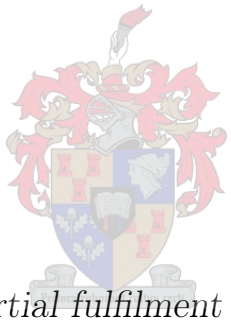
EEG artefact removal methods compared using semi-synthetic data for the analysis of ADHD EEG-data

by

Wadda Benjamin du Toit



Thesis presented in partial fulfilment of the requirements for the degree of Master of Engineering (Mechatronic) in the Faculty of Engineering at Stellenbosch University

Supervisor: Dr. M. P. Venter
Co-supervisor: Prof. D. J. van den Heever

December 2021

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: ..... December 2021. .....

Abstract

EEG artefact removal methods compared using semi-synthetic data for the analysis of ADHD EEG-data

W B du Toit

Department of Mechanical and Mechatronic Engineering,
University of Stellenbosch,
Private Bag X1, Matieland 7602, South Africa.

Thesis: MEng (Mech)

December 2021

The electroencephalogram (EEG) is a measure of the biological electrical signals that reflect the brain's functional state, allied to a person's mental condition and nervous system activity. EEG is, however, an extremely weak signal, thus easily contaminated by artefacts. Artefacts are activities that do not directly originate from the brain but are still present in the EEG data. Artefacts significantly complicate, distort and obscure the analysis of the data originating exclusively from the brain.

This thesis aimed to identify, test, and automate robust methods for removing EEG artefacts. This aim was achieved by developing a simulated dataset, based on real 'clean' EEG and artefacts, namely a semi-synthetic dataset with significant variation in types, forms, intensity and combinations of physiological artefacts. This dataset was used to test the effectiveness and efficiency of three blind source separation (BSS) techniques, namely Extended Infomax, second-order blind identification (SOBI), canonical correlation analysis (CCA) and a developed auto threshold method. Additionally, the BSS methods were fully automated, using a novel but simple approach, to prevent the preprocessing of EEG data from becoming a bottleneck for data analysis.

The semi-synthetic dataset consisted of 'clean' EEG datasets, which was contaminated separately and together by electrocardiography (ECG), electrooculography (EOG) and electromyography (EMG) artefacts varied for each EEG dataset. This thesis compared the time-series, topography, amplitude spectra, and the signal-to-noise ratio (SNR) characteristics of real 'clean' EEG and artefacts, found in the relevant literature, to the characteristics of the semi-synthetic dataset developed, for the purpose of validation.

BSS techniques were used because they were the most popular for artefact removal. Furthermore, independent component analysis (ICA) was identified as the most popular BSS subcategory. The two most popular ICA methods, namely Extended Infomax and SOBI, were tested along with CCA. In addition, an auto threshold method, using the standard deviation of the data, was created and tested on the data.

The effectiveness of the cleaning methods was determined and compared using the SNR increase of the contaminated data. The efficiency was determined and compared using the average time each BSS method took to identify the components and for the auto threshold method took identify all the artefact ranges. The SNR and time results were further analysed using boxplots and t-tests.

With the removal of EOG artefacts, CCA was the most effective. Extended Infomax was the most effective with the removal of EMG artefacts. With the removal of ECG artefacts, SOBI outperformed the other methods in terms of effectiveness. Furthermore, when combining all three artefacts, the effectiveness of the BSS methods was less distinguishable, having closer P values, with Extended Infomax being the most effective. The auto threshold method showed comparable effectiveness results to the BSS methods, but in terms of efficiency, it was about 10, 20 and 100 times faster than CCA, Extended Infomax and SOBI, respectively. Concerning the automation of the BSS methods, the fully automated and semi-automatic Extended Infomax methods showed no significant difference, based on a t-test, in the effectiveness of EOG removal.

