-window approach to identify non-stationary artifacts that produce high-amplitude activity; it does this by comparing the data to artifact-free data and can either remove the data or reconstruct the affected signal, as desired by the experimenters. After either manual removal or automated removal of the non-stationary artifacts*, Independent Component Analysis* (ICA; Makeig et al., 1995) is often applied to remove stationary artifacts. As each electrode records a mixture of signals, some being of neural and others of non-neural origin, ICA is used to separate the signal into its building blocks. Researchers can then look at the *Independent Components* (ICs), identify the ones that are non-neural, and then remove them from the signal. As above, automated procedures have also been developed to accomplish this, such as ICLabel (Pion-Tonachini, Kreutz-Delgado, & Makeig, 2019), MARA (Winkler, Haufe, & Tangermann, 2011; Winkler et al., 2014), FASTER (Nolan, Whelan, & Reilly, 2010), and ADJUST (Mognon et al., 2011).

When using ICA, the ICs that the total signal is decomposed into are not perfect segments of either neural or non-neural signals. Removal of an IC because it is deemed to be of non-neural origin means that some part of the neural signal will also be removed and there will always be some noise in the ICs that are considered to be predominantly neural signals. Additionally, judging whether an IC is of neural or non-neural origin is not always clear-cut. This is also evident when using automated methods: when using ICLabel, which consists of a neural network trained on thousands of examples of artifacts, one is given a probability that an IC is of neural or non-neural origin and the researcher has to determine themselves what they deem a reasonable threshold for retention or removal. Figure 2.3c provides an example of data that has undergone artifact removal using ASR, ICA, and ICLabel.

Lastly, to minimise noise in EEG data, researchers have often used averaging techniques based on an assumption that the noise in the signal is randomly distributed while the neural signal follows a predictable pattern. This is why tens to hundreds of trials are recorded in ERP experiments and then averaged: if a neural signal follows the same pattern in response to a given stimulus, the extraneous noise will cancel out if enough trials are averaged, leaving behind the stereotyped response consisting

of the known ERP components (Luck & Kappenman, 2011; Luck, 2014). Similarly, researchers have also grouped electrodes into *regions-of-interest* (ROIs) based on their location on the scalp. While each electrode will record a mixture of signals and artifacts, electrodes in the same area should record similar signals and, by averaging the signals from these neighbouring electrodes, the noise will, once again, be averaged out leaving behind the true neural signal.

An important point to make is that what is considered to be an artifact by one researcher can be a signal of interest to another, depending on what the research question is. Additionally, removal of artifacts generally also leads to a removal of neural signal, as the signals recorded by the electrodes are a mixture of both and it is not entirely possible to clearly separate the one from the other. The question, therefore, is how conservative or aggressive a researcher wants to be in the preprocessing of their data, as a conservative method will leave both more neural signals but also more artifacts. In the case of the research presented in this thesis, the goal was to find methods that can distinguish between PwPD and controls based on neural activity alone, which means that any signals deemed not to be of neural origin must be considered as artifacts. The EEG data that are the focus of the studies in this thesis have therefore undergone a full preprocessing procedure, which is described in Chapter 3, that included filtering, removal of aberrant electrodes, and artifact rejection, following the recommendations of Delorme and Makeig (2004) and Makoto Miyakoshi (n.d.).

2.2.3 Electroencephalography in Parkinson's disease

Before discussing the use of EEG in the investigation of PD, a question about the appropriateness of EEG for this purpose needs to be answered. As already stated in Section 2.2.1, EEG is generally measured from neurons in the cortex, but PD is often considered to be a disease that, especially early in the disease, primarily affects subcortical structures such as the SNpc. While PD does indeed affect the subcortical structures, studies indicate that cortical areas are also affected early in the disease: MRI research demonstartess atrophy of frontal areas in early PD in an asymmetrical pattern, as well as changes in a wide variety of white matter areas. Given these structural cortical changes, studies into functional changes have been undertaken and have, as we will see, found differences between PwD and controls, as well as between PwPD and people with other diseases.

Both RSEEG and ERPs have been used to investigate PD. Most of the research has focused on finding differences between controls and PwPD, but some research has also investigated the differences between PwPD and people with other neurological diseases, such as AD (Fonseca et al., 2013; Benz et al., 2014). The research into different neurological diseases and the differences between them is important, as many different diseases can lead to degeneration of similar brain areas. If this degeneration leads to similar EEG signals, it would mean that EEG was not suitable as a method for differentiating between neurodegenerative diseases. However, research has shown that there are some differences between diseases, with EEG signals differing between people with AD, PwPD with dementia, and

people with dementia with Lewy bodies (Bonanni et al., 2008). EEG has also been used to distinguish between PwPD and those with atypical Parkinsonian disorders such as corticobasal degeneration, multiple system atrophy, and progressive supra-nuclear palsy (Barcelon et al., 2019).

The major difference in RSEEG that has been observed between PwPD and people without the disease is the slowing of their brain waves (Soikkeli et al., 1991; Morita et al., 2009), which means that, relative to the sum of total power in the frequency band, a large percentage of total power is found in the lower frequencies. This slowing of oscillations seems to correlate with a variety of factors such as disease progression as measured on the Hoehn and Yahr scale (a scale that is used to quantify physical disability on a scale of 0 to 5; Goetz et al., 2004), dementia, and depression (Soikkeli et al., 1991; Morita et al., 2009; Espinoza et al., 2022). Interestingly, slowing of the EEG has also been reported in RBD, which is considered by some to be a prodromal stage of PD due to the high conversion rate from RBD to PD (Livia Fantini et al., 2003).

In a review of the association of quantitative EEG measures with outcomes of disease severity and progression in PD, Geraedts et al. (2018) found that EEG slowing, and specifically a decrease in the dominant frequency band, defined as the peak in the Fast Fourier Transform spectrum and typically between 4 and 13 Hz, as well as an increase in theta power, were correlated with cognitive impairment and were predictive of cognitive decline in the future. As EEG seemed to correlate better with non-motor features of the disease than with motor features, the authors suggested that EEG might be a candidate as a biomarker to track non-motor disease progression and could potentially also be developed as a tool for earlier diagnosis. In a longitudinal study into EEG changes and cognitive decline, Caviness et al. (2015) performed baseline measurements along with follow-up measurements an average of 3.9 years later and found that changes in delta power correlated most with the observed changes in neuropsychological testing.

In their 2016 review of ERPs and cognition in PD, Seer et al. (2016) found that, like studies into RSEEG, certain ERP components were correlated with dementia in PD and that other components were also potential markers of disease progression. Specifically, their analysis showed that the latency of P300, the component often measured in oddball experiments, and amplitude of the *mismatch negativity* (MMN), which occurs when a sequence of repetitive tones is interrupted by a deviant tone, correlated with the degree of dementia. The amplitude of the novelty P300, or P3a, which occurs when a third stimulus to which the participant should not pay attention to is introduced in an oddball experiment, was correlated with disease progression. They also concluded that ERP correlates of executive function, often tested through Go/No-Go paradigms (see Figure 2.4 for an example) may reflect dopamine-related dysfunction.

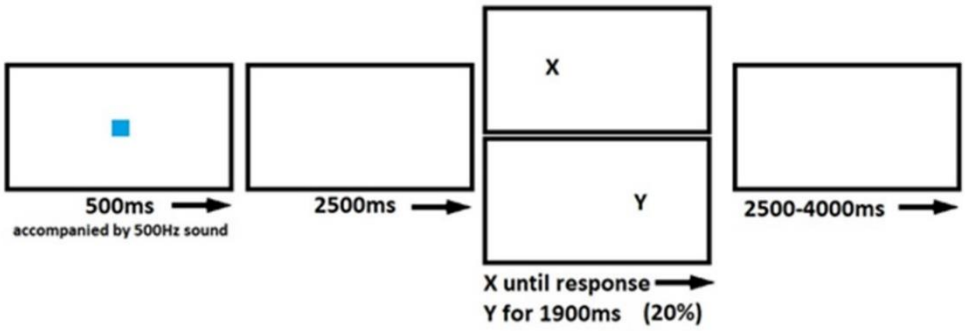


Figure 2.4. An example of a Go/No-Go task. After being shown a cue and waiting for 2500 ms, the participant is shown one of two stimuli. If an X appears, the participant is required to press a button corresponding to the side on which the X appeared. If a Y appears, they are asked to withhold any response.

2.3 Machine learning

In his book 'Machine Learning', Tom M. Mitchell (1997) provided the following definition of ML: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E". Mitchell provided several examples of problems broken down in E, T, and P, such as learning to recognise and classify handwritten words (T), the percentage of correctly classified words (P), and a database of handwritten words with labels (I), as well as playing checkers (T), playing practice games against itself (I), and the percentage of games won against opponents (P).

The term 'machine learning' was, in fact, coined by Arthur Lee Samuel in 1959 (Samuel, 1988), who was employed by IBM and created programs that could play checkers. However, the theories that ML methods have been built on go back much further than that, even as far as 1763 when Thomas Bayes' "An Essay towards solving a Problem in the Doctrine of Chances" was posthumously published (Bayes, 1763). A few years later (in 1805), Adrien-Marie Legendre described the least squares method in his "Nouvelles méthodes pour la détermination des orbites des comètes" [New Methods for the Determination of Comet Orbits] (Legendre, 1806).

2.3.1 Machine learning tasks and evaluation metrics

ML tasks can generally be divided into three groups based on how the task is accomplished: supervised learning, unsupervised learning, and reinforcement learning (Mitchell, 1997; Sidey-Gibbons & Sidey-Gibbons, 2019). In a supervised learning task, the algorithm is taught which particular outputs, be they a specific class or a value, fit with specific features from a set of examples. In unsupervised learning tasks, on the other hand, the algorithm is only given features but no output examples, so it will need to create a representation of the data structure by itself. Lastly, in reinforcement learning tasks the algorithm makes decisions after which it is provided with immediate feedback in the form of a reward.

The aim is to optimise the long-term cost or reward by identifying the set of actions necessary to accomplish this.

In this thesis, we will concern ourselves with supervised learning tasks. There are two kinds of supervised learning tasks that are commonly used: regression tasks and classification tasks. In regression tasks, the model attempts to predict a value, for example a COVID-19 risk score based on features such as age, protein values, and biomarkers related to bacterial infection (Ye et al., 2021). Performance is assessed by comparing the predictions of the model against the actual values using performance metrics such as the *Mean Squared Error* (MSE), *Root Mean Squared Error* (RMSE), and *Mean Absolute Error* (MAE; Handelman et al., 2019). The MSE is the average of squared differences between the actual values and the values that the model predicted, defined as

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \hat{x}_i\right)^2 \tag{2.1}$$

where n is the number of data points, x_i the vector of observed values, and $\hat{x}_I$ the vector of predicted values. The RMSE is an extension of the MSE that does not punish larger differences as severely as it is, as the name suggests, the square root of the MSE and is defined as

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \hat{x}_i)}{n}}. \tag{2.2}$$

The MAE is the mean of the absolute error values divided by the number of predictions, and is defined as

$$MAE = \frac{\sum_{i=1}^{n}\left|x_i - \hat{x}_i\right|}{n}. \tag{2.3}$$

To narrow the scope further, the experiments discussed in this thesis are all classification tasks. A classification task is a task in which the model attempts to predict which group a sample belongs to. For example, the goal can be to determine whether someone has a particular disease or belongs to a control group, and the success of the model can be expressed using a variety of evaluation metrics (Dalianis & Dalianis, 2018; Handelman et al., 2019). The most used metric is accuracy, defined as

$$accuracy = \frac{number\ of\ correct\ predictions}{total\ number\ of\ predictions}. \tag{2.4}$$

For healthcare and medicine, sensitivity and specificity are important metrics. Sensitivity, also called the *true positive rate* (TPR) or recall, is defined as

$$sensitivity = \frac{number\ of\ true\ positives}{total\ number\ of\ positive\ cases} \tag{2.5}$$

where a true positive is someone with a pre-specified condition. Specificity is defined as

$$specificity = \frac{number\ of\ true\ negatives}{total\ number\ of\ negative\ cases} \quad (2.6)$$

where a true negative is someone without a pre-specified condition. In diagnostic models, the sensitivity provides an estimate of the capacity of a model to identify people with a disease and to exclude those people who do not have it, while the specificity estimates the capacity of a model to identify people without the disease, and to avoid accidental exclusion of those who do have it (Dalianis & Dalianis, 2018). Accuracy, sensitivity, and specificity take values between 0 and 1, 1 being highest, and are often expressed as a percentage by multiplying by 100.

Precision, or the *positive predictive value* (PPV), is defined as

$$precision = \frac{true\ positives}{true\ positives + false\ positives} \quad (2.7)$$

where a false positive is someone identified as having a condition when, in fact, this is not the case. Accuracy, sensitivity, specificity, and precision take values between 0 and 1, 1 being highest, and are often expressed as a percentage by multiplying by 100.

Another commonly used metric is the F1-score, which is the harmonic mean of precision and sensitivity:

$$F1 = 2\frac{PPV\ x\ TPR}{PPV + TPR}. \quad (2.8)$$

Lastly, researchers might choose to present the *Area Under the Curve* (AUC) score and AUC-ROC curve (Handelman et al., 2019), where ROC stands for *receiver operating characteristic*. In a graph of an AUC-ROC curve, the TPR, or sensitivity, is plotted on the y-axis and the *false positive rate* (FPR), which is 1-specificity, is plotted on the x-axis. A perfect model would predict everything correctly, so the AUC-score would be 1.0. A model that guesses and predicts as many true positives as false positives would produce an AUC-score of 0.5. The AUC can be defined as

$$AUC = \frac{\sum y_{i=}+1, y_j=-1^{\delta(f(x_i)>f(x_j))}}{y_+ y_-} \quad (2.9)$$

where y_+ and y_- are the number of positive and negative instances, respectively, and $\delta(e)$ is either 1 or 0, depending on whether the expression is true or not (Suominen, 2009).

2.3.2 Machine learning algorithms used in Parkinson's disease classification using electroencephalography data

A wide variety of algorithms can be used for classification tasks, and many of them have been used for classification tasks concerning PD and EEG. Covering all algorithms that could be used for the purposes of classifying PwPD would require a thesis-sized manuscript in itself, so only algorithms that will be used in the studies in this thesis will be discussed. Those algorithms are the *Support Vector Machine* (SVM; Evgeniou & Pontil, 2001; Noble, 2006; Orrù et al., 2012), the *Random Forest Classifier* (RFC;

Ho, 1995; Breiman, 2001), and the *Gradient Boosting Classifier* (GBC; Friedman, 2001; Natekin & Knoll, 2013). Of these, the SVM has been used most often in medical classification tasks, and specifically in EEG classification tasks (Polat & Güneş, 2007; Yu et al., 2010; Orrù et al., 2012). Lastly, while they are not used in this thesis, *artificial neural networks* (ANNs; Jain, Mao, & Mohiuddin, 1996; Krogh, 2008) will be briefly discussed as they are very popular and have been widely used in PD classification.

The idea behind the SVM can be summarised in one sentence: draw a line through the dataset that maximises the distance between the data points of the separate classes and the line (Hearst et al., 1998; Evgeniou & Pontil, 2001; Noble, 2006). For high-dimensional data, this line is called the optimal hyperplane. The SVM derives its name from the optimisation of distance between the hyperplane and the lines that can be drawn parallel to the optimal hyperplane at the maximal distance where the nearest instances of the particular class, called support vectors, are located (Figure 2.5; Hearst et al., 1998; Noble, 2006). The SVM algorithm can do this for both linearly separable datasets and nonlinearly separable datasets. To accomplish separation of nonlinear data, SVMs make use of kernels, which are mathematical functions that transform the data into shapes that can be separated into classes (Hofmann, 2006). What this means is that a kernel can be used to map the nonlinear data from a lower-dimensional space to progressively higher dimensional spaces until the data are separable. There are many different kernels that can be used to accomplish this, such as the *radial basis function* (RBF) kernel, which uses the distance, often the Euclidean distance, between data points and a fixed, predetermined point to transform the data. This is, therefore, used when there is no prior knowledge of the data. Other kernels include the polynomial kernel and the sigmoid kernel, but it is also possible to develop customised kernels.

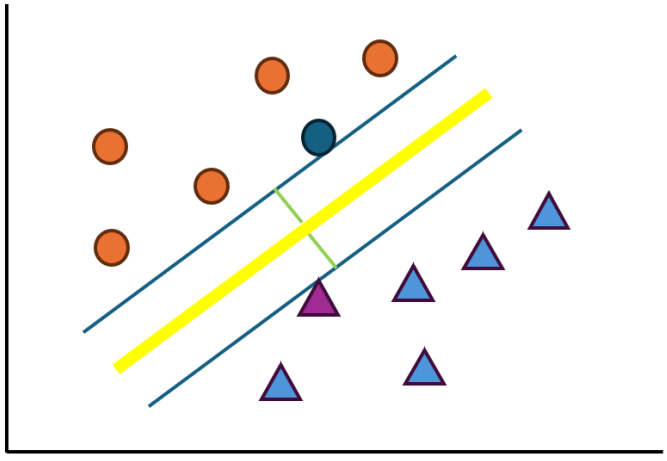


Figure 2.5. A depiction of the SVM. Two classes, triangles and circles, are separable by drawing a hyperplane, the thicker yellow line, between them. The maximum margin, indicated by the two green lines perpendicular to the hyperplane, is determined by the distance between the hyperplane and the two outermost blue lines, called the support vectors, each of which reflects the distance of the individuals closest to the hyperplane in each of the two classes, here depicted in purple for the triangles and navy for the circles.

To understand how an RFC (Ho, 1995; Breiman, 2001) works, it is important to understand what a decision tree is (Kingsford & Salzberg, 2008). A decision tree consists of a series of decision nodes connected to each other and to leaf (or terminal) nodes. The first decision node is called the root node. At each decision node, a choice is made that leads either to another decision node or to a leaf node, the latter representing the end of the decision tree. For example (Figure 2.6), if a decision tree was set up to determine whether a particular bird was a duck or not, the algorithm could start at a root node which asked the question "does it look like a duck?" If the answer was "no", the decision tree would end on a leaf node at that point, concluding that the bird was not a duck. On the other hand, if the answer was "yes", the algorithm would continue to the next decision node which asked, "does it walk like a duck?" Again, this would lead to a leaf node if the answer was "no" or to a further decision node asking "does it quack like a duck?" If the answer was "yes", this final decision node would then lead to two possible leaf nodes, one concluding that the bird was a duck and the other that it was not.

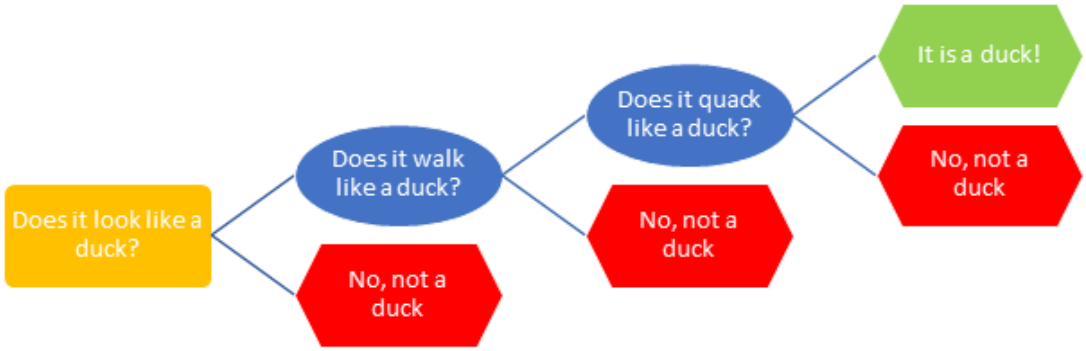


Figure 2.6. An example of a decision tree, with a root node on the far left, two further decision nodes (the ovals), and four leaf nodes (the six-sided boxes).

An RFC is an extension of the decision tree (Ho, 1995; Breiman, 2001). However, most problems are not as clear-cut as the duck example. Outcomes tend to vary based on which decisions are made at each decision node, and the order in which they are presented, RFC uses a large number of decision trees that operate in parallel. Each decision tree produces its own class prediction for each individual sample in question, and the class with the most votes across all decision trees is then chosen as the answer. The large number of decision trees operating in parallel has given rise to the word "Forest" in the term RFC.

The above only works if the individual decision trees are not all given the same features in the same order. To avoid this scenario, a procedure known as Bootstrap Aggregation, or 'bagging' for short, is employed (Breiman, 1996). Bootstrapping (Efron, 1992), referring to 'bootstrap sampling', means each of the individual decision trees is allowed to draw on a pre-specified limited number of features, with replacement of features used by subsequent decision trees. This means that, if the pre-specified number of features was four out of a total of five possible features, the first decision tree might draw features 2 and 3, and two instances of feature 4, while the second decision tree would draw features 1

and 5, and two instances of feature 2, and so on. This method ensures that the correlation between the different decision trees is low, something which is a prerequisite for an RFC to work successfully, as highly correlated trees would produce similar outcomes with similar variance and be more prone to overfitting.

While both bootstrapping (Efron, 1992) and *cross-validation* (CV) (Arlot & Celisse, 2010) are methods of resampling, bootstrapping does so with replacement (Kohavi, 1995), while CV does so without replacement (Arlot & Celisse, 2010; Berrar, 2019; see Section 2.3.3). Because it is possible for the same samples to be drawn several times during bootstrapping, getting an accurate estimate requires more computation than is required for CV (Kim, 2009). Another difference is that, while it is possible to use either method for model assessment (Kohavi, 1995), in general CV is used to obtain an assessment of model performance, while bootstrapping is used to estimate distributions of data.

A major advantage of RFCs is that they can easily extract feature importance from models (Qi, 2012). For each decision within a decision tree, it is possible to calculate the Gini importance of the feature relevant to that particular decision (Menze et al., 2009). The Gini importance provides a measure of the purity of the split generated by that particular feature, meaning that it indicates how well a feature can split the data into pure classes. The capacity to extract this information easily from the RFC means that, not only is there transparency regarding how the model's decisions were made, but the RFC itself can be used to select features for use in other algorithms. By first having the RFC calculate the Gini importance of each of the different features, it is possible to limit further use of features in other algorithms to those that score above a certain threshold.

The GBC (Friedman, 2001; Natekin & Knoll, 2013) is, like the RFC, an ensemble method that makes use of decision trees, but the manner in which it does so is quite different. A GBC starts with a base estimator, known as a decision stump, which is effectively a decision tree that makes a prediction using just one feature. Using a loss function, the difference between the predicted outcomes and the actual values of the samples is then calculated. A second estimator is then used to correct the errors made by the first estimator as far as possible. This process is continued with subsequent estimators until the goal of minimising the loss is achieved. The GBC gets its name from the fact that this minimisation of loss is achieved using gradient descent and performance is improved by building on the previous model.

As mentioned above, it is necessary to discuss ANNs briefly even though they are not used in this thesis, as they have been used frequently in PD classification tasks (Maitín, García-Tejedor, & Romero Muñoz, 2020; Maitín, Romero Muñoz, & García-Tejedor, 2022). ANNs have become popular ever since increases in processing power and the availability of large amounts of data have made it possible to develop them properly. An ANN consists of an input layer, one or more hidden layers, and an output layer (Wang & Wang, 2003). Each of these layers is comprised of a number of 'neurons', and

the connections between the neurons of different layers are assigned ‘weights’. Each of the input neurons has its own numerical value and the influence of an individual neuron on a given neuron in the next layer is determined by this value multiplied by the weight of the connection between the two neurons in question. In general, multiple neurons from one layer will connect to any given neuron in the next layer, and the inputs from each of the neurons in the first layer (i.e., their numerical values multiplied by their respective connection weights) will be summed together. Whether the neuron receiving this input will then activate (i.e., pass on further information to the next layer) is dependent on whether the total value (i.e., the sum of inputs from all the neurons contacting it from the previous layer) exceeds a pre-determined threshold for an activation function (Lederer, 2021). This activation function introduces non-linearity into the network (Wang & Wang, 2003): the summed total value is either greater than the threshold and the neuron is therefore activated, or it is not, meaning that the neuron is not activated. The process of the inputs passing through the network in this direction is called the forward pass. At the start of a training process, the weights of each of the connections between neurons are assigned at random. The weights are subsequently updated after each iteration of the forward pass through a process known as back-propagation, which modifies the weights using partial derivatives obtained from a loss function which assesses the overall performance of the network.

It generally takes considerable processing power and large amounts of data to train ANNs (Wang & Alexander, 2016). While it is possible to train a simple ANN with only one, or at most a few, hidden layers on a small amount of data, this is unlikely to outperform other algorithms like SVMs. ANNs only start to outperform other algorithms when they have access to larger datasets. Because the datasets used in this thesis are not sufficiently large to benefit from the capabilities of ANNs, all experiments here will be performed using conventional, and often simpler, algorithms that tend to be also easier to interpret.

2.3.3 Training and evaluating a machine learning model

The simplest form of training and evaluation of an ML model involves splitting the dataset into a training set and a test set, creating a model by training an algorithm on the training set, assessing that model by applying it to the test set, and then calculating the evaluation metrics (Arlot & Celisse, 2010; Berrar, 2019). The problem with this approach is that there is a possibility that the training and test sets were split in an artificially favourable way, leading to evaluation results that cannot be replicated if a different train-test split had been made. Conversely, a single train-test split might also provide evaluation results that under-represent its actual capability.

To alleviate this problem, a method called CV is used (Arlot & Celisse, 2010; Berrar, 2019). In the CV process, a dataset is divided into n parts; n-1 parts are used to train the model, and the model is then tested on the remaining part. N can be any number, but values of 5 and 10 are often used, the procedure then being called 5-fold CV or 10-fold CV. An alternative variation is leave-one-out CV, in

which the dataset is split into as many parts as there are samples, one sample being used for testing and the remaining samples for training. To obtain overall classification results, the evaluation metrics are averaged over *n* repetitions, and the standard deviation for each metric is calculated.

To get the best values for the hyperparameters, being the variables that control the learning process for the algorithms, hyperparameter optimisation is performed (Feurer & Hutter, 2019; Yang & Shami, 2020). Hyperparameter optimisation, in its simplest form, means the CV procedure is run several times with several different values and the optimal value is then chosen on the basis of which value led to the best evaluation metrics. Unfortunately, if hyperparameter optimisation is performed during the CV procedure described above, the best model cannot be reliably assessed because all data have been used to train the model and to obtain the best hyperparameters.

To overcome this problem, a dataset can be divided into training, validation (also called tuning), and test (also called evaluation) sets. It now becomes possible to use the CV procedure on the training and validation sets and then to use the hyperparameters obtained from this process to retrain the model and obtain performance metrics by testing it on the remaining holdout set. However, this creates a new problem because, once again, it is possible that the holdout set inadvertently consists of favourable samples leading to an overly optimistic evaluation (Varma & Simon, 2006).

The train/validation/test procedure can be extended to a method called nested CV, which alleviates the aforementioned problem. In a nested CV procedure (Varma & Simon, 2006), the dataset is once again divided into *n* parts (Figure. 2.7). One part is kept as the holdout set for testing and, of the remaining *n*-1 parts, another part is retained as a validation set to be used to obtain the best hyperparameters. All remaining parts are then used as the training set. The algorithm then cycles through the training and validation sets in a similar way to *n*-fold CV. This is called the inner loop. Once the inner loop has been completed, a new part is chosen as the test set. Inner loops are thus nested within an outer loop. The process is repeated until all parts have been allocated to the holdout set.

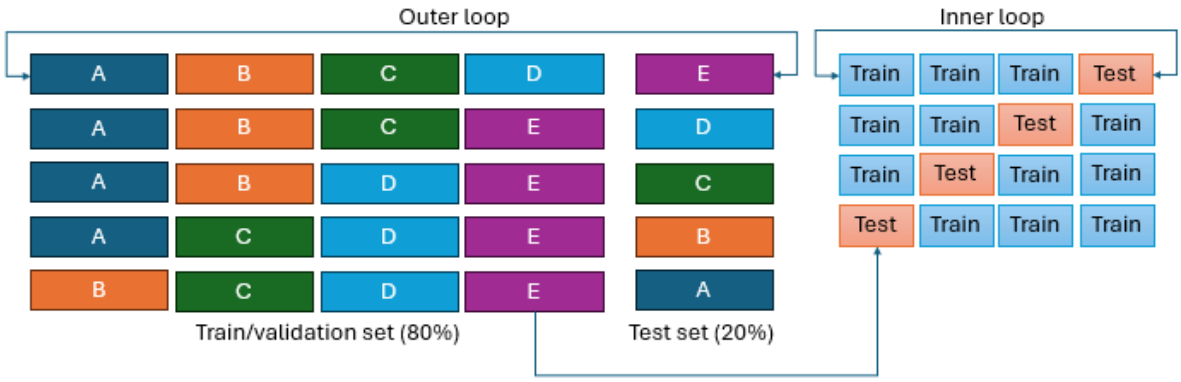


Figure 2.7. An example of nested CV. In the outer loop, the data are split into an 80% train/validation part on which CV is performed in the inner loop to obtain the best model, and a 20% part on which that model is tested. The outer loop finishes once each part has been used as the test set.

2.3.4 Statistical analysis

Let us assume a researcher ran an experiment using a SVM and an RFC to assess which algorithm performed better on a classification task. Let us say the SVM produced a mean accuracy of 75% and the RFC a mean accuracy of 77%. While tempting, to state that the RFC outperformed the SVM would be too simplistic. Unless the dataset was very uniform in its distribution, different folds in the CV process will produce different accuracy scores. Using these individual accuracy scores, one can use statistical methods to determine how significant the differences really are.

The simplest method would be to calculate the mean averages and standard deviations for both the SVM and RFC, and then use statistical tests such as a Student's t-test (Delacre, Lakens, & Leys, 2017), if the right conditions are met, to calculate a *p*-value. If this *p*-value was lower than a pre-determined *alpha level*, one could say that the result was statistically significant. If the researcher was testing more than two algorithms, they could also choose to use an *Analysis of Variance* (ANOVA; Ståhle & Wold, 1989).

If the researcher wanted to use a non-parametric test so that the specific distribution of the data did not matter, and had both the time and the computational resources, it would also be possible to run a permutation test, such as Fisher-Pitman permutation test (Berry, Mielke, & Mielke, 2002). In such a permutation test, the difference between the means of each of the two groups is first calculated, after which the groups are sampled at random and the resulting difference in means of the two smaller groups is calculated. This process is repeated multiple times and the total number of times that this difference is greater than the difference between the original two groups is counted. If this number is smaller than a pre-determined cut-off, the difference between the groups is considered to be significant.

The aforementioned statistical tests show whether the difference observed between two or more groups is likely to be a real difference or an occurrence due to chance, but it does not say anything about the magnitude of the difference (Sullivan & Feinn, 2012). Due to how significance is calculated, if the test population is large enough, even a small difference can become statistically significant. In medicine though, there are more factors to consider when determining whether something is clinically useful, such as adverse effects, time, and cost. Therefore, it is important to calculate the magnitude of an effect using what is called the effect size.

Effect sizes are generally measured using either Cohen's d (Cohen, 1969), being the difference between two means divided by the standard deviation, or Hedges' g (Hedges, 1981), which performs better when using small samples sizes. Interpretation of the resultant values are study-specific, but some guidelines have been proposed (Sawilowsky, 2009): an effect size of 0.01 is considered very small, effect sizes of 0.2, 0.5, and 0.8, small, medium, and large, respectively, an effect size of 1.2 very large, and an effect size of 2.0 huge.

2.4 Machine learning and electroencephalography in Parkinson's disease

Recently, ML researchers have started applying their knowledge to the domain of EEG in general, and to EEG datasets of PwPD and controls specifically. Maitín, García-Tejedor, & Romero Muñoz (2020) published a review on ML approaches for use in the detection of PD using EEG data and cited classification accuracies ranging from 62% to 99.62%. Nine studies were included in their review, and large differences were found between the experiments, with seven studies reporting on RSEEG and the other two reporting on motor activation EEG. The number of electrodes used by different studies ranged from 2 to 256, while preprocessing strategies varied from using raw data without any preprocessing to advanced pipelines that included filtering, electrode rejection and interpolation, and artifact rejection. Feature extraction methods and the resulting features that were used differed among studies and, in total, 11 different algorithms were used. All studies were published between 2017 and 2020.

Maitín, García-Tejedor, & Romero Muñoz followed their 2020 review with another review two years later (Maitín, Romero Muñoz, & García-Tejedor, 2022), in which they assessed 59 studies. Once again, they found great heterogeneity in preprocessing practices and a broad spectrum of extracted features, although time-frequency features dominated. SVMs and ANNs were the most commonly used algorithms, followed by *k-Nearest Neighbour* (kNN), RFC, and *convolutional neural network* (CNN) algorithms respectively.

While algorithms like the SVM can be trained using relatively small datasets, ANNs require large amounts of data for the training process to work, depending on their architecture and complexity (Wang & Alexander, 2016). Large EEG datasets are rare, so researchers have employed other methods to obtain sufficient data. Generally, in EEG research, the EEG recording of a participant is treated as just one sample and the individual features are extracted from the entire recording. Because EEG recordings, especially RSEEG recordings, often consist of several minutes of similar data, researchers in the field of ML sometimes divide the signal into smaller epochs and treat these separate epochs as independent samples (Maitín, García-Tejedor, & Romero Muñoz, 2020; Maitín, Romero Muñoz, & García-Tejedor, 2022). By doing this, a 2-minute recording can generate a total of sixty 2-second epochs, meaning that there are now 60 times more samples to use. This number can be increased even further by extracting epochs with an overlap. For example, with a 50% overlap, a researcher can extract three 2-second epochs from 4 seconds of EEG data. This method is not dissimilar from oversampling (Mohammed, Rawashdeh, & Abdullah, 2020), which is used to remedy class imbalances, except that the method here is applied to both classes to increase dataset size. In this thesis, the focus will be on conventional algorithms such as SVMs and RFCs because the datasets that will be used are small and multiple samples cannot always be extracted from the same participant for some of the comparisons.

Referring back to the 2022 review by Maitín, Romero Muñoz, & García-Tejedor, the highest classification accuracy for a non-ANN approach was reported by Yuvaraj, Rajendra Acharya and

Hagiwara (2018), who used *higher-order spectra* (HOS) to extract features from RSEEG recordings of 20 PwPD and 20 controls, each consisting of 5 minutes of data collected from 14 electrodes. After applying a 1 to 49 Hz bandpass filter, the data were segmented into epochs of 2-seconds and HOS features were extracted and these were, in turn, ranked using Student's t-tests. Features were added one by one to the classification algorithm until the highest accuracy was achieved. Using a variety of algorithms, the SVM with an RBF kernel achieved an accuracy of 99.62%.

Most studies have extracted features in the time-frequency domain (Maitín, García-Tejedor, & Romero Muñoz, 2020; Maitín, Romero Muñoz, & García-Tejedor, 2022), including the calculation of absolute power in the different frequency bands, relative power, and entropy. Features were extracted from single electrodes or from ROIs, and were extracted per participant, or, when using the oversampling method described above, by dividing each participant's EEG into smaller epochs and extracting multiple features from each epoch and effectively treating them as separate samples.

Feature selection methods varied widely between studies as well. Some studies used statistical tests, such ANOVA (Lee et al., 2021) and t-tests (Yuvaraj, Rajendra Acharya, & Hagiwara, 2018), to obtain a ranking of significant features and to undertake correlation analysis, while other ML techniques such as RFCs to obtain measurements like the Gini-importance of the features. Feature selection was performed at different points depending on the study. Some studies performed a preliminary feature ranking or used elimination as a method of dimensionality reduction before feeding the data into their chosen algorithm (Yuvaraj, Rajendra Acharya, & Hagiwara, 2018; Betrouni et al., 2019; Lee et al., 2021). Others used all features in their experiments to determine which features were most predictive of PD based on their classification outcomes (Chaturvedi et al., 2017). It should be noted that feature selection outside of CV has been shown to bias classification metrics positively in radiomics studies (Demircioğlu, 2021), the bias increasing as the number of features per sample increases, so the use of this practice in ML tasks concerning EEG from PwPD makes comparison and analysis of results difficult.

Another factor that makes comparison of results difficult relates to the manner in which the data were preprocessed. The studies discussed in both reviews by Maitín et al. vary widely in this regard. This is important because artifacts, especially muscle artifact, can have rather dramatic effects on classification outcomes. For example, Saikia et al., (2019) investigated the correlation between EEG and *electromyography* (EMG). They extracted three different features from 30 minutes of EEG recordings measured from two electrodes over the temporal and frontal regions, as well as 12 features from EMG measured from the participants' wrists. Using ANNs for classification of PwPD and controls, the researchers reported accuracies for EEG features of 62%, for EMG features of 73%, and 98.8% when using both sets of features together.

To put the effect of muscle artifact on EEG into perspective, a study by Whitham et al. (2007), set out to identify its contribution to EEG recordings during both RSEEG and cognitive tasks. Two participants had a 10-second period of RSEEG recorded with their eyes closed and a further 10-second period with their left eyes kept open to act as a baseline. Then, photic stimulation was performed for 10 seconds with eyes closed and again with the left eye open. Lastly, the participants performed an auditory oddball experiment. The cognitive tasks were performed twice, once under normal conditions and once while under complete neuromuscular blockade sparing the dominant arm. Comparing the power in the different frequency bands during the two different conditions, the researchers found that, in the normal state, EEG frequencies above 20 Hz were significantly contaminated by EMG. Specifically, there was a 6-fold reduction of EEG power in the range between 25 and 30 Hz during the paralysed state and a 100-fold reduction in the range between 40 and 100 Hz. As the authors pointed out, a 6-fold increase implied that the EEG signal measured in the normal state was effectively 84% EMG, demonstrating the need for adequate artifact removal if the intention was to investigate neural activity alone.

The above findings have important implications for classification tasks of PwPD and controls. Weyhenmeyer et al. (2014) investigated the difference in classification outcomes when using raw EEG and EEG data that had undergone full preprocessing, including artifact rejection, when classifying PwPD and controls, finding that raw EEG classified more accurately than preprocessed EEG. Using data from 9 PwPD and 10 controls recorded while performing a grasping task, the researchers used delay differential equations for classification and reported AUC-scores of 59%-86% for raw data and 57%-72% for preprocessed data. Their analysis of artifacts showed that classification using muscle artifact alone produced AUC-scores of 79%-92% for controls versus those PwPD who were not taking dopaminergic medication and 76%-90% for controls versus those who were taking dopaminergic medication.

To summarise, performance evaluation of previous experiments shows accuracies up to 99.62%, but comparison of results is difficult as the methods used vary widely. Experimental choices such as the number of electrodes used, the preprocessing steps taken to remove non-neural signals from the data, the manner in which features are created and selected, and the algorithms used for classification differ between studies. In particular, the presence of muscle artifact in certain experiments has been shown to positively influence classification outcomes.

2.5 Electroencephalography datasets

To perform the experiments described in this thesis, data was of course needed. The data used here are drawn from three separate datasets (Table 2.1): one was collected at the *Canberra Hospital* (CH) in Canberra, ACT, Australia, a second was made publicly available by the *University of New Mexico* (UNM), *United States of America* (US; Carr, 2017), and a third was made publicly available by the *University of Turku* (UTU), Finland (Suuronen et al., 2023). The use of the three datasets for the

experiments discussed in this thesis was approved by the *Australian National University* (ANU) Human Research Ethics Committee (protocol No. 2020/612).

Table 2.1. Details of the three datasets used in this study.

Dataset	Recording Duration	Sampling rate	N PwPD (female, male)	N Controls (female, male)	PwPD mean age [years]	Controls mean age [years]
CH	3 minutes	2048 Hz	21 (13, 8)	17 (9, 8)	68.2 ± 9.2	68.4 ± 8.7
UNM	1 minutes	500 Hz	28 (11, 17)	28 (11, 17)	69.8 ± 8.6	69.2 ± 9.2
UTU	2 minutes	500 Hz	20 (11, 9)	20 (12, 8)	70.0 ± 7.2	68.0 ± 6.0

For the CH dataset, PwPD were recruited through the neurology, aged care, and movement disorders clinics at CH and through Parkinson's ACT Association. Participants were English speaking, at least 18 years of age, had no diagnosis of cognitive impairment or history of developmental disability, alcohol abuse, or substance abuse. Patients had mild-to-moderate PD, no psychiatric disorders considered not to be due to PD, and no other movement disorders, while controls had no psychiatric or movement disorders. Controls were age- and sex-matched and were recruited from partners of PwPD and via community advertisement. PwPD were diagnosed by a qualified neurologist and fulfilled the *United Kingdom* (UK) Parkinson's Disease Society diagnostic criteria for PD (Hughes et al., 1992). Data were collected using a 64-electrode BioSemi Active II system, and impedances were kept below 5 Ω during recordings while two 3-minute recordings were made, one with eyes open and one with eyes closed. Collection of the CH dataset and use of all three datasets were approved by the ACT Health Human Research Ethics Committee (ETH.4.16.060) and the ANU Human Research Ethics Committee (protocol No. 2020/612). Written consent was obtained from all participants in the CH dataset.

The UTU dataset (Suuronen et al., 2023) consisted of 20 PwPD diagnosed using the UK Parkinson's Disease Society diagnostic criteria (Hughes et al., 1992) or the MDS clinical criteria (Postuma et al., 2015). The PwPD were matched with 20 controls without neurological diseases. The demographic details for both groups were similar, except that the PwPD groups showed more symptoms of depression as determined by the *Beck's Depression Inventory* (BDI; Beck, Steer, & Carbin, 1988). The EEG of participants was measured using a 64-electrode setup with a NeurOne Tesla amplifier and a 0.16-125 Hz online filter. Two 2-minute recordings were made, one while the participants had their eyes open and one while they had them closed. The eyes closed data for one patient were discarded due to a technical failure during the recording. The researchers obtained ethical approval for their experiments from the UTU.

The data in the UNM dataset (Cavanagh et al., 2018; Aljalal et al., 2022) were collected from 28 PwPD recruited from the community in Albuquerque, New Mexico, and 28 controls that had similar scores on the BDI (Beck, Steer, & Carbin, 1988) and *Mini Mental State Exam* (MMSE; Tombaugh & McIntyre, 1992); the latter is an examination of cognitive function that includes tests of attention, verbal memory, and visuospatial skills, among others. Data were collected using a 64-channel Brain Vision system and a 0.1 Hz to 100 Hz online filter and consisted of a 1-minute recording with the participants'

eyes closed followed by a 1-minute recording with their eyes open. Ethics approval for this study was obtained from the New Mexico Office of the Institutional Review Board and informed consent was acquired from participants before taking part.

2.6 A broader perspective: Parkinson's disease and Multiple Sclerosis

In Chapter 7, a literature review of the use of ERPs in the study of cognition in *Multiple Sclerosis* (MS) will be presented. This literature review is compared and contrasted with a literature review into cognition in PD by Seer et al. (2016) to put the results presented earlier in this thesis in the broader perspective of neurogenerative diseases. While PD and MS are very different diseases, the EEG methods used to investigate differences between people with either disease and controls, as well as disease progression, are similar. The hardware, software, and general preprocessing and analysis methods remain the same regardless of what disease is studied, because the methods are based on physiological processes that are the same in every healthy human being. How the results that are obtained differ between PwPD and *people with Multiple Sclerosis* (PwMS) is important to consider, but the methods themselves are the same, with protocols being adjusted to suit the peculiarities of the disease. Additionally, PD and MS have in common that diagnosis and prognostication suffer from similar problems of delay and uncertainty, as will become clear in the next paragraph.

2.7 Multiple Sclerosis

MS is a neurodegenerative disease in which the insulation, or myelin, of parts of the central nervous system is damaged, leading to visual, motor, and cognitive problems, among others. The disease can be divided into several clinical types based on the pattern of symptoms (Lublin, 2014). Most patients begin with recurrent episodes of neurological dysfunction interspersed by periods of remission which can last months or even years (*relapsing remitting MS*, RRMS). In about 50% of these cases, the initial relapsing and remitting phase transforms into a pattern of steadily progressive worsening *(secondary progressive MS*, SPMS). A small group of patients experiences steady disease progression from the outset (*primary progressive MS*, PPMS).

The current gold standard of MS diagnosis is a set of guidelines called the McDonald criteria (Thompson et al., 2018), which make heavy use of *magnetic resonance imaging* (MRI) to identify lesions in the brain and spinal cord. It can be very hard, and sometimes impossible, to make a definite diagnosis of MS, as many of the clinical features can be caused by other diseases; as a result, misdiagnosis is not uncommon (Solomon et al., 2016). The current diagnostic criteria require clinical symptoms to be present before a diagnosis of MS can be made, but, in rare cases, MRI scans performed for other reasons can suggest MS without the patient having had any symptoms at that time. Approximately a third of people with this so-called radiologically isolated syndrome go on to be diagnosed with MS in the next five years (Thompson et al., 2018).

The equivalent of the MDS-UPDRS in MS is a test called the *Expanded Disability Status Scale* (EDSS) which is used to measure disease severity. The EDSS was proposed John F. Kurtzke (1983) as a means of quantifying a person's disability and was an extension of a scale he had developed previously. The EDSS assesses eight 'functional systems' - pyramidal, cerebellar, brain stem, sensory, bowel & bladder, visual, cerebral, and other - on a scale from 0, being a normal neurological exam, to 10, being death due to MS. While considered the gold standard test to quantify disease progression in MS, the EDSS is not without criticism (for instance, see Meyer-Moock et al., 2014). A major criticism is the overemphasis on the person's ability to walk, which dominates the higher scores. It has been observed that scores tend to follow a bimodal distribution, clustering around 3.0 and 6.0 (Amato & Ponziani, 1999), and that the scores do not follow a linear trend. The difference between a step at a low score, for example between 1.0 and 1.5, and a step at a higher score, for example from 6.0 to 6.5, are different and cannot be compared. Additionally, there is significant inter-rater variability, especially in the lower scores, and scores on an individual system can vary markedly even when overall scores are comparable (Amato et al., 1988). Lastly, it has been observed that the probability of disease progression at different levels is not the same, and the EDSS has been deemed too unresponsive to have value in the evaluation of the effect of clinical interventions (Sharrack et al., 1999).

To summarise, while very different diseases in their causes and mechanisms of neurodegeneration, MS and PD have a few things in common. Both diseases can take a very long time to be properly diagnosed, and are subject to misdiagnosis, leading to late intervention and management by clinicians. Both the EDSS and MDS-UPDRS have been shown to be unreliable in tracking disease progression and cannot be used to predict disease progression. It is, therefore, important for both diseases that alternative strategies are identified to determine biomarkers of the diseases, which, as we will see, can be similar in both PD and MS.

2.8 Chapter summary and conclusion

In this chapter, PD as disease, EEG as a method of measuring brain activity, and ML as a field that can provide methods to classify PwPD and controls have been discussed. It covered how EEG data can be collected, how they can be preprocessed, and how EEG signals can differ between PwPD and people who do not have PD. It also covered how ML models are trained and evaluated, and the algorithms which are most often used. It then looked at the application of ML to PD classification using EEG data and explored why it can be difficult to compare results between different studies.

The background and the methods presented in this chapter provide a basis for the chapters that follow. They have clarified what a PD classification task looks like in general terms and what problems can be encountered when trying to classify PwPD and controls. The following chapters will follow the general order of the pipeline steps and investigate how different choices can affect evaluation results at each step, thereby determining how performance in PD classification tasks might be improved.

Lastly, MS was introduced as a neurodegenerative disease that suffers from similar problems in diagnosis and prognosis as PD. The methods of investigating cognition using ERPs in both diseases will be discussed and compared in order to put the earlier results presented in this thesis in a broader perspective of neurogenerative diseases.

Chapter 3: Preprocessing resting-state electroencephalography data and its effects on classification metrics

In this chapter, the importance of thorough preprocessing of *resting-state electroencephalography* (RSEEG) data in the context of its effect on evaluation metrics in classification tasks of *people with Parkinson's disease* (PwPD) and controls is investigated. The chapter is a continuation of, and elaboration on Sections 2.2.2 and 2.4 of Chapter 2. Preprocessing of *electroencephalography* (EEG) is often the first step that needs to be taken after collecting data because the raw data are comprised of many different signals of which many are of non-neural origins. Therefore, if one wants to analyse pure EEG data, these non-neural signals need to be removed first.

Before EEG data can be used, it generally needs to undergo preprocessing to remove non-neural signals. Section 2.2.2 provides an overview of preprocessing steps generally taken to clean EEG data. These steps include down-sampling for easier processing, high- and lowpass filtering, electrode removal and interpolation, and artifact rejection. The steps can be performed manually, or by using purpose-built functions, such as the *Artifact Subspace Reconstruction* (ASR; Mullen et al., 2015; Chang et al., 2019), Cleanline (Mullen, 2012), and ICLabel (Pion-Tonachini, Kreutz-Delgado, & Makeig, 2019) plugins for the widely used EEGLAB toolbox for EEG data preprocessing in MATLAB. The preprocessing used in *machine learning* (ML) tasks varies widely, but research into the effects that different levels of preprocessing have on performance when classifying PwPD and controls is sparse. Accordingly, these effects are investigated and reported on in this chapter.

Once data has been preprocessed and divided into epochs, one more step can be taken to reduce the amount of noise in the signal by averaging epochs. While this reduces the amount of noise, it also reduces the amount of data available, so ML researchers, especially those using neural networks, might choose to forego the averaging step. Like the effects of different levels of preprocessing, research comparing experiments that utilised averaging or not is sparce. In this chapter, the effects of preprocessing on averaged and single epochs are presented as two studies.

For both studies, the *Canberra Hospital* (CH), the *University of New Mexico* (UNM), and the *University of Turku* (UTU) datasets were preprocessed to four different levels and used as input to *Support Vector Machine* (SVM) algorithms for classification. The outcomes showed different results for the preprocessing stages. In particular, by comparing results from data that contained muscle artifact and results from data that did not contain muscle artifact, it is shown that data with muscle artifacts produced higher evaluation metrics. The research included in this chapter was presented at MedInfo

2023 and a more elaborate version of the proceeding's publication (Vlieger et al., 2024) was made available via MedRxiv (Vlieger et al., 2023a).

3.1 Evaluating effects of resting-state electroencephalography data preprocessing on a machine learning classification task for Parkinson's disease

While many studies have investigated different ML methods for the classification of PwPD and controls, the effects of their preprocessing methods on the experimental outcomes remains unclear. Most experiments used RSEEG data with a minority using task-specific data for classification, but that is often where the similarities end. This is remarkable, as the chosen preprocessing of RSEEG data has been shown to influence the results that are obtained (; Winkler et al., 2014; Ríos-Herrera et al., 2019; Yao et al., 2019; Karpiel et al., 2021).

Because preprocessing can influence results, it is important to determine what those specific effects are with specific attention being paid to the effects of muscle artifact as research has shown it to increase evaluation outcomes. What follows is an investigation of common preprocessing methods found in the literature and classification of PwPD and controls using one of the most popular algorithms, the SVM.

Acknowledgment: The work on averaged epochs discussed in Sections 3.1.1 through 3.1.5 is based on Vlieger et al., (2024), which I presented at MedInfo 2023. A more elaborate version of the presentation can be found on MedRxiv (Vlieger et al., 2023a). I came up with the initial experiments, after which they were refined in collaboration with my supervisors and co-authors Dr. Elena Daskalaki, Associate Professor Deborah Apthorp, Professor Emeritus Christian J. Lueck, and Professor Hanna Suominen. After writing the first draft, their comments were used to improve and extend the manuscript.

3.1.1 Introduction

As discussed in Chapter 2, two systematic reviews (Maitín, García-Tejedor, & Romero Muñoz, 2020; Maitín, Romero Muñoz, & García-Tejedor, 2022) were conducted into ML approaches for the detection of PD using EEG data and one of their findings was that preprocessing procedures between studies varied immensely. Based on the similarity of evaluation results, they concluded that the different preprocessing protocols did not influence results.

Because the ML and preprocessing methods used vary widely it is difficult to compare results. While the evaluation results between studies might have been similar, this does not mean that the manner in which these results were obtained can be considered equivalent. Differences exist in the frequency bands between PwPD and controls, raw data contains signals from a variety of different sources, and muscle artifact being present in the samples can lead to higher accuracies and frequencies

over 20 Hz are increasingly contaminated by muscle artifact. Therefore, if no or limited preprocessing is applied, it is hard to determine what exactly led to the successful classification.

Considering the heterogeneity in preprocessing methods found in the literature, a study was carried out to evaluate the difference in classification performance resulting from different preprocessing pipelines. Using three independent RSEEG datasets, each collected while the participants had their eyes closed, the preprocessing pipeline was divided into four stages. Features were extracted at each of these stages, and classification performed, after which differences in evaluation metrics and the features that were used during classification were analysed.

3.1.2 Methods

Eyes closed data from the CH, UNM, and UTU datasets were preprocessed (Figure 3.1) using MATLAB R2018b and the MATLAB toolbox EEGLAB 2020.0 (Delorme & Makeig, 2004), following the recommendations of the Swartz Centre of Computational Neuroscience, and its researchers (Miyakoshi, n.d). As mentioned in Section 2.5, the use of the three datasets for the experiments discussed in this thesis was approved by the Australian National University (ANU) Human Research Ethics Committee (protocol No. 2020/612). Data were rereferenced to the average of the electrodes and down sampled to 128 Hz. 1 Hz high-pass and 50 Hz low-pass filters were applied, and noisy electrodes were removed and interpolated using the EEGLAB extension 'clean_rawdata' (Mullen et al., 2015). Non-stationary artifacts were removed with the ASR function of 'clean_rawdata' (Mullen et al., 2015).

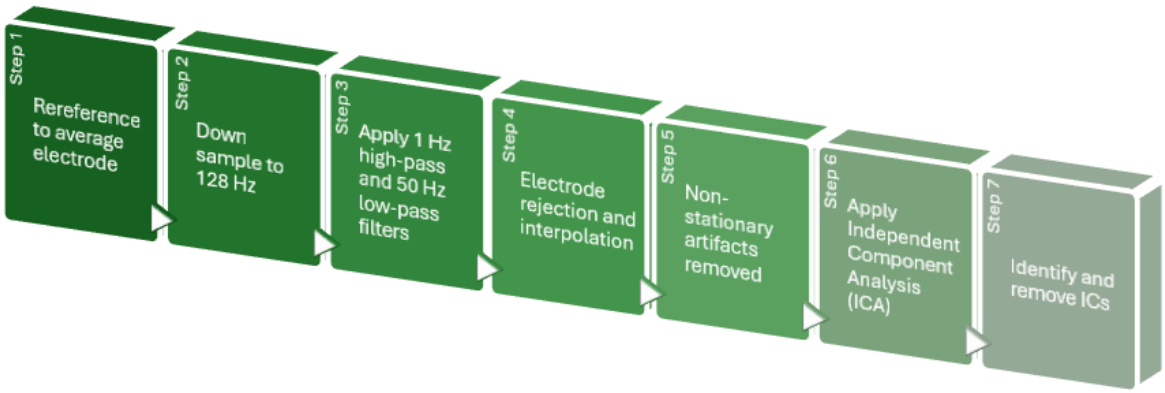


Figure 3.1. A depiction of the preprocessing pipeline described above.

When no information of electrode locations is available, 'clean_rawdata' uses statistical thresholding to determine whether an electrode should be removed (Delorme & Makeig, 2004). When channel locations are available, like in our experiments, it makes use of a method called '*random sample consensus*', or RANSAC (Miyakoshi, n.d; Bigdely-Shamlo et al., 2015). The RANSAC algorithm randomly subsamples data, which is assumed to contain both inliers and outliers, and fits a model to the data it considers inliers. This model is then tested with the full data and the data points that fit well are deemed the consensus set. This procedure is then repeated a set number of times with the aim of

producing a larger consensus set. Using this method, a consensus set of "good" electrodes is produced that is sufficiently correlated within a 5-second window and electrodes are rejected whose behaviour correlates less than a predefined threshold, set to 0.8 by default, with the time course that was predicted by the consensus set.

The ASR function of 'clean_rawdata' removes non-stationary artifacts from the data by dividing the continuous signal of all channels into 0.5-second windows with 50% overlap, applying *Principal Component Analysis* (PCA), and calculating the *standard deviation* (SD) for each *principal component* (PC; Miyakoshi et al., 2020). The SDs of these PCs is compared to the SD of the cleanest data ASR could find, called the calibration data, and if the SDs of the PCs that are being tested are *n* times higher than the SD of the calibration data the data are rejected. Threshold *n* is defined by the researcher before the experiment, which was 20 in this experiment. ASR then provides the option to correct rejected data, or to remove the rejected data, which is the recommendation of EEGLAB.

The *Independent Component Analysis* (ICA; Makeig et al., 1995) implementation 'runica' of EEGLAB and Iclabel (Pion-Tonachini, Kreutz-Delgado, & Makeig, 2019) were used to identify and remove *independent components* (ICs) that were likely to not be of neural origin. The ICLabel plugin for EEGLAB is an automated EEG IC classifier, consisting of an *artificial neural network* (ANN) architecture that was trained on a dataset of over 200,000 ICs from over 6000 EEG recordings, originating from a wide variety of experiments. The ICs were labelled by experts and other trained people with 7 different labels, being 'Brain', 'Eye', 'Muscle', 'Heart', 'Line Noise', 'Channel Noise', and 'Other'. An IC was given on average 3.8 labels, which was possible because, while most ICs are comprised of a signal primarily coming from one specific source, the ICs are composed of signals from several sources. When classifying an IC, a probability of the IC belonging to a label is produced for each of the seven categories, and the researcher can then select only the ICs that meet a desired threshold. For example, in this study only ICs that were deemed to be 70% or more likely to be 'Brain' were retained. In this way, muscle artifacts were removed from the signals and only the parts of neural origin were retained.

Six *regions-of-interest* (ROIs) over the scalp were created by averaging electrodes, placed in the locations seen in Figure 2.1b in Chapter 2, based on location (Figure 3.2), and the averaged signal for each ROI was divided into 2-second epochs. Those epochs in turn were averaged for each ROI to create 6 average epochs. Features, based on those used in the literature (Klassen et al., 2011; Chaturvedi et al., 2017; Geraedts et al., 2018; Betrouni et al., 2019), were then extracted using a Fourier Transform (Bracewell, 1986). Specifically, absolute power was extracted in frequency bands delta, theta, alpha, alpha1 (8-10 Hz), alpha2 (10-13 Hz), and beta, and relative power for these six bands was calculated by dividing the total power in one band by the total power of the delta, theta, alpha, and beta bands.

Lastly, two absolute power ratios were calculated, being the alpha1-to-theta ratio, and the high/low ratio:

$$absolute\ power\ high\ /\ low\ ratio = \frac{alpha+beta}{delta+theta}. \quad (3.1)$$

The reason for the calculation of power ratios as mentioned above was that, as explained in Chapter 2, it has been observed that there is a slowing of brain waves in PwPD compared to people without the disease (Geraedts et al., 2018), which effectively means that a larger proportion of the total power comes from lower frequencies. The ratios between power in the sum of higher and lower frequencies and in the power between a higher and a lower frequency band should therefore also differ between PwPD and controls.

To investigate the differences that differing levels of preprocessing have on classification outcomes, a total of 84 features (six absolute power values, six relative power values, and two ratio values) were extracted at four different stages: after re-referencing (Level 1), after filtering (Level 2), after running the full preprocessing pipeline without removing muscle artifact (Level 3), and after running the full pipeline (Level 4). Those Levels correspond roughly to four different manners in which preprocessing was performed in the two reviews mentioned previously (Maitín, García-Tejedor, & Romero Muñoz, 2020; Maitín, Romero Muñoz, & García-Tejedor, 2022).

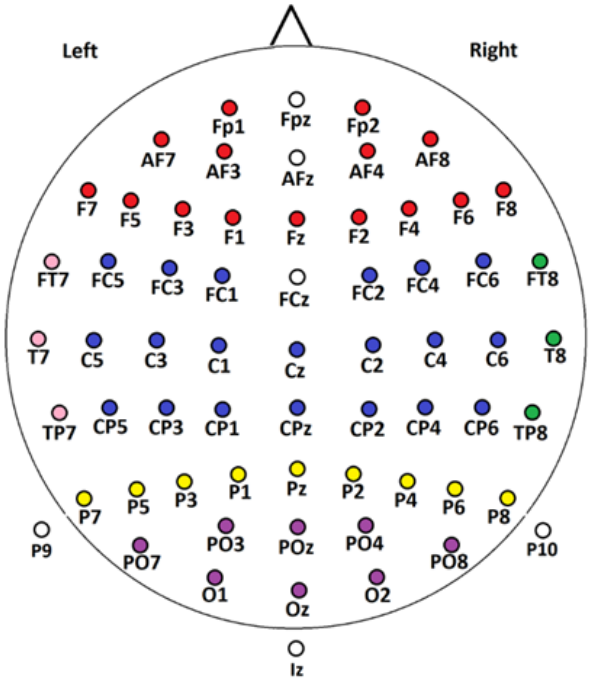


Figure 3.2. A scalp map of the 58 electrodes used in this study, colour-coded by their groupings into ROIs. Adapted from Vlieger et al., (2023b).

An SVM classifier (Noble, 2006) with a radial basis function (RBF) kernel was used for the classification. SVM algorithms have a long history in clinical tasks and are also often used in PD RSEEG classification tasks (Maitín, García-Tejedor, & Romero Muñoz, 2020; Maitín, Romero Muñoz, & García-Tejedor, 2022). It was, therefore, decided to use an SVM here as well to enable easier comparison to earlier studies (Polat & Güneş, 2007; Yu et al., 2010; Orrù et al., 2012). A log transform was used to normalise data through Scikit-learn's Power Transformer (Yeo & Johnson, 2000) which makes data more Gaussian by applying either Box-Cox transformations (Sakia, 1992) or Yeo-Johnson transformations (Raymaekers & Rousseeuw, 2021), depending on whether the data contains negative values or not. In our preliminary explorations, other methods to standardise the data were also tried, such as min-max scaling, using MinMaxScaler, and scaling based on the quartile range, using RobustScaler, but PowerTransformer outperformed those. The MinMaxScaler did not work properly as the data often contained outliers, and while the RobustScaler is supposed to handle exactly the problem of outliers, it was found that data which had undergone more normalisation provided better results.

For each of the three datasets, 10-fold CV was performed 10 times (Varma & Simon, 2006; Berrar, 2019). During each run of the CV, feature selection was performed using a *Random Forest Classifier* (RFC; Ho, 1995; Breiman, 2001), only retaining the top 30 features based on Gini-importance (Menze et al., 2009). Using grid search, hyperparameters of the SVM were optimised, while the optimal number of features was obtained by starting out with the two best features, finding the best hyperparameters, then adding the next best feature and optimising performance again until all features were included in the model. Performance was evaluated using accuracy, sensitivity, specificity, precision, and F1-score, and statistical significance tests were performed using Welch's t-tests (Delacre, Lakens, & Leys, 2017) and an alpha level of 0.05. The rather simple t-test was chosen here because simulation has indicated that even with a small number of samples, often being less than 100 samples unless dealing with extremely non-normally distributed data, the t-test is a valid statistical test (Lumley et al., 2002). Welch's t-test was chosen here as Delacre, Lakens, & Leys (2017) showed that it controls better for Type 1 errors, meaning rejection of a true null hypothesis than Student's t-test, and that it performs better when the two measured populations have unequal variance. All experiments were implemented in Python 3.8 (Van Rossum, 1995) and the Scikit-learn library was used (Pedregosa et al., 2011).

For CV, a train/validation split without a test set was used. This was necessitated by the small sizes of the datasets and warranted because the aim of the study was to evaluate the effects of preprocessing on an ML classification task rather than assessing the generalisation capabilities of the ML classifier on unseen data (i.e., on a separate test set). The problems with preprocessing would lead to models being selected that used non-neural signals, and as the test set would be preprocessed in the same manner, this problem would also occur when testing the model.

3.1.3 Results

Cleaning RSEEG and artifact removal inherently leads to a reduction in dataset size. Table 3.1 provides an overview of the reduction of total size per dataset and per preprocessing stage. While the number of participants stayed constant (39, 43, and 56 total participants in the CH, UNM, and UTU datasets, respectively), meaning the data for all participants was of sufficient quality to be used after preprocessing, removal of noisy electrodes and artifact removal meant a large reduction in all datasets for both controls and PwPD. The difference in dataset size between Level 1 (rerefenced data) and Level 4 (fully preprocessed data) was 77.63% for the CH data, 73.5% for the UNM data, and 77.4% for the UTU data, showing an equal reduction of data for all three datasets.

Table 3.1. An overview of data size per dataset and per preprocessing Level.

Dataset	Preprocessing stage	Dataset size controls [Megabyte]	Dataset size PwPD [Megabyte]
CH	Level 1	456	563
	Level 2	456	563
	Level 3	103	128
	Level 4	102	128
UNM	Level 1	404	404
	Level 2	404	404
	Level 3	110	116
	Level 4	107	115
UTU	Level 1	691	639
	Level 2	691	639
	Level 3	157	144
	Level 4	156	142

A comparison of the accuracy results of Level 4 to Level 3 using Welch's t-tests gave evidence of a significant difference for the UNM and UTU datasets ($p < 0.0001$ for both), accuracy being higher if muscle artifact was included (Table 3.2). Of interest here is the observed discrepancy, not only between different preprocessing stages, but also the datasets at the same preprocessing stage.

For the purposes of diagnosis, it is important to analyse the sensitivity and specificity of a model, as incorrectly diagnosing a healthy person with PD or not diagnosing a person who has PD can both be detrimental to that person. For both the UNM and UTU datasets, the sensitivity score is lowest for the Level 4 data, while the sensitivity score is second lowest for the CH Level 4 data. For all datasets, Level 4 data produces better scores for specificity than sensitivity, meaning all three models are better at classifying controls and PwPD. Precision is higher for all three datasets when using Level 4 data, but the F1-score is lower due to the lower sensitivity scores.

Table 3.2. Evaluation metrics from the described experiments in percentages on the validation set. The best scores for each dataset are in bold. Scores are in percentages from 0 to 100 with larger values implying better performance, and SD stands for standard deviation.

Dataset	Preprocessing stage	Accuracy ± SD [%]	Sensitivity [%]	Specificity [%]	Precision [%]	F1-score [%]
CH	Level 1	65.13±24.43	70.48	58.89	66.67	68.52
	Level 2	58.97±24.67	60.48	57.22	62.25	61.35
	Level 3	69.74±19.07	**71.43**	67.78	72.12	71.77
	Level 4	**73.85±21.20**	65.71	**83.33**	**82.14**	**73.02**
UNM	Level 1	66.96±17.36	61.79	72.14	68.92	65.16
	Level 2	**85.00±13.89**	**83.93**	**86.07**	**85.77**	**84.84**
	Level 3	79.11±15.10	75	83.21	81.71	78.21
	Level 4	66.07±20.36	57.86	74.29	69.23	63.04
UTU	Level 1	52.31±25.48	51.05	53.5	51.05	51.05
	Level 2	57.44±25.46	58.42	56.5	56.06	57.22
	Level 3	**71.03±21.52**	**63.68**	**78**	**73.33**	**68.17**
	Level 4	51.79±23.73	47.37	56	50.56	48.91

As no artifact rejection has taken place for Level 1 and Level 2 data, it is not possible to determine what specifically produced the differences in classification outcomes. In the case of Level 3 and Level 4 data, the only difference is the presence and absence of muscle artifact in the signal, so it was worthwhile to investigate from which ROIs and frequency bands the features used in classification originated as the difference can be attributed to this difference in the signal (Table 3.3). Because features selected by the RFC differed between the Level 4 data and Level 3 data, as well as between datasets, and were very diffuse, the analysis of features was done separately for location and frequency band.

Table 3.3. Distribution of features across ROIs and per frequency band, per dataset, and per preprocessing level, expressed as a percentage of the total number of features selected by the RFC.

Dataset	ROI	Level 4 [%]	Level 3 [%]		Dataset	Freq. band	Level 4 [%]	Level 3 [%]
CH	frontal	0	0		**CH**	delta	0	0
	central	0	29			theta	25	72
	parietal	25	14			alpha	25	14
	occipital	0	14			beta	0	0
	temporal left	75	29			ratio alpha1/theta	25	0
	temporal right	0	14			ratio high/low	25	14
UNM	frontal	0	10		**UNM**	delta	0	20
	central	50	17			theta	0	20
	parietal	0	20			alpha	100	37
	occipital	50	23			beta	0	20
	temporal left	0	20			ratio alpha1/theta	0	0
	temporal right	0	10			ratio high/low	0	3
UTU	frontal	31	25		**UTU**	delta	4	0
	central	7	75			theta	24	0
	parietal	14	0			alpha	41	25
	occipital	14	0			beta	17	0
	temporal left	24	0			ratio alpha1/theta	4	25
	temporal right	10	0			ratio high/low	10	50

For the CH data, 4 features were selected for the Level 4 data and 7 for Level 3 data, while for the UNM and UTU datasets these numbers were 2 and 30, and 29 and 4, respectively. In the analysis of the ROIs where the features came from, no discernible pattern was found, with ROIs from which features were selected varying widely between datasets. For the CH and UNM datasets, no features selected from the Level 4 data came from the frontal ROI, but 31% of features in the Level 4 UTU data experiment did come from the frontal area.

While ROIs from which features that improved classification were selected were diffuse, the analysis of the frequency bands these features came from was clearer. For the CH Level 4 data, 3 out of 4, or 75%, of features came from the alpha and theta frequency bands. For the UNM data, 100% of features came from the alpha band, while 69% or 20 out of 29 features came from the alpha and theta bands for the UTU Level 4 data. Compared to the Level 3 data, Level 4 data for each dataset had a larger number of features coming from the alpha band.

3.1.4 Discussion

In this study, the differences in classification performance of PwPD and controls when using different Levels of commonly performed preprocessing steps of RSEEG data, as well as the effects muscle artifact has on classification outcomes were evaluated. To enhance the validity of the observations, three different datasets were used, and the differences in features selected to attain these performance

metrics evaluated. The best (i.e., largest) metrics were achieved for the CH dataset using Level 4 (fully preprocessed) data, for the UNM dataset using Level 2 (filtered) data, and for the UTU dataset using Level 3 (with muscle artifact) data. By comparing Level 3 and Level 4 data, evidence was provided that leaving muscle artifact in the signal increases evaluation scores. It was concluded, therefore, that for the purposes of developing a tool for earlier diagnosis of PD, preferably before cardinal symptoms like tremor occur, it is important to filter such artifacts out and classify based on neural signals alone.

This study does not come without limitations. Most importantly, while the filtered data of the UNM dataset produced the highest evaluation results overall, it is difficult to provide a meaningful analysis of this result. Both the Level 1 (rereferenced) data and the Level 2 (filtered) data contain signals that are of non-neural origins, and, based on the observation that the results for these data differ from the results obtained using data containing muscle artifact and fully preprocessed data, it can be assumed that those non-neural signals do have an influence on the classification outcomes. A problem remained: there was no way to determine which signals did and did not contribute to the outcomes. The conclusion that can be drawn here is that there are non-neural signals that, when left in the data that is used for classification, will influence the evaluation results.

As expected, based on the findings in the literature (Weyhenmeyer et al., 2014; Saikia et al., 2019), for the UNM and UTU datasets, classification metrics were improved when using data retaining muscle artifact compared to the results when using fully preprocessed data. Contrary to the results from the UNM and UTU datasets, fully preprocessing the CH dataset increased performance. There are several possible explanations for this. For example, the EEG data were noisy, and the datasets were small (39, 43, and 56 total participants in the CH, UNM, and UTU datasets respectively). It is also possible that the controls in the CH dataset had more muscle artifact in their recordings compared to the other datasets, which obscured the signal and was filtered out, leading to increased performance.

The best-performing ROIs were not consistent across datasets. However, feature analysis was clearer: using fully preprocessed data, power in the theta and alpha bands contributed most for the CH set, while power in the alpha band contributed most for the UNM and UTU datasets. Features from the alpha band and theta band produced 75%, 100%, and 69% of features used for classification for the THC, UNM, and UTU datasets respectively. It is important to note that the cut-off points between frequency bands are still being debated (Olejarczyk, Bogucki, & Sobieszek, 2017; Zalewska, 2020), so selected features being spread over two adjacent frequency bands was not too surprising. Grouping frequency bands together would not be unreasonable and this suggests that a more detailed analysis of frequency in future research is worthwhile. To summarise, the alpha band produced the most features used in the classification process with some additional theta band features, but because these bands are right next to each other, other frequency splits might be worthwhile to investigate.

3.1.5 Conclusion

The results of this study indicate that, while evaluation results may be similar, the manner in which preprocessing is performed is of influence on these results. Specifically, the results show muscle artifact can lead to higher evaluation results compared to data that has undergone preprocessing that involves artifact rejection. The removal of artifacts is essential if the intention is to classify subjects based on neural activity, and, when this is done, theta and alpha features contribute most to classification accuracy. Further research is needed to determine which specific features are necessary for accurate classification.

3.2 Preprocessing without averaging epochs

In Section 3.1, one of the steps taken before extracting features was the averaging of epochs, a common technique that aims to reduce the noise in the signal. The assumption is that noise is random while the signal is not, so averaging epochs maintains the signal and removes noise. Because many ML algorithms work better when supplied with abundant data, researchers sometimes forgo the averaging procedure and treat every epoch as its own sample. In this manner, the subject does not have just one epoch but many more, inflating the total number of samples available for training. Because this method provides more data for training but also contains more noise than when averaging epochs, the experiments reported on in Section 3.1 were repeated without averaging to gauge how this difference affected classification outcomes.

3.2.1 Introduction

In the study reported on in Section 3.1, classification was performed on a per-participant basis: the data for each participant was divided into 2-second epochs, the epochs were averaged, and features extracted, leading to one feature set for each participant. Another method that is often encountered in the literature (Maitín, García-Tejedor, & Romero Muñoz, 2020; Maitín, Romero Muñoz, & García-Tejedor, 2022) involves the use of each epoch as separate samples, which leads to a much larger dataset. It is unclear whether this method would produce different results than the results obtained with the averaged epochs when using the same preprocessing steps.

3.2.2 Methods

All three datasets, being the CH, UNM (Carr, 2017), and UTU datasets (Suuronen et al., 2023), were used here and preprocessed to the same levels as explained in Section 3.1.2. Similar to the experiments reported on in Section 3.1, the use of the three datasets for the experiments discussed in this thesis was approved by the ANU Human Research Ethics Committee (protocol No. 2020/612). Instead of extracting features after averaging epochs, features were extracted from each 2-second epoch. As Table

3.5 shows, this produced many more samples than were available in the previous experiments reported on in Section 3.1.

Table 3.5. Comparison of the number of samples in each dataset per preprocessing step when averaging epochs and when using single epochs as samples.

Dataset	Preprocessing stage	N samples using averaged epochs	N samples using single epochs
CH	Level 1	38	2280
	Level 2	38	2280
	Level 3	38	1689
	Level 4	38	1681
UNM	Level 1	56	1680
	Level 2	56	1680
	Level 3	56	1117
	Level 4	56	1069
UTU	Level 1	40	2955
	Level 2	40	2955
	Level 3	40	2150
	Level 4	40	2122

While performing the experiments described here, care was taken to ensure data from a particular participant was only present in either the train or validation set and not both. The samples from one participant are not independent, so having samples of a participant in both the train and validation set would be a form of data leakage as the model would have been provided with information about the validation set during training, and, accordingly, would perform better than when the samples of one participant would be in either set but not both.

3.2.3 Results

Looking at Table 3.5 once more, the differences between the experiments reported on in Section 3.1 and those reported on here extend further than just the number of samples. The number of samples in Section 3.1 were equal to the number of participants and, therefore, the same for each preprocessing stage. Here, the number of samples varies between preprocessing stages, which was to be expected, because one of the steps of preprocessing was rejection of noisy data and epochs. This did mean that the ML algorithm had varying amounts of data to learn from depending on the preprocessing stage.

As with the results from the averaged epochs, the results for the single epochs differed per dataset (Table 3.6). The highest accuracies were obtained using rereferenced data from the CH dataset, using filtered data from the UNM dataset, and using fully preprocessed data from the UTU dataset. For each dataset, the highest accuracy achieved using single epochs was lower than when using averaged epochs. When comparing Level 4 data to Level 3 data using Welch's t-tests (Delacre, Lakens, & Leys,

2017), there was a statistically significant difference for the CH and UNM datasets (p = 0.0236 for the CH data, and p = 0.0077 for the UNM data).

Table 3.6. Evaluation metrics from the described experiments in percentages on the validation set. The best scores for each dataset are in bold. Scores are in percentages from 0 to 100 with larger values implying better performance, and SD stands for standard deviation.

Dataset	Preprocessing stage	Accuracy ± SD [%]	Sensitivity [%]	Specificity [%]	Precision [%]	F1-score [%]
CH	Level 1	**66.15±9.24**	**73.81**	56.69	**67.79**	**70.67**
	Level 2	64.36±14.88	71.33	55.74	66.56	68.87
	Level 3	64.97±17.11	68.08	**61.4**	67.01	67.54
	Level 4	59.77±15.05	59.38	60.24	64.08	61.64
UNM	Level 1	74.51±14.28	74.21	74.8	74.65	74.43
	Level 2	**75.93±9.73**	76.24	**75.62**	**75.77**	76
	Level 3	75.36±9.79	**82.83**	66.76	74.17	**78.26**
	Level 4	70.71±14.19	71.9	69.3	73.53	72.71
UTU	Level 1	65.21±15.64	**60.73**	69.35	64.74	62.67
	Level 2	62.94±16.02	56.21	69.17	62.81	59.33
	Level 3	64.97±16.72	55.36	**73.88**	66.3	60.34
	Level 4	**66.41±19.25**	59.46	72.89	**67.19**	**63.09**

The results in the single epoch experiments indicate that muscle artifact can increase evaluation results just as was observed in the averaged epoch experiments. In general, evaluation results were lower when using single epochs than when using averaged epochs, and the features were distributed more diffusely across both scalp locations and frequency bands. While this is true, when looking at the Level 4 data, which is the data of most interest because it contains the least non-neural signals, the evaluation results when using single epoch data were higher than when using averaged epochs for the UNM and UTU datasets. Additionally, sensitivity and specificity were more similar in values when using single epochs than when using averaged epochs for the UNM and CH datasets.

For the same reason as explained in Section 3.3, analysis focused on Level 3 and Level 4 data were analysed based on the ROIs and frequency bands. Similarly, as when averaged epochs were used, the ROIs that the selected features came from in the single epoch experiments were very diffuse and did not seem to show a specific pattern. Contrary to the results from the averaged epoch experiments, the division by frequency bands does not show a pattern either (Table 3.7).

Tables 3.7. Distribution of features across ROIs and per frequency band, per dataset, and per preprocessing stage, expressed as a percentage of the total number of features selected by the RFC.

Dataset	ROI	Level 4	Level 3		Dataset	Freq. band	Level 4	Level 3
		[%]	[%]				[%]	[%]
CH	frontal	10	20		CH	delta	5	7
	central	25	13			theta	30	33
	parietal	10	17			alpha	15	17
	occipital	20	20			beta	45	30
	temporal left	20	10			ratio alpha1/theta	0	7
	temporal right	15	20			ratio high/low	5	7
UNM	frontal	13	33		UNM	delta	7	0
	central	20	0			theta	17	33
	parietal	13	0			alpha	47	67
	occipital	33	67			beta	23	0
	temporal left	13	0			ratio alpha1/theta	7	0
	temporal right	7	0			ratio high/low	0	0
UTU	frontal	67	17		UTU	delta	0	10
	central	0	27			theta	33	30
	parietal	33	13			alpha	0	20
	occipital	0	13			beta	67	20
	temporal left	0	10			ratio alpha1/theta	0	3
	temporal right	0	20			ratio high/low	0	17

3.2.4 Discussion

In the experiments described here, it was investigated whether the results reported on in Section 3.1, where averaged epochs were used, would be similar when using single epochs. The results provide evidence that, indeed, muscle artifact leads to higher evaluation results, but no clear pattern was found in the ROIs or frequency bands from which features that best helped classify originated from. For the UNM and UTU datasets, Level 4 data produced better results for single epochs than when using averaged epochs, which is an important finding as this is the data that contains the least non-neural signals and would, therefore, be used to develop models for earlier PD diagnosis.

A possible reason for the diffuse features could be the presence of outliers in the data. An average number of approximately 52 epochs was used to create the averaged epochs, which would generally mean that an outlier in one specific point of one epoch would be attenuated by the other 51 epochs. Even if each epoch had several deviant points, unless these deviant points occurred at the exact same point in multiple epochs, they would be averaged out. In the single epochs, all these deviant points would be present and influence the classification outcomes.

Similar to the use of averaged epochs, a limitation of the work presented here is that it is not possible to determine which non-neural signals contributed to the increased evaluation metrics for the Level 1 and 2 data. Especially non-stationary artifacts could play a large role here, because, as

previously mentioned, no averaging would take place. This could lead to them having an even larger effect than when the averaging took place, because while care is taken to handle artifacts with well-established methods, some of these cannot be completely removed.

For the UNM and UTU datasets, the evaluation results when using fully preprocessed single epoch data were higher than when using averaged epochs. Given that our focus is on data that contains as few signals from non-neural origin as possible, single epoch approaches are to be investigated. As outliers might be a problem when using the single epoch approach, outlier handling should be investigated too. Here ROIs and eyes closed data was used, but single electrodes and eyes open data could also be used. Whether choices for ROIs or single electrodes and eyes open or eyes closed data influence classification outcomes are questions yet to be answered. To summarise, going forward the following should be investigated regarding single epoch approaches:

1) Is outlier handling necessary when not using averaging of epochs, and, if so, which method is best?
2) Does the use of ROIs or single electrodes lead to better results?
3) Is it better to use data collected with the eyes open or with the eyes closed?

The research here was performed with the intention of clarifying the role of preprocessing of RSEEG data in a ML classification task of PwPD and controls. The preprocessing methods here are not specific to PD, as they are guidelines for EEG research in general that are included in most EEG preprocessing toolboxes and recommended by their developers (Miyakoshi, n.d; Oostenveld et al., 2011). Therefore, the results here can be placed in a broader context of neurodegenerative diseases and would only require recognition of factors that might be specific to a disease, like tremor is in PD, or at most adjustment of parameters for data cleaning for methods such as ASR and ICA.

3.2.5 Conclusion

As with averaged epochs, the results of this study indicate that the manner in which preprocessing is performed when using single epochs is of influence on the results. Muscle artifact can also lead to higher evaluation results compared to data that has undergone preprocessing that involves artifact rejection when using single epochs, similarly to results obtained when using averaged epochs. When using fully preprocessed single epochs, higher evaluation scores were obtained for two out of three datasets compared to when averaged epochs were used. Single epoch approaches may therefore be better but may need additional outlier handling because of the fact that noise will not be averaged out.

3.3 Chapter summary and conclusion

In this chapter, the effects of preprocessing steps on evaluation results were investigated by preprocessing three datasets to four different Levels, using an RFC for feature selection, and classifying controls and PwPD using an SVM. The results from both the experiments with averaged epochs and the experiments with single epochs show that muscle artifact can lead to higher evaluation results. In the averaged epoch experiments, the features that were selected by the RFC primarily came from the alpha and theta bands, but the ROIs these features came from did not show a pattern.

In the single epoch experiments, evaluation results generally were lower for each dataset, but were higher for the UNM and UTU datasets when using fully preprocessed (Level 4) data. As fully preprocessed data are the focus, using single epoch data instead of averaged epoch data should be investigated, especially paying attention to the possible role outliers might play when no averaging of epochs is employed. As for where the ROIs and frequency bands features in single epoch experiments came from, no clear pattern emerged, which might be due to outliers in the data.

In both the averaged epoch and single epoch experiments, comparisons of three datasets were made, which showed that results can vary based on the data source. All three datasets were collected on different equipment, for different durations, and a different sampling rates. While all datasets were preprocessed in the exact same manner, the results do show heterogeneity, which can be explained by the heterogeneity of the data. PD is a disease that produces a wide variety of symptoms and progresses at different rates in different individuals, and the datasets all contained a rather small number of participants. It is therefore important to ensure that, apart from making sure that proper preprocessing is carried out, the dataset itself is sufficiently large to capture the breadth of PD progression. Standardisation of acquisition methods should also be encouraged, to allow for easier combination of data from different sources into larger datasets.

In the experiments in this chapter, ROIs were used instead of single electrodes and the data was obtained while the participants had their eyes closed. Whether single electrodes and eyes open data would lead to different, or possible better, results, remains to be investigated. The results presented here provide an indication of how such experiments could be performed: by using fully preprocessed data that has undergone artifact rejection, and by using single epoch data instead of averaged epoch data, as the results here indicate that leads to better outcomes. By using the results presented here in this way, the results can be placed in a clearer context in relation to previous experiments, and steps can be taken to develop RSEEG into a clinical tool for PD.

Chapter 4: Effects of using single or averaged epochs and electrodes from different resting-state electroencephalography datasets

As discussed in Section 2.2.2, researchers studying *electroencephalography* (EEG) will often average either epochs or electrodes to improve the *signal-to-noise ratio* (SNR) under the assumption that the noise is randomly distributed and will average out. The averaging techniques significantly reduce the size of the dataset, something that can be problematic when researchers wish to apply *machine learning* (ML) techniques that often require a sufficiently large dataset to train on. Instead, ML experiments using *resting-state EEG* (RSEEG) data will often divide the recording from a participant into several epochs, sometimes with overlap between those epochs, and treat each epoch as its own sample to increase dataset size. In Chapter 3, the results indicated that the choice of using single epochs or averaged epochs has an effect on the outcomes of experiments. Specifically, when using fully preprocessed data, single epochs performed better than averaged epochs.

RSEEG datasets often consist of two recordings per participant, one recorded with their *eyes open* (EO) and one with their *eyes closed* (EC: Carr, 2017; Suuronen et al., 2023; Khare, Bajaj, & Acharya, 2021). The majority of studies use data collected with EC, with few studies comparing the different recordings (Suuronen et al., 2023; Jennings et al., 2022). It has been shown that there are differences between EO and EC recordings of healthy participants that persist as they age (Barry & De Blasio, 2017), and that a combination of features derived from EO and EC recordings can be used for classifying different subtypes of dementia, including *Parkinson's disease* (PD) dementia (Jennings et al., 2022).

As with the preprocessing methods discussed in the previous chapter, methods of treating electrodes and epochs in previous studies are very heterogeneous. This makes comparison of results very difficult, so experiments clarifying the effects of treating electrodes and epochs differently before extracting features will help put previous results in a clearer perspective and give insight into which methods are preferable when aiming to develop a clinical tool for the diagnosis of PD.

In this chapter, results from Vlieger et al., (2023c) will be discussed that were presented at the 45th Annual International Conference of the Institute of Electrical and Electronics Engineers (IEEE) Engineering in Medicine and Biology Society (EMBC). Experiments will be presented that use data from EO and EC experiments and different combinations of single epochs, averaged epochs, single electrodes, and *regions-of-interest* (ROIs). Building on the knowledge presented in Chapter 3, these combinations will be investigated by fully preprocessing both EO and EC datasets to obtain data that

contain as little non-neural signal as possible, extracting features from data of which electrodes and epochs are either averaged or not, and then classifying *people with PD* (PwPD) and controls using *Gradient Boosting Classifiers* (GBCs, see Section 2.3.2 for an explanation).