Uittreksel

Metodes vir die verwydering van EEG-artefakte vergelyk met behulp van semi-sintetiese data vir die ontleding van ADHD EEG-data

W B du Toit

Departement Meganiese en Megatroniese Ingenieurswese,
Universiteit van Stellenbosch,
Privaatsak X1, Matieland 7602, Suid Afrika.

Tesis: MIng (Meg)

Desember 2021

Die elektroencefalogram (EEG) is 'n maatstaf van die biologiese elektriese seine wat die brein se funksionele toestand weerspieël, wat verband hou met 'n persoon se geestestoestand en aktiwiteit van die senuweestelsel. EEG is egter 'n uiters swak sein, wat dus maklik deur artefakte gekontamineer kan word. Artefakte is aktiwiteite wat nie direk uit die brein afkomstig is nie, maar steeds in die EEG-data voorkom. Artefakte bemoeilik, verdraai en verduister die analise van die data wat uitsluitlik uit die brein afkomstig is.

Hierdie tesis het ten doel gehad om robuust metodes vir die verwydering van EEG-artefakte te identifiseer, te toets en te outomatiseer. Hierdie doel is bereik deur 'n gesimuleerde datastel te ontwikkel, gebaseer op werklike 'skoon' EEG en artefakte, naamlik 'n semi-sintetiese datastel met 'n beduidende variasie in tipes, vorms, intensiteit en kombinasies van fisiologiese artefakte. Hierdie datastel is gebruik om die doeltreffendheid en tyddoeltreffendheid van drie Blinde bron skeiding (BBS) tegnieke te toets, naamlik Uitgebreide Infomax, tweede-orde blinde identifikasie (TOBI), kanonieke korrelasie-analise (KKA) en 'n ontwikkelde outomatiese drumpel metode. Die BBS-metodes is volledig geoutomatiseer, met 'n nuwe maar eenvoudige benadering, om te voorkom dat die voorafverwerking van EEG-data 'n 'bottleneck' word vir data-analise.

Die semi-sintetiese datastel bestaan uit 'skoon' EEG-datastelle, wat afsonderlik en saam gekontamineer is deur elektrokardiografie (EKG), elektrookulografie (EOG) en elektromyografie (EMG) artefakte wat vir elke EEG-datastel gevarieer is. Hierdie tesis vergelyk die tydreekse, topografie, amplitude-spektra

en die sein-tot-geraas-verhouding (SGV) kenmerk van die 'skoon' EEG en artefakte in die relevante literatuur met die kenmerke van die semi-sintetiese datastel wat ontwikkel is vir validasie doeleindes.

BBS-tegnieke is gebruik omdat dit die gewildste is vir die verwydering van artefakte. Verder is onafhanklike komponent analise (OKA) geïdentifiseer as die gewildste BBS-subkategorie. Die twee gewildste OKA-metodes, naamlik Extended Infomax en TOBI, is saam met KKA getoets. Boonop is 'n outomatiese drumpel metode, met behulp van die standaardafwyking van die data, geskep en getoets op die data.

Die doeltreffendheid van die skoonmaakmetodes is bepaal en vergelyk met behulp van SGV-toename van die gekontamineerde data. Die doeltreffendheid is bepaal en vergelyk deur gebruik te maak van die gemiddelde tyd wat elke BBS-metode geneem het om die komponente te identifiseer, en vir die outomatiese drumpelmetode al die artefakreekse te identifiseer. Die SGV- en tydresultate is verder geanaliseer met behulp van boksplotte en t-toetse.

Met die verwydering van EOG-artefakte was KKA die doeltreffendste. Uitgebreide Infomax was die doeltreffendste met die verwydering van EMG-artefakte. Met die verwydering van EKG-artefakte het TOBI beter gevaar as die ander metodes wat doeltreffendheid betref. Boonop was die doeltreffendheid van die BBS-metodes by die kombinasie van al drie artefakte minder onderskeibaar, met nader P-waardes, met Extended Infomax die doeltreffendste. Die outomatiese drumpel metode het vergelykbare doeltreffendheid resultate getoon met die BBS-metodes, maar wat tyddoeltreffendheid betref, is dit ongeveer 10, 20 en 100 keer vinniger as onderskeidelik KKA, Extended Infomax en TOBI. Met die BBS-metodes outomatiser, het die volledig outomatiese en semi-outomatiese uitgebreide Infomax-metodes geen beduidende verskil getoon op grond van 'n t-toets in die doeltreffendheid van EOG verwydering nie.

Acknowledgements

I would like to express my sincere gratitude to the following people: My Supervisor, Dr. M.P. Venter, for taking me under his wing under short notice and providing invaluable guidance. My Co-supervisor, Prof. D.J. van den Heever, who understood where the support mattered most and provided it without question. To Louise Van der Westhuizen, who exceeded her role as an academic mentor, to become my saviour in dire straits. To my father, who never doubted me for a second. Finally, if only tears could write, I would be able to express my gratitude to my mother.

Dedications

This thesis is dedicated to my family with their unquivering love and support

Contents

List of Figures

List of Tables

Nomenclosure

Abbreviations

ADHD	Attention-Deficit Hyperactivity Disorder
AMI	Auto Mutual Information
BCI	Brain-Computer Interface
BSS	Blind Source Separation
CCA	Canonical Correlation Analysis
CLT	Central Limit Theorem
CNS	Central Nervous System
EC	Eyes Closed
EO	Eyes Open
ECG	Electrocardiography
EEG	Electroencephalography
EMD	Empirical Mode Decomposition
EMG	Electromyography
EOG	Electrooculography
ERD	Event-Related Desynchronisation
ERF	Event-Related Field
ERP	Event-Related Potential
HEOG	Horizontal-EOG
HOS	Higher-Order Statistics
ICA	Independent Component Analysis
IQR	Interquartile range
LAMICA	Lagged Auto-Mutual Information clustering
LPM	Linear Programming Machine
NFB	Neurofeedback
NMD	Nonlinear Mode Decomposition
OCD	Obsessive-Compulsive Disorder
PCA	Principal Component Analysis
PDR	Posterior Dominant Rhythm

PNS	Peripheral Nervous System
PSD	Power Spectral Density
REM	Rapid Eye Movement
ROI	Regions of Interest
SOBI	Second-Order Blind Identification
SNR	Signal to Noise Ration
SOS	Second-Order Statistics
SWT	Synchrosqueezed Wavelet Transform
TDSEP	Temporal Decorrelation Source Separation
VEOG	Vertical-EOG
WNN	Wavelet Neural Networking

Symbols

A	Mixing Matrix
CMD	Combined Artefacts
CMD_EEG	Combined Artefacts
m	Gradient
P	Percentage
S	Estimated Sources
V	Voltage
W	Unmixing Matrix
X	Time-Domain Signals

Greek letters

α	Alpha
β	Beta
δ	Delta
γ	gamma
θ	Theta
μ	Micro

Superscripts

a	Artefacts
c	Contaminated EEG
k	Cleaned EEG
s	Clean EEG

Subscripts

f	Frequency Index
i	Channel Index
j	Sample Index
M	Total Components
n	Total Channels
N	Total Samples
15	15% Index

Chapter 1

Introduction

1.1 Problem statement

The electroencephalogram (EEG) is a measure of the biological electrical signals that reflect the brain's functional state. These measurements can be allied to a person's mental condition and nervous system activity. EEG research is mainly used to evaluate and study neurological disorders and functions of the brain. The research is traditionally conducted in clinical and laboratory environments. EEG is an essential signal in aiding researchers and medical practitioners to extract vital information for diagnosis and monitoring a patient's health relating to different brain conditions [1–5]. It has been repeatedly demonstrated that EEG signals are closely related to cerebral diseases, such as cerebrovascular diseases, migraines, and epilepsy [5]. A systematic review by McVoy et al. [6] shows that most research involving EEG is used as a diagnostic tool rather than a treatment tool and that almost half of the diagnostic tool research is dedicated to the diagnosis of attention-deficit hyperactivity disorder (ADHD).