Acknowledgment: The work discussed in this Chapter is based on Vlieger et al., (2023c), which was presented by me at EMBC 2023. The initial experiments were conceived by me, after which they were refined in collaboration with my supervisors and co-authors Dr. Elena Daskalaki and Professor Hanna Suominen. Associate Professor Deborah Apthorp and Professor Emeritus Christian J. Lueck provided their expertise on the acquisition of RSEEG data, and all four co-authors were involved in the refinement of the manuscript after I had written the first draft.

4.1 Introduction

In EEG studies, different techniques can be applied to improve the SNR (Congedo, Gouy-Pailler, & Jutten, 2008; Hajra et al., 2021). Researchers can divide the signal from an electrode into smaller time windows, called epochs, and average those, under the assumption that noise is random. Similarly, researchers can group electrodes in ROIs and average them, as electrodes in close vicinity may pick up similar signals (Holsheimer & Feenstra, 1977; Congedo, Gouy-Pailler, & Jutten, 2008). In contrast, it is not uncommon in ML to treat epochs from a participant as separate samples instead of treating each participant as a sample (Maitín, García-Tejedor, & Romero Muñoz, 2020; Maitín, Romero Muñoz, & García-Tejedor, 2022); the RSEEG datasets are often rather small, and treating each epoch as a sample provides the ML algorithm with more data to learn from. This is often necessary for researchers using deep learning (DL) algorithms, as these methods require even larger amounts of data to train models than their more traditional counterparts in ML.

While the results in Chapter 3 showed clearly that preprocessing influences classification results, they also provided evidence for differences between single epochs and averaged epochs. These results were obtained from EC data using ROIs, but other studies have used EO data and single electrodes for the aforementioned reason of providing the algorithms with more data to learn from. While both using single epochs and electrodes and averaging occur in the literature, it is unclear what effects the choice for these particular methods has on outcomes and what effects their combinations have on outcomes. Additionally, the differences in methods in both EO and EC data has received minimal attention.

The choices to use ROIs or single electrodes and single epochs or averaged epochs not only influence the evaluation metrics, but also influence the available amount of data that can be used and consequently influence the available options for further experimentation. For example, they also influence decisions such as the algorithm (DL methods require more data than conventional ML

methods) and the *cross-validation* (CV) method, as a smaller number of samples may not be sufficient for adequate training after also creating validation and test sets.

Because of the uncertainty around the differences in methods and their effects when using EC and EO data, these effects on classification metrics when classifying PwPD and controls were investigated using EO and EC recordings from three different research groups. The following four methods were used: (1) single electrodes and averaged epochs; (2) single electrodes and single epochs; (3) ROIs and averaged epochs; and, (4) ROIs and single epochs. Using a nested CV approach, GBCs were used to classify PwPD and controls. Performance expressed in accuracy indicated that (1) ROIs outperform single electrodes and (2) EO recordings deliver better outcomes, but much of the data are lost in the preprocessing, making EC clinically more useful.

4.2 Methods

The *Canberra Hospital* (CH), the *University of New Mexico* (UNM; Carr, 2017), and the *University of Turku* (UTU; Suuronen et al., 2023) datasets were preprocessed to Level 4 as described in Section 3.1.2. To form four averaging and data increase methods, electrodes were either treated separately or combined into ROIs (see Section 3.1.2, Figure 3.1), and 2-second epochs were extracted and either treated as separate samples or averaged for each participant. In Method 1, single electrodes were used, and each extracted epoch treated as a single sample, while in Method 2, single electrodes were used, and the epochs were averaged for each participant. In Method 3, electrodes were combined into ROIs, and epochs were treated as separate samples. Finally, in Method 4, ROIs were used, and epochs were averaged for each participant. Table 4.1 summarises the methods used in this study. Absolute and relative power, as well as power ratios were extracted as described in Section 3.1.2.

The creation of ROIs reduces the amount of data, which begs the question why this method of data reduction was used and not, for example, principal component analysis (PCA), a method that has been successfully applied in ERP research (Dien et al., 2007; Dien, 2010a; Dien, 2010b), or independent component analysis (ICA). This was done for the same reasons the choice was made to extract power values, namely that the chosen methods closely resemble methods used in the literature as seen in the reviews by Maitín et al. (2020; 2022), and therefore allow for easier comparison to earlier results of experiments applying ML to EEG data.

Table 4.1. An overview of the four methods of electrode and epoch handling used in this study.

Method	Electrodes	Epochs
1	Single	Single
2	Single	Averaged
3	Averaged	Single
4	Averaged	Averaged

In the experiments reported on in the previous chapter, only train and validation sets were created. This was necessary because the three individual datasets were small and simply dividing them into train, validation, and test sets would produce quite small folds. Additionally, the aim in the previous chapter was to show the difference between different levels of preprocessing levels, for which only train and validation sets were necessary. In contrast, in the experiments discussed here the aim was to investigate the effects of methods on actual classification outcomes on test sets. To accomplish that, enough data was necessary to create train, validation, and test sets, so the three datasets were combined. As the results presented in Chapter 3 indicated, the results of experiments with the separate datasets were heterogeneous. To alleviate this problem, the data of the three data sets was made more uniform by through a rescaling process. Data of the UNM and UTU datasets were rescaled to the CH datasets by dividing the median of individual features in the UNM and UTU datasets by the median in the same feature in the CH dataset, and each feature value in the UNM and UTU dataset then divided by that number. The outer loop of the nested procedure was then created and consisted of 5-fold CV (see Figure 2.5 of Chapter 2). Data were split based on participant IDs to ensure samples from the same person were only present in either the train/validation data, to be used for hyperparameter optimisation, or the test data, but not both. ID numbers of controls and PwPD were kept separate, randomised, and then divided into five groups, after which the sets were created by combining the data from the controls and PwPD that were assigned to a particular group. This ensured that there were similar numbers of controls and PwPD in each set.

The GBC (Friedman, 2001; Natekin & Knoll, 2013) implementation of Scikit-Learn (Pedregosa et al., 2011) was used here based on the excellent performance of Boosted algorithms in ML challenges and its handling of outliers. In the experiments discussed in Chapter 3, Support Vector Machines (SVMs) were used for classification, but conventional SVMs are known to be sensitive to outliers (Xulei, Qing, & Cao, 2005), while GBCs have been shown to handle outliers relatively well (Hodge & Austin, 2018). Because outliers would not be averaged out in the experiments using single epochs, the GBC is a better choice than an SVM. Indeed, in our preliminary explorations, SVMs did not perform better than chance when using single epoch data.

Data were normalised to zero mean, unit variance using 'PowerTransformer' (Yeo & Johnson, 2000) of Scikit-Learn. Like the train/validation and test split, care was taken to make sure that the data of a participant were only present in either the training or validation data, and not both, as this would be considered data leakage. During the inner loop of 10 times repeated 10-fold cross-validation, hyperparameter values for learning rate, number of estimators, subsample, and maximum depth were acquired using grid search (Table 4.2). In Methods 1 and 3, median values for each feature were calculated on the train set, and values larger than three times the median set to 0. If more than 10 of such outliers were detected in a sample, the sample was removed from the train set. This was done because data inspection showed that, even after preprocessing, single epoch data would occasionally

contain outliers that were several times larger than can be expected of neural signals. Measurements of accuracy, sensitivity, specificity, precision, and F1-score were obtained by retraining on the total 80% used in training and validation, and classifying on the test set. Because of the stochastic nature of GBCs, this was done 10 times for each fold. The main evaluation metric used to assess performance was accuracy. All experiments were implemented in Python 3.8 (Van Rossum, 1995). Like the statistical testing in Chapter 3, Welch's t-test was used here to test whether there was a statistically significant difference between methods.

Table 4.2. Values used during the hyperparameter optimisation process.

	Hyperparameter values				
Learning rate	0.01	0.1	1		
N estimators	50	100	250		
Subsample	0.4	0.5	0.6	0.7	
Max depth	2	3	4	5	6

As with the experiments in Chapter 3, all experiments reported on in this chapter were approved by the Australian National University (ANU) Human Research Ethics Committee (protocol No. 2020/612).

4.3 Results

As nested CV was used, it was first ascertained that the test sets sufficiently represented the data that was used for training and validation. Three of the test folds contained 13 controls and 14 PwPD, while two test folds contained 13 controls and 13 PwPD. This meant that each fold should have on average 3.4, 5.6, and 4 controls, and 4.2, 5.6, and 3.8 PwPD for the CH, UNM, and UTU datasets respectively. Checking the test sets, all five contained controls and PwPD from each of the three datasets. With standard deviations of 0.94, 1.57, and 1.41 for controls, and 1.48, 1.47, and 1.88 for the PwPD for the CH, UNM, and UTU datasets respectively. All datasets were sufficiently represented in each test set.

The first major difference between the EO and EC datasets became apparent after preprocessing and feature extraction. The number of extracted epochs after applying Methods 1 and 3 differed greatly depending on whether the participants had their EO or EC (Table 4.3). In particular, the difference between the number of controls was stark: EO conditions had 58.5% fewer trials than EC conditions. This was also the case for Methods 2 and 4 using averaged epochs: the data of three control participants did not survive the pre-processing stage.

Table 4.3. The number of samples per method.

	Eyes Closed Samples		Eyes Open Samples	
Methods	**N controls**	**N PwPD**	**N controls**	**N PwPD**
1, 3	2351	2518	977	1909
2, 4	65	68	62	69

Method 3, using single epochs and ROIs, performed best when using EC data (Table 4.4), attaining an accuracy that was 5.79%, 6.71%, and 7.46% higher than those achieved by Methods 1, 2, and 4 respectively. This difference was attributable to a higher sensitivity, with the second highest Method attaining a sensitivity that was 10.16% lower. A t-test showed a significant difference for an alpha level of 0.05 between Method 3 and Method 1, the method with the second highest accuracy ($p < 0.0001$).

Table 4.4. Values of evaluation metrics in EC experiments for the four methods. The best scores are in bold.

Method	Accuracy [%]	Sensitivity [%]	Specificity [%]	Precision [%]	F1-score [%]
1	56.05	58	54.96	57.89	59.96
2	55.13	57.71	52.46	56.78	56.7
3	**61.84**	**68.16**	55.55	**62**	**64.66**
4	54.38	53.16	**55.85**	55.74	53.93

For experiments using EO data, Method 4 performed best (Table 4.5), attaining the highest accuracy across all experiments. Method 3 attained a comparable accuracy only being 2.16% lower; however, its specificity was as low as 3.57% (but with 91.81% sensitivity and 64.63% precision). This problem also occurred using Method 2, with a specificity of 7.46% (but with 94.10% sensitivity and 53.15% precision). Apart from Method 4, all methods in EO experiments attained specificities that were at least 21.54% lower than when the same methods were used in EC experiments.

Table 4.5. Values of evaluation metrics in EO experiments for the four methods. The best scores are in bold.

Method	Accuracy [%]	Sensitivity [%]	Specificity [%]	Precision [%]	F1-score [%]
1	57.96	70.42	33.42	**66**	67.41
2	53.06	**94.1**	7.46	53.15	67.72
3	62.16	91.81	3.57	64.63	**75.43**
4	**64.32**	70.37	**57.4**	64.41	66.66

4.4 Discussion

In this study, the interactions of different data size increase and averaging methods on a RSEEG classification task of PwPD and controls were investigated. Using thorough preprocessing, GBCs, and nested CV, the results showed a reduction in the number of participants and trials in EO datasets compared to EC datasets, and a tendency for high sensitivity scores in EO experiments at the cost of specificity scores. For the EC data, the best scores were obtained using a method of ROIs and single epochs, while for EO data the best method was one using ROIs and averaged epochs.

The observation that ROIs produced the best outcomes for both EO and EC data could be explained by the rationale that leads to the use of ROIs in the first place, namely that each electrode measures a mixture of neural and noise, and that this noise is randomly distributed while the neural signal is not. Consequently, the combination of several electrodes into ROIs leads to the noise being averaged out and the neural signal remaining. Especially in conditions where researchers can expect a high level of noise in the data, for example when measuring from a population of people with diminished motor control like PwPD, an approach that includes the use of ROIs seems logical.

Single epochs performed better than averaged epochs for the EC data, which is an encouraging finding when which the algorithm and CV method can be considered for use. DL algorithms generally need a larger amount of data than conventional algorithm, such as the one used here. Using DL methods would therefore be harder to utilise if the findings presented here demonstrated that averaged epochs produced the best results. Extracting several samples from the same participant, a method leading to a total of 4869 samples in the study presented here as can be seen in Table 4.3, has been used in previous studies to increase the data size (Yuvaraj et al., 2018; Khare et al., 2021; Lee et al., 2021), and the results from the study presented here provide support for this practice with the caveat that samples from the same participant should only be present in either the train, validation, or test set to avoid data leakage. Similarly, the single epoch approach also allows for CV procedures, such as the nested approach used here, which was not possible in, for example, the study presented in Section 3.1 due to the shortage of data to create the splits.

For a test to be clinically relevant, it needs to be able to reliably distinguish between people with a disease and people without the disease. Therefore, the first step towards this aim should be for the method to deliver reliable data that can be interpreted. The EO data failed that requirement in the above experiments: the data of three control participants did not survive the preprocessing stage due to excessive noise and artifacts, and there was a reduction of 58.4% in total trials in the EO single epoch experiments compared to the same in the EC condition. This measured reduction in number of trials has been observed by other researchers as well, with 33% of participants being removed due to loss of data (Jennings et al., 2022).

While this reduction of trials in EO experiments is found elsewhere (Jennings et al., 2022), in this study there is also a possibility that the trials were rejected because the preprocessing process is overly aggressive. The current preprocessing procedure, based on the steps and parameters found in the literature and specifically chosen to mimic those in previous research, is a general one, in the sense that it was set up to preprocess PwPD and controls as the two groups under investigation. The three datasets are diverse, containing data from people of different sexes, at different stages of PD, and of different ages. It might well be possible that a preprocessing pipeline tailored to a more specific group instead of a whole cohort would yield different results, possibly enabling use of EO data.

The second problem with the EO data is the low specificity scores. The specificity scores in EC experiments were rather low as well, but all were higher than the scores when using equivalent methods for the EC data, except for Method 4. Once again, this is a major problem if the data are to be used to make clinically useful decisions. However, the fact that the EO experiment using Method 4 produced higher accuracy, sensitivity, and specificity scores than the EC experiments shows this problem can be alleviated and concluding EO data are inferior to EC data would be premature, as other researchers found benefit in combining EO and EC data (Jennings et al., 2022).

A limitation of the study is the problem of data scarcity. Here datasets from three different research groups were combined to obtain a sufficient number of participants for the experiments, but these datasets were all collected on different equipment with the recordings also being of different lengths. Although the same preprocessing procedures were used for all three datasets, it is not inconceivable that there are differences between the datasets that persist after preprocessing. How these potential differences affect the results is unclear. As can be seen from the participant numbers reported in Table 4.6, data scarcity is a problem that has plagued most ML research using RSEEG data, and that research in general would be helped by the collection of larger datasets.

Table 4.6. Experimental details in the literature.

Reference	N PwPD/controls	On/off medication	EC/EO	Epoch length	Algorithm
Yuvaraj et al., 2018	20/20	On	EC	2 seconds	SVM-RBF
Lee et al., 2021	20/21	On and off	Unknown	1 second	SVM-RBF
Waninger et al., 2020	21/25	Off	EC	2 seconds	Linear DFA
Guo et al., 2021	9/11	Unknown	EO	60 seconds	Linear SVM
Khare et al., 2021	15/16	Off	EO	2 seconds	Least squares SVM
Khare et al., 2021	15/16	On	EO	2 seconds	Least squares SVM

Abbreviations: SVM is ‘support vector machine’, RBF is ‘radial basis function’, DFA is ‘discriminant function analysis’.

The test accuracies and other metrics presented here are lower than those reported elsewhere in studies classifying PwPD and controls using RSEEG (Maitín, Romero Muñoz, & García-Tejedor,

2022), with other studies able to achieve accuracies over 95% as can be seen in Table 4.7. Once again, comparison of results is difficult, as medication status, experimental parameters, participant numbers, epoch length, and algorithms vary widely (Table 4.6). There are also two specific reasons why our results might deviate from those found in the literature. First, our data underwent a preprocessing procedure that included rejection of muscle artifacts, which we saw in Chapter 3 influences classification outcomes. The studies included in Table 4.7 preprocessed their data to differing levels. Second, it is not uncommon for previous studies to have performed some sort of feature selection on their datasets before feeding it to their classification algorithm of choice (Yuvaraj, Rajendra Acharya, & Hagiwara, 2018; Maitín, García-Tejedor, & Romero Muñoz, 2020; Maitín, Romero Muñoz, & García-Tejedor, 2022). This has been shown to positively bias results (Demircioğlu, 2021) due to data leakage.

Table 4.7. Results found in the literature.

Reference	Accuracy [%]	Sensitivity [%]	Specificity [%]	Precision [%]	F1-score [%]
Yuvaraj et al., 2018	99.62	100.00	99.25	99.38	98.00
Lee et al., 2021	95.40	94.30		96.30	95.30
Waninger et al., 2020	95.24	94.74	95.65		
Guo et al., 2021	87.54	86.19	88.89		
Khare et al., 2021	96.13				
Khare et al., 2021	97.65				

The results of using Method 3 in EC experiments and Method 4 in EO experiments indicate that RSEEG has potential as a tool for diagnosis in PD. The datasets used in this study were mixes of male and female participants, with a large spread of different ages and different *Movement Disorder Society-Unified Parkinson's Disease Rating Scale* (MDS-UPDRS) scores. An interesting direction for future research would be to investigate the differences these participant characteristics may produce in classification tasks, as an improved understanding could prove invaluable for precision medicine in the development of more personalised diagnostic models.

Lastly, the experiments here are based on the principle that all classification should take place on neural signals as much as possible, and the preprocessing was designed accordingly. For the purposes of fair comparison, the methods were the same for all four Methods used here, but future research might find that a pipeline specifically tailored to EO data produces good results, and research into possible scenarios where this would be the case would be valuable.

4.5 Conclusion

To conclude, the results in this study indicate that using ROIs is preferable to using single electrodes. There was no clear difference between the results obtained using single epochs or averaged epochs. For

EC data, single epochs produced better results, but for EO data, averaged epochs worked better. The results presented here place the results found in the literature in a clearer perspective, as they show how certain decisions that have to be made before experimentation influence the outcomes, and may help guide other researchers in the design of their studies going forward. Based on the results here, it can be concluded that a combination of ROIs and single epochs will work best.

4.6 Chapter summary and conclusion

Building on the findings presented on the effects of preprocessing on averaged and single epoch data in Chapter 3, in this chapter the interactions between epoch handling and electrode handling were investigated in both EC and EO data by either using averaged or single epochs and single electrodes or ROIs. Because EO data suffered from the problem of data loss during the preprocessing stage, the EC approach using single epochs and ROIs is the preferable method.

In Chapter 3, experiments were conducted on three separate datasets, and no independent test set was used, which resulted in slightly different results between the results and those presented here in Chapter 4. The average accuracy across the three datasets in Chapter 3 was 63.90%, while the accuracy here, attained on a combination of the three datasets, was 61.84%. These results are easily explained by the aforementioned combination of the datasets, and the fact that an individual test set was used in this chapter.

Once again, the heterogeneity of methods makes it difficult to compare the results presented here to those in the literature (Maitín, García-Tejedor, & Romero Muñoz, 2020; Maitín, Romero Muñoz, & García-Tejedor, 2022). Having covered the differences in preprocessing and now the differences in datasets, epoch handling, and electrode handling, there are two more questions about methods that need answering. First, it is a common occurrence that researchers perform either feature selection or feature ranking before the CV process, but, as Demircioğlu (2021) showed, this introduces data leakage. For now, it is unclear how much this influences results in the context of RSEEG research. Second, due to the difficulty of acquiring large datasets, each of the three datasets used here are mixes of participants of different ages, sex, and MDS-UPDRS scores. It is unclear how this influences results and if results could be improved by separating participants by certain characteristics.

With the questions discussed in the previous paragraph unanswered, the following chapters will investigate them. Chapter 5 will delve into the effects of feature selection and CV methods found in the literature. Similar to the results presented in Chapters 3 and 4, the results of the experiments presented in Chapter 5 will help to place previous research into a clearer perspective and create a better understanding of what is necessary to design successful experiments. In Chapter 6, the influence of different participant characteristics will be investigated, which is a subject that deserves more attention but has not received that attention due to a lack of large enough datasets. By combining the three

datasets, as has been done in this Chapter, it is possible to look at patient characteristics at a more granular level, and will therefore give researchers acquiring data and working with large enough datasets some guidelines on what characteristics can be taken into account to increase model performance.

Chapter 5: Effects of sample randomisation and the moment of feature selection

After selecting whether to use *eyes closed* (EC) or *eyes open* (EO) data, how to process the data, and whether or not to average epochs and create *regions-of-interest* (ROIs), another important step in the attempt to classify *people with Parkinson's disease* (PwPD) and controls is the selection of the features that best separate the two groups. Researchers have used different methods to do this, such as statistical tests (Yuvaraj, Rajendra Acharya, & Hagiwara, 2018; Lee et al., 2021) and the use of *machine learning* (ML) methods such as *Random Forest Classifiers* (RFCs), which provide measures of feature importance. Researchers have also employed these algorithms at different stages of the classification pipeline, something that can influence the results. Specifically, in some cases researchers performed feature selection before the *cross-validation* (CV) process (Yuvaraj, Rajendra Acharya, & Hagiwara, 2018; Betrouni et al., 2019; Lee et al., 2021), a common practice that should be avoided as it introduces data leakage (Demircioğlu, 2021).

A second form of data leakage that has been reported on before in other medical classification tasks (Tampu, Eklund, & Haj-Hosseini, 2022; Ge et al., 2023) and *electroencephalography* (EEG) classification tasks in other diseases (Barcelon et al., 2019) involves the use of multiple samples of the same participant. When using single epochs, as described in Section 4.3, care must be taken to ensure that samples from the same participant are either in the test or train set, but not both, as this would mean information about that particular participant is present in both sets. There is no hard evidence that this has occurred in ML studies investigating *Parkinson's disease* (PD) using *resting-state EEG* (RSEEG), however, the potential effects in the context of this research requires investigation, at the minimum, to provide an indication of what results would look like if this has happened.

In this chapter, results of an experiment into the use of ERPs for classification of PwPD and controls will be presented in which data leakage led to increased, or overly optimistic, evaluation outcomes, as a prelude to a more comprehensive investigation of data leakage in classification of PwPD and controls. Results are presented of experiments using RSEEG data into the effects of data leakage due to feature selection being performed outside of CV, the effects of samples of the same participant being present in the train and test set, and the effects of both methods together. The hyperparameters of the *Support Vector Machine* (SVM) were kept constant and the effects of data leakage assessed by performing feature selection outside of CV, randomising the whole dataset when using single epochs, combining these two methods, and comparing them to experiments where feature selection was part of CV, and randomisation was performed on participants instead of samples.

5.1 The use of event-related potentials and machine learning to improve diagnostic testing and prognostication in Parkinson's disease

Acknowledgment: The work discussed in Sections 5.1.1–5.1.5 is based on Vlieger et al., (2021), which I presented at Nursing Informatics 2021. The data used was collected as part of a larger project headed by Associate Professor Deborah Apthorp and Professor Emeritus Christian J. Lueck. The initial ML experiments were conceived by me, Dr. Elena Daskalaki, and Professor Hanna Suominen. After I wrote the first draft, all four co-authors were involved in the refinement of the manuscript.

5.1.1 Introduction

It has been observed in a variety of neurodegenerative diseases that certain cognitive domains, such as working memory, attention, and executive function, may get affected as the disease progresses. For example, in *Multiple Sclerosis* (MS), between 43% and 70% of people develop cognitive impairments (Chiaravalloti & DeLuca, 2008), while in PD, depending on the criteria used, prevalence is estimated to be between 20% and 41% (Gonzalez-Latapi et al., 2021). Research has, therefore, looked into methods to quantify the difference between people with neurodegenerative diseases and controls for the purposes of diagnosis and prognosis.

A method that has been used in the context of quantifying differences in cognition between people with neurodegenerative diseases and controls is the use of *event-related potentials* (ERPs; Luck & Kappenman, 2011; Luck, 2014). In ERP experiments, a participant is given a task and their EEG is measured during this task. This method has been used in a wide variety of neurodegenerative diseases, such as MS (Newton et al., 1989), *Huntington's disease* (HD; Turner et al., 2015), and *Alzheimer's disease* (AD; Jackson & Snyder, 2008), while using a broad range of different experimental tasks, such as the oddball paradigm mentioned in the introduction, but also paradigms such as the Stroop task, the n-back task, and the Go/No-Go task.

The EEG-based biomarkers for HD are of particular interest in the context of PD because HD is an inherited disease and people destined to develop the disease can be identified pre-symptomatically based on their genetic profile. Turner et al. (2015) devised an experiment to evaluate the differences in ERPs between healthy controls, people destined to develop HD (pre-HD), and people with symptomatic HD. They analysed the *Contingent Negative Variation* (CNV; Tecce, 1972), a slow-building ERP component, and results demonstrated differences in peak CNV amplitudes between people with HD and controls, but also between pre-HD groups and controls. This strongly indicates that the CNV has the potential to detect disease existence, even before symptoms have become apparent. While PD is not an inherited disease as HD is, it does have a prodromal stage in which symptoms are not yet overt, which means the paradigm used by Turner et al. could be used as a marker of disease development in PD.

In the study, it was investigated using the Turner et al. (2015) paradigm whether the CNV can be used to distinguish between PwPD and controls, extracting amplitude values from the CNV, performing feature selection using an RFC, and classifying the participants using an SVM. The results indicate a potential for CNV to be used as a marker of disease status in PD.

5.1.2 Methods

Participants were 21 PwPD (8 females and 13 males, mean age 68.6 ± 9.92 years) of which 15 were right-handed and 2 ambidextrous, and 16 controls (7 females and 9 males, mean age 68.33 ± 8.68 years) of which 9 were right-handed. Handedness of 4 controls was unknown. PwPD were recruited through the neurology, aged care, and movement disorders clinics at the Canberra Hospital in Canberra, Australia, and through the Parkinson's ACT Association. Participants were English speaking, at least 18 years of age, had no diagnosis of cognitive impairment or history of developmental disability, alcohol abuse, or substance abuse. Patients had mild-to-moderate PD (Martinez-Martin et al., 2013), no psychiatric disorders considered not to be due to PD, and no other movement disorders, while controls had no psychiatric or movement disorders. Controls were age- and sex-matched and were recruited from partners of PwPD and via community advertisement. PwPD were in the ON state during testing, meaning their medication was effective in alleviating symptoms, but there were no limitations placed on medication timing relative to the session. PwPD were diagnosed by a qualified neurologist and fulfilled the *United Kingdom* (UK) Parkinson's Disease Society diagnostic criteria for PD (Hughes et al., 1992). The study was approved by the ACT Health Human Research Ethics Committee (ETH.4.16.060) and written consent was obtained from all participants.

Before starting ERP experiments, participants provided clinical information regarding their medical history, including the time of the last medication dose, specifics about the onset and diagnosis of PD, and demographic information. PwPD also underwent a neurological examination and were assessed using the *Movement Disorder Society-Unified Parkinson's Disease Rating Scale* (MDS-UPDRS; Goetz et al., 2008). Two trained clinicians performed the MDS-UPDRS assessment on the first few patients to calibrate their scoring; all other patients were assessed by a single clinician with the second clinician validating the scores several times during the data collection.

ERPs were acquired with a 64-channel Ag/AgCl BioSemi™ Active II system (www.biosemi.com) at a sampling rate of 2,048 Hz. Participants were fitted with an EEG cap containing 64 Ag/AgCl electrodes arranged according to the international 10-20 system (Jasper, 1958) and electrode impedances were kept below 5 Ω. Participants sat in a chair at a distance of 290 *centimetres* (cm) from a 195 x 145 cm projector screen in a room that was entirely black and with the lights turned off. Participants kept a RESPONSEPixx 5-button response box in their lap during the experiment.

Participants were shown a fixation cross in the middle of the screen. To start a trial, a blue square subtending a visual angle of 3 x 3° was presented for 500 *milliseconds* (ms). 2,500 ms later, an

X or a Y would appear on either the left or right side of the screen. If an X appeared the participant had to press the left or right button on the button box with their right index finger depending on the position of the letter. The X was the 'Go'-stimulus. If a Y appeared, on the other hand, the participant had to withhold any response; the Y was the 'No-Go'-stimulus. X stimuli remained visible on the screen until the participant responded with a button press and Y stimuli remained on the screen for 1900 ms. The disappearance of the stimuli was followed by randomised inter-trial interval with a duration between 2,500 ms and 4,000 ms. Participants began each session with 6 practice trials followed by feedback on their performance. Thereafter, they performed 90 trials without feedback, 80% of which were 'Go' trials and 20% were 'No-Go' trials. Figure 5.1a shows a schematic of the experiment, seen earlier in Chapter 2.

a. b.

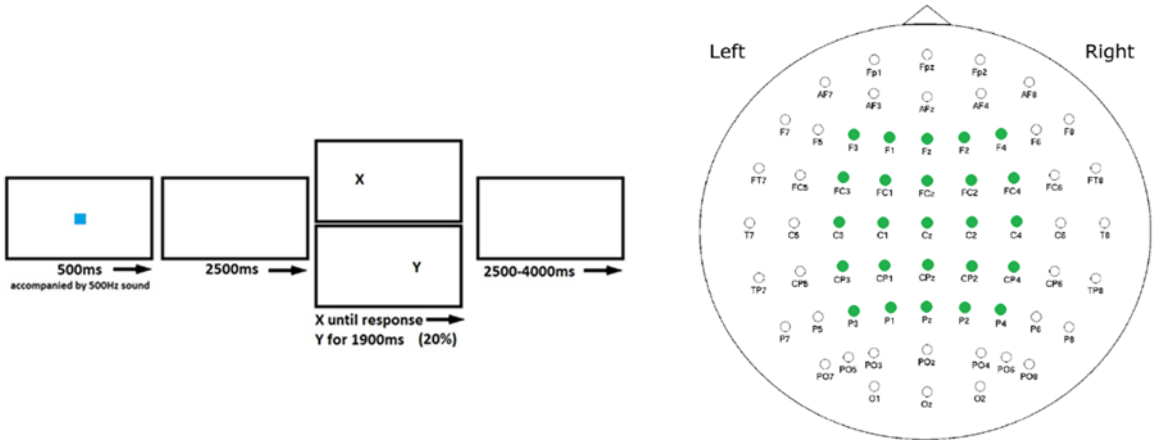


Figure 5.1. (a) A schematic of the Go/No-Go experiment. (b) A schematic of the 25 electrodes retained (green) after removing 37 electrodes often affected by artifacts.

ERP preprocessing was performed in MATLAB R2018b using the MATLAB toolbox EEGLAB 2020.0 (Delorme & Makeig, 2004), following recommendations on the wiki of Swartz Centre of Computational Neuroscience. Because tremor is a cardinal feature of PD, 37 electrodes that were frequently contaminated by muscle artifact were excluded, retaining 25 electrodes (Figure 5.1b). The pipeline used here was similar to the pipeline described before, except for two differences. The filter used here was a 0.03-35 Hz 8th order Butterworth bandpass filter. This is a similar filter as is used elsewhere in the literature and is used because the CNV is a slow-building potential and using a filter with a customary 1.0 Hz cut-off would possibly filter out important slow frequencies that contribute to the signal. Secondly, instead of using ICLabel (Pion-Tonachini, Kreutz-Delgado, & Makeig, 2019) for the removal of non-neural *independent components* (ICs), the 'Multiple Artifact Rejection Algorithm' (MARA; Winkler, Haufe, & Tangermann, 2011; Winkler et al., 2014) was used, which performs a similar task but does not require the statistical thresholding ICLabel requires.

Data were epoched from -3,000 ms before until 1,000 ms after the Go/No-Go stimulus using ERPLAB. An example of a CNV measured in the experiment is visualised in Figure 5.2. The CNV was quantified by subtracting the mean amplitude in the 2000–1500ms interval from the 500–0ms interval pre-stimulus. The resultant features were used as input to an RFC and electrodes selected that had Gini-indices of (2/number of features) and had occurred in more than 30% of the 100 times the RFC was run. All selected electrodes were used as input to a SVM for ML-based classification. Starting with all selected electrodes, they were sequentially removed, beginning with the electrode that occurred the fewest time in the RFC runs, until the minimum number of electrodes for the highest classification accuracy was reached. *Leave-one-out CV* (LOOCV) was used for classification evaluation and performance was evaluated using accuracy, sensitivity, specificity, and precision. To see if the results were dependent on the algorithm, the same input that was provided to the SVM was also provided to a *k-Nearest Neighbours* (kNN) algorithm.

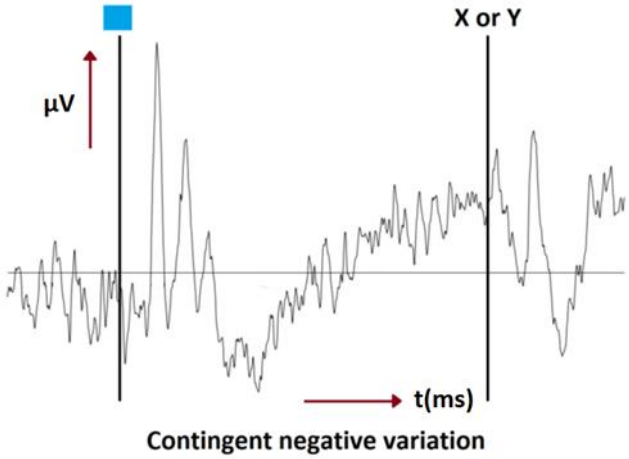


Figure 5.2. An example of a CNV in microvolts measured during the experiment. The CNV slowly builds during the 2500 ms after the presentation of a cue in anticipation of a response to a stimulus.

5.1.3 Results

The combination of the SVM with LOOCV produced an accuracy of 72.97%, with sensitivity of 71.43%, specificity of 75%, and precision of 78.95%. In absolute numbers, this meant that 15 PwPD and 12 controls were correctly classified, while 6 PwPD and 4 controls were incorrectly classified. The electrodes that occurred in more than 30% of runs of the RFC were Fz, F1, F2, F4, FC3, C3, and CP4. The greatest accuracy was achieved by the SVM using 3 electrodes FC3, Fz, and F4.

Successful classification was very much dependent on both the algorithm used and the electrodes that were selected. After running the SVM experiments, the same input was provided to the kNN, but this only led to an accuracy of 51.35% with sensitivity of 33.33% and specificity of 75.00%. Similarly, making different electrodes available to the SVM would often lead to results that were on chance level.

5.1.4 Discussion

Because the early CNV differs between people with other neurodegenerative diseases and controls, similar differences were expected to be present between people with PD and controls. The electrodes selected by the RFC corresponded with those found elsewhere in the literature and the combination of the RFC feature selection and the SVM for classification provided evidence that the CNV can be used for classification of PwPD and controls.

This study had a number of limitations, including the relatively small number of participants and the fact that it studied individuals whose diagnosis was already known, that is, the PwPD participating in this study were all reliably diagnosed by qualified clinicians. Before any further conclusions can be drawn about the role of CNV as a biomarker of PD, several issues must be resolved. First, the performance of the models must be tested on other, independent datasets. Then, to assess the potential that CNV might have in diagnosis, a prospective trial looking at as-yet undiagnosed people is necessary. This might be difficult in practice considering the necessary number of participants, but the study sample might be enriched by considering people who are predisposed to develop PD such as people with *rapid eye movement sleep behaviour disorder* (RBD).

Electrodes identified by the RFC were in agreement with the location of the CNV reported in the literature, which describe the CNV as having an early, frontally-dominant wave and a more central late wave that can be detected prior to the execution of a motor response (Tecce, 1972). The CNV is generally found to be bilaterally symmetrical and most prominent over the vertex, consistent with the findings in this study in which all frontal electrodes (F1 to F4) were represented. The fact that FC3 and C3 were found to be significant could be explained by their proximity to the left motor cortex since most of the participants were predominantly right-handed.

The methodology used here leaves room for improvement. The CNV is a slow-building potential and mean amplitudes may not be the best to capture subtle changes. EEG data are noisy and different electrodes often pick up on the same signal to varying degrees. Dimensionality and noise reduction techniques, such as *Principal Component Analysis* (PCA) which has been used before to investigate the CNV, might be able to improve classification accuracy. Therefore, different quantification methods, as well as dimensionality and noise reduction techniques are avenues for future work.

5.1.5 Conclusion

The results of the experiment reported on in this Chapter indicate that controls and PwPD can be differentiated based on the CNV when using ML algorithms using conventional measurements of amplitude. Future research might be able to improve on the results of this study by using methods for dimensionality reduction and signal-to-noise ratio improvement. By utilising these methods and taking

into account the differences in the CNV between PwPD and controls, the CNV is a probable candidate as a biomarker of PD development.

5.2 When the science catches up

"If I have seen farther, it is by standing on the shoulders of giants", a remark arguably most often, possibly wrongly, attributed to Sir Isaac Newton (Turnbull, 1959), might capture the endeavour of scientific discovery in its purest form. One of the main components of academic writing is properly citing the sources used to build one's own research. Unfortunately, those ideas are sometimes superseded by better ideas. This is exactly what happened with the piece of research presented in Section 5.1.

When I set up my ML experiment, I looked at the EEG literature and the methods that were used. Several studies would perform a form of feature selection or ranking and then use the outcome as input to their preferred classification algorithm (Yuvaraj, Rajendra Acharya, & Hagiwara, 2018; Betrouni et al., 2019; Maitín, García-Tejedor, & Romero Muñoz, 2020; Lee et al., 2021). In a study into feature selection in radiomics studies, Demircioğlu (2021) investigated studies in which feature selection was performed before CV, using the whole dataset, and then performing train-test splits, to methods in which feature selection was solely performed on the train set. The results showed that data leakage caused by feature selection being performed on the whole dataset before CV produces biased results, with a 17% increase in accuracy compared to when feature selection is part of the CV process. The bias was larger in high dimensional datasets compared to low dimensional datasets: datasets with more features per sample tend to produce larger biases than datasets with less features per sample.

If this method of feature selection and ranking has occurred several times, how much would the results differ if they were compared to a manner of in-CV feature selection? If we performed a similar experiment as (Demircioğlu, 2021) did, adapted to our EEG data, what would be the outcomes? While performing the experiments again on ERP data would make comparison easy, only this one CNV dataset with a limited number of participants is available. Instead, the experiments were performed using the three RSEEG datasets previously used.

5.2.1 Introduction

As seen in Chapters 2, 3, and 4, researchers have employed a variety of data cleaning methods and classification algorithms in an effort to create models for the classification of PwPD and controls. While features are most often extracted in the time-frequency domain, with studies using absolute power and relative power often, and a minority of studies utilising entropy features (Maitín, García-Tejedor, & Romero Muñoz, 2020; Maitín, Romero Muñoz, & García-Tejedor, 2022), feature selection methods differ. They include ranking features based on statistical tests, such as Student's t-tests and *Analysis of Variance* (ANOVA), correlation analysis, such as Pearson's correlation, and ML methods, such as RFCs

Chaturvedi et al., 2017), where the contribution to classification of a feature is quantified as the gain in accuracy or Gini-importance.

As mentioned in Section 5.2, the point in the pipeline at which feature selection takes place is important, as it is generally advised to only perform feature selection on the train set after the train-test split has been made to avoid data leakage (Demircioğlu, 2021) with the effect being larger if there are less samples per feature. Feature selection and feature ranking outside of CV has been reported in ML studies into RSEEG of PwPD and controls. It is unclear, however, how much this affects evaluation metrics, as data cleaning methods and classification algorithms vary widely.

Data leakage due to the use of multiple samples from the same participant and these samples ending up in both the test and train set has been reported on both in other medical classification tasks (Tampu, Eklund, & Haj-Hosseini, 2022) and other EEG classification tasks (Bhalerao & Pachori, 2022). As was mentioned in Chapter 4, care must be taken to ensure that samples from the same participant are either in the test or train set, but not both, as this would mean information about that particular participant is now present in both (Suominen, Pahikkala, & Salakoski, 2008). While there is no hard evidence this has occurred in ML studies investigating PD using RSEEG, the fact it has been reported on before in other medical domains indicates that the effects require study in the context of EEG research. Specifically, the variety of preprocessing levels used, combined with the different points at which features are selected and the manner in which samples are randomised, make that an investigation of effects will help put previous results in a clearer perspective and enable easier comparison of studies.

In this study, the effects of data leakage due to feature selection being performed outside of CV, the effects of samples of the same participant being present in the train and test set, and the effects of both methods together were investigated. SVM hyperparameters were kept constant and the effects of data leakage assessed by performing feature selection outside of CV, randomising the whole dataset when using several samples from the same participant, combining these two methods, and comparing them to experiments where feature selection was part of CV, and randomisation was performed on subjects instead of samples.

5.2.2 Methods

The *Canberra Hospital* (CH), the *University of New Mexico* (UNM; Carr, 2017), and the *University of Turku* (UTU; Suuronen et al., 2023) datasets were preprocessed to the four different Levels as described in Section 3.1.2, as these correspond to the different methods found in the literature. After preprocessing, 2-second epochs were extracted that were either averaged or treated as separate samples, similar to the Method used in Section 4.2. This created a total of 48 sets to use for classification: three datasets were preprocessed to four different Levels for four epoch handling Methods. Datasets from different research groups were used individually to ascertain that results would be equivalent regardless of the source of the data. Also, the datasets were used individually, as opposed to being combined like

in Chapter 4, because it is often the case that other studies used participant numbers equivalent to those in the individual datasets (see Table 4.5 for participant numbers in other studies).

All experiments were coded in Python 3.8 (Van Rossum, 1995) using Scikit-Learn (Pedregosa et al., 2011). For all experiments, an SVM with a *radial basis function* (RBF) kernel was used, as this is an algorithm that has often been used in medical classification in general and RSEEG classification for PD specifically to make comparison to results from previous studies easier. No hyperparameter optimisation was performed: regularisation parameter C was set to 1.0 and gamma was set to automatically scale to 1/n_features. 10-fold CV was used to obtain accuracy, sensitivity, specificity, precision, and F1-score.

For experiments using single samples, four methods were used. In Method 1, care was taken to make sure samples from the same participant were either in the train or test set by randomising participant codes, and then creating train and test sets. Feature selection was performed in the CV on the train set using an RFC with 100 estimators, selecting only the features that had Gini-coefficients higher than 2/n_features. In Method 2, features were selected similarly, but all samples had been randomised throughout the train and test sets. In Method 3, feature selection was performed before CV, running the same RFC 100 times and once again selecting only those features with Gini-coefficients higher than 2/n_features. Train and test sets were created similarly as in Method 1. In Method 4, feature selection was performed before CV and samples were randomised throughout the train and test sets. For mean epochs, only two methods were used, one with feature selection occurring during CV (Method A), and one with feature selection before CV (Method B).

As with the previous experiments, the experiments described here were approved by the *Australian National University* (ANU) Human Research Ethics Committee (protocol No. 2020/612).

5.2.3 Results

The experiments using single samples are best compared between datasets by looking at the four Levels introduced in Section 3.1.2 and the four Methods introduced in Section 4.2. Full results for the individual datasets can be found in Supplement 1.