EEG plays an essential role in identifying brain activity and behaviour. It is, however, a weak signal, thus easily contaminated by electrical artefacts. In this context, artefacts are activities that do not directly from the brain but are still present in the measured EEG data [5]. Some artefacts can imitate cognitive or pathological activity and become a significant problem, resulting in misleading visual interpretations and diagnosis of diseases such as sleep disorders, Alzheimer's disease, etc. [2, 3].

Artefacts complicate, distort and obscure the analysis of the data originating exclusively from the brain [3, 7, 8]. Artefacts originate from various sources and significantly and detrimentally affect EEG due to large variations in temporal and spectral contamination [9–12]. There are numerous types of artefacts, each with its own characteristics, each of which must be addressed independently. Physiological artefacts such as electrocardiography (ECG), electrooculography (EOG) and electromyography (EMG) have a significant effect

on EEG and cause severe problems for EEG analysis [10–12]. Furthermore, due to the weakness of EEG signals, contamination from extraneous sources such as electrical, magnetic, sound, optical and electromagnetic waves etc. are limitless [2, 9, 13].

The extent of artefact contamination depends on the acquisition devices, the setup and participant compliance to preventative guidelines [3, 14]. Participants that are non-compliant to artefact preventive guidelines, such as those with ADHD, increase the likelihood of physiological artefacts contaminating the EEG data [3, 10, 15].

Traditionally, EEG research and medical applications involve the use of expensive and medical-grade EEG equipment and procedures that emphasise preventative measures to reduce artefacts [1, 16]. Furthermore, traditional EEG research requires complex equipment, assembly, and extensive application time that inadvertently inflicts discomfort and distress on the participants [16]. These characteristics and the high cost of medical-grade EEG equipment decrease participation attractiveness and use in research [16]. In response to this, researchers are shifting to varying forms of consumer-grade EEG equipment, enabled by the development of more affordable advanced EEG related hardware [1]. These EEG devices are advertised as portable, affordable, effortless, and marketed for personal and everyday applications at home and school despite technical limitations [1, 16, 17].

The transition from expensive and medical-grade to consumer-grade EEG devices for research has led to decreased preventative measures, increased data collection, and an increased likelihood of artefact contamination [1, 14, 16, 17]. Therefore, effective and practical cleaning methods are essential for current EEG research and applications. It is also essential to automate these effective methods so that the preprocessing of the EEG does not become a bottleneck before applying quality analysis.

1.2 Aim and objectives

With the need for robust cleaning methods being established, this project aims to identify, automate and evaluate adequate methods for removing artefacts from EEG data. To achieve this aim, the researcher must accomplish the following objectives:

1. Identify the most relevant artefacts to remove from EEG for non-compliant participants
2. Identify and deploy the most effective methods for cleaning EEG in the research context
3. Identify and deploy a metric for evaluating the performance of these methods

4. Develop a semi-synthetic dataset to test the performance of the employed methods

5. Automate the employed methods for dealing with greater EEG data demands

1.3 Thesis overview

From this thesis, the reader will gain insight into the effectiveness of the most popular blind source separation (BSS) methods in cleaning electrocardiography (ECG), electrooculography (EOG) and electromyography (EMG) and a combination of these artefacts. The fully automated BSS methods will be presented and will show comparable results with cleaning all 50 datasets for each type of artefact in about three minutes, compared to the semi-automatic methods, which require a few hours due to the manual input. Based on the standard deviation of data, the auto threshold method showed comparable results and could clean all 50 datasets for each type of artefact in about 2.5 seconds.

The thesis is structured in three main chapters: firstly, the Literature Review, secondly, Materials and Methods and thirdly, the Results and Discussion. In the Literature Review chapter, the reader will find a discussion of EEG, artefacts such as EOG, ECG and EMG and popular cleaning methods such as Extended Infomax, second-order blind identification (SOBI), canonical correlation analysis (CCA). In the Materials and Methods chapter, the methodology followed and materials used to create the semi-synthetic dataset for testing the cleaning methods, and the methodology for the cleaning methods will be discussed. In the Results and Discussion chapter, the reader will find a thorough validation of the semi-synthetic contaminated data and the results of cleaning the contaminated data using the BSS methods, the fully automated BSS methods and the auto threshold method. These results will be communicated using signal-to-noise ratio (SNR) and cleaning time as performance measures for effectiveness and efficiency.

A thoroughly validated semi-synthetic dataset will support the claims made about the different cleaning methods. Furthermore, the maximum variation within realistic limits will be imposed for all the different contaminations to ensure that the artefact removal methods are thoroughly tested. Finally, the claims made about the effectiveness of the methods tested will be based on a statistical analysis of quantitative results. It is hypothesised that using synthetically augmented measured EEG data rather than either purely synthetic or purely measured EEG data provides a simple yet more justifiable datum for comparison of various cleaning methods.

Chapter 2

Literature Review

2.1 Electroencephalography

2.1.1 The nervous system

The human nervous system is categorised into two different systems, the central nervous system (CNS) and the peripheral nervous system (PNS). The CNS contains the brain and spinal cord. The PNS consists mainly of nerves, which are long fibres that connect the central nervous system to every other part of the body. The spinal cord is connected to the brain via the brainstem and conducts all the electrical signals between the brain and the PNS. The brain can be grouped into three main parts, the cerebrum, cerebellum and brainstem, as shown in Figure 2.1 [18–23].

As the largest part of the brain, the cerebrum consists of a left and right hemisphere connected by the corpus callosum. The cerebrum's outer layer, the cerebral cortex, consists of 2 to 4 mm of grey matter and underlying white matter and is made up of numerous folds (gyri) and grooves (sulci), referred to as convolutions. A high density of neurons (estimated 10^{10} neurons) and its close proximity to the electrodes make the cerebral cortex the most significant region in EEG studies [19, 23, 24].

Figures 2.1 show four of the five lobes in each cerebral hemisphere, separated by the central, lateral, and parieto-occipital sulcus. The four lobes visible from the surface are the frontal, parietal, temporal and occipital lobes, covering the insular lobe [19, 23–25].

Figure 2.1 shows the frontal lobe as the anterior portion of each cerebral hemisphere. Voluntary motor control of skeletal muscles is the primary function of the frontal lobes. Additionally, frontal lobes are associated with higher levels of intellectual functioning such as planning, decision making, and verbal communication. Disorders such as depression, anxiety, poor executive planning, lack of motivation, migraines, personality disorders, ADHD and obsessive-compulsive disorder (OCD) are associated with the frontal lobe [19, 23, 26–29].

As shown in Figure 2.1, the parietal lobe is separated from the frontal lobe by a deep fissure called the central sulcus. A primary function of the parietal lobe is somaesthetic interpretation, for example, of skin and muscular sensations. The parietal lobe is further associated with formulating words, understanding speech and expressing thoughts and emotions. Disorders associated with the parietal lobes include learning disorders, poor spatial awareness, anxiety, depression and poor abstract comprehension [18, 19, 23, 29].

As shown in Figure 2.1, the temporal lobes are located below both the frontal and parietal lobes of the cerebral hemispheres. The temporal lobes are close to the hippocampus and are very important for memory creation, especially long-term memory and further associated with the interpretation of auditory and visual information. Disorders related to the temporal lobes are poor emotional control, rage, anger, learning disorders, poor memory and amusia [19, 23, 26, 27].