For all three datasets, the results for Method 1, the Method without data leakage, were best when using data preprocessed to Level 2, being filtered data. Once again, the Level of preprocessing seemed to influence the results, as all Levels showed different evaluation results for each of the three datasets. Similar to the Chapter 3 results, the values of the different evaluation results differed between the three datasets. Also, Methods 2 and 4, in which samples were randomised throughout the training and test set, had much smaller standard deviations than Methods where this was not the case. Method 4 showed slightly increased scores compared to Method 2. This indicates that, while Method 3 showed

less increases than Method 2, it did lead to increases in evaluation metrics even when using randomisation.

Having looked at the differences between Method 1 and the other Methods within datasets, a comparison of the average difference between Methods across datasets was the next step. Table 5.1 therefore provides an overview of the average difference per Method and preprocessing stage for the three datasets. The average presented is an average difference that can be expected at a particular Level of preprocessing regardless of the dataset. Level 4 data are what would be preferred when trying to classify on signals that are of neural origin as much as possible, and it is this Level shown to be most affected by the problem of data leakage. An average across preprocessing Levels is also provided to give an indication of what kind of difference can be expected regardless of preprocessing Level.

Table 5.1. Average differences between evaluation results between Method 1 and the other methods per preprocessing stage for the three datasets using single samples.

	Preprocessing	**Accuracy [%]**	**Acc_std [%]**	**Sensitivity [%]**	**Specificity [%]**	**Precision [%]**	**F1-score [%]**
Method 2	Level 1	17.75	10.95	16.43	19.72	16.89	16.74
	Level 2	11.14	11.49	12.52	9.69	9.95	11.18
	Level 3	12.09	11.54	11.34	12.88	12.73	11.9
	Level 4	22.94	12.51	18.76	27.56	22.87	20.86
	Average across Levels	15.98	11.62	14.76	17.46	15.61	15.17
Method 3	Level 1	3.88	0.41	4.2	3.63	3.74	3.98
	Level 2	4.09	0.53	5.2	2.96	3.44	4.32
	Level 3	4.36	1.22	3.67	5.22	4.85	4.18
	Level 4	8.67	1.73	8.3	9.29	7.74	7.99
	Average across Levels	5.25	0.97	5.34	5.28	4.94	5.12
Method 4	Level 1	18.35	11.56	16.61	20.75	17.63	17.19
	Level 2	11.25	11.9	12.37	10.05	10.1	11.17
	Level 3	12.66	11.4	12.39	13.08	13.13	12.64
	Level 4	23.73	13.3	20.13	27.67	23.27	21.74
	Average across Levels	16.5	12.04	15.38	17.89	16.03	15.69

Similar to the results seen when using single samples, average samples were affected by the different Methods. Once again, full results for individual datasets can be found in Supplement 1. To summarise, for each dataset and for each preprocessing level Method B showed higher evaluation scores than Method A. In addition to increased evaluation metrics, the standard deviation of the accuracy also decreased in Method B, reinforcing the observation that feature selection outside of CV positively

biases outcomes. Compared to Method 3, the method using single epochs in which feature selection was performed outside the CV process, the equivalent Method B for averaged epochs produced larger differences. For each dataset and each preprocessing level all, evaluation scores increased, and the standard deviation of the accuracy decreased.

Table 5.2 shows the average difference per preprocessing stage for all three datasets when using averaged epochs. Compared to the difference between Method 1 and Method 3, the difference between Method A and Method B was larger on average. Unlike the results of the experiments using single samples, the difference in Level 1 results was the largest, while Level 4 results produced the second largest difference. The results implied that, regardless of how the samples are handled, the moment of feature selection positively biases classification outcomes.

Table 5.2. Average differences between evaluation metrics between Method 1 and the other methods per preprocessing stage for the three datasets using averaged samples.

	Accuracy [%]	**Acc_std [%]**	**Sensitivity [%]**	**Specificity [%]**	**Precision [%]**	**F1-score [%]**
Level 1	13.49	4.04	15.58	11.54	12.83	14.15
Level 2	6.99	1.45	6.55	7.38	6.38	6.45
Level 3	6.36	2.65	5.52	7.67	7.05	6.18
Level 4	10.24	2.22	7.31	13.85	13.52	9.55
Average across Levels	9.27	2.59	8.74	10.11	9.95	9.08

Lastly, a difference in the effects of feature selection methods was observed based on the number of samples. Namely, noting that the number of samples was much larger when using single samples instead of averaged samples, a comparison between the two was made. Table 5.3 provides an overview of the difference in scores between average mean sample metrics and single sample metrics across all datasets per preprocessing stage. For each preprocessing Level, evaluation scores for mean samples showed larger increases due to data leakage than evaluation scores for single samples.

Table 5.3. Overview of the difference in scores between average mean sample metrics and single sample metrics across all datasets per preprocessing stage.

	Preprocessing	**Accuracy [%]**	**Acc_std [%]**	**Sensitivity [%]**	**Specificity [%]**	**Precision [%]**	**F1-score [%]**
Difference between averaged and single samples	**Level 1**	9.61	3.63	11.38	7.91	9.09	10.18
	Level 2	2.90	0.92	1.34	4.42	2.94	2.14
	Level 3	2.00	2.32	1.85	2.45	2.19	2
	Level 4	1.57	0.50	-0.99	4.56	5.78	1.56
	Average across Levels	4.02	1.84	3.40	4.84	5	3.97

5.2.4 Discussion

In this study, the effects of two methods that cause data leakage, feature selection outside of CV, and randomisation of samples through the train and test set were investigated. The effects were investigated at different stages of preprocessing and when using single samples and averaged samples. The results indicated that data leakage positively biases results for all datasets at all preprocessing Levels and for all tested Methods. For single samples, the effects of data leakage were largest for data that were preprocessed at Level 4, which is the Level of preprocessing preferred if the objective is to classify based on neural signals alone. Lastly, as observed elsewhere in the literature (Aldehim & Wang, 2017; Demircioğlu, 2021), the effects of performing feature selection outside of CV were larger when the number of samples per feature were smaller.

When comparing Methods 2, 3, and 4 to Method 1, and Method B to Method A, for each dataset and for each preprocessing stage, data leakage led to inflated scores for each and every evaluation metric. What also stood out is that the standard deviation of the accuracy became much smaller when samples were completely randomised compared to methods where samples from the same participant were assigned to either the train or test set. Large differences existed between the individual datasets, but what was true for all three was that randomising single epochs through the train and test set caused higher inflation of evaluation metrics than feature selection outside of CV. The lowest accuracy difference for randomised single epochs was 8.76% and 2.55% for mean samples, both observed using the UNM dataset.

A limitation of this study is that it was performed on three small datasets. The problem of feature selection outside of CV diminishes as the number of samples per feature increases, so there might not be much, if any difference if the dataset is large enough. That being said, the current reality of many researchers that apply ML to medical problems is that they only have access to small datasets, so care should be taken not to improperly select features. The randomisation of data from the same participant through the train and test set is a much bigger problem that leads to much larger inflation effects. This is something that every researcher that wants to increase the dataset by using single samples should be very careful about as to avoid overly optimistic models.

Looking at Methods 1 and A, similar to the results in Section 2.2.3, while not always being the highest score for the particular dataset, leaving in muscle artifact produces higher evaluation metrics than filtering all artifacts out for the CH and UNM datasets when using single samples, and for all three datasets when using mean samples. This underlines, once again, the importance of diligent preprocessing to avoid classification on non-neural activity. Looking at the other evaluation metrics, the highest scores achieved for single datasets almost always come from data that has not undergone complete preprocessing, with only sensitivity and F1-score for single epochs of the UTU dataset and specificity for mean epochs of the CH dataset being the highest for these datasets.

Comparing the results from single samples experiments to the results from averaged samples experiments, which means a comparison of only the difference in feature extraction method, the differences for averaged samples are larger than for single samples for every dataset: 11.86% and 7.84% for the CH dataset, 7.81% and 2.55% for the UNM dataset, and 5.36% and 8.14% for the UTU dataset. This observation is in line with other findings in the literature that the differences are larger when the number of samples per feature is smaller (Aldehim & Wang, 2017; Demircioğlu, 2021). This observation still holds when we break down the results by preprocessing stage. The difference between averaged sample and single sample accuracy is 9.61% when using rereferenced data, 2.90% when using filtered data, 2.00% when using data containing muscle artifact, and 1.57% when data are fully filtered.

The results presented here concerning RSEEG data are in line with findings from earlier ML studies in other fields. Moreover, the problem of data leakage is by no means a new topic. The methods that were tested here have been employed in studies in recent years, which makes it difficult to compare studies as the methods of feature selection and randomisation vary per study. By employing and comparing methods that have been used before, this study attempts to shed light on previous results and enable comparison. Taking the results of the experiments presented here into account, future studies should (1) make sure that feature selection is part of the CV process, and (2) ensure that, when using single samples, all data from a participant are only present in one set, whether it is a train, validation, or test set.

5.2.5 Conclusion

This study shows that both feature selection outside of CV and randomising samples from the same participant through the train and test set lead to inflated evaluation metrics, with the latter leading to larger inflation effects. The inflation effects for feature selection outside of CV diminish when there are more samples per features, as stated in the literature, but given most EEG datasets are small, researchers should attempt to avoid this approach. This study helps put the results of previous studies into a clearer perspective and encourages investigators to pay extra attention to methods that may introduce data leakage. They should attempt to minimise any data leakage and, where possible, use the largest dataset available.

5.3 Chapter summary and conclusion

This chapter started off by looking into the usefulness of the CNV for the classification of PwPD and controls. Given the work that has been done by researchers such as Turner et al. (2015), who showed that the CNV can be used to distinguish between people with HD, people predisposed to develop HD, and controls, the research presented here investigated whether it was possible to use the CNV to classify PwPD and controls. The outcomes did show that, using a conventional quantification of the CNV, it

was possible to do this. Unfortunately, the outcomes were positively biased, as feature selection had taken place before the CV process.

As the CNV dataset was very small, the three datasets used throughout this thesis were used to investigate the effects of data leakage in EEG classification experiments. Building on the work presented in Chapters 3 and 4, the work in Section 5.2 investigated the effect of data leakage at different preprocessing levels and for both single and averaged epochs. In the case of single epochs, the method that was used in subsequent experiments in this thesis based on the results in Chapter 4, it was clear that randomisation of samples through the train and test set led to increased evaluation scores, while the increase was smaller but still there for experiments in which feature selection was performed outside of the CV process.

With the knowledge gained from the experiments in Chapters 3, 4, and 5, a clearer picture of what is necessary for reliable classification of PwPD and controls based on RSEEG is starting to form. First, the results in Chapter 3 indicated that data need to undergo preprocessing that includes artifact rejection. Second, Chapter 4 provided evidence that electrodes need to be combined into ROIs. Third, Chapter 4 also showed features need to be extracted from single epochs instead of averaged epochs. Lastly, the current chapter gave evidence that feature selection needs to be integrated in the CV process, and that samples from the same participant should only be presented in the train, validation, or test set to avoid data leakage.

Chapter 6: Effect of participant characteristics on classification outcomes

In the previous chapters the focus has been on what to do with data once they have been collected. Questions that were investigated include how data should be preprocessed, how data should be treated after preprocessing, at what stage features should be extracted, what kind of data, and which algorithms, produce the best results. What is still missing is an examination of the data acquisition itself and whether or not the manner in which data are generally obtained can be improved upon. In this chapter, the results of experiments looking into the effects of modifying the composition of data based on different participant characteristics are presented.

Due to the difficulty of collecting *resting-state electroencephalography* (RSEEG) data, most datasets in the *Parkinson's disease* (PD) literature are small and consist of people of differing age, of both sexes, and at different stages of their disease. Differences between the *electroencephalography* (EEG) recordings of men and women, as well as between those of people with different *Movement Disorder Society-Unified Parkinson's Disease Rating Scale* (MDS-UPDRS) scores have been reported in the literature, so it was logical to devise experiments that investigated the effect of reanalysing the data as a result of changes in these characteristics. In the following study, the differences between men and women, as well as between people at different stages of PD, as measured by the MDS-UPDRS, were investigated by combining the datasets from three different research groups, splitting the data by the aforementioned characteristics, and performing classification.

6.1 Introduction

While it is known that differences which could affect EEG recordings exist both between and within men and women, such as different hairstyles (Greischar et al., 2004), a different average skull thickness, and different skull sizes (Li et al., 2007), studies into the use of RSEEG for the classification of *people with Parkinson's disease* (PwPD) and controls tend to ignore the possibility that these differences might influence results. Experiments are therefore run on mixed-sex datasets (Maitín, García-Tejedor, & Romero Muñoz, 2020; Maitín, Romero Muñoz, & García-Tejedor, 2022), mostly out of necessity because RSEEG datasets tend to be quite small due to the time-consuming nature of the recording methods. The literature shows differences between men and women, both in RSEEG recordings and in PD development (Cerri, Mus, & Blandini, 2019). However, it is unclear whether or not these differences also influence evaluations metrics in *machine learning* (ML) classification tasks.

It is well documented that RSEEG parameters change as PD progresses. Overall slowing of brain oscillations is one of the most often reported (Soikkeli et al., 1991; Morita et al., 2009; Benz et al., 2014), but RSEEG datasets often include PwPD who are at different stages of their disease. For example, a dataset can be described as including people with mild-to-moderate PD, expressed as their MDS-UPDRS score, but the range of scores can still vary widely.

In this study, the effects of participant sex and MDS-UPDRS score on evaluation metrics when classifying PwPD and controls were investigated. This was accomplished by combining three independent RSEEG datasets that, together, provided adequate numbers of participants which made it possible to create meaningful subsets that could be used to answer our questions. After preliminary exploration of the data, first a conventional experiment was run which included all the data to obtain baseline performance metrics. Then further experiments were performed which involved splitting the data according to sex and MDS-UPDRS score. After this, further experiments were undertaken which involved comparing men and women, and then low versus moderate MDS-UPDRS scores. For the last two experiments the most discriminating features were extracted. It was concluded that predictive outcomes were improved by separating men from women, and by separating people with low MDS-UPDRS scores from those with moderate scores.

6.2 Methods

The *eyes closed* (EC) data from the *Canberra Hospital* (CH), the *University of New Mexico* (UNM; Carr, 2017), and the *University of Turku* (UTU; Suuronen et al., 2023) datasets were used here, preprocessed to Level 4 (full preprocessing including artifact rejection) as described in Sections 3.1.2 and 3.2.2, using single epochs. The absolute and relative power features, as well as power ratios, were extracted from six *region-of-interest* (ROIs), as described in Section 3.1.2. Features from the UTU and

UNM datasets were rescaled to the CH dataset and then combined to create one large dataset, as described in Section 4.3.

From the larger combined dataset, separate data splits were created for males, females, PwPD with MDS-UPDRS scores up to and including, 24, and PwPD with MDS-UPDRS scores over 24. The value of 24 was chosen here because it produced a median split. All PwPD in these four data splits were age-matched with a maximum difference of three years, and sex-matched, if the dataset contained both males and females, with controls. Two more data splits were made for an experiment classifying males with PD and females with PD against each other, and an experiment classifying the 26 PwPD with the lowest MDS-UPDRS against the 26 PwPD with the highest MDS-UPDRS scores. For the last experiment, the former group will be referred to as the "low MDS-UPDRS group", and the latter as the "high MDS-UPDRS group". In the male PwPD vs female PwPD experiment, males and females were matched by both age and MDS-UPDRS-score. In the low vs high MDS-UPDRS score experiment, the low MDS-UPDRS group had an average score of 13.85±4.60, and the high MDS-UPDRS group had an average score of 38.75 ± 9.81. An overview of the participant numbers for each split can be found in Table 6.1. In total, seven experiments were performed: classification using the full dataset for benchmarking, a female-only experiment, a male-only experiment, a low MDS-UPDRS-only experiment, a high MDS-UPDRS-only experiment, a female PwPD vs male PwPD experiment, and a low vs high MDS-UPDRS experiment. As with the previous experiments, the experiments described here were approved by the *Australian National University* (ANU) Human Research Ethics Committee (protocol No. 2020/612).

Table 6.1. An overview of participant numbers and age for the full dataset and per data split. SD means standard deviation.

Data split	N PwPD/N Controls	Average age ± SD PwPD	Average age ± SD controls
Full data	68/65	68.54±8.20	69.24±8.45
Female	22/22	66.59±7.69	66.95±8.40
Male	31/31	70.68±6.94	70.87±6.81
PwPD with UPDRS ≤ 24	31/31	67.94±7.06	68.13±7.34
PwPD with UPDRS > 24	32/32	69.56±8.02	69.75±8.27
	N females/ N males	**Average age ± SD females**	**Average age ± SD males**
Female vs male	26/26	69.88±6.71	70.19±6.39
	N Low UPDRS participants /N moderate UPDRS participants	**Average age low ± SD UPDRS participants**	**Average age ± SD moderate UPDRS participants**
Low vs high UPDRS	26/26	69.04±7.64	71.15±8.53

Additionally, to ensure a fair comparison that was not influenced by the difference in the number of male and female participants, a male-only experiment was run with 22 PwPD and 22 controls. The average age in this experiment was 68.36 years ± 5.81 for PwPD and 68.00 years ± 6.10

for controls, which was not statistically different from the age of the equivalent groups in the female-only experiment.

Classification was performed using a *Gradient Boosting Classifier* (GBC; Friedman, 2001; Natekin & Knoll, 2013) and data were log normalised. Nested *cross-validation* (CV) was performed as described in Section 4.3: hyperparameter optimisation was performed using grid search during the inner loop consisting of 10 times 10-fold CV, while the outer loop consisted of 5-fold CV. In both the inner and outer loops, participants were randomized based on ID number to avoid data from a participant being present in more than one set, as described in Section 4.3. All experiments were implemented in Python 3.6 (Van Rossum, 1995) using Scikit-Learn (Pedregosa et al., 2011). To quantify performance, accuracy, sensitivity, specificity, precision, and F1-score were obtained, and Gini-importances were used to find the most discriminative features for both ROIs and frequency bands in the female vs male and low vs high MDS-UPDRS experiments. Additionally, because multiple samples were classified for each participant, accuracies were also obtained for each individual participant when using the full dataset for an exploratory analysis of commonalities in MDS-UPDRS score, sex, and age based on these accuracies.

6.3 Results

An exploratory full-data experiment was performed to find commonalities between people that were classified correctly more often than others. Because each participant had multiple samples, it was possible to derive a percentage score of how many of a participant's samples were correctly classified. It was found that the group of people for whom more than 50% of samples were correctly classified were on average 3.38 years older than the people for whom less than 50% of samples were correctly classified. The difference in MDS-UPDRS score between PwPD who were correctly classified more than 50% but less than 75%, with an average of 23.54, and the group classified correctly more than 75%, with an average of 29.24, is 5.70 points. This indicates that samples from PwPD with higher MDS-UPDRS scores were generally more accurately classified than samples from PwPD with lower MDS-UPDRS scores.

Next, classification metrics were compared between data splits, showing that all evaluation results improved compared to those obtained when the full dataset was used. Table 6.2 shows the full dataset yielded a test accuracy of 58.94%, which was 11.02 percentage points lower than the female-only experiment, which was the data split with the lowest accuracy. The data splits also reduced the disparity between specificity and sensitivity. The difference between the two values was 11.64 percentage points for the full dataset and 7.40 percentage points for the female-only split, with all other data splits showing much smaller discrepancies between sensitivity and specificity.

Table 6.2. Evaluation results for all data splits and the full dataset.

Data split	Accuracy [%]	Sensitivity [%]	Specificity [%]	Precision [%]	F1-score [%]
full dataset	58.94	64.63	52.99	59.48	61.72
female	69.96	66.17	73.57	69.89	67.07
male	74.35	76.87	74.09	76.15	75.29
PwPD with UPDRS ≤ 24	76.67	77.18	77.84	80.48	77.13
PwPD with UPDRS > 24	75.06	78.20	73.13	74.02	74.56

As mentioned, the number of male participants was greater than the number of female participants, so for a fairer comparison an experiment with equal numbers was performed. As Table 6.3 shows, the reduced number of participants in the male-only experiment did not cause a decrease in evaluation results, rather they improved.

Table 6.3. Evaluation results for male and female experiments when equal numbers of participants are used in both.

Data split	Accuracy [%]	Sensitivity [%]	Specificity [%]	Precision [%]	F1-score [%]
male	83.44	84.77	84.47	82.98	82.74
female	69.96	66.17	73.57	69.89	67.07

In the two last experiments, only data from PwPD were used. In the first, females with PD and males with PD were classified, while in the second, 26 PwPD with the lowest MDS-UPDRS scores and 26 PwPD with the highest MDS-UPDRS scores were classified. Table 6.4 provides and overview of accuracies when performing these group vs group experiments, as well as the accuracies per group. The results show that differences between males and females, as well as between people with low and high MDS-UPDRS scores are large enough to classify them based on RSEEG alone. The next question to answer is then what features of the data makes this classification possible.

Table 6.4. Evaluation metrics from the experiments classifying females and males, and people with low MDS-UPDRS scores and people with moderate MDS-UPDRS scores.

Experiment	Accuracy [%]	Group 1 accuracy [%]	Group 2 accuracy [%]
female vs male	68.59	67.08	71.42
low vs high MDS-UPDRS	71.47	69.77	72.94

To clarify what contributed to correct classification in the female vs male and low vs high MDS-UPDRS experiments, features were analysed by ROI and frequency band. Selected features were relatively evenly distributed across the scalp and the number of features selected was also relatively stable, being 9 for the female vs male experiment, and 10 for the low MDS-UPDRS vs moderate MDS-UPDRS experiment. Except for beta in the low vs high UPDRS experiment, which is relative beta, all features were absolute power features (Figure 6.1). Based on the frequency bands and ROIs from which

features were derived, it seems differences between sexes and disease scores come about due to global differences in activity and not due to specific areas or frequencies bands being very different.

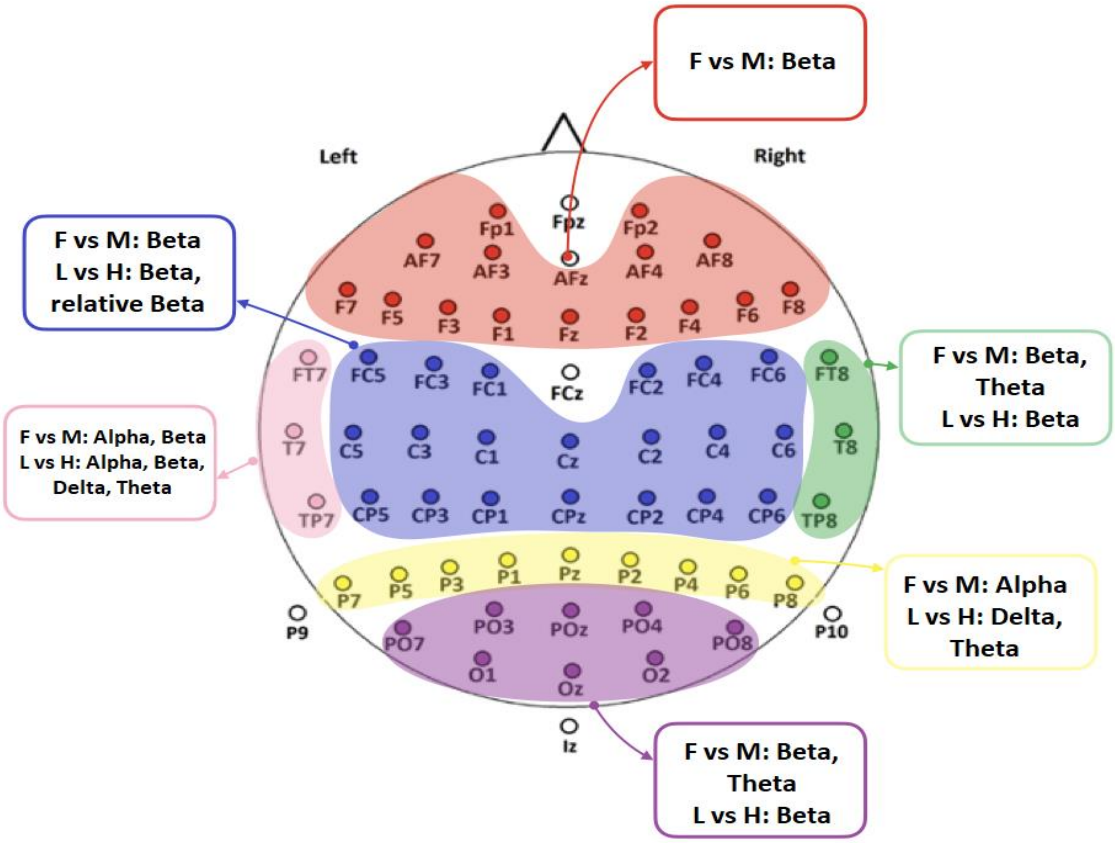


Figure 6.1. An overview of features per ROI by frequency band. “F vs M” refers to the female vs male experiment, and “L vs H” to the low vs high MDS-UPDRS experiment.

6.4 Discussion

In this study, we investigated the effects of participant characteristics on classification outcomes by splitting datasets by sex and MDS-UPDRS score. The experiments provided evidence that splitting data by sex and by MDS-UPDRS both improve classification metrics compared to a dataset that uses a mix of genders and disease stages. Male participants were more accurately classified than female participants regardless of whether or not the numbers of females and males in the experiments were the same. Running two experiments classifying females vs males and people with low MDS-UPDRS scores vs high MDS-UPDRS scores provided evidence of differences in EEG features between females and males with PD, and between those with low MDS-UPDRS and high MDS-UPDRS. These differences were global rather than restricted to a specific frequency band or ROI.

The highest accuracy, specificity, precision, and F1-score out of all data splits was achieved when classifying people with MDS-UPDRS scores of up to and including 24 against controls. This was an encouraging finding, because if RSEEG was to be used for the diagnosis of PD, it is necessary to be

able to classify people that would score low on the MDS-UPDRS, as people early on in their disease generally score lower than people that have lived with PD for longer. This data split also produced the most balanced results, with a difference of only 0.66 percentage point between sensitivity and specificity.

The data split using MDS-UPDRS scores over 24 also produced greatly improved metrics, with a sensitivity higher than in the data split with MDS-UPDRS scores up to and including 24, but slightly lower scores for all other evaluation metrics. One would expect that evaluation results would improve when comparing experiments with participants with higher to those with lower MDS-UPDRS scores, so it is surprising that this is not the case here. A possible reason for the discrepancy could be the distribution of males and females. In the data split with MDS-UPDRS scores up to and including 24, 16 out of 31 participants were female. In the data split using MDS-UPDRS scores over 24, 11 out of 32 participants were female. Both male-only experiments delivered better outcomes than the female-only experiments, so it is possible the different number of females caused a difference in results.

The low MDS-UPDRS vs high MDS-UPDRS experiment produced a test accuracy of 71.47%, showing that the model was able to distinguish between the two groups. This is encouraging if RSEEG is to be used for the purposes of prognosis, as the results show there are differences between people at different stages of their disease. The analysis of the location and frequency bands of the features showed that they are relatively equally distributed across locations. This could be because degeneration tends to become more global as PD progresses rather than being confined to specific brain structures. It is then not surprising to find that the features that lead to the best classification are also distributed across frequency bands and ROIs.

The study presented here has several limitations. First, while combining the three datasets enabled creating data splits of participants by sex and MDS-UPDRS and the results provided evidence that those splits improved classification outcomes, it was not possible to split the dataset by sex and MDS-UPDRS. Such a split would lead to four rather small data splits that might not be very informative, which is exactly the problem that was avoided in this study by combining the three datasets. Second, the literature has often found a slowing of frequencies when comparing PwPD to controls, with differences observed in the theta and delta bands. Comparisons between PwPD at different stages of their disease and between sexes like they are presented here have not been made, which makes it difficult to interpret the results concerning the ROIs and frequency bands that distinguishing features come from.

This study also provides avenues for future studies. The EEG literature provides evidence for the physiological differences between males and females (Greischar et al., 2004; Li et al., 2007), as well as for differences in disease progression (Cerri, Mus, & Blandini, 2019). When looking at characteristics of the people who generally participate in PD EEG experiments, it turns out that participants have an

average age of around 70 years old. Many males will have developed some form of male pattern baldness at this point (Hamilton, 1951), while most females, while their hair might have thinned, will have more hair. Given that anything which is placed between the electrode and neural cells might influence the signal, and, given that research shows different hairstyles and hair thickness contribute to differences in EEG recordings (Choy, Baker, & Stavropoulos, 2022), it seems that this should be explored in the specific context of PD research.

The results presented here seem to indicate that there are indeed differences between males and females, as the female vs male experiment shows that the groups could be classified with an accuracy of 68.59%, but why these differences exist should be investigated. Studies of such differences in the context of ML classification tasks have been performed using other kinds of data, such as voice data (Wang et al., 2020) with the process of twinning, meaning matching people by age and sex, being successful in evaluating features and uncovering differences between participants. Similar approaches would also be worthwhile of investigation in the context of RSEEG in PD. Lastly, it would be of interest to replicate methods such as these described by Prichep et al. (2012), based on earlier research into aging of the brain and its effect on EEG (John et al., 1980; John et al., 1988; Thatcher et al., 2009), to regress out the influence of the age differences between participants, as age differences can be quite large in PD datasets.

Another issue that is yet to be resolved is the nature of features and ROIs. PD has been shown to develop differently in males and females (Cerri, Mus, & Blandini, 2019), so it is possible that there are fundamental differences between males and females on the electrophysiological level that require investigation; it could be the case that there are other feature quantification methods that would work better for classification of females with PD. Similarly, as degeneration tends to become more global as PD progresses, rather than being confined to specific brain structures, it could be possible that measurements such as global power or different ROIs might lead to better results. The study presented here investigated if splitting data by factors that can be identified beforehand would improve evaluation metrics. To answer which features in the identified subgroups are necessary for reliable classification would require separate studies with larger cohorts with similar characteristics. At present, differences in EEG data, specifically between males and females with PD, are sparsely investigated and provide an avenue of future research in order to develop separate diagnostic and prognostic models for men and women.

This study provides evidence that differences exist between people at different stages of PD progression and between males and females. These results provide evidence of the benefit of splitting data based on these participant characteristics, which can help researchers that aim to develop diagnostic methods for PD create more specific models. Additionally, the results of this study provide evidence

that RSEEG is not only useful to distinguish PwPD from controls, but also PwPD at different stages of their disease. This indicates the potential of RSEEG as a prognostic tool.

The present study only looked at PwPD and controls, dividing people by sex and age, but this is of course not the only classification work that can be undertaken. As mentioned in Section 2.2.3, EEG has been shown to allow differentiation between PwPD and other diagnoses (Bonanni et al., 2008; Barcelon et al., 2019), so a worthwhile research endeavour might be to extend this line of research and investigate differential diagnoses in females and males separately. Additionally, dementia is a common feature of PD (Hely et al., 2008), but might affect different PwPD at different rates, so a more in-depth investigation of the probability of developing dementia would also be an avenue for future research.

The collection of sufficient amounts of data to perform any kind of classification experiment is difficult enough without putting any extra constraints on the sex and MDS-UPDRS score of the participants, but the experiments here show that there are differences between the groups that influence classification outcomes. In this study, this problem was partially solved by combining datasets, but a problem with that approach was that the different datasets were collected using different equipment and slightly different conditions (see Section 2.5). More uniformity in the equipment used to collect data, the number of electrodes used, and the length and acquisition conditions would be a step forward in this respect. Additionally, collection of data from PwPD is difficult, but creation of a larger database of control participants would be slightly easier and would benefit from a collective effort.

In a broader context, the study presented here shows the importance of large enough datasets to isolate specific characteristics of participants that may influence results. The study of other neurodegenerative diseases through RSEEG is hindered by the same difficulty of collecting data, so the approach taken here is worthwhile to considered for other diseases as well.

6.5 Conclusion

This study gives evidence that splitting data by sex and by MDS-UPDRS score improves the performance of classification between PwPD and controls. The female vs male experiment indicates that there are differences between males and females which warrant splitting data by sex to produce better models. The low MDS-UPDRS vs controls experiment is promising from a diagnosis standpoint, as the results show that even early in the disease it is possible to distinguish between PwPD and controls. The low MDS-UPDRS vs moderate MDS-UPDRS experiment provides evidence that RSEEG can potentially be used for the purpose of prognosis, as the differences between the two groups are large enough for reliable classification. Based on the results, it is clear that splitting the data is important for the development of better models, and that more data should be collected to investigate interactions such as MDS-UPDRS score and sex, which was not possible here due to data scarcity.

6.6 Chapter summary and conclusion

In this chapter, the last remaining question regarding the use of RSEEG for the classification of PwPD and controls was investigated: what are the effects of different participant characteristics on classification outcomes and should data be split based on certain characteristics? Building on the results presented in the previous chapters, data from three datasets were fully preprocessed, features were extracted and the data from the different datasets combined, different data splits based on sex and disease status were created, and PwPD and controls classified using GBCs. The results show that splitting data by sex and MDS-UPDRS score improves classification outcomes. Given these results, larger datasets are a necessity to develop RSEEG into a diagnostic and prognostic tool for PD, and a focus on models for separate subsets of PwPD will most likely produce the best results. Interestingly, the outlier rejection that was necessary to achieve results on the mixed data in Chapter 4 was not necessary to produce good results when splitting the data as was done in this chapter.

Chapter 7: Event-related potentials in neurodegenerative diseases

Most of research in this thesis so far has focused on *resting-state electroencephalography* (RSEEG), but *event-related potentials* (ERPs), measured during tasks, have been a topic of investigation for decades. Using a wide variety of cognitive paradigms, researchers have attempted to make clear the differences between healthy controls and people with neurodegenerative diseases. The variety of paradigms has enabled researchers to investigate different aspects of cognition, such as working memory and attention, but it also makes it more difficult to compare results. Electrode setup, preprocessing, number of stimuli, stimulus modality, and latency and amplitude quantification can all differ depending on the aim of the study. In this chapter, methods used in ERP research into cognition in *Multiple Sclerosis* (MS) and *Parkinson's disease* (PD) will be investigated to come to a better understanding of ERP research in neurodegenerative diseases.

After an introduction of the neurodegenerative disease MS and an explanation of why this disease was chosen, a systematic review of the use of EPRs in the investigation of cognitive performance in people with MS is presented. The systematic review set out to answer five questions related to the use of ERPs in distinguishing people with MS (PwMS) from controls, namely: (1) which stimulus modality (i.e., visual or auditory) discriminates best; (2) which experimental paradigm produces the best discrimination; (3) which individual electrode(s) generate(s) the best results; (4) which electrophysiological component(s) of the ERP discriminate best; and, (5) whether amplitude or latency is a better measure.

The MS review is compared to and contrasted with a literature review by Seer et al. (2016) examining the use of ERPs in the study of PD. This review investigated ERPs in the context of cognitive dysfunction and decline in idiopathic PD. The review by Seer et al. and the MS review are compared to clarify the differences and commonalities between research in MS and PD in order to gain a better understanding of ERP research in neurodegenerative diseases in general.

7.1 The use of event-related potentials in the investigation of cognitive performance in people with Multiple Sclerosis: systematic review

As mentioned in Section 2.8, while PD and MS are very different diseases, the methods used to investigate differences between people with either disease and controls, as well as disease progression, are similar. The hardware, software, and general preprocessing and analysis methods remain the same regardless of what disease is studied as the methods are based on physiological processes that are the same in every healthy human being. How the results that are obtained differ between PwPD and PwMS is important to consider, but the methods themselves are the same, with protocols being adjusted to the peculiarities of the disease.

As will become clear later, the same ERP paradigms are also used to study both MS and PD. This provides us with an opportunity to compare and contrast these methods in order to gain a better understanding of ERP research into neurodegenerative diseases. For example, the results in both diseases can inform us on what paradigms are generally used, which paradigms produce the best results in the context of distinguishing between a control and a person with a disease, and which stimulus modality produces the most reliable results. In this manner, we can extract general recommendations for future ERP investigations in neurodegenerative diseases.

Just as some researchers have investigated RSEEG, other researchers have asked themselves the question of what changes can be observed when people with neurological diseases are compared to controls when performing specific tasks and measuring their brain activity during those tasks. ERPs are the electrophysiological activity of the brain during cognitive tasks and are used to quantify domains of cognition, such as attention, short-term memory, and processing speed. In the following systematic review, we investigated how ERPs have been measured in the quantification of cognition in MS, and answered questions about the best ways in which to do so.

Acknowledgment: The systematic review presented in Sections 7.1.1 through 7.1.5 was published in ‘Brain Research’. The study was conceived by me in collaboration with my supervisors and co-authors Dr. Elena Daskalaki, Associate Professor Deborah Apthorp, Professor Emeritus Christian J. Lueck, and Professor Hanna Suominen. Dianne Walton-Sonda helped me in the development of the search strategy and extraction of studies from databases. Dr. Duncan Austin acted as the second reviewer when assessing studies for inclusion and validated the extracted data. Dr. Artem Lenskiy assisted me in the extraction of data from the selected studies. The initial manuscript and its subsequent versions, refined using comments from my co-authors, were written by me. I presented a preliminary version of the results at the 2022 ACT MS Symposium.

7.1.1 Introduction

At present, disease severity in Multiple Sclerosis (MS) is most often measured clinically using the Expanded Disability Status Scale (EDSS; Kurtzke, 1983). However, this tool is recognised to lack precision and is dominated by assessment of motor activity, particularly at its more severe end (Meyer-Moock, Feng, Maeurer, Dippel, & Kohlmann, 2014). The focus on motor activity means that the EDSS does not necessarily provide a holistic assessment of disability. For example, cognitive deficits are reported in 43%-70% of People with MS (PwMS) (Chiaravalloti & DeLuca, 2008). These cognitive deficits can often be detected in people with very mild disease, for example, a radiologically isolated syndrome (Lebrun et al., 2010) or a clinically isolated syndrome (Anhoque et al., 2012).

Several tests of cognitive function are already in clinical use, including the Symbol Digit Modalities Test (SDMT; Smith, 1973) and the Paced Auditory Serial Addition Test (PASAT; Gronwall, 1977). These tests are not without criticism: for example, the SDMT, while being a sensitive test, measures cognition in a non-specific manner (Sandry et al., 2021; Berrigan et al., 2022), and the PASAT has been observed to suffer from learning effects (Tombaugh, 2006; Nagels et al., 2008). In an effort to provide a more complete assessment of cognition, test batteries consisting of tests assessing individual cognitive domains have been designed. Examples of such test batteries are the Minimal Assessment of Cognitive Function in MS (MACFIMS; Benedict et al., 2006), TRACK-MS (Taranu et al., 2022), and the Brief Repeatable Battery of Neuropsychological Tests (BRBNT; Boringa et al., 2001). While providing a more complete overview of cognition, the results of these tests are still dependent on responses made by the participants. ERPs, while often requiring a participant to generate a response, provide a direct measurement of brain activity itself, thereby providing an additional dimension to the quantification of cognitive functioning and offering a potential advantage over more conventional neuropsychological testing.

Accordingly, it has been suggested that the event-related potential (ERP), an electrophysiological brain response time-locked to a stimulus, might provide a more objective measure in this way (Newton et al., 1989, Vazquez-Marrufo, 2017). ERPs have been studied in several different neurodegenerative diseases, and many abnormalities have been described (Tanaka et al., 2000; Turner et al., 2015; Covey et al., 2017). The relative ease with which ERPs can be obtained (measurements can be made from just a small number of electrodes), together with research providing evidence that ERPs can detect compensatory mechanisms before cognitive impairment becomes apparent, indicate that ERPs might act as a clinically useful measure of disease severity and/or prognosis in PwMS. Specifically, combining ERPs with individual tests such as the SDMT and PASAT or the aforementioned test batteries could lead to a more holistic and objective assessment of cognition in MS and, very possibly, provide evidence of abnormality before it becomes detectable by the existing tests.

ERPs are measured using scalp electrodes that are most often embedded in an electrode cap. The stimuli used can be cognitive, visual, auditory, or tactile in nature, and the resulting ERPs are fairly stereotyped, with several components that are labelled according to their polarity (i.e., whether the component has a negative or positive peak) and their approximate latency. For example, P300 refers to a positive potential occurring at around 300 ms after a stimulus (Luck & Kappenman, 2011). This differs from the analysis of electroencephalographic data in the frequency domain (Cohen, 2014), which has been used both in clinical settings as well as to investigate cognition. In such scenarios characteristics of the signal such as power in particular frequency bands and connectivity are analysed (Leocani et al., 2010; Sjøgård et al., 2021; Jamoussi et al., 2023), rather than looking at specific responses in the time domain. Evoked potential (EP) studies in clinical use, on the other hand, do analyse data in the time domain but these tests, although sometimes used to provide information on broader cognitive functioning (Hansch et al., 1982; O'Donnell et al., 1987), generally focus on sensory pathway functioning rather than on specific cognitive processes. The components measured during ERP experiments can be related to specific processes, such as focusing attention on a stimulus or making a decision about the stimulus, enabling the study of different cognitive domains.

The current gold standard method to assess structural cortical changes in MS is magnetic resonance imaging (MRI), with functional MRI (fMRI) used as a method to measure cortical activity during tasks (Filippi et al, 2010; Filippi & Rocca, 2013; Benedict et al., 2020). The use of (f)MRI has two major drawbacks. First, the high cost of MRI equipment, often running in the millions of dollars, and the necessity for MRI scanners to be housed in purpose-built facilities, make that it is generally only available in areas with a sufficiently large population. EEG equipment, on the other hand, is available for a fraction of the cost of MRI scanners and can be set up in almost any room. Second, fMRI is based on the haemodynamic response to stimuli, which is slower than the underlying neural processes in response to the stimuli. This can be partially alleviated by the experimental designs, but does not reach the temporal specificity of ERP measurements (Glover, 2011).

Most research into degenerative diseases has looked at differences in the latency and amplitude of ERPs. A good example involves Huntington's Disease (HD) where differences in ERPs have been found between people with clinical HD, people who are asymptomatic HD gene carriers, and age-matched controls (Turner et al., 2015). ERPs have also been studied in PwMS. Collectively, these studies have used many different types of stimuli, required different types of responses, and measured both amplitudes and latencies of ERPs in many different ways. This means that it is difficult to determine from individual studies how ERPs might best be used to study MS, and to determine the potential role of ERPs in the diagnosis, assessment, and management of MS.

Accordingly, this systematic review set out to answer five questions related to the use of ERPs in distinguishing PwMS from controls, namely: (1) which stimulus modality (i.e., visual or auditory) is

most discriminatory, (2) which experimental paradigm produces the most discriminatory results, (3) which individual electrodes generate the best results, (4) which electrophysiological component(s) of the ERP are most discriminatory and, (5) whether amplitude or latency is a better measure.

7.1.2 Methods

This review was carried out according to the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) guidelines for systematic reviews (Page et al., 2021), and was registered under the international prospective register of systematic reviews (PROSPERO) number CRD42020166633. Studies were retrieved from Medline, Embase, and Cumulative Index of Nursing and Allied Health Literature (CINAHL) databases in May 2023 using the search strategy "Multiple Sclerosis" AND ("contingent negative variation" OR "event-related potential" OR "readiness potential" OR "bereitschaftspotential" OR "bereitschafts potential" OR "event-related brain potential" OR "event-related auditory evoked potential"). Retrieved studies were screened for duplicates and uploaded to Covidence (Babineau, 2014). After title and abstract screening, a full-text review was carried out by two reviewers with relevant and complementary experience; any conflicts were resolved through mutual discussion or adjudication by a third reviewer. Validity of the studies was assessed using the Risk of Bias Assessment tool for Non-randomized Studies (RoBANS).

Studies were included if they met the following criteria:

1. Participants were 18 years of age or older.
2. A comparison was made between a control group and PwMS, or between different subgroups of PwMS.
3. ERPs were measured during a cognitive task.
4. Specific values for latencies and amplitudes were reported or could be reasonably accurately derived from published figures.

Studies into EPs, meaning those studies involving detection of an electroencephalographic response to a stimulus without requiring any cognitive processing of the stimulus, were excluded. However, because the terms ERPs and EPs have sometimes been used interchangeably, studies looking at 'cognitive EPs', meaning EPs that were measured during a cognitive task, were included. Other studies that were excluded were studies that were not original research, such as review articles, studies that did not use ERPs as the comparator, such as studies into the effects of drug interventions, studies in which ERP processing or the study design could not be extracted from the text, and studies that provided insufficient data for further analysis, such as conference abstracts. Studies that did not measure ERP component and did not report the temporal occurrence of responses were excluded. If a study into rehabilitation or drug intervention included baseline data, these baseline data were included. No limitations were placed on date of publication or language of publication.

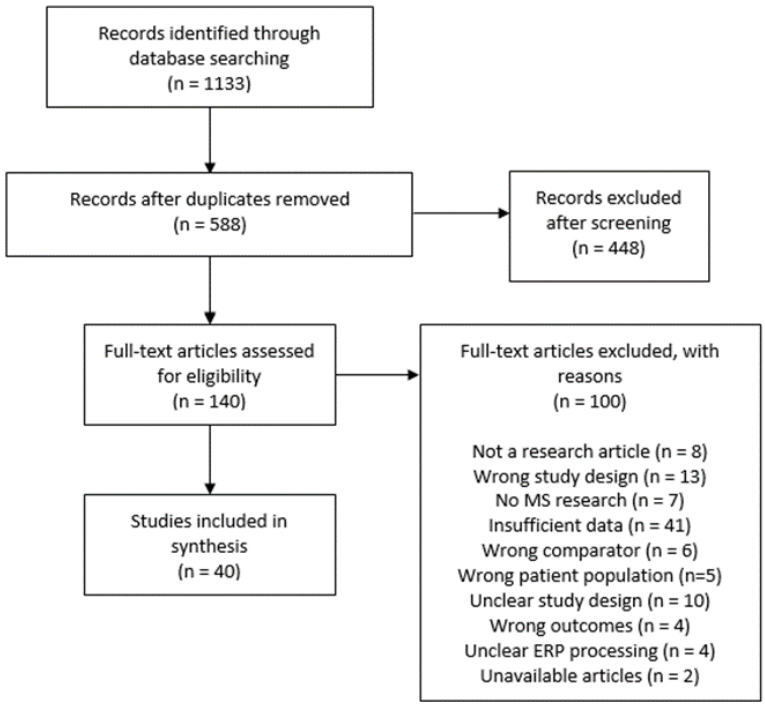
For each study, the following information was extracted: demographic information on the PwMS and controls, the cognitive paradigm used, the neuropsychological tests that were administered, the number of electrodes, the reference, and the sample frequency used, and the latency and amplitude measures for each component of the ERP. In the PwMS group, the EDSS score and MS subtypes relapsing-remitting MS (RRMS), primary progressive MS (PPMS), and secondary progressive MS (SPMS) were recorded.

Pooled effect sizes and confidence intervals were calculated for each component and displayed as forest plots using RevMan 5.4 (The Cochrane Collaboration, 2020). Effect sizes quantify the magnitude of the difference between two groups. They have been used in medical research to describe treatment effects, for example, and it is argued they provide a better assessment of these effects than p-values alone (McGough & Faraone, 2009). Hedges' g was used to assess effect sizes: an effect size of 0.2 or smaller was considered small, an effect size of 0.2–0.5 medium, and an effect size of 0.8 or more large (Sullivan & Feinn, 2012). A list of reviewed studies can be found in Supplement 2.

7.1.3 Results

In total, 1133 studies were retrieved, of which 588 remained after removal of duplicates. 448 studies were removed based on title and/or abstract, and another 100 were excluded after a full-text review. Data were extracted from the remaining 40 studies (Figure 1).

Figure 1. PRISMA diagram from the 1133 identified to 40 included studies.



In all, the 40 studies reported on a total of 62 experiments. Of these experiments, 33, 8, and 6 reported on people with RRMS, SPMS, and PPMS, respectively, and 4 experiments included people with Clinically Isolated Syndrome (CIS). Eleven studies included PwMS with different subtypes of MS but did not distinguish between the subtypes when analysing the results. Unfortunately, demographic data on people with PPMS were not provided by the specific studies so this information was not available for inclusion in Table 1. Study population sizes ranged from 19 to 172, with an average of 32 PwMS and 26 controls per study. In the studies that reported on the sex of participants, 62.5% of PwMS were female compared to 63% of controls. Unfortunately, while most, but not all, studies reported characteristics such as sex, age, disease duration, and EDSS score, characteristics such as ethnicity were absent. Participant demographics are provided in Table 1 and details per experiment can be found in Supplement 3.

Table 1. Breakdown of participants in the 62 experiments. Ranges for all values across studies are given in brackets.

	CIS	RRMS	PPMS	SPMS	BMS	PwMS	Controls
average number of participants	20.5 [7-44]	28.6 [11-72]	N/A	16 [-]	9.7 [9-10]	32.2 [9-101]	26.3 [7-89]
average number of female participants	16.3 [6-27]	18.2 [7-51]	N/A	11 [-]	6 [-]	18.6 [6-72]	17.9 [4-51]
average number of male participants	8.7 [4-17]	10.4 [0-32]	N/A	5 [-]	3.7 [3-4]	9.9 [0-32]	10.0 [0-38]
average age (years)	31.6 [27.4-40.4]	37.3 [32.5-45.6]	N/A	43.8 [-]	41.1 [38.7-42.3]	38.9 [27.4-51.0]	37.2 [277 -46.5]
average disease duration (years)	3.3 [0.4-4.9]	6.8 [1.3-11.6]	N/A	13.9 [-]	12.2 [12.1-12.4]	8.2 [0.4-17.7]	-
average EDSS score	1.7 [1.4-3.0]	2.3 [0.9-3.1]	N/A	5.8 [-]	1.8 [1.6-1.9]	3.0 [0.87-5.78]	-

CIS: clinically isolated syndrome; EDSS: expanded disability status scale; PwMS: People with MS; RRMS: relapsing remitting MS; PPMS: primary progressive MS; SPMS: secondary progressive MS; BMS: benign MS. Note: average numbers for female and male participants may not correspond to averages for all participants because the number of studies reporting numbers for female and male participants differ.

Five experimental paradigms, namely the oddball paradigm, Posner paradigm, n-back tasks, choice reaction experiments, and sensorimotor integration paradigm (see Supplement 4 for general examples of these paradigms), were used in more than one experiment. Table 2 provides an overview of these paradigms, but a full overview of paradigms and experimental settings used per study can be found in Supplement 5. Most studies reported on oddball experiments, in which participants are presented with a string of stimuli of which some differ. The participants have to either count these ‘oddballs’ or respond to them through, for example, a button press. While the aim of the paradigm is the same, there were

large differences in individual paradigms. There was considerable variation in the number of stimuli per experiment, ranging from 100 to 900 for auditory stimuli and 100 to 400 for visual stimuli. In addition, the frequency of the tone used in auditory experiments was also variable, with the oddball tone ranging from 1000 Hz to 2000 Hz and the standard tone from 500 Hz to 1000 Hz. Even though differing in their specifics, most oddball experiments used auditory stimuli and asked participants to count the number of oddballs.

Table 2. Overview of the experimental paradigms that were used in more than one experiment.

Experimental paradigm	N experiments		N visual and auditory experiments
Oddball experiments	42		
		visual	8
		auditory	34
Posner paradigms	4		
		visual	4
Choice reaction experiments	3		
		visual	2
		auditory	1
N-back tasks	2		
		visual	2
Sensorimotor integration paradigm	2		
		multimodal	2

In the Posner paradigm experiments (Posner, 1980), the paradigms were more uniform. A participant was first shown a cue in the form of an arrow, followed by a target or standard stimulus. Participants had to withhold responses to the standard stimulus and press the left or right button depending on the side on which a target stimulus appeared. (Gonzalez-Rosa et al., 2006; Vázquez-Marrufo et al. 2009, Gonzalez-Rosa et al., 2011) For the choice reaction experiments, (Sundgren et al. 2015, Cooray et al. 2020), participants were presented with two different auditory or visual stimuli and asked to press a button that corresponded to the presented stimulus with their left or right index finger. In the n-back tasks, (Covey et al. 2017) participants had to compare the current letter that was being presented to a letter presented n stimuli earlier. They pressed two buttons with their thumbs if they detected a match, and two different buttons if there was no match. The study also included a 0-back experiment where participants had to press a button the moment a stimulus appeared, effectively making it a choice-reaction experiment. In the sensorimotor integration experiments (Uysal et al. 2014),

participants were presented with a warning tone, followed by stimulus 2 seconds later to which they had to respond as quickly as possible by pressing a button with their dominant hand. Supplement 5 provides an overview of experimental parameters per experiment.

The many differences in experimental paradigms required analysis from several different angles. In order to address the five questions listed in the introduction, the results of the relevant studies had to be collapsed across different paradigms, electrodes, modalities, etcetera, even though the data were very heterogeneous.

Comparison of Auditory and Visual Stimuli. Overall, studies using visual stimuli generated greater effect sizes (Figure 2) and there was a larger difference for latency than for amplitude. Pooled effect sizes for latency measurements were 0.3 CI: 0.15, 0.46] and 0.68 CI: 0.55, 0.81] for auditory and visual stimuli, respectively, while the same measurements for amplitude were 0.15 CI: 0.05, 0.26] and 0.52 CI: 0.42, 0.62].

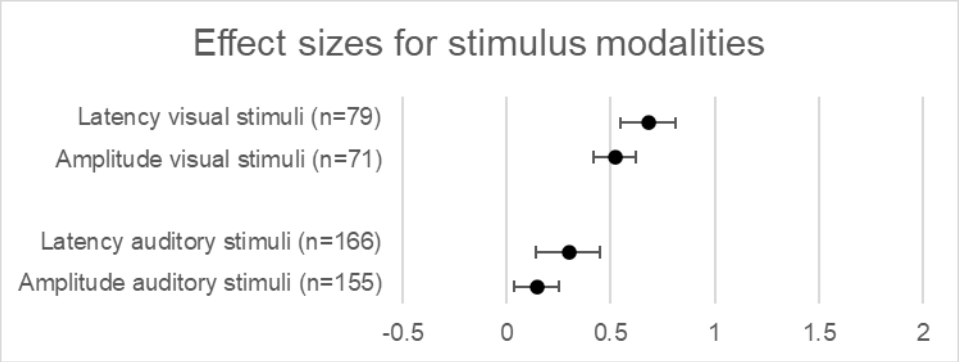
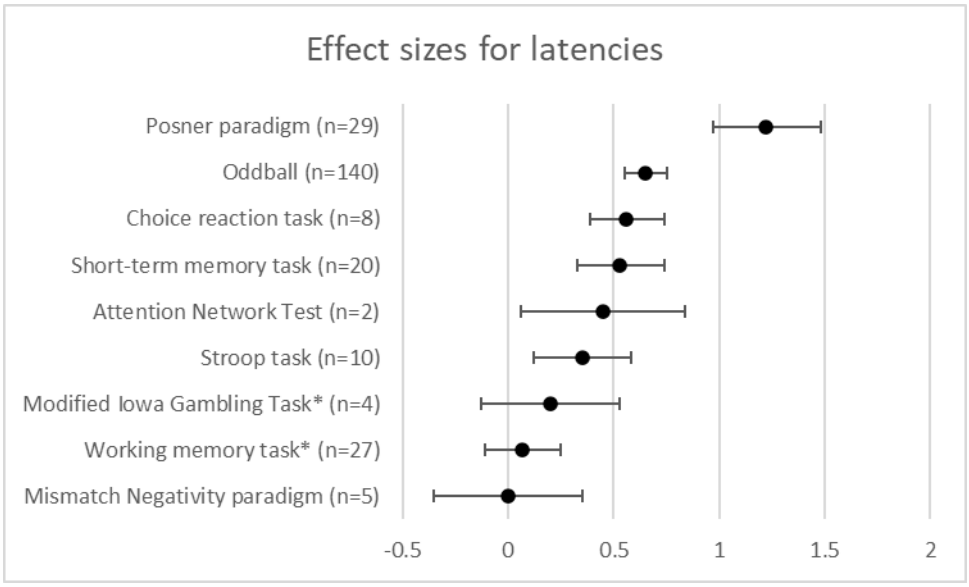


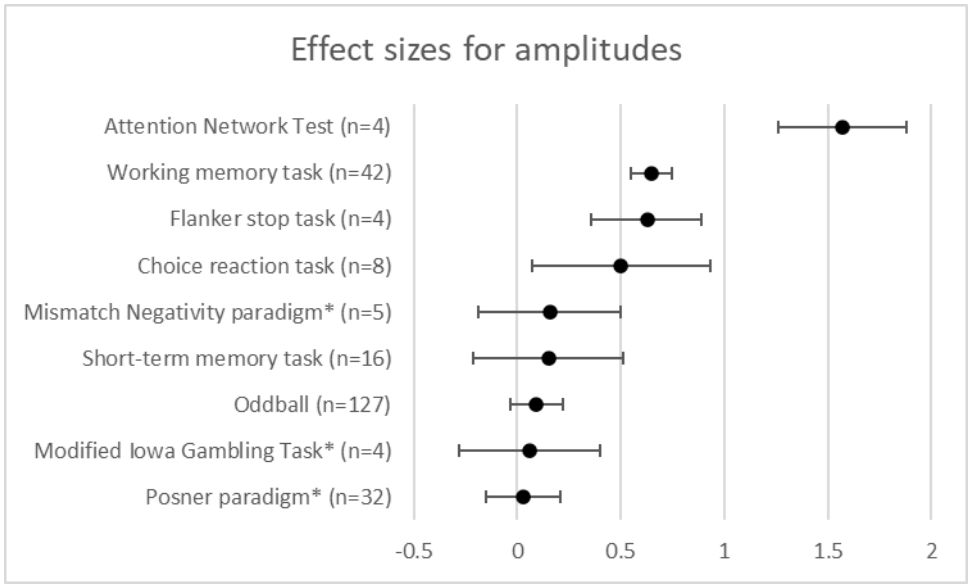
Figure 2. Pooled effect sizes and 95% confidence intervals comparing visual and auditory stimuli for latency and amplitude. 'n' refers to the total number of measurements made across all experiments, because multiple measurements from different electrodes were often made in any given experiment.

Comparison of experimental paradigms. Oddball experiments made up approximately 63% of experiments but the Posner paradigm produced the largest pooled effect size for latency (Figure 3a) and the Attention Network Test (ANT) produced the largest effect size for amplitude (Figure 3b). Latencies were longer for PwMS than controls except for the working memory tasks and modified Iowa Gambling Tasks (marked with an asterisk).

a)



b)



Figures 3a and 3b. Pooled effect sizes and 95% confidence intervals for the experimental paradigms looking at latency (a) and amplitude (b). 'n' refers to the total number of measurements made across all experiments, because multiple measurements from different electrodes were often made in any given experiment. In measurements with an asterisk, measurements for PwMS were larger than those of controls, while all other measurements were larger for the control group.

Comparison of electrodes. Almost all experiments reported measurements from electrodes Fz, Cz, and Pz. (Figure 4). Latency measurements were largest for electrode Cz (ES = 0.84 0.62, 1.05]), while the pooled effect size for Fz was noticeably smaller. Pooled effect sizes for all three electrodes were smaller for amplitude measurements than for latency measurements.

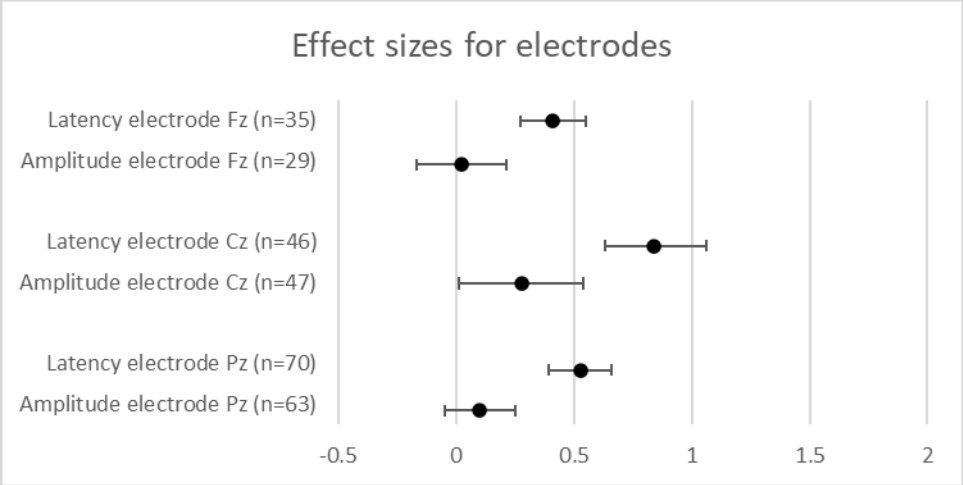
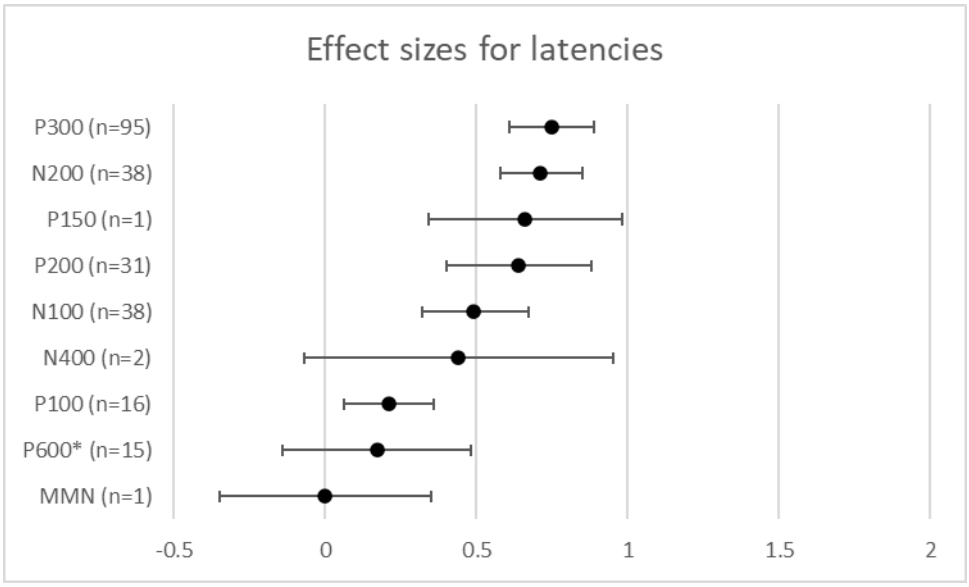


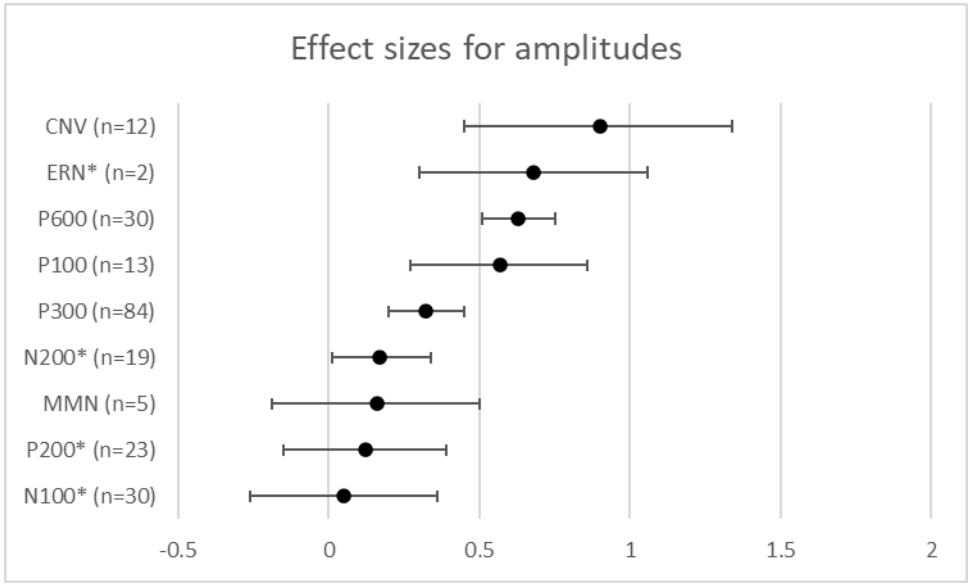
Figure 4. Pooled effect sizes and 95% confidence intervals for measurements derived from electrodes Fz, Cz, and Fz for latency and amplitude. 'n' refers to the total number of measurements made across all experiments, because multiple measurements from different electrodes were often made in any given experiment.

Comparison of ERP components. The precise component of the ERP that was measured in each experiment depended on the paradigm used. As expected from the large number of oddball experiments, the component that was most often measured was P300, followed by N100 and N200 (Figure 5). The N200-P300 complex was frequently measured as a single entity: this measure refers to the peak-to-peak difference in amplitude between the N200 to P300 components, as these two potentials immediately follow each other. For latency measurements, P300 and N200 produced the largest pooled effect sizes, followed by P150, but this was based on only 1 experiment, resulting in a large confidence interval. In auditory experiments, the largest pooled effect sizes for amplitude were observed at the contingent negative variation (CNV, g = 0.9 0.46, 1.35]). The Error-Related Negativity (ERN, g = 0.68 0.3, 1.06]), which is the potential that occurs when a participant commits an error in a cognitive task, and P600 (g = 0.63 0.51, 0.75]), generated the second and third largest pooled effect sizes.

a)



b)



Figures 5a and 5b. Pooled effect sizes and 95% confidence intervals for the different ERP components looking at latency (a) and amplitude (b). 'n' refers to the total number of measurements made across all experiments, because multiple measurements from different electrodes were often made in any given experiment. In measurements with an asterisk, measurements for PwMS were larger than those of controls, while all other measurements were larger for the control group.

Comparison of Latency and Amplitude. Effect sizes for individual latency measurements ranged considerably, with a value of 1.51 for shorter latencies in PwMS at one extreme to 5.02 for shorter latencies in controls at the other. In total, there were 34 measurements in which latencies in PwMS were shorter than those in controls compared to 211 measurements in which latencies were shorter in controls. Regarding amplitude measurements, effect sizes ranged from 1.47 higher in PwMS to 22.98 higher amplitudes in controls. In total, there were 83 measurements in which amplitudes in PwMS were higher than those in controls compared to 143 measurements in which amplitudes were higher in controls.

Overall, latency produced larger pooled effect sizes than amplitude when comparing experiments with auditory and visual stimuli. When pooling based on electrode location, latency produced larger pooled effect sizes for all three locations. For the modified Iowa Gambling Tasks (mIGT), the short-term memory tasks, the oddball experiments, and the Posner paradigm experiments, latency produced larger pooled effect sizes, while the pooled effect sizes produced by measuring amplitude were larger for the n-back tasks, working memory tasks, and the ANTs. Interestingly, the largest individual pooled effect size occurred for the amplitude measurements in the ANTs. When calculating pooled effect sizes for components, the latency pooled effect sizes were larger for components N100, N200, P200, and P300, while amplitude pooled effect sizes were larger for P100 and P600.

Direction of effects. As mentioned above, not all effect sizes were in the same direction. For most paradigms, latencies were longer and amplitudes lower for PwMS compared to controls, but this was not the case in experiments using the mIGT and working memory tasks for latencies, and Posner paradigms, the mIGT, and mismatch negativity tasks for amplitudes. Components, locations and stimulus modality leading to these results vary. P300 was measured from electrode Pz at four different moments during a visual mIGT, with latencies being longer for controls than PwMS for all four, but amplitude only being larger for PwMS than controls in one measurement, and P600 showed longer latencies in an auditory working memory task at 9 of 15 measured locations, mainly over frontal and central electrodes. 3 out of 5 measurements in electrode Fz during a mismatch negativity experiment showed larger amplitudes for PwMS than controls, while the majority of measurements during the Posner paradigm for components N100, N200, and P200 were larger for PwMS than controls in unspecified electrodes. Lastly, two measurements of the ERN during a flanker test also showed larger amplitudes for PwMS than controls.

ERP components and neuropsychological tests. Of the forty studies included in this systematic review, 26 also reported on the results of neuropsychological tests. Of the 26 studies, 12 employed the SDMT and 11 employed the PASAT (see Table 3 for details on the versions used) but not all studies reported on the relation between the results of the tests of cognitive function and the ERPs. Supplement 6 provides an overview of the individual tests used in the various studies.

Table 3. An overview of the versions of the SDMT and PASAT as described by the authors in case such details were provided.

SDMT		PASAT	
Reference	**Version**	**Reference**	**Version**
Artemiadis et al., 2018	Oral SDMT	Covey et al., 2017	2-second PASAT 3-second PASAT
Bissonnette et al., 2023	Oral SDMT	Gerschlager et al., 2000	Easy PASAT Hard PASAT
Chinnadurai et al., 2016	Oral 60 second modified SDMT Oral 180 second modified SDMT	Kocer et al., 2008	
Covey et al., 2017	Rao adaptations	López-Góngora et al., 2015	2-second PASAT 3-second PASAT
López-Góngora et al., 2015	Oral SDMT	Magnié et al., 2007	3-second PASAT
Paolicelli et al., 2021		Paolicelli et al., 2021	3-second PASAT 5-second PASAT
Pokryszko-Dragan et al., 2016a		Pokryszko-Dragan et al., 2016a	3-second PASAT
Pokryszko-Dragan et al., 2016b		Pokryszko-Dragan et al., 2016b	
Vázquez-Marrufo et al., 2008		Uysal et al., 2014	3-second PASAT
Vázquez-Marrufo et al., 2014		Vázquez-Marrufo et al., 2014	3-second PASAT
Vázquez-Marrufo et al., 2019		Vázquez-Marrufo et al., 2019	3-second PASAT
Waliszewska-Prosół et al., 2018			

For the SMDT, two studies (Chinnadurai et al., 2016; Artemiadis et al., 2018) found a difference between controls and PwMS, four studies (López-Góngora et al., 2015; Covey et al., 2017; Waliszewska-Prosół et al., 2018; Paolicelli et al., 2021; Bissonnette et al., 2023) found no difference, and five (Vázquez-Marrufo et al., 2008; Vázquez-Marrufo et al., 2014; Pokryszko-Dragan et al., 2016a; Pokryszko-Dragan et al., 2016b; Vázquez-Marrufo et al., 2019) did not report on the comparison between controls and PwMS. Nine of these studies (Vázquez-Marrufo et al., 2008; Vázquez-Marrufo et al., 2014; López-Góngora et al., 2015; Chinnadurai et al., 2016; Pokryszko-Dragan et al., 2016b; Covey et al., 2017; Waliszewska-Prosół et al., 2018; Artemiadis et al., 2018) found differences in their ERP measurements between controls and PwMS. Of these, two (Vázquez-Marrufo et al., 2014; Artemiadis et al., 2018) reported a correlation between ERP measurements and the SDMT results with correlations between worse SDMT performance and CNV amplitude, longer P300 latency, and reduced P300 amplitude. One study (Covey et al., 2017) reported no significant group differences in the SDMT between controls and PwMS, but did find more variability in PwMS, as well as different relationships between ERPs and SDMT score, with P300 latency on a 2-back task being predictive of SDMT performance in controls. Two studies (López-Góngora et al., 2015; Waliszewska-Prosół et al., 2018) found differences in ERP but no corresponding difference in SDMT.

For the PASAT, three studies (Gerschlager et al., 2000; Uysal et al., 2014; Paolicelli et al., 2021) found a difference between controls and PwMS, three studies (Kocer et al., 2008; López-Góngora et al., 2015; Covey et al., 2017) found no difference, and five (Magnié et al., 2007; Vázquez-Marrufo et al., 2014; Pokryszko-Dragan et al., 2016a; Pokryszko-Dragan et al., 2016b; Vázquez-Marrufo et al., 2019) did not report comparisons between controls and PwMS. Eight studies (Magnié et al., 2007; Kocer et al., 2008; Uysal et al., 2014; Vázquez-Marrufo et al., 2014; López-Góngora et al., 2015; Pokryszko-Dragan et al., 2016a; Pokryszko-Dragan et al., 2016b; Covey et al., 2017) reported

differences in ERPs between controls and PwMS. Two studies (Magnié et al., 2007; Uysal et al., 2014) found a correlation between PASAT and ERP: one study reported a correlation between PASAT and the P300 amplitude in a visual oddball experiment while the other reported that an increase CNV amplitude correlated with a better PASAT performance. One study (Covey et al., 2017) reported P300 amplitude on the 2-back task in electrode Pz was predictive of performance of PwMS in the 2- and 3-second PASAT. Additionally, reduced P100 amplitude in electrode Pz on the 2-back task was associated with better performance by PwMS on the 2-second PASAT.

7.1.4 Discussion

This study reviewed the existing literature to determine which stimulus modality, experimental paradigm, scalp electrode, ERP component, and parameter were most discriminatory between PwMS and controls in cognitive ERP experiments, with a view to trying to standardise future experiments using this technique. Overall, the largest pooled effect sizes were seen for visual stimuli, the Posner and ANT paradigms, and electrode Cz. The ERN and P300 were the best-performing ERP components, and latency was more discriminatory than amplitude.

Collapsing across the various studies, visual stimuli produced larger pooled effect sizes than auditory stimuli, regardless of whether they were expressed in terms of latency or amplitude. This difference was especially noticeable for amplitude measurements where visual stimuli produced an overall pooled effect size of 0.52 while auditory stimuli resulted in a pooled effect size of only 0.15. Interestingly, there were more measurements from experiments using auditory stimuli than from experiments using visual stimuli, mostly because of the large number of auditory oddball experiments. Given the more discriminatory results produced by the experiments using visual stimuli, this seems to be the better modality to use in future experiments.

An important consideration is that some of the difference found between visual and auditory stimuli may relate to fact that visual disturbance, particularly in the form of optic neuritis (Chan, 2002), is common in PwMS. Involvement of the auditory pathways is relatively rare. Indeed, in some of the studies reviewed here, it was observed that the latency of N100, a component associated with attention, was prolonged with visual stimuli (Vázquez-Marrufo et al., 2014). Research elsewhere (Covey et al., 2021) has shown that there may be an impairment of integration of visual information. To clarify this issue, future research will need to look into the correlation between visual function and the ERP.

Analysis of experimental paradigms revealed that oddball paradigms were by far the most frequently used paradigm, particularly as an auditory version. However, this paradigm did not produce the largest pooled effect sizes, either for latency or amplitude. Other paradigms, like the Posner paradigm (measuring latency) or the ANTs (measuring amplitude), produced much larger effect sizes than the oddball paradigm. While this finding could relate to the issue of visual involvement in PwMS

discussed above, the lack of available information about participants' visual function made it impossible to comment on this issue here. The ANT results of g = 1.57 (CI = 1.26, 1.88]) would need further investigation, as these are generally only observed for very noticeable differences between the groups that are being compared (Hilgard, 2021). For example, the difference in height between men and women born in Spain in 1980 produces a Hedges' g of 1.87, based on data reported by (Garcia & Quintana-Domeque, 2007).

The vast majority of experiments reported results from the three midline electrodes Fz, Cz, and Pz, with only few reporting results in other electrodes. This observation held true across all experiments regardless of the paradigm used. In this regard, the positioning of a midline electrode does not relate well to specific underlying brain areas and, therefore, it is difficult to draw any 'anatomical' functional conclusions from the results provided by these electrodes. However, placement of midline electrodes is relatively easy, making them potentially more convenient for future clinical use.

Overall, latency measurements produced larger effect sizes for all electrodes than amplitude measurements, Cz produced the largest pooled effect sizes for both latency and amplitude measurements, and Pz produced larger pooled effect sizes than Fz. Based on these results it would be worth exploring central and posterior electrodes, possibly using regions of interest (ROIs) instead of single electrodes or independent components, if high-density electrode setups with 128 or more electrodes are available.

The exact component that was reported by each of the studies appeared to depend on which experimental paradigm was employed. For example, P300 and N200 were most frequently measured during oddball experiments. In fact, the overall effect sizes for latency at both P150 and N400 were larger than those at P300, and P600 performed better when measuring amplitude. These larger effect sizes were, however, based on a relatively small number of measurements, often from different sections of the same experiment. It was therefore not possible to determine conclusively which EEG component was best at distinguishing PwMS from controls.

In general, latency measurements produced the largest effect sizes, though the largest effect size overall was seen for amplitude in the ANTs. However, this latter result was derived from only four individual experiments, compared with 85 for oddball experiments, making comparison difficult. The same was true for many of the other measurements, meaning that further investigation is required before drawing a definitive conclusion.

The direction of effects was not the same in each study and for each component. Unfortunately, the results that produced these differences come from very different studies, hindering comparison. For example, the two studies producing shorter latencies for PwMS than controls measured P300 and P600, respectively, from different electrodes and using visual stimuli in one experiment and auditory stimuli in the other. For the amplitude measurements that were larger for PwMS than controls, the study

providing most of the results did not specify which electrodes were used. This heterogeneity in methods and the fact information for comparison is missing means the differences can be noted but not further analysed.

This study has the following two main limitations. First, the literature consisted of reports of many different experimental arrangements in each individual study, meaning that the numbers pertinent to any one group were often very small. The way this was dealt with here was to collapse across paradigms, electrodes, etc. to try to generate an overall 'summary' picture. This means that the heterogeneity for any comparison was very large. It is likely that more meaningful (and clinically useful) results would be obtained if researchers adopted a more standardised methodology in the future. To assist with this, we have provided recommendations for future research in Table 3.

Table 3. An overview of recommendations for future studies.

1) Visual stimuli are preferable to auditory stimuli
2) Paradigms other than the oddball paradigm are likely to be more discriminatory
3) Sufficiently large cohorts are required to generate meaningful results
4) Reporting of measured latencies and amplitude in ms and μV, respectively, as such measurements permit easier comparison
5) The precise subtype of MS being studied should be clarified

A second limitation of our research is that the results presented here treat PwMS as a single entity. 'PwMS' actually comprises several different clinical subtypes such as RRMS, PPMS and SPMS, and clinically or radiologically isolated syndromes. While some information about subtype was available, the numbers were too small to allow meaningful comparison. With this in mind, however, it is important for future investigators to clarify the precise subgroup of MS being studied.

Another important consideration is that there may have been an effect arising from technical issues such as the type of electrode used, the method of electrode application, or the location of the reference electrode. Granted the large variation in the information provided by the various studies (where this was available) this issue could not be commented on here. However, it is worth pointing out that it would be important to assess this in future studies.

The review presented here synthesised results from research over a period of more than 30 years, from 1992 (Filippi et al. 1992; Giesser et al. 1992; Honig et al. 1992) until April 2023. (Bissonnette et al. 2023) In doing so, this review provides an overview of the most prevalent paradigms and methods, similar to reviews that synthesized results in other neurodegenerative diseases (Seer et

al., 2016), and, by doing so and in combination with other reviews, can help inform decision making about experimental setups not only in MS but also other diseases.

More broadly, research into cognitive performance of PwMS, would benefit from a more uniform and shared approach to data collection, analysis, and reporting methods, and results should be reported for the different subtypes of PwMS. In particular, equipping research into ERPs in the investigation of cognitive performance of people with MS with more standardised methods, would facilitate the potential of ERPs to be developed into a clinical tool in MS. Such more uniform and shared approaches have been successful in driving transformational research in other medical informatics contexts (Chapman et al. 2011, Huang & Lu 2016) and are hence likely to benefit studying clinical markers of MS as well. If this can be accomplished, ERPs may become useful in the management of MS, by providing a more objective measure of cognitive function. This would potentially assist PwMS, who are often uncertain about what the future holds for them (Hossain et al., 2022), by providing greater clarity on their individual rate of disease progression, as well as by providing a useful measure of the effectiveness (or not) of disease-modifying medication (Capone et al., 2020; Gerschlager et al., 2000). At this point, however, considerable further work is required before the ERP can be considered clinically useful.

7.1.5 Conclusion

This study has synthesised a substantial body of existing work with a view to assisting future studies to use the most discriminatory experimental setups, i.e. those that distinguish best between PwMS and controls. Our results show visual stimuli are preferable over auditory stimuli, and that paradigms other than the oddball paradigm (which has been most frequently used in the past) are likely to provide better results. To derive more conclusive outcomes, larger cohorts are needed, and any results should be divided into specific MS subtypes and reported in terms of raw numbers. In summary, future work looking at the role of ERPs in PwMS would benefit from a more systematic and uniform approach to collection, analysis, and reporting methods. However, the results show that, with more standardised methods, ERPs have significant potential to contribute to the clinical assessment of PwMS, possibly as an objective biomarker of cognitive decline in the disease. This would provide additional information to that provided by the EDSS, generate a more holistic assessment of patients and, therefore, offer an improved health experience to PwMS.

7.2 A comparison to event-related potential research in Parkinson's disease

Just as the use of ERPs in the investigation of cognitive performance in people with MS was investigated in the review reported on above, other researchers have carried out similar reviews into the use of ERPs in PD. A review into the study of cognition in PD using ERPs was carried out by Seer et al. (2016), focusing on four components: the mismatch negativity, P300, N200, and the error-related negativity. This, of course, has an influence on the paradigms included, as for example P300 is mostly studied using oddball paradigms. This is different to the approach in the MS review, as the components to focus on had not been determined a priori in that review. Another difference is that Seer et al. did not report statistical meta-analyses as we did, calculating effect sizes; they only reported differences in amplitude and latency and the direction of the differences.

The largest group of experiments reported on by Seer et al. (2016) consisted of studies utilising oddball experiments. They included 65 studies reporting on 74 oddball experiments. 51 of these used auditory stimuli, 17 used visual stimuli, two used somatosensory stimuli, three used mixed stimuli, and for one experiment it was unclear which stimuli where used. The number of controls ranged from eight to 55 with an average of approximately 21, while the number of people with PD (PwPD) ranged from 6 to 61 with an approximate average of 22. 61 experiments used a standard oddball paradigm, 12 used a 3-stimulus paradigm, and one experiment employed a mix. The percentage of oddball stimuli varied from 7% to 33% of total stimuli.

Comparing the findings of Seer et al. (2016) to the MS review, there are some striking similarities. Even in the use of a well-established task like the oddball paradigm there can be large differences between experimental setups. In the case of the MS review, most studies would combine people with different subtypes of MS and report findings as one MS group. While PD does not have demarcated subtypes like MS, there are two differences that require special attention. First, studies can be performed in PwPD who are either in an ON state or an OFF state, meaning that they are on dopaminergic medication, or are not on medication. For the latter, PwPD either have never been medicated or underwent a washout period varying from 12 hours to 14 days. Second, differences have been found between PwPD with dementia and without dementia.

Seer et al. (2016) address both of the issues mentioned in the previous paragraph. The studies that looked into the effects of dopaminergic medication on P300 showed inconsistent latency findings, and Seer et al. (2016) raised several methodological issues. Given that one of the issues was small sample size of individual studies, calculation of a pooled effect size might have been desirable to provide more clarity. For now, the conclusion is that P300 does not crucially depend on dopaminergic pathways based on the inconclusive findings.

The research concerning the second issue, namely the differences between PwPD with and without dementia is more conclusive. Amplitude findings showed that of the 54 studies that quantified amplitude, 39 did not observe a significant difference between PwPD and controls. For latency measurements, 39 or 53% of experiments produced longer latencies for PwPD, with only one reporting short latencies. Digging further into the latency measurements, Seer et al. (2016) found that when excluding studies into dementia in PD only 19 out of 50 studies produced latency differences, while 10 out of 14 studies that did not distinguish between PwPD with and without dementia found latency differences, and all studies that only included PwPD with dementia reported latency prolongation in the patient group. They concluded there is reliable evidence for a prolongation of P300 latency in PwPD with dementia, but not for those without dementia.

The finding by Seer et al. (2016) is of interest when taking into account the observation in the MS review that there might be differences between people with different subtypes of MS that need to be investigated, which is not possible at the moment since results are generally not reported by MS subtype. It also adds one more dimension to the possible future research that can follow the results presented in Chapter 4. If differences exist between males and females with PD, and between PwPD with and without dementia, an interesting avenue of research would be the interaction of sex, disease progression, and cognitive decline. As the MS review showed that studies into MS did not distinguish between sexes either, such a question is also of relevance in MS research.

Seer et al. (2016) concluded their review in a similar manner to the MS review. They noted the heterogeneity of stimuli, recording procedures, and analysis approaches, and suggested standardisation of methods. They also noted the small sample size of many studies, which hampers statistical power and makes it difficult to obtain reliable results. They argued that larger, multi-site studies are necessary for ERPs in PD to be developed into a tool that can be used for the prediction of individual differences. Given the findings in the MS review and those reported in Chapter 4, I agree.

7.3 Chapter summary and conclusion

In this chapter a systematic review of the use of ERPs in the investigation of cognitive performance in people with MS was first presented and then compared to a review carried out by Seer et al. (2016) into ERPs and cognition in PD. While PD and MS are very different diseases, both can take a very long time to accurately diagnose, and the tools used to assess disease status and predict progression, such as the MDS-UPDRS in PD and the EDSS in MS, suffer from similar deficiencies, namely variability in results and an inability to provide accurate predictions of disease progression.

The MS review and the review performed by Seer et al. (2016) differed in methodology. The MS review did not focus on specific components, while the PD review focused on four components, being the mismatch negativity (MMN), P300, N200, and the ERN. In the MS review, effect sizes were

calculated for each experiment and pooled effect sizes for paradigms, stimulus modalities, electrodes, etc., while Seer et al. (2016) reported differences between PwPD and controls, or PwPD at different time points, stating the direction of the difference.

However, both the conclusions of the MS review and PD review were similar. To develop ERPs into a useful clinical tool that can accurately describe individual differences, studies with larger cohorts using standardised methods are necessary. One of the recommendations of the MS review was to report findings by subtypes, instead of merging people with different MS subtypes into one MS group. Seer et al. (2016) found that P300 latency differences were a marker of dementia in PD and that results in PwPD without dementia varied. This is interesting, as it has been suggested there are subtypes of PD, in which either motor or cognitive features dominate. In that light it might be necessary to divide PwPD in ERP studies similarly to the groupings suggested in the MS review in MS studies.

The results by Seer et al. (2016) and the MS results show that heterogeneity of methods and small sample sizes in ERP experiments are problems across neurodegenerative diseases. The results in both reviews show ERPs show promise to be developed into a clinically useful tool for diagnosis and prognostication of both MS and PD. The experimental paradigms, preprocessing, and quantification of ERP components in both diseases can be largely standardised as they are not dependent on the disease itself. If this can be accomplished and experiments can be run with larger cohorts enabling division by disease subtype in MS, and people with and without dementia in PD, further research should be able to find biomarkers of disease status and disease progression in both diseases.

Machine learning (ML) and data science might be valuable in standardising methods, as data quality, data curation, and standardisation are concepts that are generally necessary in these fields to come to satisfactory results. An example of the possibilities ML and DS might offer was mentioned several times throughout this thesis: the ICLabel implementation (Pion-Tonachini, Kreutz-Delgado, & Makeig, 2019) used to removed non-neural signals in Chapters 3, 4, 5, and 6 consists of an *artificial neural network* (ANN) trained on a curated dataset. Consequently, ML and DS might be valuable at other points in the diagnosis and prognosis pipeline when using RSEEG and ERP data.

Chapter 8: Conclusion

In this interdisciplinary PhD thesis predominantly in computer science, as well as clinical neuroscience, five studies were reported that look at the effects of varying preprocessing methods, sample randomisation methods, the moment of feature selection, *machine learning* (ML) classification algorithms, and participant characteristics on performance metrics in classification tasks of *people with Parkinson's disease* (PwPD) and controls using *electroencephalography* (EEG) data. Specifically, four research questions were answered:

1) What are the effects of different preprocessing methods on classification outcomes?