The occipital lobes can be found at the anterior regions of each cerebral hemisphere, as seen in Figure 2.1. The primary function of this lobe is to process vision, coordinate eye movements and the conscious perception of sight. The disorders associated with the occipital lobes are visual agnosia (inability to perceive and draw complete objects), simultaneous agnosia (inability to see multiple things simultaneously), and learning disorders related to visual processing [19, 23, 26, 27].

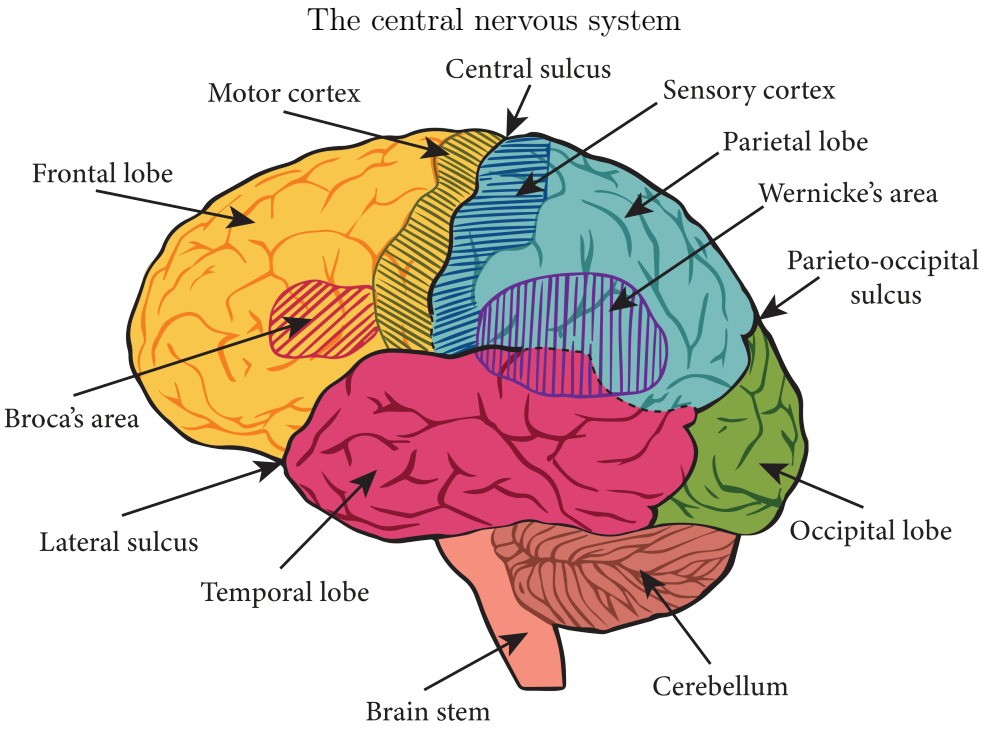


Figure 2.1: The brain adapted from [23].

A crucial area to consider is the sensorimotor (sensory and motor) cortex, which is the hatched region overlapping the frontal and parietal lobes, as seen in Figure 2.1. Disorders associated with the sensorimotor cortex are paralysis, stroke, seizure, poor handwriting, ADHD, depression, and anxiety [19, 23, 26–28, 30].

2.1.2 Electroencephalography generation

EEG is the measurement of synchronous or asynchronous communication through the electrical activity of large groups of neurons measured mainly from the cerebral cortex [17, 19, 31]. The term used for groups of organised cells in the nervous system is nervous tissue. Nervous tissue consists of neurons and glial cells. The neurons are the basic units of the nervous system and form an intricate system responsible for the generation and conduction of electrical events. The glial cells support the nervous system and are mainly found in the brain and spinal cord (CNS) [19, 23, 32, 33].

Figure 2.2 shows that each neuron consists of three parts: the dendrites, a cell body, and an axon. The dendrites are considerably branched cytoplasmic extensions of the cell body and receive inputs from other neurons or receptor cells. The cell body contains the nucleus, which is the metabolic centre of the cell. The axon is a single cytoplasmic extension of the cell body that conducts the nerve impulses from the cell body to other neuron or effector (muscle or gland) cells. An axon's length can range from millimetres to a meter [19, 23].