2) How do classification outcomes differ when using single or averaged epochs or electrodes, and do they differ between data obtained while participants had their *eyes open* (EO) or *eyes closed* (EC)?

3) What are the effects of feature selection at different points in the classification pipeline and different methods of sample randomisation?

4) How do participant characteristics, such as age, sex, and *Movement Disorder Society-Unified Parkinson's disease rating Scale* (MDS-UPDRS) score, influence classification outcomes?

The aim was to provide insight into the precise effects of using different methods and, ultimately, to make recommendations for processing data when using EEG to develop diagnostic and prognostic models for *Parkinson's disease* (PD) in the future, with the possible outcome of developing methods that are the gold standard or part of the gold standard. These studies investigated methods of data cleaning, outlier removal, featurisation and feature selection, as well as clarifying the effects of demographics such as MDS-UPDRS score and sex on classification metrics. The key outcomes and recommendations were:

1) A full preprocessing pipeline that includes thorough artifact rejection is necessary to ensure classification is based on neural signals and not artifacts.

2) EC data are preferable over EO data due to the latter containing more non-neural signals and most data being lost in the preprocessing stage. *Regions-of-interest* (ROIs) formed by combining neighbouring electrodes led to better classification outcomes compared to single electrodes.

3) Feature selection outside of *cross-validation* (CV), leads, as expected, to overly optimistic results, especially when using a lower ratio of samples-to-features. An even larger effect is observed when samples from the same participant are present in both the train and test data.

4) Both splitting data by MDS-UPDRS score and sex led to improved classification results, indicating that diagnostic models should be divided for men and women, and that differences between people with higher and lower MDS-UPDRS scores exist that might be leveraged for prognostic purposes.

In addition, a literature review into the use of *event-related potentials* (ERPs) to investigate cognitive tasks in *people with Multiple Sclerosis* (PwMS) was included to begin to expand the conclusions to neurodegenerative diseases in general. In fact, the initial plan for this PhD thesis was to include experiments investigating both PwPD and PwMS, but the COVID-19 pandemic made data collection from PwMS impossible. The literature review was included here with a view to compare and contrast the results of EEG research in *Multiple Sclerosis* (MS) with those of EEG research in PD.

The key findings of the literature review into MS were that a wide variety of paradigms are used in the investigation of cognition in MS, and that this makes comparison results difficult. To develop ERPs into a useful clinical tool, research would benefit from a more unified approach and the following recommendations were made:

1) Visual stimuli are preferable over auditory stimuli.
2) Paradigms other than the oddball paradigm are likely to be more discriminatory and should be investigated.
3) Sufficiently large cohorts are required to generate meaningful results.
4) Reporting of raw latency and amplitude measurements permits easier comparison.
5) The precise subtype of MS being studied should be clarified.

Comparing research in MS to PD, it was found that the same issues plague PD research, namely heterogeneity of stimuli, and analysis approaches, which could be alleviated by standardisation of methods. Moreover, cohorts were often small, hampering statistical power. The recommendations for MS studies therefore also hold for PD studies.

8.1 Thesis summary

As mentioned in the introduction, the main objective of the work presented here was to investigate methods for the use of EEG data in the classification of people with neurodegenerative diseases and controls using ML, resulting in the proposal of an end-to-end processing pipeline. After introducing the thesis topic in Chapter 1, the background and methods that would be used in this thesis were introduced in Chapter 2. Both PD and MS were discussed, as well as EEG as a method of recording brain activity. Methods of preprocessing EEG data as recommended by the developers of EEGLAB were introduced, ML methods that were going to be utilised, such as *Support Vector Machines* (SVMs), *Random Forest*

Classifiers (RFCs), and *Gradient Boosting Classifiers* (GBCs), were discussed, and statistical analysis using significance tests and effect sizes were explored.

In Chapter 3, the question of the effects of different preprocessing steps was answered by performing four different levels of preprocessing on three separate datasets. Different studies into the classification of PwPD and controls using resting-state EEG (RSEEG) and ML had used varying levels of preprocessing (Maitín, García-Tejedor, & Romero Muñoz, 2020; Maitín, Romero Muñoz, & García-Tejedor, 2022), but it was unclear how these varying levels of preprocessing influenced results. By keeping all other steps the same and using four different levels of preprocessing, ranging from using rereferenced data to fully preprocessed data, the results showed it was necessary to use the latter, and that muscle artifact positively biased results (Vlieger et al., 2023). This is undesirable when the goal is to classify on neural activity alone, which was the goal of this thesis.

In Chapter 4, the effects of averaged and single epochs and electrodes were investigated. Once again, methods used in previous studies varied widely, which makes comparison of results difficult. Underpinning the findings from Chapter 3, namely deploying a full preprocessing pipeline that includes artifact rejection (Vlieger et al., 2023a), four combinations of average and single electrodes and average and single epochs were explored to create features from datasets with EC and EO data from three separate research groups. The results indicated that using EO data leads to a loss of data that makes EO data unsuitable for clinical purposes for now, and that ROIs created by combining neighbouring electrodes outperformed the use of single electrodes (Vlieger et al., 2023b).

In Chapter 5, the effects of data leakage on experimental outcomes were elaborated on. Data leakage is commonly found in the literature, concerning RSEEG as well as other medical fields, that feature selection is performed outside of the CV process (Yuvaraj, Rajendra Acharya, & Hagiwara, 2018; Betrouni et al., 2019; Demircioğlu, 2021; Lee et al., 2021), which has been shown to cause data leakage (Kapoor & Narayanan, 2023). Additionally, uncertainty exists about how samples were randomised, with some randomisation methods leading to data leakage (Tampu, Eklund, & Haj-Hosseini, 2022; Kapoor & Narayanan, 2023). In this chapter, both cases were investigated using three datasets, showing that both can cause increased evaluation metrics, with the effects for incorrect randomisation being larger when the sample-to-feature ratio was lower.

In Chapter 6, datasets were created based on sex and MDS-UPDRS score to investigate their effects on classification outcomes. Due to the difficulty of collecting RSEEG data, many datasets are small (Carr, 2017; Suuronen et al., 2023), and dividing participants by specific characteristics is often nearly impossible because this would lead to insufficient data to train on. This limited availability of data is unfortunate, as the literature shows different RSEEG patterns for males and females and differing patterns for people at different stages of PD. In this chapter, combining three datasets made it possible to divide people by sex and MDS-UPDRS score, with experiments showing splitting males and females

improves classification for both. Additionally, PwPD with low MDS-UPDRS scores were distinguishable from those with higher MDS-UPDRS scores, giving credence to the idea RSEEG could be used for prognostication.

In Chapter 7, a systematic literature review into the use of ERPs in the investigation of cognition in MS was presented. Five questions about classification of PwMS and controls were answered, being (1) which stimulus modality (i.e., visual or auditory) is most discriminatory, (2) which experimental paradigm produces the most discriminatory results, (3) which individual electrodes generate the best results, (4) which electrophysiological component(s) of the ERP are most discriminatory and, (5) whether amplitude or latency is a better measure. The results imply that visual stimuli are preferable over auditory stimuli, that oddball paradigms were the most prevalent but did not produce the largest effect sizes, and that electrode Cz produced the largest effect size out of electrodes Fz, Cz, and Pz. Results also indicated that methods were very heterogeneous, that results were often presented for PwMS instead of distinguishing between MS subtypes, and that cohort sizes varied widely.

The results of the review were compared and contrasted with those from a review of cognitive ERPs in PD by Seer et al. (2016), finding similar heterogeneity in methods, the majority of experiments being oddball paradigms. While MS has distinct subtypes, PD does not. That being said, it has been suggested there might be difference between PwPD with dementia and those without. Seer et al. found that there is evidence there is a prolongation of P300 latency in those with dementia, while no such prolongation seemed to occur in PwPD without dementia. Similar to the findings in the MS review, they argued for more standardised experimental methods and larger cohorts to make it possible to develop ERPs into a clinical tool.

8.2 Thesis contributions

The main contributions in this thesis can broadly be divided into four categories, namely the effects of preprocessing steps, the effects of different data handling and feature selection methods, the importance of participant characteristics, and the broader context of the findings reported on in this thesis for neurodegenerative diseases in general. These contributions will be elaborated on in this section and placed within the context of the existing literature.

8.2.1 Preprocessing pipeline

Whether studying ERPs or RSEEG, the first step researchers generally take is to preprocess their data (Miyakoshi, n.d; Luck & Kappenman, 2011; Cohen, 2014; Luck, 2014). The scalp electrodes used in EEG research do not distinguish between muscle activity, heart beats, neural activity, or activity from any other source, and will simply record everything. To remove everything but the neural signals from an EEG recording is, therefore, a field in and of itself with new methods being proposed, tested, and implemented regularly.

Several preprocessing packages exist (Oostenveld et al., 2011; Gramfort et al., 2014), with the MATLAB package EEGLAB (Delorme & Makeig, 2004) being widely used. Makoto Miyakoshi, an EEG specialist and researcher who worked on EEGLAB produced a widely used unofficial manual, called "Makato's preprocessing pipeline" (Miyakoshi, n.d), with explanations for each step. Roughly, the pipeline consists of the following parts: rereferencing, filtering, data cleaning through electrode removal and interpolation, and artifact rejection. How each step is carried out and which parameters are used depends on the particular problem that is being researched, and studies can be found that investigate each separate step.

One of the first steps, if not the first step, is the rereferencing of recordings. Not all EEG data are recorded with the same reference electrode, so it is often necessary to rereference data when comparing data from different sources. This rereferencing has been shown to influence data, and it is possible to get different analysis outcomes depending on which reference is used (Ríos-Herrera et al., 2019; Yao et al., 2019; Zhang et al., 2020). Similarly, what kind of filters are use and what the parameters are, such as the order of the filter and the filter frequency, have also been shown to influence results (Karpiel et al., 2021). Lastly, methods to handle data that is contaminated with non-neural signals vary widely. The most traditional manner to do this is to visually inspect the data and remove parts of the data and sometimes electrodes that are deemed to contain non-neural signals (Kaya, 2019). This includes removal of epochs by visually inspecting them, and visually inspecting *independent components* (ICs) resulting from the *independent component analysis* (ICA) procedure to determine their origin. Nowadays it has become more common to use automated methods, but which method is chosen, and the combination of methods can influence results as well (Nolan, Whelan, & Reilly, 2010; Mognon et al., 2011; Winkler, Haufe, & Tangermann, 2011).

While much has been written about the effects different preprocessing methods have on the analysis of EEG data, less has been published on how the differences in methods would affect the outcomes in ML classification experiments of PwPD and controls. The research presented in Chapter 3 looked systematically at the effects of preprocessing methods by dividing the complete preprocessing pipeline into four Levels (rereferencing data, filtering data, using the full preprocessing pipeline with muscle artifact left in, and the full preprocessing pipeline). The choice to leave muscle artifact in at Level 3 was a choice specific to research into PD.

In 2007, Whitham et al. (2007) published a study that showed that measurements of frequencies over 20 Hz in healthy participants were progressively more contaminated with muscle artifact. Additionally, the study showed that there was a difference in raw EEG data and cleaned EEG data collected a second before trials of a grasping task, and their analysis showed the difference to be attributable to muscle artifact. Based on these findings, it was decided to investigate what the contribution of muscle artifact was on the RSEEG classification tasks that made up our experiments.

As the results showed, muscle artifact does indeed have the capacity to increase evaluation results and should, therefore, be removed from the data if the aim is to classify on neural signal alone (Vlieger et al., 2023).

To summarise, the main contributions of the studies included in Chapter 3 were that it made clear what the effects of different levels of preprocessing were on classification outcomes. Because methods found in the literature are so heterogeneous, the results here help put other studies in a clearer perspective. Additionally, the results showed the necessity of using a full preprocessing pipeline that involves thorough artifact removal. In this way, the research also provided a recommendation to researchers planning to perform ML classification tasks of PwPD and controls using RSEEG data.

8.2.2 Data handling and feature selection

After preprocessing data, choices must be made about what to do with the time-series and electrodes, which features to extract, and at what point in the classification pipeline this should happen. EEG datasets often consists of two recordings per participant, one EC and one EO recording. This complicates matters further, as these recordings are not the same and each of the aforementioned choices can have a different effect on the recordings, and the resultant classification results. Little research has focused on this problem, especially in a manner where each particular choice, like whether to use EO or EC data (Suuronen et al., 2023) being one study to do so), has been systematically reviewed.

The first step after preprocessing is often to decide how to deal with the electrodes and the time-series. There are generally two ways in which electrodes and epochs are treated in the literature: either some form of averaging is used to improve the *signal-to-noise ratio* (SNR; Congedo, Gouy-Pailler, & Jutten, 2008; Hajra et al., 2021), or each electrode or epoch is treated separately. Averaging of epochs leads to a higher SNR because noise is generally randomly distributed and will therefore cancel out when averaged, and averaging nearby electrodes similarly increases the SNR as nearby electrodes tend to record from similar neural populations and the noise is random. The disadvantage of the averaging method is that this reduces the amount of data available for training, while the use of single electrodes and epochs has the disadvantage that outliers have a larger effect on the training process.

Once it is decided how electrodes and epochs are treated, features are extracted. In the majority of studies, those are time-frequency features. It is here that some previous studies have made choices that can lead to overly optimistic results. Namely, both feature selection outside of CV and randomisation of samples from the same participant through the train and test set lead to data leakage, meaning the trained model has information about the test set it would normally not possess. For example, Demircioğlu (2021) showed that, using radiomics data, performing feature selection outside of the CV process, produced results that were overly optimistic and did not hold up when the feature selection was performed during the CV process. This effect was larger when there were less samples

per feature than when more samples were available. As the datasets used in PD classification experiments were often small, the effect cannot be underestimated.

To first clarify the data handling choices regarding electrodes and epochs in EC and EO data, a study was presented in Chapter 4 that systematically investigated the effects of using single and averaged electrodes and epochs from EO and EC data when classifying PwPD and controls. This was done by creating four Methods, namely: (1) single electrodes and single epochs; (2) single electrodes and averaged epochs; (3) ROIs, created by averaging electrodes, and single epochs; and, (4) ROIs and averaged epochs for EC and EO data from three datasets.

In Chapter 5, a study into the effects of different sample randomisation methods and the importance of feature selection in the CV process was presented. This was done by performing feature selection as part of the CV process and outside of the CV process, and by randomising samples through the train and test set, and by making sure all samples were in either set. Because preprocessing methods often differ between studies, this was done for all four Levels of preprocessing first presented earlier in Chapter 3.

There were some differences between EO and EC results that deserved attention. First, while it produced the best evaluation results, EO data suffered from data loss, with an inordinate number of samples being lost when using single epochs and loss of all data from specific participants when using averaged epochs. This is not the first time this has been reported in the literature, and it would require additional research to solve the problem of data loss for EO data to be clinically useful. Second, EO data often produced an imbalance between sensitivity and specificity, meaning almost all participants got classified as having PD. While the overall accuracy might be around 60%, having the vast majority of controls being classified as having PD is of course unacceptable when the intention is to develop a diagnostic model.

Looking at the results of the experiments in Chapter 4, in general, ROIs outperformed single electrodes for EO and EC data. As explained in Section 2.2.2, the rationale of averaging electrodes is that neighbouring electrodes will record from similar populations of neurons and that the noise is randomly distributed, so averaging electrodes would lead to clearer neural signals. In the case of the experiments in Chapter 4, six ROIs were created from 58 electrodes and 14 features extracted from each ROI. A benefit of this approach, apart from the improved SNR, is that the total of features is 84, 14 features times 6 ROIs, instead of 812 when using 58 electrodes and 14 features. This is not an issue with the rather small datasets used here, but could significantly speed up computation with larger datasets. While ROIs definitely were a better option than single electrodes, there seemed not to be such a clear-cut difference between using single or averaged epochs.

The results in Chapter 4 were lower than those generally found in the literature. There are three potential reasons for these differences. First, as the results presented in Chapter 3 showed, the level of

preprocessing can have a large effect on the classification outcomes. Preprocessing steps in the literature vary widely, and so do the resultant classification outcomes. In the experiments in Chapter 4, all data was fully preprocessed, which can explain the difference between these results and the results reported in the literature when no preprocessing or preprocessing that did not include artifact rejection was utilised.

Secondly, the point and the manner in which features were selected in some previous studies may have led to overly optimistic results. As mentioned, Demircioğlu (2021) showed this for radiomics data, but it was unclear what the effects would be for RSEEG data. Additionally, in our experiments, the IDs of participants were randomised and divided over the train, validation, and test sets, but as has been mentioned elsewhere in the literature, it is not always clear if other studies were carried out in a similar manner (Tampu, Eklund, & Haj-Hosseini, 2022; Kapoor & Narayanan, 2023). If not, and all samples were randomised, this would constitute data leakage, and results would be increased.

The effects of data leakage were discussed in Chapter 5. The two methods of data leakage were applied to data from all three aforementioned separate datasets at the four different Levels of preprocessing described in Chapter 3. The results showed that, indeed feature selection outside of the CV process increases evaluation metrics, which is attenuated when there are more samples per feature. Randomising samples produced even further increased evaluation results and cannot be alleviated in anyway, so should be always avoided.

The contribution of the research presented here on data handling was that it made clear the differences between EO and EC data, as well as between the choices that can be made about electrode and epoch handling. By using different combinations of averaged and single electrodes and epochs for both EO and EC data, the experiments provide evidence of preferable combinations of methods, in this case being ROIs and single epochs, as well as clarifying the usefulness of EO and EC data. By doing this, it also places previous research in a clearer context, and it makes comparison of results less difficult. Similarly, the results reported on in Chapter 5 provide evidence of the importance of the avoidance of data leakage and place the results of previous studies that might have inadvertently introduced data leakage in their classification pipeline in a clearer perspective.

8.2.3 The importance of participant characteristics

Due to the difficulty of collecting large quantities of EEG data, it is often unclear which participant characteristics affect experimental outcomes. Data collection can be time-intensive, inconvenient due to application of gel to the head and the necessity to often go to a hospital or research facility, and, in some cases, boring. This all together leads to individual EEG datasets generally being rather small, and an inability to split data by characteristics such as sex or age.

The aforementioned problems with data collection are compounded when researching neurodegenerative diseases. While PD is one of the most common neurological diseases (Nutt & Wooten, 2005), there are only a limited number of PwPD and most of them will not be in the vicinity of where RSEEG research is taking place. This means that researchers cannot be very picky about which PwPD to include in their research, as that would even further diminish the number of participants. EEG datasets are therefore made up of PwPD that are of different ages, different sexes, and have different MDS-UPDRS scores.

The differences between participants pose potential problems. For example, differences between PwPD at different stages of their disease have been observed in their RSEEG recordings (Soikkeli et al., 1991; Morita et al., 2009; Benz et al., 2014). If one would devise a classification task of PwPD and controls, but some PwPD have MDS-UPDRS scores of 20 and others have MDS-UPDRS scores of 60, meaning they have progressed much further, there is reason to believe this could cause problems. Additionally, differences in RSEEG signals have been reported between males and females. It is known there are differences between how PD develops in males and females (Cerri, Mus, & Blandini, 2019), but that is not the only possible origin of these discrepancies. For example, the thickness and size of the skull are also contributing factors (Li et al., 2007).

In Chapter 6, three separate datasets were combined that made it possible to create sufficiently large groups of men and women, as well as PwPD with different MDS-UPDRS scores, to run classification experiments. Using the optimal methods found in Chapters 3, 4, and 5, a female-only experiment, a male-only experiment, a low MDS-UPDRS-only experiment, and a moderate MDS-UPDRS-only experiment were performed and compared to an experiment that used the full dataset. The results implied that splitting data by sex and MDS-UPDRS score improved classification metrics. The reliable classification in the female vs male and low vs moderate MDS-UPDSRS experiments provided evidence that there are differences between PwPD of different sexes, and that differences between different MDS-UPDRS scores exist.

The contribution of the research reported on in Chapter 6 is that, by combining three separate datasets, an investigation into the effects of sex and disease progression, expressed in the MDS-UPDRS score, was possible, resulting in confirmation that creating datasets of only males and only females increased classification outcomes for both groups. To provide more evidence that there are differences between males and females with PD that cannot be ignored, an experiment was performed that showed that males with PD and females with PD can be classified into their specific sex group based on their RSEEG data. Similarly, experiments where people with different MDS-UPDRS scores were separated showed that this increased outcomes, followed by an experiment where 26 PwPD with the lowest MDS-UPDRS scores were classified against 26 PwPD with the highest MDS-UPDRS scores. This showed that participants can be classified based on the MDS-UPDRS score and should, therefore, not be

grouped together. Additionally, this experiment provided evidence that RSEEG can be used to track disease progression as the differences in RSEEG signals between people with lower and higher MDS-UPDRS scores can be used for classification.

8.2.4 The broader context of neurodegenerative diseases

As EEG recording methods are based on physiological processes that are the same in every human being, they do not differ when researching different neurodegenerative diseases. It is only when it comes to the data analysis that different choices are made based on the particular characteristics of a disease. It was therefore possible to put the results presented in the previous Chapters in the broader perspective of neurodegenerative diseases, by comparing research in PD to research in MS.

Chapter 7 contained a systematic literature review of the use of ERPs in the investigation of cognition in MS, which was then compared to and contrasted with a systematic review into ERPs in the investigation of cognition in PD (Seer et al., 2016). Similar as with the PD review, the contribution of the MS review was the synthesis of heterogeneous results into an overview of the study of cognition using ERPs and recommendations that flowed from this synthesis. The results of the MS review indicate that methods were very heterogeneous, that results were often presented for PwMS instead of distinguishing between MS subtypes, and that cohort sizes varied widely. The PD review found similar heterogeneity of methods and small cohort sizes, leading to both recommending standardisation of methods and the creation of larger datasets.

A main difference between the two reviews was the manner in which the data was synthesised. In the MS review, pooled effect sizes were used to provide a quantification of results. This was not the case for the PD review, which did not report statistical synthesis of results. Differences notwithstanding, both reviews found similar issues in the study of the respective diseases and came to similar conclusions on where gains could be made in research. First and foremost, the heterogeneity of methods was a hindrance to the development of ERPs as a clinical tool. A wide variety of experimental equipment, analysis methods, experimental paradigms, and cohort sizes were used, which makes comparison of results difficult. Standardisation of experimental methods would make comparison of results much easier. Second, while it is an issue that has been mentioned many times and in different contexts, cohort size is important. While some studies use sizeable participant groups, others perform experiments with small numbers of participants. This lack of statistical power can lead to spurious results that would not have been reported for larger cohort sizes.

The combined findings from the two literature reviews show that similar questions that were asked in this thesis about RSEEG for the classification of PwPD can be asked about other neurodegenerative diseases. While the experimental findings in this thesis all pertained to PD, the methods themselves are independent of the disease being investigated. Whether one is investigating PD, MS, Huntington's disease, or Alzheimer's disease, the principles of thorough preprocessing,

avoidance of data leakage and ascertaining that the participant cohort is not too diverse, still hold true. Similarly, methods used for the acquisition and analysis of ERP and RSEEG data will follow similar conventions and mostly differ in which features are extracted, with RSEEG data often being analysed in the time-frequency domain, with features such as power and entropy being extracted, and ERP data being analysed in the time domain, with features of latency and amplitude being extracted.

The contribution of the literature review into the study of cognition in MS using ERPS and the comparison of that review with the PD review is that it makes clear which issues exist particular to the diseases, and which issues exist in the broader context of ERP research into neurodegenerative diseases. In this manner, several recommendations can be made that can help develop ERPs into a clinical tool. Specific to MS, the evidence suggests visual stimuli are preferred to auditory stimuli, and that data should be separated by MS subtype. In general, concerning both MS and PD studies, cohorts tended to be small, leading to a lack of statistical power. The easiest way to alleviate this problem is to either combine datasets or create larger datasets. For the former, standardisation of methods is paramount, as the equipment and parameters used to collect data vary widely between studies, making it difficult to compare results. In the MS review, it became clear that the majority of experiments were oddball experiments, something that was also the case in the PD review, while other paradigms might be better suited to investigate differences in cognition. These should, therefore, be researched. Lastly, to make comparison of results easier, it is recommended to report raw latencies and amplitudes.

8.3 Limitations

The biggest limitation of the work presented here is that all results are based on three small datasets. While combining these into one larger dataset made it possible to perform some research that has not been performed previously, the number of participants is still limited. The problem with combining the three datasets is that all three were acquired using different experimental equipment (Cavanagh et al., 2018; Suuronen et al., 2023). While attempts were made to make the data as uniform as possible, by using the exact same preprocessing parameters for all datasets and by rescaling the features, there may still be differences that cannot be corrected. Additionally, the recordings were of different lengths (Vlieger et al., 2023), which means that one dataset would provide more samples than another when using single epochs in experiments, and that averaged epochs for other experiments were created by averaging more epochs for one dataset than for another.

Another limitation concerns the development of a recommended pipeline. While recommendations were made in this thesis, these recommendations are a product of their time. Both the fields of EEG and ML progress fast, so methods that are commonplace or even state-of-the-art today might be outdated tomorrow. While there are particular best practice guidelines that might universally hold, the manner in which they are implemented might change. For example, artifact rejection has always been a staple of EEG preprocessing, but where this used to be done by hand, it is how often

done using automated methods. This does not change the recommendation that artifact rejection should take place but does change what is considered the best method to do so. So, while the recommendations made in this thesis may contain some universal truths about how a PD or MS classification task using EEG data should be performed, the specifics are a product of their time. Additionally, neurodegenerative diseases tend to develop very at varying rates and with a wide variety of symptoms in different people, and EEG measurements themselves can be quite varied as well. If different subgroups of people with a particular disease are identified, the pipeline might need to be tailored to the particular subgroup. To take MS as an example, it might be that parameters for people with relapsing-remitting MS would be different than for people with secondary progressive MS.

One of the major problems with neurodegenerative diseases is that it can take years or even decades to come to an accurate diagnosis (Fernández et al., 2010; Rossi, Perez-Lloret, & Merello, 2021). This is because many diseases might either not produce features immediately, as is the case with PD (Hawkes, Del Tredici, & Braak, 2010), where cardinal features often do not appear until the disease has been progressing for a decade or more, or because the disease features can be attributed to a variety of diseases. In the research here, PwPD were compared to controls and the results show that PwPD can reliably be distinguished from controls if researchers use the methodology described in this thesis. What this does not show is that PwPD can be distinguished from people that have other neurodegenerative diseases. In the real world, the problem of diagnosis is much more difficult and the boundaries between healthy and not healthy, or between one disease and another, are much murkier. To determine whether PD can be reliably distinguished from other similar neurodegenerative diseases would, therefore, require experiments with cohorts with a wide variety of neurodegenerative diseases.

Lastly, one of the aims of this thesis was to lay the groundworks for EEG as a gold standard of diagnosis and prognosis of PD. The problem here is that the current gold standards are insufficient, which means that benchmarking EEG against them is not useful: if the current gold standard does not measure what it is supposed to measure, then comparing a method with the aim of showing that the new method is better does not make sense. For example, comparing the prognostic power of the MDS-UPDRS to that of RSEEG would not be very useful, as the former has been shown to be lacking in that aspect. Fortunately, as will be discussed in the next section, there might be other ways to deal with this problem.

8.4 Future work

The research presented in this thesis was built on research into RSEEG in PD going back decades and more recent ML studies that used RSEEG data from PwPD and controls. By using methods that have been used in the literature and asking questions about the used methodologies, it was made clearer how results can be interpreted, what choices lead to particular results, and what is necessary for reliable classification. Because the goal was to scrutinise methods and clarify the effects of these methods, it

was never the intention here to produce optimal results. It is, therefore, entirely possible that other averaging techniques, different algorithms, or newer artifact removal methods would produce better results. What the results presented here do provide is a methodological basis for future researchers to investigate these methods and relate their own results to those of previous research.

While Chapter 5 showed methods that should not be used for feature selection, other methods can be used if one wants to select features a priori. As was reported on in Chapter 2, a vast amount of literature exists on what features differ between PwPD and controls (Soikkeli et al., 1991; Morita et al., 2009; Caviness et al., 2015; Geraedts et al., 2018), as well as between PwPD and people with other diseases (Bonanni et al., 2008; Barcelon et al., 2019). Researchers can use this literature to either decide which features are worthwhile to extract or, if features have already been extracted, which features to include in their experiments. Basing feature selection on previous literature may help improve generalisability and reduce the chances of overfitting. Taking the results of Chapter 6 into account, a search for features may be most worthwhile when performed in more uniform subgroups than a search in a more diverse group than PwPD in general would be.

Concerning the development of EEG as a gold standard, comparison to current gold standards, such as the MDS-UPDRS for prognostication, might not be useful. The results in this thesis, and specifically those in Chapter 6 concerning PwPD with different MDS-UPDRS scores, provide an insight into what might be possible. While the MDS-UPDRS is not useful as a prognostic tool, it is the current best instrument to determine disease status at a certain moment. The classification experiment of PwPD at different stages of their disease as determined by the MDS-UPDRS implied that there are differences in the RSEEG data of people at different stages of their disease, an outcome supported by findings in the literature that show a slowing of the EEG as the disease develops. Furthermore, the findings regarding *rapid eye movement sleep behaviour disorder* (RBD), show that this phenomenon also takes place in a disease of which many patients go on to develop PD (Livia Fantini et al., 2003). Based on this information, and to verify that this holds true over a longer period of time, it would be necessary to perform a longitudinal study. This could be similar to the study by Caviness et al. (2015), but at a larger scale and involving both PwPD and people suffering from RBD.

When it comes to the findings about the effects of participant characteristics influencing results, the next step would be to create larger separate cohorts of male PwPD and female PwPD to understand the differences within the sexes in disease progression. Additionally, for RSEEG to become a diagnostic tool it is necessary to create cohorts of people with a wide variety of neurodegenerative diseases to investigate whether it is possible to distinguish one disease from another using RSEEG. This would also be best done by taking participant characteristics such as age and sex into consideration.

8.5 Broader impact

In this thesis, the main focus of experiments has been squarely on PD, but as the MS review and the comparison to the PD review by Seer et al. (2016) showed, issues that concern EEG research in MS also show up in PD research. As mentioned before, the physiological basis of EEG does not change with the neurodegenerative disease that is being studied, and the general methods are therefore the same, save for certain specific parameters that can be changed to address certain problems that are particular to a disease, for example the presence of involuntary muscle activity in PwPD. Given that the methods are generally the same, the experiments presented in this thesis can rather easily be replicated for other diseases, and the recommendations can serve as a starting point for researchers investigating other neurodegenerative diseases.

Apart from recommendations concerning experimental methods, the experiments and review included in this thesis also clarified what is necessary from an experimental and data collection perspective if EEG is to become a clinical tool for PD or MS. In both PD and MS, task-specific EEG recordings are made in such a variety of manners that comparison becomes very difficult. Even the three RSEEG datasets used throughout this thesis were acquired from widely different equipment with recordings being of different lengths. By showing these issues persist across the study of different diseases, it highlights the necessity of standardisation of experimental and reporting methods.

As mentioned in the introduction, this thesis is interdisciplinary in nature, predominantly in computer science, as well as in clinical neuroscience. This required the interrogation of methods that are sometimes common knowledge in one field, but not in the other. For example, preprocessing of EEG data is widely studied in the field of (clinical) neuroscience, in the context of general EEG experiments as well as the specific case of PD, but methods of preprocessing in the context of ML classifications tasks were less well understood. Similarly, the effects of data leakage caused by feature selection and sample randomisation methods are not new in computer science, yet the methods used in the context of PD classification tasks did show evidence of data leakage. By investigating these topics in an interdisciplinary context, the research in this thesis did specifically what Bammer (2013) described, namely: dealing “with complex real-world problems by bringing together disciplinary and stakeholder knowledge and explicitly dealing with remaining unknowns, in order to use that integrated research to support policy and practice change.”

8.6 Final remarks

The aim of the research presented here has been to provide insight into the precise effects of using different preprocessing methods, to investigate how differences in datasets, such as eyes open and eyes closed recording conditions lead to different outcomes, and clarify how participant characteristics influence classification outcomes. Through the eight chapters in this thesis, the whole classification

pipeline was scrutinised, and recommendations made on how reliable classification of PwPD and controls can be accomplished. Having done that, hopefully this thesis can help researchers in their investigations of neurodegenerative diseases using RSEEG and ML.

References

Aldehim, G., & Wang, W. (2017). Determining appropriate approaches for using data in feature selection. *International Journal of Machine Learning and Cybernetics*, 8(3), 915-928.

Aljalal, M., Aldosari, S. A., Al Sharabi, K., Abdurraqeeb, A. M., & Alturki, F. A. (2022). Parkinson's disease detection from resting-state EEG signals using common spatial pattern, entropy, and machine learning techniques. *Diagnostics*, 12(5), 1033.

Amato, M. P., & Ponziani, G. (1999). Quantification of impairment in MS: discussion of the scales in use. *Multiple Sclerosis Journal*, 5(4), 216-219.

Amato, M. P., Fratiglioni, L., Groppi, C., Siracusa, G., & Amaducci, L. (1988). Interrater reliability in assessing functional systems and disability on the Kurtzke scale in multiple sclerosis. *Archives of Neurology*, *45*(7), 746-748.

Anhoque, C. F., Biccas Neto, L., Domingues, S. C. A., Teixeira, A. L., & Domingues, R. B. (2012). Cognitive impairment in patients with clinically isolated syndrome. *Dementia & Neuropsychologia, 6*, 266-269.

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection.

Artemiadis, A., Anagnostouli, M., Zalonis, I., Chairopoulos, K., & Triantafyllou, N. (2018). Structural MRI correlates of cognitive function in multiple sclerosis. *Multiple sclerosis and related disorders*, *21*, 1-8.

Babineau, J. (2014). Product review: Covidence (systematic review software). *Journal of the Canadian Health Libraries Association/Journal de l'Association des bibliothèques de la santé du Canada, 35*(2), 68-71.

Bammer, G. (2013). *Disciplining interdisciplinarity: Integration and implementation sciences for researching complex real-world problems.* ANU Press.

Barcelon, E. A., Mukaino, T., Yokoyama, J., Uehara, T., Ogata, K., Kira, J.-I., & Tobimatsu, S. (2019). Grand Total EEG Score Can Differentiate Parkinson's Disease From Parkinson-Related Disorders. *Frontiers in Neurology, 10*, 398.

Barry, R. J., & De Blasio, F. M. (2017). EEG differences between eyes-closed and eyes-open resting remain in healthy ageing. *Biological Psychology*, 129, 293-304.

Bayes, T. (1763). LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFR S. *Philosophical transactions of the Royal Society of London* (53), 370-418.

Beck, A. T., Steer, R. A., & Carbin, M. G. (1988). Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation. *Clinical Psychology Review*, 8(1), 77-100.

Benedict, R. H., Cookfair, D., Gavett, R., Gunther, M., Munschauer, F., Garg, N., & Weinstock-Guttman, B. (2006). Validity of the minimal assessment of cognitive function in multiple sclerosis (MACFIMS). *Journal of the International Neuropsychological Society*, *12*(4), 549-558.

Benedict, R. H., Amato, M. P., DeLuca, J., & Geurts, J. J. (2020). Cognitive impairment in multiple sclerosis: clinical management, MRI, and therapeutic avenues. *The Lancet Neurology*, *19*(10), 860-871.

Benz, N., Hatz, F., Bousleiman, H., Ehrensperger, M. M., Gschwandtner, U., Hardmeier, M., . . . Monsch, A. U. (2014). Slowing of EEG background activity in Parkinson's and Alzheimer's disease with early cognitive dysfunction. *Frontiers in Aging Neuroscience, 6*, 314.

Berrar, D. (2019). Cross-Validation. In *Encyclopedia of Bioinformatics and Computational Biology*, 1, 542-545.

Berrigan, L. I., LeFevre, J. A., Rees, L. M., Berard, J. A., Francis, A., Freedman, M. S., & Walker, L. A. (2022). The symbol digit modalities test and the paced auditory serial addition test involve more than processing speed. *Multiple Sclerosis and Related Disorders*, *68*, 104229.

Berry, K. J., Mielke, P. W., Jr., & Mielke, H. W. (2002). The Fisher-Pitman permutation test: an attractive alternative to the F test. *Psychological Reports*, 90(2), 495-502.

Betrouni, N., Delval, A., Chaton, L., Defebvre, L., Duits, A., Moonen, A., . . . Dujardin, K. (2019). Electroencephalography-based machine learning for cognitive profiling in Parkinson's disease: Preliminary results. *Movement Disorders, 34*(2), 210-217.

Bhalerao, S. V., & Pachori, R. B. (2022). Sparse spectrum based swarm decomposition for robust nonstationary signal analysis with application to sleep apnea detection from EEG. *Biomedical Signal Processing and Control*, 77, 103792.

Biasiucci, A., Franceschiello, B., & Murray, M. M. (2019). Electroencephalography. *Current Biology, 29*(3), R80-R85.

Bigdely-Shamlo, N., Mullen, T., Kothe, C., Su, K. M., & Robbins, K. A. (2015). The PREP pipeline: standardized preprocessing for large-scale EEG analysis. *Frontiers in Neuroinformatics*, 9, 16.

Bissonnette, J. N., Pimer, L., Francis, A. M., Hull, K. M., Leckey, J., MacGillivray, M., Berrigan, L. I., & Fisher, D. J. (2023). An investigation of auditory processing in relapsing-remitting multiple sclerosis. *Experimental Brain Research*, 241(5), 1319-1327.

Bonanni, L., Thomas, A., Tiraboschi, P., Perfetti, B., Varanese, S., & Onofrj, M. (2008). EEG comparisons in early Alzheimer's disease, dementia with Lewy bodies and Parkinson's disease with dementia patients with a 2-year follow-up. *Brain, 131*(3), 690-705.

Boringa, J. B., Lazeron, R. H., Reuling, I. E., Ader, H. J., Pfennings, L. E., Lindeboom, J., ... & Polman, C. H. (2001). The brief repeatable battery of neuropsychological tests: normative values allow application in multiple sclerosis clinical practice. *Multiple Sclerosis Journal, 7*(4), 263-267.

Braak, H., Del Tredici, K., Rüb, U., De Vos, R. A., Steur, E. N. J., & Braak, E. (2003). Staging of brain pathology related to sporadic Parkinson's disease. *Neurobiology of aging, 24*(2), 197-211.

Bracewell, R. N. (1986). The Fourier transform and its applications (Vol. 31999). McGraw-Hill New York.

Breiman, L. (1996). Bagging predictors. *Machine learning, 24*, 123-140.

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32.

Capone, F., Motolese, F., Falato, E., Rossi, M., & Di Lazzaro, V. (2020). The potential role of neurophysiology in the management of multiple sclerosis-related fatigue. *Frontiers in Neurology, 11*, 531272.

Carr, S. (2017). UNM researchers design patient repository for more effective EEG diagnoses. Retrieved from http://news.unm.edu/news/unm-researchers-design-patient-repository-for-more-effective-eeg-diagnoses

Cavanagh, J. F., Kumar, P., Mueller, A. A., Richardson, S. P., & Mueen, A. (2018). Diminished EEG habituation to novel events effectively classifies Parkinson's patients. *Clinical Neurophysiology*, 129(2), 409-418.

Caviness, J. N., Hentz, J. G., Belden, C. M., Shill, H. A., Driver-Dunckley, E. D., Sabbagh, M. N., . . . Adler, C. H. (2015). Longitudinal EEG changes correlate with cognitive measure deterioration in Parkinson's disease. *Journal of Parkinson's Disease, 5*(1), 117-124.

Cerri, S., Mus, L., & Blandini, F. (2019). Parkinson's Disease in Women and Men: What's the Difference? *Journal of Parkinson's Disease, 9*(3), 501-515.

Chan, J. W. (2002). Optic neuritis in multiple sclerosis. *Ocular Immunology and Inflammation, 10*(3), 161-186.

Chang, C.-Y., Hsu, S.-H., Pion-Tonachini, L., & Jung, T.-P. (2019). Evaluation of artifact subspace reconstruction for automatic artifact components removal in multi-channel EEG recordings. *IEEE Transactions on Biomedical Engineering, 67*(4), 1114-1121.

Chapman, W. W., Nadkarni, P. M., Hirschman, L., D'avolio, L. W., Savova, G. K., & Uzuner, O. (2011). Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. *Journal of the American Medical Informatics Association*, 18(5), 540-543.

Chaturvedi, M., Hatz, F., Gschwandtner, U., Bogaarts, J. G., Meyer, A., Fuhr, P., & Roth, V. (2017). Quantitative EEG (QEEG) measures differentiate Parkinson's disease (PD) patients from healthy controls (HC). *Frontiers in Aging Neuroscience*, 9, 3.

Chiaravalloti, N. D., & DeLuca, J. (2008). Cognitive impairment in multiple sclerosis. *Lancet Neurology*, 7(12), 1139-1151.

Chinnadurai, S. A., Venkatesan, S. A., Shankar, G., Samivel, B., & Ranganathan, L. N. (2016). A study of cognitive fatigue in Multiple Sclerosis with novel clinical and electrophysiological parameters utilizing the event related potential P300. *Multiple Sclerosis and Related Disorders, 10,* 1-6.

Choy, T., Baker, E., & Stavropoulos, K. (2022). Systemic racism in EEG Research: considerations and potential solutions. *Affective Science, 3*(1), 14-20.

Claassen, J. A. (2005). The gold standard: not a golden standard. *British Medical Journal, 330*(7500), 1121.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. Academic Press.

Cohen, M. X. (2014). *Analyzing neural time series data: theory and practice*: MIT Press.

Congedo, M., Gouy-Pailler, C., & Jutten, C. (2008). On the blind source separation of human electroencephalogram by approximate joint diagonalization of second order statistics. *Clinical Neurophysiology*, 119(12), 2677-2686.

Cooray, G. K., Sundgren, M., & Brismar, T. (2020). Mechanism of visual network dysfunction in relapsing-remitting multiple sclerosis and its relation to cognition. *Clinical Neurophysiology*, Counter, S. A., Olofsson, A., Grahn, H. F., & Borg, E. (1997). MRI acoustic noise: sound pressure and frequency analysis. *Journal of Magnetic Resonance Imaging, 7*(3), 606-611.131(2), 361-367.

Covey, T. J., Shucard, J. L., & Shucard, D. W. (2017). Event-related brain potential indices of cognitive function and brain resource reallocation during working memory in patients with Multiple Sclerosis. *Clinical Neurophysiology, 128*(4), 604-621.

Currier, R. D. (1996). Did John Hunter give James Parkinson an idea? *Archives of Neurology, 53*(4), 377-378.

Dalianis, H., & Dalianis, H. (2018). Evaluation metrics and evaluation. *Clinical text mining: secondary use of electronic patient records*, 45-53.

Delacre, M., Lakens, D., & Leys, C. (2017). Why psychologists should by default use Welch's t-test instead of Student's t-test. *International Review of Social Psychology*, 30(1).

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods, 134*(1), 9-21.

Demircioğlu, A. (2021). Measuring the bias of incorrect application of feature selection when using cross-validation in radiomics. *Insights into Imaging, 12*(1), 1-10.

Dien, J., Khoe, W., & Mangun, G. R. (2007). Evaluation of PCA and ICA of simulated ERPs: Promax vs. Infomax rotations. *Human brain mapping*, *28*(8), 742-763.

Dien, J. (2010a). The ERP PCA Toolkit: An open source program for advanced statistical analysis of event-related potential data. *Journal of neuroscience methods*, *187*(1), 138-145.

Dien, J. (2010b). Evaluating two-step PCA of ERP data with geomin, infomax, oblimin, promax, and varimax rotations. *Psychophysiology*, *47*(1), 170-183.

Efron, B. (1992). Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics: Methodology and distribution* (pp. 569-593). Springer.