Glial cells do not conduct impulses. Instead, they bind neurons together, modify the extracellular environment of the nervous system, and influence the nourishment, restoration and electrical activity of neurons. When a person learns new things, the learning process is accompanied by structural changes directed and reinforced by glial cells. Three of the most important types of glial cells are the oligodendrocytes, astrocytes, and microglia glial cells. The oligodendrocytes glial cells are used to increases the myelin sheath layers, where the increase of myelin sheath layers corresponds to learning. The astrocytes glial cells repair and provide nutrients to brain cells and maintain the equilibrium of neuronal functions. The microglia glial cells are used to remove waste and also respond to injury and infection [19, 23].

The functional point of connection between two neurons, enabling communication, is called the synapse. The axons of a neuron are always either close to or in contact with the cell of another neuron. An action potential from the cell body of a neuron travels through the axon to the axon terminal bundle and through the synapse to either directly or indirectly stimulate or inhibit the cell of another neuron. With some exceptions, the action potential of the neuron ends at the axon terminal, where it stimulates the release of a chemical neurotransmitter that affects the cell of the next neuron [19, 23].

With time, connections between neurons strengthen due to the increasing thickness of the myelin sheaths covering the axons, as shown in Figure 2.2. The

myelin growth process first starts when a human is born. Therefore babies have a low posterior dominant rhythm (PDR) between 1 and 4 Hz. At the age of about 14, when extra layers of myelin have been added, the PDR converges to the higher alpha frequency. Lower frequency PDRs can be recognised with young ADHD adults due to reduced or delayed development of these myelin layers [19, 33–35].

EEG data represent the electrochemical events that originate from the synaptic electrochemical action potentials. These chemical synapses can be categorised into slow and fast synapses. Fast synapses involve glutamate and gamma-aminobutyric acid, while slow synapses involve dopamine, serotonin, acetylcholine, and norepinephrine. These synaptic potentials give rise to local field potentials, which influence the firing of action potentials of the pyramidal neurons. The EEG data is acquired by measuring the electrochemical reactions from the pyramidal neurons by placing electrodes on the participant's scalp [19, 36].

The neural network

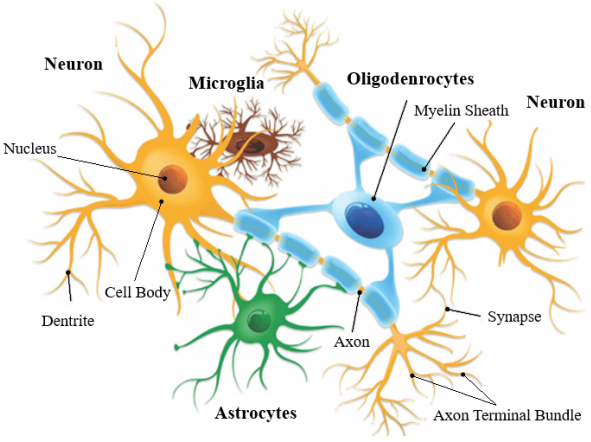


Figure 2.2: The neural network adapted from [19].

2.2 Electroencephalography in practice

2.2.1 Standard Electroencephalography electrode placements and associations

The International Federation of Clinical Neurophysiology proposed the International 10-20 system for Electroencephalography and Clinical Neurophysiology in 1958 [37]. Since it has sufficient electrodes and high resolution for most clinical applications, the International 10-20 system is most commonly used in these contexts. The system is often used as a standard for various transcranial mapping methods [38], is efficient and has a good balance between time,

effort, cost, and accuracy. It is adequate unless disorders such as epilepsy are considered due to localisation becoming too critical [19, 26, 27, 37].