Espinoza, A. I., May, P., Anjum, M. F., Singh, A., Cole, R. C., Trapp, N., ... & Narayanan, N. S. (2022). A pilot study of machine learning of resting-state EEG and depression in Parkinson's disease. *Clinical Parkinsonism & Related Disorders, 7,* 100166.

Evers, L. J., Krijthe, J. H., Meinders, M. J., Bloem, B. R., & Heskes, T. M. (2019). Measuring Parkinson's disease over time: the real-world within-subject reliability of the MDS-UPDRS. *Movement Disorders, 34*(10), 1480-1487.

Evgeniou, T., & Pontil, M. (2001). Support vector machines: Theory and applications. In *Machine Learning and Its Applications: Advanced Lectures* (pp. 249-257): Springer.

FDA-NIH Biomarker Working Group. (2016). *BEST (Biomarkers, endpoints, and other tools) resource*

Fearnley, J. M., & Lees, A. J. (1991). Ageing and Parkinson's disease: substantia nigra regional selectivity. *Brain, 114*(5), 2283-2301.

Fernández, O., Fernández, V., Arbizu, T., Izquierdo, G., Bosca, I., Arroyo, R., García Merino, J. A., & de Ramón, E. (2010). Characteristics of multiple sclerosis at onset and delay of diagnosis and treatment in Spain (the Novo Study). *Journal of Neurology*, *257*(9), 1500-1507.

Feurer, M., & Hutter, F. (2019). Hyperparameter optimization. In *Automated Machine Learning* (pp. 3-33): Springer, Cham.

Filippi, M., Medaglini, S., Mammi, S., Martinelli, v., Lia, C., Campi, A., Sirabian, G., & Comi, G. (1992). Correlations between auditory event-related potentials (ERPs), brain MRI and neuropsychological tests in multiple sclerosis. [Italian] [Correlazioni tra p300 uditiva, risonanza magnetica dell'encefalo e test neuropsicologici nella sclerosi multipla.]. *Rivista di Neurobiologia*, 38(5), 369-375.

Filippi, M., Rocca, M. A., Benedict, R. H. B., DeLuca, J., Geurts, J. J. G., Rombouts, S. A. R. B., ... & Comi, G. (2010). The contribution of MRI in assessing cognitive impairment in multiple sclerosis. *Neurology*, *75*(23), 2121-2128.

Filippi, M., & Rocca, M. A. (2013). Present and future of fMRI in multiple sclerosis. *Expert review of neurotherapeutics*, *13*(sup2), 27-31.

Fonseca, L. C., Tedrus, G. M., Carvas, P. N., & Machado, E. C. (2013). Comparison of quantitative EEG between patients with Alzheimer's disease and those with Parkinson's disease dementia. *Clinical Neurophysiology, 124*(10), 1970-1974.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 1189-1232.

Garcia, J., & Quintana-Domeque, C. (2007). The evolution of adult height in Europe: a brief note. *Economics & Human Biology*, 5(2), 340-349.

Ge, W., Lueck, C., Suominen, H., & Apthorp, D. (2023). Has machine learning over-promised in healthcare?: A critical analysis and a proposal for improved evaluation, with evidence from Parkinson's disease. *Artificial Intelligence in Medicine*, 139, 102524.

Geraedts, V. J., Boon, L. I., Marinus, J., Gouw, A. A., van Hilten, J. J., Stam, C. J., . . . Contarino, M. F. (2018). Clinical correlates of quantitative EEG in Parkinson disease: A systematic review. *Neurology, 91*(19), 871-883.

Gerschlager, W., Beisteiner, R., Deecke, L., Dirnberger, G., Endl, W., Kollegger, H., ... & Lang, W. (2000). Electrophysiological, neuropsychological and clinical findings in multiple sclerosis patients receiving interferon β-1b: a 1-year follow-up. *European neurology*, *44*(4), 205-209.

Giesser, B. S., Schroeder, M. M., LaRocca, N. G., Kurtzberg, D., Ritter, W., Vaughan, H. G., & Scheinberg, L. C. (1992). Endogenous event-related potentials as indices of dementia in multiple sclerosis patients. *Electroencephalography & Clinical Neurophysiology*, 82(5), 320-329.

Glover, G. H. (2011). Overview of functional magnetic resonance imaging. *Neurosurgery Clinics, 22*(2), 133-139.

Goetz, C. G., Poewe, W., Rascol, O., Sampaio, C., Stebbins, G. T., Counsell, C., Giladi, N., Holloway, R. G., Moore, C. G., & Wenning, G. K. (2004). Movement Disorder Society Task Force report on the Hoehn and Yahr staging scale: status and recommendations of the Movement Disorder Society Task Force on rating scales for Parkinson's disease. *Movement Disorders*, 19(9), 1020-1028.

Goetz, C. G., Tilley, B. C., Shaftman, S. R., Stebbins, G. T., Fahn, S., Martinez-Martin, P., Poewe, W., Sampaio, C., Stern, M. B., & Dodel, R. (2008). Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results. *Movement Disorders: Official Journal of the Movement Disorder Society*, 23(15), 2129-2170.

Gonzalez-Latapi, P., Bayram, E., Litvan, I., & Marras, C. (2021). Cognitive Impairment in Parkinson's Disease: Epidemiology, Clinical Profile, Protective and Risk Factors. *Behavioral Sciences*, 11(5).

Gonzalez-Rosa, J. J., Inuggi, A., Riccitelli, G., Rocca, M. A., Filippi, M., Comi, G., & Leocani, L. (2012). Brain reorganisation and temporal activation patterns in early multiple sclerosis patients revealed by estimation of EEG cortical sources and individual MRI. *Journal of Neurology*, (1), S196.

Gonzalez-Rosa, J. J., Vazquez-Marrufo, M., Vaquero, E., Duque, P., Borges, M., Gamero, M. A., Gomez, C. M., & Izquierdo, G. (2006). Differential cognitive impairment for diverse forms of multiple sclerosis. *BMC Neuroscience*, 7, 39.

Gonzalez-Rosa, J. J., Vazquez-Marrufo, M., Vaquero, E., Duque, P., Borges, M., Gomez-Gonzalez, C. M., & Izquierdo, G. (2011). Cluster analysis of behavioural and event-related potentials during a contingent negative variation paradigm in remitting-relapsing and benign forms of multiple sclerosis. *BMC Neurology*, 11, 64.

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., . . . Hämäläinen, M. S. (2014). MNE software for processing MEG and EEG data. *NeuroImage, 86*, 446-460.

Greischar, L. L., Burghy, C. A., van Reekum, C. M., Jackson, D. C., Pizzagalli, D. A., Mueller, C., & Davidson, R. J. (2004). Effects of electrode density and electrolyte spreading in dense array electroencephalographic recording. *Clinical Neurophysiology, 115*(3), 710-720.

Gronwall, D. M. A. (1977). Paced auditory serial-addition task: a measure of recovery from concussion. *Perceptual and motor skills, 44*(2), 367-373.

Guo, G., Wang, S., Wang, S., Zhou, Z., Pei, G., & Yan, T. (2021). Diagnosing Parkinson's Disease Using Multimodal Physiological Signals. *Human Brain and Artificial Intelligence*, Singapore.

Haas, L. F. (2003). Hans Berger (1873–1941), Richard Caton (1842–1926), and electroencephalography. *Journal of Neurology, Neurosurgery & Psychiatry, 74*(1), 9-9.

Hajra, S. G., Liu, C. C., Fickling, S. D., Pawlowski, G. M., Song, X., & D'Arcy, R. C. N. (2021). Event Related Potential Signal Capture Can Be Enhanced through Dynamic SNR-Weighted Channel Pooling. *Sensors*, 21(21).

Hamilton, J. B. (1951). Patterned loss of hair in man: types and incidence. *Annals of the New York Academy of Sciences*, *53*(3), 708-728.

Handelman, G. S., Kok, H. K., Chandra, R. V., Razavi, A. H., Huang, S., Brooks, M., . . . Asadi, H. (2019). Peering into the black box of artificial intelligence: evaluation metrics of machine learning methods. *American Journal of Roentgenology, 212*(1), 38-43.

Hansch, E. C., Syndulko, K., Cohen, S. N., Goldberg, Z. I., Potvin, A. R., & Tourtellotte, W. W. (1982). Cognition in Parkinson disease: an event-related potential perspective. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, *11*(6), 599-607.

Hawkes, C. H., Del Tredici, K., & Braak, H. (2010). A timeline for Parkinson's disease. *Parkinsonism & Related Disorders, 16*(2), 79-84.

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and Their Applications, 13*(4), 18-28.

Hedges, L. V. (1981). Distribution Theory for Glass's Estimator of Effect size and Related Estimators. *Journal of Educational Statistics*, 6(2), 107-128.

Hely, M. A., Reid, W. G., Adena, M. A., Halliday, G. M., & Morris, J. G. (2008). The Sydney multicenter study of Parkinson's disease: the inevitability of dementia at 20 years. *Movement Disorders, 23*(6), 837-844.

Hilgard, J. (2021). Maximal positive controls: A method for estimating the largest plausible effect size. *Journal of Experimental Social Psychology*, 93, 104082.

Ho, T. K. (1995). *Random decision forests.* Paper presented at the *Proceedings of the 3rd International Conference on Document Analysis and Recognition*.

Hodge, V. J., & Austin, J. (2018). An evaluation of classification and outlier detection algorithms. arXiv preprint arXiv:1805.00811.

Hofmann, M. (2006). Support vector machines-kernels and the kernel trick. *Notes, 26*(3), 1-16.

Holsheimer, J., & Feenstra, B. (1977). Volume conduction and EEG measurements within the brain: a quantitative approach to the influence of electrical spread on the linear relationship of activity measured at different locations. *Electroencephalography and Clinical Neurophysiology*, 43(1), 52-58.

Honig, L. S., Ramsay, R. E., & Sheremata, W. A. (1992). Event-related potential P300 in multiple sclerosis. Relation to magnetic resonance imaging and cognitive impairment. *Archives of Neurology*, 49(1), 44-50.

Hossain, M. Z., Daskalaki, E., Brüstle, A., Desborough, J., Lueck, C. J., & Suominen, H. (2022). The role of machine learning in developing non-magnetic resonance imaging based biomarkers for multiple sclerosis: a systematic review. *BMC Medical Informatics and Decision Making*, *22*(1), 242.

Huang, C. C., & Lu, Z. (2016). Community challenges in biomedical text mining over 10 years: success, failure and the future. *Briefings in bioinformatics*, *17*(1), 132-144.

Hughes, A. J., Daniel, S. E., Kilford, L., & Lees, A. J. (1992). Accuracy of clinical diagnosis of idiopathic Parkinson's disease: a Parkin-pathological study of 100 cases. *Journal of Neurology Neurosurgery & Psychiatry,* 55(3), 181-184.

Jackson, C. E., & Snyder, P. J. (2008). Electroencephalography and event-related potentials as biomarkers of mild cognitive impairment and mild Alzheimer's disease. *Alzheimer's & Dementia*, 4(1), S137-S143.

Jain, A. K., Mao, J., & Mohiuddin, K. M. (1996). Artificial neural networks: A tutorial. *Computer, 29*(3), 31-44.

Jamoussi, H., Ali, N. B., Missaoui, Y., Cherif, A., Oudia, N., Anane, N., ... & Fredj, M. (2023). Cognitive impairment in multiple sclerosis: Utility of electroencephalography. *Multiple Sclerosis and Related Disorders*, *70*, 104502.

Jankovic, J. (2008). Parkinson's disease: clinical features and diagnosis. *Journal of Neurology, Neurosurgery & Psychiatry, 79*(4), 368-376.

Jasper, H. H. (1958). Report of the committee on methods of clinical examination in electroencephalography: 1957. *Electroencephalography and Clinical Neurophysiology*, 10, 370-375.

Jennings, J. L., Peraza, L. R., Baker, M., Alter, K., Taylor, J.-P., & Bauer, R. (2022). Investigating the power of eyes open resting state EEG for assisting in dementia diagnosis. *Alzheimer's Research & Therapy*, 14(1), 1-12.

John, E. R., Ahn, H., Prichep, L., Trepetin, M., Brown, D., & Kaye, H. (1980). Developmental equations for the electroencephalogram. *Science*, 210(4475), 1255-1258.

John, E. R., Prichep, L. S., Fridman, J., & Easton, P. (1988). Neurometrics: computer-assisted differential diagnosis of brain dysfunctions. *Science, 239*(4836), 162-169.

Kalia, L. V., & Lang, A. E. (2015). Parkinson's disease. *The Lancet, 386*(9996), 896-912.

Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*.

Karpiel, I., Kurasz, Z., Kurasz, R., & Duch, K. (2021). The Influence of Filters on EEG-ERP Testing: Analysis of Motor Cortex in Healthy Subjects. *Sensors,* 21(22).

Kaufman, D. M., & Milstein, M. J. (2012). *Kaufman's Clinical Neurology for Psychiatrists E-book*: Elsevier Health Sciences.

Kaya, I. (2019). A brief summary of EEG artifact handling. *Brain-Computer Interface*.

Khare, S. K., Bajaj, V., & Acharya, U. R. (2021). Detection of Parkinson's disease using automated tunable Q wavelet transform technique with EEG signals. *Biocybernetics and Biomedical Engineering*, 41(2), 679-689.

Kim, J.-H. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics & Data Analysis*, 53(11), 3735-3745.

Kingsford, C., & Salzberg, S. L. (2008). What are decision trees? *Nature Biotechnology, 26*(9), 1011-1013.

Klassen, B., Hentz, J., Shill, H., Driver-Dunckley, E., Evidente, V., Sabbagh, M., Adler, C., & Caviness, J. (2011). Quantitative EEG as a predictive biomarker for Parkinson disease dementia. *Neurology*, 77(2), 118-124.

Kocer, B., Unal, T., Nazliel, B., Biyikli, Z., Yesilbudak, Z., Karakas, S., & Irkec, C. (2008). Evaluating sub-clinical cognitive dysfunction and event-related potentials (P300) in clinically isolated syndrome. *Neurological Sciences, 29*, 435-444.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *International Joint Conference on Artificial Intelligence*, 14.

Krogh, A. (2008). What are artificial neural networks? *Nature Biotechnology, 26*(2), 195-197.

Kurtzke, J. F. (1983). Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). *Neurology, 33*(11), 1444-1444.

Lebrun, C., Blanc, F., Brassat, D., Zephir, H., de Seze, J., & CFSEP, b. o. (2010). Cognitive function in radiologically isolated syndrome. *Multiple Sclerosis Journal, 16*(8), 919-925.

Lederer, J. (2021). Activation functions in artificial neural networks: A systematic overview. *arXiv* preprint arXiv:2101.09957.

Lee, S., Hussein, R., Ward, R., Jane Wang, Z., & McKeown, M. J. (2021). A convolutional-recurrent neural network approach to resting-state EEG classification in Parkinson's disease. *Journal of Neuroscience Methods, 361*, 109282.

Lees, A. J. (2007). Unresolved issues relating to the shaking palsy on the celebration of James Parkinson's 250th birthday. *Movement Disorders: Official Journal of the Movement Disorder Society, 22*(S17), S327-S334.

Legendre, A. M. (1806). *Nouvelles méthodes pour la détermination des orbites des comètes* [New Methods for the Determination of Comet Orbits]

Leocani, L., Gonzalez-Rosa, J. J., & Comi, G. (2010). Neurophysiological correlates of cognitive disturbances in multiple sclerosis. *Neurological Sciences*, *31*(Suppl 2), 249-253.

Li, H., Ruan, J., Xie, Z., Wang, H., & Liu, W. (2007). Investigation of the critical geometric characteristics of living human skulls Parkinson medical image analysis techniques. *International Journal of Vehicle Safety, 2*(4), 345-367.

Livia Fantini, M., Gagnon, J. F., Petit, D., Rompré, S., Décary, A., Carrier, J., & Montplaisir, J. (2003). Slowing of electroencephalogram in rapid eye movement sleep behavior disorder. *Annals of Neurology, 53*(6), 774-780.

López-Góngora, M., Escartín, A., Martinez-Horta, S., Fernandez-Bobadilla, R., Querol, L., Romero, S., ... & Riba, J. (2015). Neurophysiological evidence of compensatory brain mechanisms in early-stage multiple sclerosis. *PloS one*, *10*(8), e0136786.

Lublin, F. D. (2014). New multiple sclerosis phenotypic classification. *European Neurology*, 72(Suppl. 1), 1-5.

Luck, S. J. (2014). *An introduction to the event-related potential technique*: MIT Press.

Luck, S. J., & Kappenman, E. S. (2011). *The Oxford handbook of event-related potential components*: Oxford University Press.

Lumley, T., Diehr, P., Emerson, S., & Chen, L. (2002). The importance of the normality assumption in large public health data sets. *Annual Review of Public Health*, 23(1), 151-169.

Magnié, M. N., Bensa, C., Laloux, L., Bertogliati, C., Faure, S., & Lebrun, C. (2007). Contribution of cognitive evoked potentials for detecting early cognitive disorders in multiple sclerosis. *Revue Neurologique*, *163*(11), 1065-1074.

Maitín, A. M., García-Tejedor, A. J., & Romero Muñoz, J. P. (2020). Machine learning approaches for detecting Parkinson's disease from EEG analysis: A systematic review. *Applied Sciences, 10*(23), 8662.

Maitín, A. M., Romero Muñoz, J. P., & García-Tejedor, Á. J. (2022). Survey of machine learning techniques in the analysis of EEG signals for Parkinson's disease: A systematic review. *Applied Sciences, 12*(14), 6967.

Makeig, S., Bell, A., Jung, T.-P., & Sejnowski, T. J. (1995). Independent component analysis of electroencephalographic data. *Advances in Neural Information Processing Systems,* 8.

Martinez-Martin, P., Rodriguez-Blazquez, C., Alvarez-Sanchez, M., Arakaki, T., Bergareche-Yarza, A., Chade, A., . . . Martinez-Castrillo, J. C. (2013). Expanded and independent validation of the Movement Disorder Society–Unified Parkinson's disease rating scale (MDS-UPDRS). *Journal of Neurology, 260*(1), 228-236.

McGough, J. J., & Faraone, S. V. (2009). Estimating the size of treatment effects: moving beyond p values. *Psychiatry (Edgmont),* 6(10), 21.

Menze, B. H., Kelm, B. M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., & Hamprecht, F. A. (2009). A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics, 10*(1), 213. Doi:10.1186/1471-2105-10-213.

Meyer-Moock, S., Feng, Y.-S., Maeurer, M., Dippel, F.-W., & Kohlmann, T. (2014). Systematic literature review and validity evaluation of the Expanded Disability Status Scale (EDSS) and the Multiple Sclerosis Functional Composite (MSFC) in patients with multiple sclerosis. *BMC Neurology,* 14(1), 1-10.

Mitchell, T. M. (1997). *Machine learning* (Vol. 1): McGraw-Hill New York.

Miyakoshi, M. (n.d). Makoto's preprocessing pipeline. Retrieved from https://sccn.ucsd. edu/wiki/Makotos_preprocessing_pipeline.

Miyakoshi, M., Jurgiel, J., Dillon, A., Chang, S., Piacentini, J., Makeig, S., & Loo, S. K. (2020). Modulation of Frontal Oscillatory Power during Blink Suppression in Children: Effects of Premonitory Urge and Reward. *Cerebral Cortex Communications,* 1(1).

Mognon, A., Jovicich, J., Bruzzone, L., & Buiatti, M. (2011). ADJUST: An automatic EEG artifact detector based on the joint use of spatial and temporal features. *Psychophysiology, 48*(2), 229-240.

Mohammed, R., Rawashdeh, J., & Abdullah, M. (2020). *Machine learning with oversampling and undersampling techniques: overview study and experimental results.* Paper presented at the *2020 11th International Conference on Information and Communication Systems*.

Morita, A., Kamei, S., Serizawa, K., & Mizutani, T. (2009). The relationship between slowing EEGs and the progression of Parkinson's disease. *Journal of Clinical Neurophysiology, 26*(6), 426-429.

Mullen, T. R. (2012). CleanLine EEGLAB plugin. *San Diego, CA: Neuroimaging Informatics Toolsand Resources Clearinghouse (NITRC)*.

Mullen, T. R., Kothe, C. A., Chi, Y. M., Ojeda, A., Kerth, T., Makeig, S., . . . Cauwenberghs, G. (2015). Real-time neuroimaging and cognitive monitoring using wearable dry EEG. *IEEE Transactions on Biomedical Engineering, 62*(11), 2553-2567.

Munn, Z., Moola, S., Lisy, K., Riitano, D., & Murphy, F. (2015). Claustrophobia in magnetic resonance imaging: a systematic review and meta-analysis. *Radiography*, *21*(2), e59-e63.

Nagels, G., D'hooghe, M. B., Kos, D., Engelborghs, S., & De Deyn, P. P. (2008). Within-session practice effect on paced auditory serial addition test in multiple sclerosis. *Multiple Sclerosis Journal*, *14*(1), 106-111.

Natekin, A., & Knoll, A. (2013). Gradient boosting machines: a tutorial. *Frontiers in Neurorobotics, 7*, 21.

Newton, M., Barrett, G., Callanan, M., & Towell, A. (1989). Cognitive event-related potentials in multiple sclerosis. *Brain*, 112(6), 1637-1660.

Niedermeyer, E., & da Silva, F. L. (2005). *Electroencephalography: Basic principles, Clinical Applications, and Related Fields*: Lippincott Williams & Wilkins.

Noble, W. S. (2006). What is a support vector machine? *Nature Biotechnology, 24*(12), 1565-1567.

Nolan, H., Whelan, R., & Reilly, R. B. (2010). FASTER: fully automated statistical thresholding for EEG artifact rejection. *Journal of Neuroscience Methods, 192*(1), 152-162.

Nutt, J. G., & Wooten, G. F. (2005). Diagnosis and initial management of Parkinson's disease. *New England Journal of Medicine, 353*(10), 1021-1027.

O'Donnell, B. F., Squires, N. K., Martz, M. J., Chen, J. R., & Phay, A. J. (1987). Evoked potential changes and neuropsychological performance in Parkinson's disease. *Biological Psychology*, *24*(1), 23-37.

Olejarczyk, E., Bogucki, P., & Sobieszek, A. (2017). The EEG split alpha peak: phenomenological origins and methodological aspects of detection and evaluation. *Frontiers in Neuroscience*, 11, 506.

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.

Orrù, G., Pettersson-Yeo, W., Marquand, A. F., Sartori, G., & Mechelli, A. (2012). Using Support Vector Machine to identify imaging biomarkers of neurological and psychiatric disease: A critical review. *Neuroscience & Biobehavioral Reviews,* 36(4), 1140-1152.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., . . . Moher, D. (2021). Updating guidance for reporting systematic reviews: development of the PRISMA 2020 statement. *Journal of Clinical Epidemiology, 134*, 103-112.

Paolicelli, D., Manni, A., Iaffaldano, A., Tancredi, G., Ricci, K., Gentile, E., ... & Trojano, M. (2021). Magnetoencephalography and high-density electroencephalography study of acoustic event related potentials in early stage of multiple sclerosis: a pilot study on cognitive impairment and fatigue. *Brain sciences*, *11*(4), 481.

Parkinson, J. (2002). An essay on the shaking palsy. *The Journal of Neuropsychiatry and Clinical Neurosciences,* 14(2), 223-236.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.

Pion-Tonachini, L., Kreutz-Delgado, K., & Makeig, S. (2019). ICLabel: An automated electroencephalographic independent component classifier, dataset, and website. *NeuroImage,* 198, 181-197.

Pohjolainen, S., & Heiliö, M. (2016). *Mathematical Modelling*.

Pokryszko-Dragan, A., Dziadkowiak, E., Zagrajek, M., Slotwinski, K., Gruszka, E., Bilinska, M., & Podemski, R. (2016a). Cognitive performance, fatigue and event-related potentials in patients with clinically isolated syndrome. *Clinical Neurology and Neurosurgery, 149*, 68-74.

Pokryszko-Dragan, A., Zagrajek, M., Slotwinski, K., Bilinska, M., Gruszka, E., & Podemski, R. (2016b). Event-related potentials and cognitive performance in multiple sclerosis patients with fatigue. *Neurological Sciences, 37*, 1545-1556.

Polat, K., & Güneş, S. (2007). Breast cancer diagnosis using least square support vector machine. *Digital Signal Processing,* 17(4), 694-701.

Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, 32(1), 3-25.

Postuma, R. B., Berg, D., Stern, M., Poewe, W., Olanow, C. W., Oertel, W., Obeso, J., Marek, K., Litvan, I., & Lang, A. E. (2015). MDS clinical diagnostic criteria for Parkinson's disease. *Movement Disorders*, 30(12), 1591-1601.

Prichep, L. S., Jacquin, A., Filipenko, J., Dastidar, S. G., Zabele, S., Vodencarevic, A., & Rothman, N. S. (2012). Classification of traumatic brain injury severity using informed data reduction in a series of binary classifier algorithms. *IEEE transactions on neural systems and rehabilitation engineering, 20*(6), 806-822.

Qi, Y. (2012). Random forest for bioinformatics. In *Ensemble Machine Learning: Methods and Applications* (pp. 307-323): Springer.

Ramaker, C., Marinus, J., Stiggelbout, A. M., & Van Hilten, B. J. (2002). Systematic evaluation of rating scales for impairment and disability in Parkinson's disease. *Movement Disorders: Official Journal of the Movement Disorder Society, 17*(5), 867-876.

Raymaekers, J., & Rousseeuw, P. J. (2021). Transforming variables to central normality. *Machine Learning,* 1-23.

Ríos-Herrera, W. A., Olguín-Rodríguez, P. V., Arzate-Mena, J. D., Corsi-Cabrera, M., Escalona, J., Marín-García, A., Ramos-Loyo, J., Rivera, A. L., Rivera-López, D., & Zapata-Berruecos, J. F. (2019). The influence of EEG references on the analysis of spatio-temporal interrelation patterns. *Frontiers in Neuroscience,* 13, 941.

Rossi, M., Perez-Lloret, S., & Merello, M. (2021). How much time is needed in clinical practice to reach a diagnosis of clinically established Parkinson's disease? *Parkinsonism and Related Disorders*, *92*, 53-58.

Van Rossum, G. (1995). Python reference manual. *Department of Computer Science [CS](R 9525)*.

Saikia, A., Hussain, M., Barua, A. R., & Paul, S. (2019). EEG-EMG correlation for Parkinson's disease. *International Journal of Engineering and Advanced Technology,* 8(6), 1179-1185.

Sakia, R. M. (1992). The Box-Cox transformation technique: a review. *Journal of the Royal Statistical Society: Series D (The Statistician),* 41(2), 169-178.

Samuel, A. L. (1988). Some studies in machine learning using the game of checkers. II—recent progress. *Computer Games I*, 366-400.

Sandry, J., Simonet, D. V., Brandstadter, R., Krieger, S., Sand, I. K., Graney, R. A., ... & Sumowski, J. F. (2021). The Symbol Digit Modalities Test (SDMT) is sensitive but non-specific in MS: Lexical access speed, memory, and information processing speed independently contribute to SDMT performance. *Multiple sclerosis and related disorders*, *51*, 102950.

Sawilowsky, S. S. (2009). New effect size rules of thumb. Journal of Modern Applied Statistical Methods, 8(2), 26.

Schenck, C. H., Bundlie, S. R., & Mahowald, M. W. (1996). Delayed emergence of a parkinsonian disorder in 38% of 29 older men initially diagnosed with idiopathic rapid eye movement sleep behavior disorder. *Neurology,* 46(2), 388-393.

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), 1-7.

Seer, C., Lange, F., Georgiev, D., Jahanshahi, M., & Kopp, B. (2016). Event-related potentials and cognition in Parkinson's disease: An integrative review. *Neuroscience & Biobehavioral Reviews,* 71, 691-714.

Sharrack, B., Hughes, R. A., Soudain, S., & Dunn, G. (1999). The psychometric properties of clinical rating scales used in multiple sclerosis. *Brain*, 122(1), 141-159.

Sidey-Gibbons, J. A., & Sidey-Gibbons, C. J. (2019). Machine learning in medicine: a practical introduction. *BMC Medical Research Methodology, 19*(1), 1-18.

Sjøgård, M., Wens, V., Van Schependom, J., Costers, L., D'hooghe, M., D'haeseleer, M., ... & De Tiège, X. (2021). Brain dysconnectivity relates to disability and cognitive impairment in multiple sclerosis. *Human brain mapping*, *42*(3), 626-643.

Smith, A. (1973). Symbol digit modalities test. *The Clinical Neuropsychologist*.

Soikkeli, R., Partanen, J., Soininen, H., Pääkkönen, A., & Riekkinen Sr, P. (1991). Slowing of EEG in Parkinson's disease. *Electroencephalography and Clinical Neurophysiology, 79*(3), 159-165.

Solomon, A. J., Bourdette, D. N., Cross, A. H., Applebee, A., Skidd, P. M., Howard, D. B., Spain, R. I., Cameron, M. H., Kim, E., & Mass, M. K. (2016). The contemporary spectrum of multiple sclerosis misdiagnosis: a multicenter study. *Neurology*, 87(13), 1393-1399.

Squires, N. K., Squires, K. C., & Hillyard, S. A. (1975). Two varieties of long-latency positive waves evoked by unpredictable auditory stimuli in man. *Electroencephalography and Clinical Neurophysiology,* 38(4), 387-401.

Ståhle, L., & Wold, S. (1989). Analysis of variance (ANOVA). *Chemometrics and Intelligent Laboratory Systems*, 6(4), 259-272.

Stern, G. (1989). Did parkinsonism occur before 1817? *Journal of Neurology, Neurosurgery & Psychiatry,* 52(Suppl), 11.

Strimbu, K., & Tavel, J. A. (2010). What are biomarkers? *Current Opinion in HIV and AIDS,* 5(6), 463.

Sullivan, G. M., & Feinn, R. (2012). Using Effect Size-or Why the P Value Is Not Enough. *Journal of Graduate Medical Education*, 4(3), 279-282.

Sundgren, M., Nikulin, V. V., Maurex, L., Wahlin, A., Piehl, F., & Brismar, T. (2015). P300 amplitude and response speed relate to preserved cognitive function in relapsing-remitting multiple sclerosis. *Clinical Neurophysiology*, 126(4), 689-697.

Suominen, H. (2009). Machine Learning and Clinical Text: Supporting Health Information Flow (Publication Number 125) University of Turku.

Suominen, H., Pahikkala, T., & Salakoski, T. (2008). Critical points in assessing learning performance via cross-validation. *Proceedings of the 2nd International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning*, Helsinki University of Technology.

Suuronen, I., Airola, A., Pahikkala, T., Murtojärvi, M., Kaasinen, V., & Railo, H. (2023). Budget-based classification of Parkinson's disease from resting state EEG. *IEEE Journal of Biomedical and Health Informatics*.

Tampu, I. E., Eklund, A., & Haj-Hosseini, N. (2022). Inflation of test accuracy due to data leakage in deep learning-based classification of OCT images. *Scientific Data*, 9(1), 580.

Tanaka, H., Koenig, T., Pascual-Marqui, R. D., Hirata, K., Kochi, K., & Lehmann, D. (2000). Event-related potential and EEG measures in Parkinson's disease without and with dementia. *Dementia and geriatric cognitive disorders,* 11(1), 39-45.

Taranu, D., Tumani, H., Holbrook, J., Tumani, V., Uttner, I., & Fissler, P. (2022). The TRACK-MS test battery: a very brief tool to track multiple sclerosis-related cognitive impairment. *Biomedicines*, *10*(11), 2975.

Tazegul, G., Etcioglu, E., Yildiz, F., Yildiz, R., & Tuney, D. (2015). Can MRI related patient anxiety be prevented?. *Magnetic resonance imaging*, *33*(1), 180-183.

Tecce, J. J. (1972). Contingent negative variation (CNV) and psychological processes in man. *Psychological Bulletin*, 77(2), 73.

The Cochrane Collaboration. (2020). Review Manager (RevMan) (Version 5.4).

Thatcher, R. W., & Lubar, J. F. (2009). History of the scientific standards of QEEG normative databases. *Introduction to quantitative EEG and neurofeedback: Advanced theory and applications*, *2*, 29-59.

Thompson, A. J., Banwell, B. L., Barkhof, F., Carroll, W. M., Coetzee, T., Comi, G., Correale, J., Fazekas, F., Filippi, M., & Freedman, M. S. (2018). Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. *The Lancet Neurology*, 17(2), 162-173.

Tombaugh, T. N., & McIntyre, N. J. (1992). The mini-mental state examination: a comprehensive review. *Journal of the American Geriatrics Society*, 40(9), 922-935.

Trost, Z., Jones, A., Guck, A., Vervoort, T., Kowalsky, J. M., & France, C. R. (2017). Initial validation of a virtual blood draw exposure paradigm for fear of blood and needles. *Journal of anxiety disorders*, *51*, 65-71.

Tudor, M., Tudor, L., & Tudor, K. I. (2005). [Hans Berger (1873-1941)—the history of electroencephalography]. *Acta Med Croatica*, 59(4), 307-313. (Hans Berger (1873-1941)—povijest elektroencefalografije.)

Turnbull, H. W. (1959). The correspondence of Isaac Newton: 1661-1675 (Vol. 1). In: London, UK Published for the Royal Society at the University Press.

Turner, L. M., Croft, R. J., Churchyard, A., Looi, J. C., Apthorp, D., & Georgiou-Karistianis, N. (2015). Abnormal electrophysiological motor responses in Huntington's disease: evidence of premanifest compensation. *PloS One*, 10(9), e0138563.

Uysal, U., Idiman, F., Idiman, E., Ozakbas, S., Karakas, S., & Bruce, J. (2014). Contingent negative variation is associated with cognitive dysfunction and secondary progressive disease course in multiple sclerosis. *Journal of Clinical Neurology*, 10(4), 296-303.

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics, 7*(1), 91.

Vazquez-Marrufo, M., Gonzalez-Rosa, J. J., Vaquero, E., Duque, P., Escera, C., Borges, M., ... & Gómez, C. M. (2008). Abnormal ERPs and high frequency bands power in multiple sclerosis. *International Journal of Neuroscience*, *118*(1), 27-38.

Vázquez-Marrufo, M., González-Rosa, J., Vaquero-Casares, E., Duque, P., Borges, M., & Izquierdo, G. (2009). Cognitive evoked potentials in remitting-relapsing and benign forms of multiple sclerosis. *Revista de Neurologia*, 48(9), 453-458.

Vazquez-Marrufo, M., Galvao-Carmona, A., González-Rosa, J. J., Hidalgo-Munoz, A. R., Borges, M., Ruiz-Peña, J. L., & Izquierdo, G. (2014). Neural correlates of alerting and orienting impairment in multiple sclerosis patients. *PLoS One*, *9*(5), e97226.

Vazquez-Marrufo, M. (2017). Event-related potentials for the study of cognition. *Event-Related Potentials and Evoked Potentials*, *1*.

Vázquez-Marrufo, M., Galvao-Carmona, A., Caballero-Díaz, R., Borges, M., Paramo, M. D., Benítez-Lugo, M. L., ... & Izquierdo, G. (2019). Altered individual behavioral and EEG parameters are related to the EDSS score in relapsing-remitting multiple sclerosis patients. *PLoS One*, *14*(7), e0219594.

Versi, E. (1992). "Gold standard" is an appropriate term. *British Medical Journal, 305*(6846), 187.

Vlieger, R., Daskalaki, E., Apthorp, D., Lueck, C.J., & Suominen, H. (2021). The use of event-related potentials and machine learning to improve diagnostic testing and prediction of disease progression in Parkinson's disease. *Studies in Health Technology and Informatics*, 284, 333-335.

Vlieger, R., Daskalaki, E., Apthorp, D., Lueck, C., & Suominen, H. (2023a). Evaluating Effects of Resting-State Electroencephalography Data Pre-Processing on a Machine Learning Task for Parkinson's Disease. *medRxiv*, 2023.2003.2006.23286826.

Vlieger, R., Suominen, H., Apthorp, D., Lueck, C. J., & Daskalaki, E. (2023b). Evaluating methods of oversampling and averaging resting-state electroencephalography data in classifying Parkinson's disease. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (pp. 1-5). IEEE.

Vlieger, R., Daskalaki, E., Apthorp, D., Lueck, C. J., & Suominen, H. (2024). Evaluating Effects of Resting-State Electroencephalography Data Pre-Processing on a Machine Learning Task for Parkinson's Disease. *Studies in health technology and informatics*, *310*, 1480–1481.

Waliszewska-Prosół, M., Nowakowska-Kotas, M., Kotas, R., Bańkowski, T., Pokryszko-Dragan, A., & Podemski, R. (2018). The relationship between event-related potentials, stress perception and personality type in patients with multiple sclerosis without cognitive impairment: A pilot study. *Advances in Clinical and Experimental Medicine*, *27*(6), 787-794.

Wang, L., & Alexander, C. A. (2016). Machine learning in big data. *International Journal of Mathematical, Engineering and Management Sciences,* 1(2), 52-61.

Wang, M., Ge, W., Apthorp, D., & Suominen, H. (2020). Robust feature engineering for Parkinson disease diagnosis: New machine learning techniques. *JMIR Biomedical Engineering*, 5(1), e13611.

Wang, S.-C., & Wang, S.-C. (2003). Artificial neural network. Interdisciplinary computing in java programming, 81-100.

Waninger, S., Berka, C., Stevanovic Karic, M., Korszen, S., Mozley, P. D., Henchcliffe, C., Kang, Y., Hesterman, J., Mangoubi, T., & Verma, A. (2020). Neurophysiological Biomarkers of Parkinson's Disease. *Journal of Parkinson's Disease*, 10(2), 471-480.

Weyhenmeyer, J., Hernandez, M. E., Lainscsek, C., Sejnowski, T. J., & Poizner, H. (2014). Muscle artifacts in single trial EEG data distinguish patients with Parkinson's disease from healthy individuals. Paper presented at the *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*.

Whitham, E. M., Pope, K. J., Fitzgibbon, S. P., Lewis, T., Clark, C. R., Loveless, S., . . . Lillie, P. (2007). Scalp electrical recording during paralysis: quantitative evidence that EEG frequencies above 20 Hz are contaminated by EMG. *Clinical Neurophysiology,* 118(8), 1877-1888.

Winkler, I., Brandl, S., Horn, F., Waldburger, E., Allefeld, C., & Tangermann, M. (2014). Robust artifactual independent component classification for BCI practitioners. *Journal of Neural Engineering,* 11(3), 035013.

Winkler, I., Haufe, S., & Tangermann, M. (2011). Automatic classification of artifactual ICA-components for artifact removal in EEG signals. *Behavioral and Brain Functions, 7*(1), 1-15.

Wu, Y., Le, W., & Jankovic, J. (2011). Preclinical biomarkers of Parkinson disease. *Archives of Neurology,* 68(1), 22-30.

Xulei, Y., Qing, S., & Cao, A. (2005, 31 July-4 Aug. 2005). Weighted support vector machine for data classification. Proceedings. *2005 IEEE International Joint Conference on Neural Networks*, 2005.

Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing,* 415, 295-316.

Yao, D., Qin, Y., Hu, S., Dong, L., Bringas Vega, M. L., & Valdés Sosa, P. A. (2019). Which Reference Should We Use for EEG and ERP practice? *Brain Topography*, 32(4), 530-549.

Ye, J., Hua, M., & Zhu, F. (2021). Machine Learning Algorithms are Superior to Conventional Regression Models in Predicting Risk Stratification of COVID-19 Patients. *Risk Management and Healthcare Policy,* 14, 3159-3166.

Yeo, I. K., & Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, 87(4), 954-959.

Yu, W., Liu, T., Valdez, R., Gwinn, M., & Khoury, M. J. (2010). Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. *BMC Medical Informatics and Decision Making,* 10(1), 1-7.

Yuvaraj, R., Rajendra Acharya, U., & Hagiwara, Y. (2018). A novel Parkinson's Disease Diagnosis Index using higher-order spectra features in EEG signals. *Neural Computing and Applications,* 30(4), 1225-1235.

Zalewska, E. (2020). Is So Called "Split Alpha" in EEG Spectral Analysis a Result of Methodological and Interpretation Errors? *Frontiers in Neuroscience,* 14, 608453.

Zhang, L., Wang, P., Zhang, R., Chen, M., Shi, L., Gao, J., & Hu, Y. (2020). The Influence of Different EEG References on Scalp EEG Functional Network Analysis During Hand Movement Tasks. *Frontiers in Human Neuroscience*, 14, 367.

Supplement 1

Table S1.1. An overview of evaluation results for experiments using single samples and per preprocessing stages for the Canberra Hospital dataset.

CH	Preprocessing	Accuracy [%]	Acc_std [%]	Sensitivity [%]	Specificity [%]	Precision [%]	F1-score [%]
Method 1	Level 1	58.68	16.01	68.15	46.98	61.36	64.58
	Level 2	59.13	15.63	66.9	49.54	62.09	64.4
	Level 3	57.69	15.78	54.81	60.99	61.81	58.1
	Level 4	47.83	16.86	55.12	39.12	51.96	53.5
Method 2	Level 1	89.79	2.38	91.4	87.78	90.24	90.82
	Level 2	75.49	2.87	84.57	64.26	74.51	79.22
	Level 3	70.53	3.28	64.88	77.04	76.5	70.21
	Level 4	74.96	4.66	75.26	74.59	77.97	76.59
Method 3	Level 1	62.73	14.97	71.88	51.42	64.64	68.07
	Level 2	66.03	13.97	74.63	55.4	67.4	70.83
	Level 3	64.99	16.99	59.58	71.22	70.45	64.56
	Level 4	60.95	17.46	65.16	55.92	63.85	64.5
Method 4	Level 1	91.79	1.8	93.06	90.22	92.16	92.6
	Level 2	76.62	2.42	85.84	65.23	75.3	80.23
	Level 3	71.39	3.51	65.37	78.33	77.65	70.98
	Level 4	77.43	2.82	78.8	75.8	79.55	79.17

Table S1.2. An overview of evaluation results for experiments using single samples and per preprocessing stages for the University of New Mexico dataset.

UNM	Preprocessing	Accuracy [%]	Acc_std [%]	Sensitivity [%]	Specificity [%]	Precision [%]	F1-score [%]
Method 1	Level 1	71.57	11.75	69.81	73.33	72.36	71.06
	Level 2	73.3	13.06	70.96	75.63	74.44	72.66
	Level 3	71.5	15.14	75.35	67.05	72.49	73.89
	Level 4	64.59	14.89	69.5	58.77	66.66	68.05
Method 2	Level 1	81.77	3.5	82.27	81.27	81.46	81.86
	Level 2	78.15	3.07	76.95	79.36	78.85	77.89
	Level 3	77.72	3.3	83.08	71.54	77.08	79.97
	Level 4	78.33	3.82	81.98	74.01	78.91	80.42
Method 3	Level 1	74.89	12.2	72.63	77.15	76.07	74.31
	Level 2	74.42	12.48	72.77	76.06	75.25	73.99
	Level 3	72.69	14.86	76.22	68.61	73.67	74.92
	Level 4	69.16	13.11	74.05	63.35	70.56	72.26
Method 4	Level 1	81.77	2.72	81.8	81.74	81.75	81.77
	Level 2	78.02	2.88	76.05	79.99	79.17	77.58
	Level 3	78.1	3.97	82.86	72.62	77.71	80.2
	Level 4	77.9	4.15	82.14	72.86	78.21	80.13

Table S1.3. An overview of evaluation results for experiments using single samples and per preprocessing stages for the University of Turku dataset.

UTU	Preprocessing	Accuracy [%]	Acc_std [%]	Sensitivity [%]	Specificity [%]	Precision [%]	F1-score [%]
Method 1	Level 1	61.62	13.95	56.78	66.1	60.81	58.72
	Level 2	61.98	14.82	56.92	66.66	61.26	59.01
	Level 3	55.64	13.68	53.12	57.97	53.99	53.55
	Level 4	60.68	17.13	63.74	57.83	58.52	61.02
Method 2	Level 1	73.55	2.97	70.34	76.51	73.51	71.89
	Level 2	74.18	3.1	70.83	77.29	74.29	72.52
	Level 3	72.83	3.4	69.34	76.07	72.9	71.08
	Level 4	88.64	2.88	87.4	89.8	88.89	88.14
Method 3	Level 1	65.88	13.3	62.81	68.73	65.04	63.91
	Level 2	66.23	15.46	62.98	69.23	65.47	64.2
	Level 3	60.22	16.4	58.48	61.84	58.72	58.6
	Level 4	69.02	13.13	74.05	64.32	65.96	69.77
Method 4	Level 1	73.35	2.51	69.73	76.71	73.5	71.56
	Level 2	73.52	2.51	70.01	76.77	73.63	71.77
	Level 3	73.31	2.91	72.24	74.3	72.3	72.27
	Level 4	88.98	2.0	87.81	90.06	89.19	88.5

Table S1.4. Differences between evaluation results between Method 1 and the other methods per preprocessing stage for the Canberra Hospital dataset.

CH	Preprocessing	Accuracy [%]	Acc_std [%]	Sensitivity [%]	Specificity [%]	Precision [%]	F1-score [%]
Method 2	Level 1	31.11	13.63	23.25	40.8	28.88	26.24
	Level 2	16.36	12.77	17.67	14.73	12.42	14.82
	Level 3	12.85	12.5	10.07	16.05	14.69	12.11
	Level 4	27.12	12.2	20.14	35.46	26	23.09
Average difference		21.86	12.77	17.78	26.76	20.5	19.07
Method 3	Level 1	4.05	1.04	3.73	4.44	3.28	3.49
	Level 2	6.9	1.66	7.74	5.86	5.31	6.43
	Level 3	7.31	1.21	4.77	10.23	8.65	6.46
	Level 4	13.12	0.6	10.04	16.8	11.89	11.0
Average difference		7.84	0.22	6.57	9.33	7.28	6.85
Method 4	Level 1	33.11	14.2	24.9	43.24	30.8	28.03
	Level 2	17.49	13.22	18.94	15.69	13.22	15.83
	Level 3	13.71	12.27	10.55	17.34	15.84	12.88
	Level 4	29.6	14.04	23.68	36.68	27.59	25.68
Average difference		23.47	13.43	19.52	28.23	21.86	20.6

Table S1.5. Differences between evaluation results between Method 1 and the other methods per preprocessing stage for the University of New Mexico dataset.

UNM	Preprocessing	Accuracy [%]	Acc_std [%]	Sensitivity [%]	Specificity [%]	Precision [%]	F1-score [%]
Method 2	Level 1	10.2	8.25	12.46	7.94	9.1	10.8
	Level 2	4.86	9.99	5.99	3.73	4.41	5.23
	Level 3	6.22	11.84	7.73	4.49	4.59	6.07
	Level 4	13.74	11.07	12.48	15.24	12.25	12.37
Average difference		8.76	10.29	9.67	7.85	7.59	8.62
Method 3	Level 1	3.32	0.46	2.82	3.82	3.71	3.25
	Level 2	1.12	0.58	1.81	0.43	0.81	1.33
	Level 3	1.19	0.28	0.87	1.56	1.18	1.03
	Level 4	4.57	1.78	4.55	4.58	3.9	4.21
Average difference		2.55	0.54	2.51	2.6	2.4	2.46
Method 4	Level 1	10.2	9.03	11.99	8.4	9.39	10.71
	Level 2	4.72	10.17	5.08	4.36	4.73	4.92
	Level 3	6.61	11.17	7.51	5.57	5.22	6.31
	Level 4	13.3	10.74	12.64	14.09	11.55	12.08
Average difference		8.71	10.28	9.3	8.11	7.72	8.5

Table S1.6. Differences between evaluation results between Method 1 and the other methods per preprocessing stage for the University of Turku dataset.

UTU	Preprocessing	Accuracy [%]	Acc_std [%]	Sensitivity [%]	Specificity [%]	Precision [%]	F1-score [%]
Method 2	Level 1	11.93	10.98	13.57	10.41	12.7	13.17
	Level 2	12.21	11.71	13.91	10.63	13.02	13.5

	Level 3	17.2	10.28	16.22	18.1	18.91	17.53
	Level 4	27.96	14.25	23.66	31.97	30.36	27.12
Average difference		17.32	11.81	16.84	17.78	18.75	17.83
Method 3	Level 1	4.27	0.64	6.04	2.63	4.24	5.19
	Level 2	4.25	0.65	6.06	2.57	4.21	5.19
	Level 3	4.59	2.72	5.36	3.87	4.73	5.05
	Level 4	8.34	4.0	10.31	6.49	7.43	8.75
Average difference		5.36	0.32	6.94	3.89	5.15	6.04
Method 4	Level 1	11.74	11.44	12.95	10.61	12.69	12.84
	Level 2	11.54	12.3	13.08	10.12	12.37	12.76
	Level 3	17.67	10.77	19.12	16.33	18.31	18.72
	Level 4	28.29	15.13	24.07	32.23	30.67	27.48
Average difference		17.31	12.41	17.31	17.32	18.51	17.95

Table S1.7. An overview of evaluation metrics for experiments using averaged samples and per preprocessing stages for the Canberra Hospital dataset.

CH	Preprocessing	Accuracy [%]	Acc_std [%]	Sensitivity [%]	Specificity [%]	Precision [%]	F1-score [%]
Method A	Level 1	55.38	24.66	65.24	43.89	57.56	61.16
	Level 2	50.26	24.55	55.71	43.89	53.67	54.67
	Level 3	65.9	24.76	68.57	62.78	68.25	68.41
	Level 4	61.28	22.64	58.57	64.44	65.78	61.96
Method B	Level 1	70.51	20.36	80.0	59.44	69.71	74.5
	Level 2	63.59	23.05	69.52	56.67	65.18	67.28
	Level 3	73.08	21.03	69.05	77.78	78.38	73.42
	Level 4	73.08	21.84	58.57	90.0	87.23	70.09

Table S1.8. An overview of evaluation metrics for experiments using averaged samples and per preprocessing stages for the University of New Mexico dataset.

UNM	Preprocessing	Accuracy [%]	Acc_std [%]	Sensitivity [%]	Specificity [%]	Precision [%]	F1-score [%]
Method A	Level 1	62.32	21.57	61.07	63.57	62.64	61.84
	Level 2	78.75	16.27	80.71	76.79	77.66	79.16
	Level 3	71.43	15.74	70	72.86	72.06	71.01
	Level 4	60	20.95	60	60	60	60
Method B	Level 1	75.36	17.19	73.57	77.14	76.3	74.91
	Level 2	81.25	13.56	82.86	79.64	80.28	81.55
	Level 3	78.21	18.07	76.07	80.36	79.48	77.74
	Level 4	68.93	16.95	70.36	67.5	68.4	69.37

Table S1.9. An overview of evaluation metrics for experiments using averaged samples and per preprocessing stages for the University of Turku dataset.

UTU	Preprocessing	Accuracy [%]	Acc_std [%]	Sensitivity [%]	Specificity [%]	Precision [%]	F1-score [%]
Method A	Level 1	44.62	25.94	40.0	49.0	42.7	41.3
	Level 2	45.13	25.87	45.79	44.5	43.94	44.85
	Level 3	66.15	22.73	62.63	69.5	66.11	64.32
	Level 4	47.44	23.28	43.16	51.5	45.81	44.44
Method B	Level 1	56.92	22.49	59.47	54.5	55.39	57.36
	Level 2	50.26	25.74	49.47	51.0	48.96	49.21
	Level 3	71.28	20.83	72.63	70.0	69.7	71.13
	Level 4	57.44	21.41	54.74	60.0	56.52	55.61

Table S1.10. Differences between evaluation metrics between Methods A and B per preprocessing stage for the Canberra Hospital dataset when using averaged samples.

CH	Accuracy [%]	Acc_std [%]	Sensitivity [%]	Specificity [%]	Precision [%]	F1-score [%]
Level 1	15.13	4.3	14.76	15.56	12.15	13.34
Level 2	13.33	1.5	13.81	12.78	11.51	12.61
Level 3	7.18	3.74	0.48	15.0	10.13	5.01
Level 4	11.79	0.8	0.0	25.56	21.46	8.12
Average difference	11.86	2.58	7.26	17.22	13.81	9.77

Table S1.11. Differences between evaluation metrics between Methods A and B per preprocessing stage for the University of New Mexico dataset when using averaged samples.

UNM	Accuracy [%]	Acc_std [%]	Sensitivity [%]	Specificity [%]	Precision [%]	F1-score [%]
Level 1	13.04	4.38	12.5	13.57	13.66	13.06
Level 2	2.5	2.7	2.14	2.86	2.61	2.39
Level 3	6.79	2.33	6.07	7.5	7.42	6.72
Level 4	8.93	4.0	10.36	7.5	8.4	9.37
Average difference	7.81	3.35	7.77	7.86	8.02	7.89

Table S1.12. Differences between evaluation metrics between Method 1 and the other methods per preprocessing stage for the University of Turku dataset when using averaged samples.

UTU	Accuracy [%]	Acc_std [%]	Sensitivity [%]	Specificity [%]	Precision [%]	F1-score [%]
Level 1	12.31	3.46	19.47	5.5	12.7	16.06
Level 2	5.13	0.14	3.68	6.5	5.02	4.37
Level 3	5.13	1.9	10.0	0.5	3.59	6.81
Level 4	10.0	1.88	11.58	8.5	10.71	11.17
Average difference	8.14	1.84	11.18	5.25	8.0	9.6

Supplement 2

Reference list of all studies from which data was extracted.

[1]	Amato, N., et al. (2016). "Stroop event-related potentials as a bioelectrical correlate of frontal lobe dysfunction in multiple sclerosis." Multiple Sclerosis and Demyelinating Disorders 1(1): 1-10.
[2]	Artemiadis, A. K., et al. (2018). "Structural MRI correlates of cognitive event-related potentials in multiple sclerosis." Journal of Clinical Neurophysiology **35**(5): 399-407.
[3]	Ates, H., et al. (2001). "The contribution of P300 test for cognitive evaluation in Multiple Sclerosis. [Turkish]." Ondokuz Mayis Universitesi Tip Dergisi **18**(2): 87-96.
[4]	Azcárraga-Guirola, E., et al. (2017). "Electrophysiological correlates of decision making impairment in multiple sclerosis." European Journal of Neuroscience **45**(2): 321-329.
[5]	Bissonnette, J. N., et al. (2023). "An investigation of auditory processing in relapsing–remitting multiple sclerosis." Experimental Brain Research **241**(5): 1319-1327.
[6]	Boose, M. A. and J. L. Cranford (1996). "Auditory event-related potentials in multiple sclerosis." Otology & Neurotology **17**(1): 165-170.
[7]	Casanova-González, M., et al. (1999). "Neurophysiological assessment in patients with clinically defined multiple sclerosis with special reference to P300 wave study." Revista de Neurologia **29**(12): 1134-1137.
[8]	Chinnadurai, S. A., et al. (2016). "A study of cognitive fatigue in Multiple Sclerosis with novel clinical and electrophysiological parameters utilizing the event related potential P300." Multiple Sclerosis and Related Disorders **10**: 1-6.
[9]	Covey, T. J., et al. (2017). "Event-related brain potential indices of cognitive function and brain resource reallocation during working memory in patients with Multiple Sclerosis." Clinical Neurophysiology **128**(4): 604-621.
[10]	Filippi, M., et al. (1992). "Correlations between auditory event-related potentials (ERPs), brain MRI and neuropsychological tests in multiple sclerosis. " Rivista di Neurobiologia 38(5): 369-375.
[11]	Folmer, R. L., et al. (2012). Electrophysiological measures of auditory processing in patients with multiple sclerosis. Seminars in Hearing, Thieme Medical Publishers.
[12]	Gerschlager, W., et al. (2000). "Electrophysiological, neuropsychological and clinical findings in multiple sclerosis patients receiving interferon β-1b: a 1-year follow-up." European neurology **44**(4): 205-209.
[13]	Giesser, B. S., et al. (1992). "Endogenous event-related potentials as indices of dementia in multiple sclerosis patients." Electroencephalography and clinical neurophysiology **82**(5): 320-329.
[14]	Gil, R., et al. (1993). "Event-related auditory evoked potentials and multiple sclerosis." Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section **88**(3): 182-187.

[15]	Gonzalez-Rosa, J. J., et al. (2006). "Differential cognitive impairment for diverse forms of multiple sclerosis." BMC neuroscience **7**: 1-11.
[16]	Gonzalez-Rosa, J. J., et al. (2011). "Cluster analysis of behavioural and event-related potentials during a contingent negative variation paradigm in remitting-relapsing and benign forms of multiple sclerosis." BMC neurology **11**: 1-19.
[17]	Honig, L. S., et al. (1992). "Event-related potential P300 in multiple sclerosis: relation to magnetic resonance imaging and cognitive impairment." Archives of neurology **49**(1): 44-50.
[18]	Kimiskidis, V. K., et al. (2016). "Cognitive event-related potentials in multiple sclerosis: Correlation with MRI and neuropsychological findings." Multiple Sclerosis and Related Disorders **10**: 192-197.
[19]	Kocer, B., et al. (2008). "Evaluating sub-clinical cognitive dysfunction and event-related potentials (P300) in clinically isolated syndrome." Neurological Sciences **29**: 435-444.
[20]	Li, X. (2006). "Combined application of three evoked potentials to improve the diagnosis of multiple sclerosis." Neural Regeneration Research **1**(6): 569-572.
[21]	López-Góngora, M., et al. (2015). "Neurophysiological evidence of compensatory brain mechanisms in early-stage multiple sclerosis." PLoS One **10**(8): e0136786.
[22]	Magnié, M., et al. (2007). "Contribution of cognitive evoked potentials for detecting early cognitive disorders in multiple sclerosis." Revue Neurologique **163**(11): 1065-1074.
[23]	Paolicelli, D., et al. (2021). "Magnetoencephalography and high-density electroencephalography study of acoustic event related potentials in early stage of multiple sclerosis: a pilot study on cognitive impairment and fatigue." Brain Sciences **11**(4): 481.
[24]	Papageorgiou, C. C., et al. (2007). "Changes in LORETA and conventional patterns of P600 after steroid treatment in multiple sclerosis patients." Progress in Neuro-Psychopharmacology and Biological Psychiatry **31**(1): 234-241.
[25]	Piras, M. R., et al. (2003). "Longitudinal study of cognitive dysfunction in multiple sclerosis: neuropsychological, neuroradiological, and neurophysiological findings." Journal of Neurology, Neurosurgery & Psychiatry **74**(7): 878-885.
[26]	Pokryszko-Dragan, A., et al. (2009). "Neuropsychological testing and event-related potentials in the assessment of cognitive performance in the patients with multiple sclerosis—a pilot study." Clinical Neurology and Neurosurgery **111**(6): 503-506.
[27]	Pokryszko-Dragan, A., et al. (2016). "Cognitive performance, fatigue and event-related potentials in patients with clinically isolated syndrome." Clinical Neurology and Neurosurgery **149**: 68-74.
[28]	Pokryszko-Dragan, A., et al. (2016). "Event-related potentials and cognitive performance in multiple sclerosis patients with fatigue." Neurological Sciences **37**: 1545-1556.
[29]	Sandroni, P., et al. (1992). "'Fatigue'in patients with multiple sclerosis: motor pathway conduction and event-related potentials." Archives of neurology **49**(5): 517-524.

[30]	Sfagos, C., et al. (2003). "Working memory deficits in multiple sclerosis: a controlled study with auditory P600 correlates." Journal of Neurology, Neurosurgery & Psychiatry **74**(9): 1231-1235.
[31]	Sundgren, M., et al. (2015). "P300 amplitude and response speed relate to preserved cognitive function in relapsing–remitting multiple sclerosis." Clinical Neurophysiology **126**(4): 689-697.
[32]	Triantafyllou, N., et al. (1992). "Cognition in relapsing-remitting multiple sclerosis: a multichannel event-related potential (P300) study." Acta neurologica scandinavica **85**(1): 10-13.
[33]	Uysal, U., et al. (2014). "Contingent negative variation is associated with cognitive dysfunction and secondary progressive disease course in multiple sclerosis." Journal of Clinical Neurology **10**(4): 296-303.
[34]	Van Dijk, J., et al. (1992). "What is the validity of an “abnormal” evoked or event-related potential in MS?: Auditory and visual evoked and event-related potentials in multiple sclerosis patients and normal subjects." Journal of the neurological sciences **109**(1): 11-17.
[35]	Vazquez-Marrufo, M., et al. (2008). "Abnormal ERPs and high frequency bands power in multiple sclerosis." International Journal of Neuroscience **118**(1): 27-38.
[36]	Vázquez-Marrufo, M., et al. (2009). "Cognitive evoked potentials in remitting-relapsing and benign forms of multiple sclerosis." Revista de Neurologia **48**(9): 453-458.
[37]	Vázquez-Marrufo, M., et al. (2014). "Neural correlates of alerting and orienting impairment in multiple sclerosis patients." PLoS One **9**(5): e97226.
[38]	Vázquez-Marrufo, M., et al. (2019). "Altered individual behavioral and EEG parameters are related to the EDSS score in relapsing-remitting multiple sclerosis patients." PLoS One **14**(7): e0219594.
[39]	Waliszewska-Prosół, M., et al. (2018). "The relationship between event-related potentials, stress perception and personality type in patients with multiple sclerosis without cognitive impairment: A pilot study." Advances in Clinical and Experimental Medicine **27**(6): 787-794.
[40]	Zeng, Q., et al. (2017). "Cognitive impairment in Chinese IIDDs revealed by MoCA and P300." Multiple Sclerosis and Related Disorders **16**: 1-7.

Supplement 3

Overview of 62 experiments from which data was extracted. Reference number refers to the reference in Supplement 1. In experiments where the subtype is 'MS', the researchers either did not specify a subtype or added groups of different subtypes together.

A	Reference
B	Paradigm
C	Subtype of patients
D	N patients
E	N female patients
F	N male patients
G	Mean age patients
H	Standard deviation age patients
I	Mean disease duration patients
J	Standard deviation disease duration patients
K	Mean EDSS score patients
L	Standard deviation EDSS score patients
M	N controls
N	N female controls
O	N male controls
P	Mean age controls
Q	Standard deviation age controls

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Stroop task	MS	29	15	14	40	8					16	9	7	36	10
2	oddball	RRMS	58	41	17	41.5	10.5	11.64	7.08	2.2	1.8	51	41	10	38.7	9.2
3	oddball	MS	50	33	17			6.9	5.2	3.34	1.53	50	25	25	36.71	9.2
4	modified Iowa Gambling Task	MS	16	10	6	39.4	12.6	8.9	7			19	13	6	38.3	10.3
5	mismatch negativity paradigm	RRMS	13	13	0	45.6	13.1			2	1.1	13	13	0	46.5	12.6
6	oddball	MS	30	25	5	45.1		11.8				30	25	5	45	
7	oddball	MS	26	22	4	38.03	11.44	9.46	7.98			26			`	
8	oddball	MS	50	35	15	33.6	10.6	6	7.4	4.6	1.9	50	34	16	33.6	9.6
8	oddball	MS	50	35	15	33.6	10.6	6	7.4	4.6	1.9	50	34	16	33.6	9.6
9	0-back task	MS	25	14	11	44.12	8.78	9.05	7.93	2.79	1.51	18	12	6	45.89	9.51
9	1-back task	MS	25	14	11	44.12	8.78	9.05	7.93	2.79	1.51	18	12	6	45.89	9.51
9	2-back task	MS	25	14	11	44.12	8.78	9.05	7.93	2.79	1.51	18	16	6	45.89	9.51
10	oddball	MS	54	39	15	39.7	9.2	11.1	7.2	3.9	1.6	57				
11	oddball	MS	20	15	5	51	9.8	17.7	11.6	3.8	1.7	20	12	8	45.9	12.5
12	oddball	RRMS	14	11	3	37.5	7.74	10	7.29	3.14	0.74	14				
13	double oddball	MS	12	9	3	36.5	9.53	9.55	6.57	4.8	1.9	7	4	3	32	5
13	double oddball	MS	12	9	3	36.5	9.53	9.55	6.57	4.8	1.9	7	4	3	32	5
14	oddball	MS	101	72	29	39.9	11.2	7.78				71	39	32	42.1	13.5
15	Posner paradigm	RRMS	17	12	5	38.88	9.04	4.67	4.13	1.6	0.9	18	15	3	36.54	8.73
15	Posner paradigm	BMS	10	6	4	42.3	7.21	12.09	4.77	1.6	0.9	18	15	3	36.54	8.73
16	Posner paradigm	RRMS	17	12	5	38.88	9.04	4.67	4.13	1.6	0.9	18	15	3	36.54	8.73
16	Posner paradigm	BMS	10	6	4	42.3	7.21	12.09	4.77			18	15	3	36.54	8.73
17	oddball	MS	31	20	11	40	10.7	10.2	6.5	4.92	2.2	32	15	17	33.7	13
18	oddball	MS	59			37.82	1.38	6.76	5.3	3.77	2.14	26			41.42	15.39
19	oddball	CIS	20	16	4	27.35	5.4	4.92	5.31			20	14	6	27.65	5.75
20	oddball	MS	25	8	17	39						20	5	15		
21	modified flanker stop task	RRMS	27	16	11	34.5	7.5	1.3	0.8	0.87	0.91	31	19	12	37.5	8.9
21	3-stimulus oddball	RRMS	27	16	11	34.5	7.5	1.3	0.8	0.87	0.91	31	19	12	37.5	8.9

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
22	oddball	MS	20			38.35	5.73	3.64	1.04			10			32.6	8.2
22	oddball	MS	20			38.35	5.73	3.64	1.04			10			32.6	8.2
22	oddball	MS	10			34.9	3.7	3.47	1.1			10			32	4.5
22	oddball	MS	10			34.9	3.7	3.47	1.1			10			32	4.5
23	oddball	MS	16	14	2	36.5	7.5	0.95	1.66	2	1	19	13	6	36	16
24	digit span Wechsler test	RRMS	18	12	6	34.8	9.73	4.12	5.675	3.11	1.15	16	10	6	32	6.07
25	oddball	RRMS	12	10	2	32.5	8	7.7	6.3			21	15	6		
25	oddball	RRMS	12	10	2	32.5	8	7.7	6.3			21	15	6		
26	oddball	MS	21	17	4	31.09		4.09		2.8		21	15	6		
27	oddball	CIS	44	27	17	31.4				1.39		45	32	13	30.7	
28	oddball	MS	86	62	24	39.55		8.57	8.4	3.03	1.42	40	28	12	38.8	
29	oddball	MS	10	8	2	39.3	4.4									
29	short-term verbal memory test	MS	10	8	2	39.3	4.4									
29	short-term verbal memory test	MS	10	8	2	39.3	4.4									
30	digit span Wechsler test	RRMS	22	14	8	40.77	9.38	9.31	9.62	4.29	1.92	20			38.7	7.1
31	choice reaction task	RRMS	72	51	21	37.9	10	9.3	6.5	2.7	1.5	89	51	38	38.2	11.5
31	choice reaction task	RRMS	72	51	21	37.9	10	9.3	6.5	2.7	1.5	89	51	38	38.2	11.5
32	oddball	RRMS	24	8	16	34.9	10.6	6.5	7	1.9	0.4	24	11	13	34.4	9.4
32	oddball	RRMS	23	7	16	36.6	9.9	7.4	6.4	3.7	0.8	24	11	13	34.4	9.4
32	oddball	RRMS	47	15	32	35.7	10.2	6.9	6.7	2.8	1.1	24	11	13	34.4	9.4
33	Sensorimotor integration paradigm	RRMS	50	33	17	39.52	10.41	9.08	6.65	1.83	1.07	40	27	13	37.03	10.43
33	Sensorimotor integration paradigm	SPMS	16	11	5	43.81	7.47	13.94	5.8	5.78	1.08	40	27	13	37.03	10.43
34	oddball	MS	30	15	15	48.4	13.5	17.1	10	3.9	2.2	19	13	6	44.9	14.2
34	oddball	MS	30	15	15	48.4	13.5	17.1	10	3.9	2.2	19	13	6	44.9	14.2
34	oddball	MS	30	15	15	48.4	13.5	17.1	10	3.9	2.2	19	13	6	44.9	14.2
34	oddball	MS	30	15	15	48.4	13.5	17.1	10	3.9	2.2	19	13	6	44.9	14.2
35	oddball	RRMS	11	7	4	37	8					11				

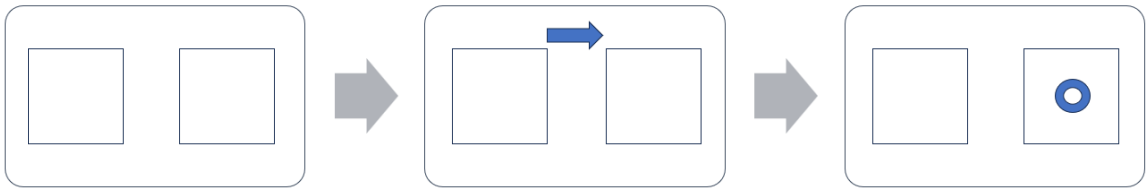
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
36	oddball	RRMS	17	12	5	37.7	7.7	5.5	4.4	1.6	0.8	19	12	7	33.3	9.1
36	oddball	BMS	9	6	3	38.7	5.6	12.4	4.4	1.94	1.2	19	12	7	33.3	9.1
37	Attention Network Test	RRMS	26	16	10	34.42	6	7.15	4.35	2.4		26	11	15	30.31	9.3
38	oddball	RRMS	20	7	13	41.5	9.86					10	4	6	41.6	10.12
39	oddball	RRMS	30	26	4	34.9		6.2	4.3	1.8		26	22	4	34.6	
40	oddball	MS	24	18	6	36.54	10.08	4.56	4.68	3.17	1.45	13	10	3	31.62	11.42
40	oddball	CIS	11	6	5	40.36	8.58	0.38	0.62	2.95	1.81	13	10	3	31.62	11.42

Supplement 4

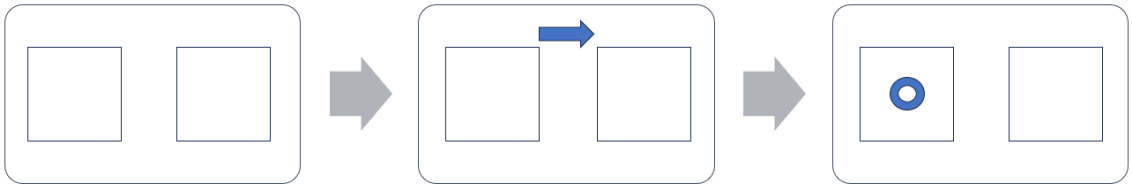
General examples of the oddball paradigm, Posner paradigm, n-back task, and choice reaction experiment.

Posner paradigm

a)

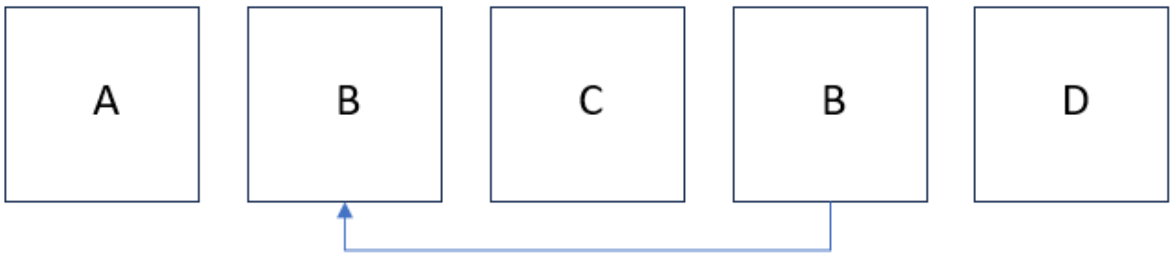


b)



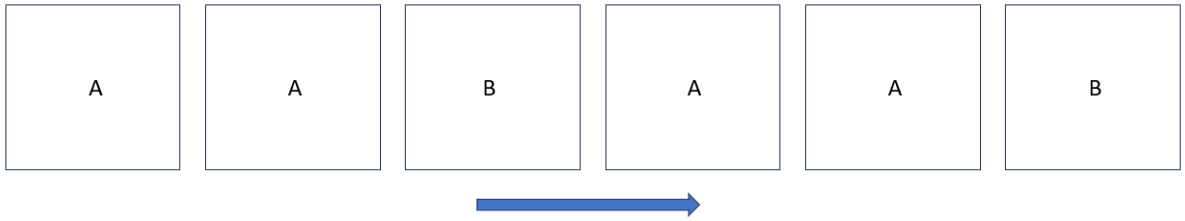
The participant is presented with either a valid trial (a) or where the cueing arrow points to the location of the following stimulus, or and invalid trial (b), where the arrow points in the opposite direction. If a valid stimulus is presented, the participant presses the button corresponding to that side, and if an invalid trial is presented they withhold a response.

N-back task



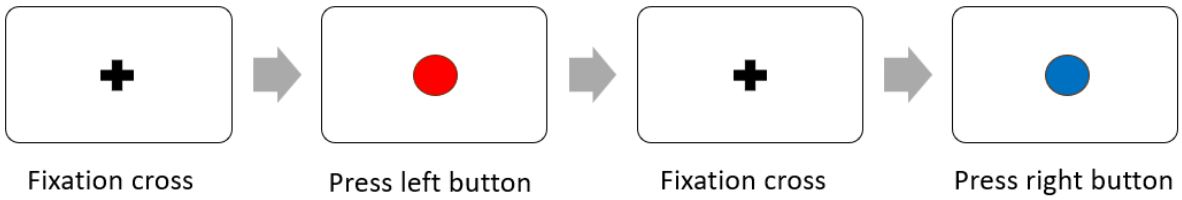
In the n-back task, a participant compares the current stimulus to a stimulus presented n stimuli back. In the case presented here, the current stimulus is a B which is being compared to the stimulus 2 stimuli back. In this case, the participant should press a button that indicates the stimuli are the same. If this was not the case, the participant should push another button.

Oddball experiment



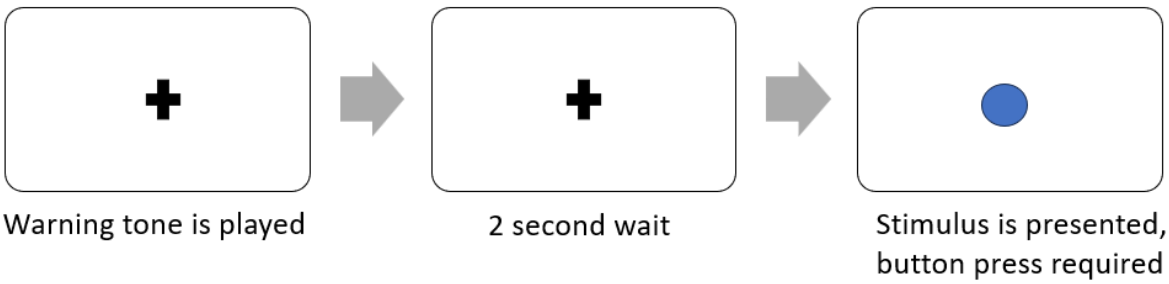
In an oddball experiment, the participant is presented with a string of stimuli, of which the majority are standard stimuli, the letter A in this example, and the minority 'oddballs', the letter B in this example. The participant has to either respond to the oddballs through, for example, a button press, or count them and report the number afterwards. The example here is a visual experiment, but the stimuli can also be substituted for auditory stimuli of different frequencies. The paradigm can also be expanded by including more standard or oddball stimuli.

Choice reaction experiment



In the choice reaction experiment, the participant is asked to respond to the occurrence of a stimulus by, for example, pressing the button that corresponds to that particular stimulus.

Sensorimotor experiment



In a sensorimotor integration experiment, the participant is presented with a warning stimulus, in this case a warning tone, which indicates a stimulus that requires a response will be presented soon. After a small wait, in this case 2 seconds, a stimulus is presented, and the participant responds. In this experiment, the wait time can be made variable, and the stimuli can be expanded to include stimuli that require a response or require the participant to withhold response. This is known as a Go/No-Go paradigm

Supplement 5

An overview of experimental parameters for the 62 experiments reported on in this review.

Ref.	Paradigm	Subtype of patients	Stimuli	Response	N stimuli	N electrodes	References	Low frequency filter	High frequency filter	Sampling rate
1	Stroop task	MS	visual	Button press	120	29	binaural	DC	50Hz	250Hz
2	oddball	RRMS	visual	Button press	300	4	binaural	0.01Hz	35Hz	285Hz
3	oddball	MS	auditory		200					
4	modifed Iowa Gambling Task	MS	visual	Button press	300	19	binaural	0.5Hz	30Hz	500Hz
5	mismatch negativity paradigm	RRMS	auditory		615	32	nose	0.01Hz	20Hz	500Hz
6	oddball	MS	auditory	Button press	200	1	binaural	1Hz	30Hz	
7	oddball	MS	visual		125	4				
8	oddball	MS	auditory		300	1	monaural	1Hz	25Hz	
8	oddball	MS	auditory		900	1	monaural	1Hz	25Hz	
9	0-back task	MS	visual	Button press	150	16	binaural	0.1Hz	25Hz	250Hz
9	1-back task	MS	visual	Button press	150	16	binaural	0.1Hz	25Hz	250Hz
9	2-back task	MS	visual	Button press	150	16	binaural	0.1Hz	25Hz	250Hz
10	oddball	MS	auditory		100	16		1Hz	100Hz	
11	oddball	MS	auditory		250	15	nose			1000Hz
12	oddball	RRMS	auditory	finger lift	420	4	linked mastoid	DC	100Hz	250Hz
13	double oddball	MS	auditory	finger lift	142	14	binaural	0.1Hz	35Hz	
13	double oddball	MS	auditory	finger lift	142	14	binaural	0.1Hz	35Hz	
14	oddball	MS	auditory		200	1	linked mastoid	1Hz	100Hz	
15	Posner paradigm	RRMS	visual	Button press	200	13	linked mastoid	0.01Hz	100Hz	500Hz
15	Posner paradigm	BMS	visual	Button press	200	13	linked mastoid	0.01Hz	100Hz	500Hz
16	Posner paradigm	RRMS	visual	Button press	200	13	linked mastoid	0.01Hz	100Hz	500Hz
16	Posner paradigm	BMS	visual	Button press	200	13	linked mastoid	0.01Hz	100Hz	500Hz

Ref.	Paradigm	Subtype of patients	Stimuli	Response	N stimuli	N electrodes	References	Low frequency filter	High frequency filter	Sampling rate
17	oddball	MS	auditory			20	linked cheeks	1Hz	70Hz	500Hz
18	oddball	MS	auditory			2	binaural	0.1Hz	50Hz	250Hz
19	oddball	CIS	auditory			3	binaural	0.1Hz	50Hz	
20	oddball	MS	auditory			1	mastoid			
21	modified flanker stop task	RRMS	visual	Button press	200	19	linked mastoid	0.1Hz	35Hz	250Hz
21	3-stimulus oddball	RRMS	auditory	Button press		19	linked mastoid	0.1Hz	35Hz	250Hz
22	oddball	MS	visual	Button press	150	28	binaural	0.01Hz	30Hz	256Hz
22	oddball	MS	auditory	Button press		28	binaural	0.01Hz	30Hz	256Hz
22	oddball	MS	visual	Button press	150	28	binaural	0.01Hz	30Hz	256Hz
22	oddball	MS	auditory			28	binaural	0.01Hz	30Hz	256Hz
23	oddball	MS	auditory	Button press	300	64	linked mastoids	0.1Hz	35Hz	1024Hz
24	digit span Wechsler test	RRMS	auditory		26	15	both earlobes	0.05Hz	35Hz	500Hz
25	oddball	RRMS	visual			30	binaural	1Hz	70Hz	
25	oddball	RRMS	auditory			30	binaural	1Hz	70Hz	
26	oddball	MS	auditory			3	binaural	0.3Hz	70Hz	
27	oddball	CIS	auditory			3	binaural	0.3Hz	70Hz	
28	oddball	MS	auditory			3	binaural	0.3Hz	70Hz	
29	oddball	MS	auditory	Button press	300	7	binaural	0.1Hz	100Hz	200Hz
29	short-term verbal memory test	MS	auditory	Button press	40	7	binaural	0.1Hz	100Hz	200Hz
29	short-term verbal memory test	MS	auditory							
30	digit span Wechsler test	RRMS	auditory	Verbal	26	15	binaural	0.05Hz	35Hz	500Hz
31	choice reaction task	RRMS	auditory	Button press	60	23	nose	0.5Hz	70Hz	256Hz
31	choice reaction task	RRMS	visual	Button press	240	23	mastoid	0.5Hz	70Hz	512Hz
32	oddball	RRMS	auditory			16	linked cheeks	1Hz	30Hz	128Hz
32	oddball	RRMS	auditory			16	linked cheeks	1Hz	30Hz	128Hz
32	oddball	RRMS	auditory			16	linked cheeks	1Hz	30Hz	128Hz
33	Sensorimotor integration paradigm	RRMS	multimodal	Button press		3	linked mastoid	0.03Hz	100Hz	

Ref.	Paradigm	Subtype of patients	Stimuli	Response	N stimuli	N electrodes	References	Low frequency filter	High frequency filter	Sampling rate
33	Sensorimotor integration paradigm	SPMS	multimodal	Button press		3	linked mastoid	0.03Hz	100Hz	
34	oddball	MS	auditory	Button press	250	3	binaural	1Hz	40Hz	512Hz
34	oddball	MS	auditory	Button press	300	3	binaural	1Hz	40Hz	512Hz
34	oddball	MS	visual	Button press	100	3	binaural	1Hz	40Hz	512Hz
34	oddball	MS	visual	Button press	100	3	binaural	1Hz	40Hz	512Hz
35	oddball	RRMS	auditory		200	13	linked mastoid	0.01Hz	100Hz	250Hz
36	oddball	RRMS	auditory		200	13	left mastoid	0.01Hz	100Hz	250Hz
36	oddball	BMS	auditory	Button press	200	13	left mastoid	0.01Hz	100Hz	250Hz
37	Attention Network Test	RRMS	visual	Button press	144	58	average	0.01Hz	100Hz	500hz
38	oddball	RRMS	visual	Button press	200	64	average	0.01Hz	100Hz	500Hz
39	oddball	RRMS	auditory			3	binaural	0.3Hz	70Hz	
40	oddball	MS	auditory	finger lift		12	binaural			
40	oddball	CIS	auditory	finger lift		12	binaural			

Supplement 6

An overview of neuropsychological tests used in each study.

Reference	Neuropsychological tests
1	Stroop test, Tower of Hanoi, Dual task, Wisconsin Card Sorting, semantic and phonemic verbal fluency tests
2	Symbol Digits Modalities Test (SDMT), the California Verbal Learning Test II (CVLT-II), and the Brief Visuospatial Memory Test-Revised (BVMT-R)
3	Cross Cultural Cognitive Examination (CCCE)
4	Backwards Digit Span Test, the Wisconsin Card Sorting Test (WCST), and the Tower of London Test (TOL)
5	Symbol digit modalities test (SDMT)
6	
7	
8	modified Stroop test, modified SDMT, serial addition test
9	Rao adaptations of the Symbol Digit Modalities Test (SDMT), oral version, and the Paced Auditory Serial Addition Test (PASAT), National Adult Reading Test (NART)
10	
11	
12	Verbal Selective Reminding Test (SRT), a concentration endurance test, the 7/24 Spatial Recall Test (SRT 7/24), the Paced Auditory Serial Addition Test (PASAT), Verbal fluency test
13	Blessed Test, Benton Visual Retention Test (recognition version), Buschke's Selective Reminding Test and the Category Retrieval Test
14	
15	
16	
17	Mini-Mental State Examination (MMSE), the Dementia Score (DS), and the Information-Memory-Concentration Test (IMCT)
18	Stroop Test, Trail Making Test, Raven Progressive Matrices Sets: A, B, C, D and E for adults, Wisconsin Card Sorting Test, logical memory and subtest combinational memory subtests of the Wechsler Memory Scale Form II, Verbal Fluency Test
19	Mini-Mental State Examination (MMSE), Serial Digit Learning Test (SDLT), Boston Naming Test (BNT), Auditory Verbal Learning Test (AVLT), Non-verbal Cancellation Test (NVCT), Benton Judgement of Line Orientation Test (BJLOT), Rey Complex Figures Test (RCFT), Benton Facial Recognition Test (BFRT), 10-Point Clock Drawing Test, Stroop Interference Test (SIT), Paced Auditory Serial Addition Test (PASAT), Controlled Oral Word Association Test (COWAT), Standard Raven's Progressive Matrices Test (SRPMT)
20	
21	Selective Reminding Test (SRT), 10/36 Spatial Recall Test, Symbol Digit Modalities Test (SDMT), Paced Auditory Serial Addition Test (PASAT), Word List Generation (WLG), phonetic verbal fluency test
22	Mini-Mental State Examination (MMSE), Paced Auditory Serial-Addition Task (PASAT), formal and categorical verbal fluency tests of the Brief Repeatable Battery (BRB-N), Stroop test, Trail Making Test (TMT), Selective Reminding Test (SRT), 10/36 Spatial Recall Test, Boston Naming Test, code subtests and WAIS-R cubes and the visual-construction test of the BEC 96
23	Selective Reminding Test—Long-Term Storage (SRT-LTS), SRT—Consistent Long-Term Retrieval (SRT-CLTR), SRT—Delayed recall (SRT-D), 10/36 Spatial Recall Test, Spatial Recall Test—Delayed recall, Symbol Digit Modality Test (SDMT) 3 and 5s Paced Auditory Serial Addition Tests (PASAT-3 and PASAT-5), Word List Generation (WLG), Trail Making Test (TMT)
24	
25	Italian version of Wechsler adult intelligence scale (WAIS), Raven's progressive coloured matrices, visual search test, corsi span, digit span from WAIS verbal subtest, verbal supraspan—selective reminding test, learning spatial supraspan, aphasia test
26	Wechsler Adult Intelligence Scale (WAIS-r), the Rey Auditory Verbal Learning Test (AVLT 1 and 2), and the Trail Making Test (TMT)
27	Selective Verbal Reminding Test (SVRT), Spatial Recall Test (SpaRT), Symbol Digit Modalities Test (SDMT), Paced Auditory Serial Additive Test (PASAT) and Word List Generation (WLG)
28	Selective Verbal Reminding Test (SVRT), Spatial Recall Test (SpaRT), Symbol Digit Modalities Test (SDMT), Paced Auditory Serial Additive Test (PASAT) and Word List Generation (WLG)
29	
30	
31	Benton visual retention test, Vocabulary test, controlled oral word association test, Digit span test, Trail making test, Stroop test, Block design test, Digit symbol coding test, Symbol search test

Reference	Neuropsychological tests
32	
33	Paced Serial Addition Test (PASAT), Stroop test, Controlled Oral Word Association Test (COWAT), Rey Auditory Verbal Learning Test (RAVLT)
34	
35	Stroop test and Symbol Digit Modality test (SDMT)
36	
37	Paced Auditory Serial Addition Test, 3 seconds (PASAT-3s) and Symbol Digit Modality Test (SDMT)
38	Paced Auditory Serial Addition Test, 3 seconds (PASAT-3s) and Symbol Digit Modality Test (SDMT)
39	Symbol Digit Modality Test (SDMT)
40	Mini-Mental State Examination (MMSE), Montreal Cognitive Assessment (MoCA)