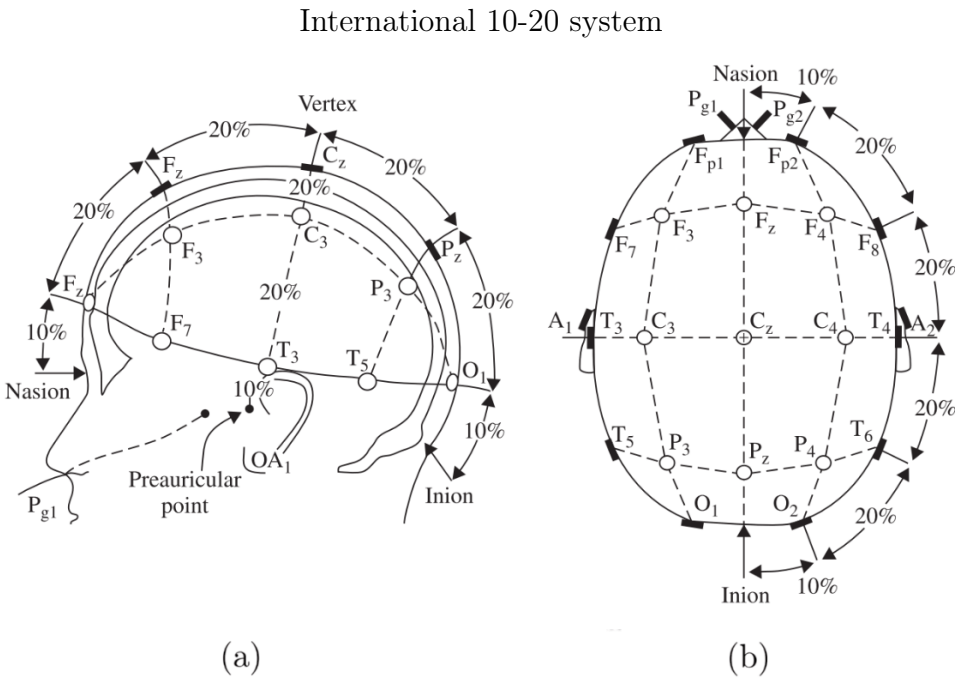


Figure 2.3: (a) Left view of 10-20 system. (b) Top view of 10-20 system adapted from [22].

The International 10-20 system determines electrode positions based on fractions (10% or 20%) of the distance between the nasion-inion position and the pre-auricular positions, as shown in Figure 2.3. The labelling of the electrodes is based on an anatomical convention used to locate the various brain regions [39]. The letters 'F', 'T', 'P', 'O' or 'A' correspond with the electrodes positioned over the frontal, temporal, parietal, occipital or auricular regions respectively. The letter 'C' corresponds to the central electrodes. The subscripts of the labels are either odd or even numbers, corresponding to the left or right hemisphere. The z subscripts correspond to electrodes located on the midline of the brain [19, 26, 27, 37, 39].

2.2.2 Standard Electroencephalography frequency bands and associations

The amplitude spectra of EEG data have been categorised into different frequency bands or ranges to analyse brain function more effectively for diagnoses and treatment. Abnormally high or low amplitudes at these frequencies have various associations and may lead to certain diagnoses. There are five standard frequency bands, namely the delta (δ), theta (θ), alpha (α), beta (β),

and gamma (γ) bands. Gamma bands include the fastest waves in terms of frequency and delta the slowest. The exact ranges of these frequency bands differ slightly depending on the research or clinical context. Due to this research being focused on physiological artefacts, the bands are consistent with those used with participants who show non-compliant behaviour, such as ADHD participants [19, 26, 40, 41].

The frequency of the delta band is under 4 Hz. The delta band is the slowest frequency and is also associated with hyperactivity and impulsivity. The frequency of the theta band is in the region of 4 to 7 Hz and is considered one of the lower frequencies associated with hyperactivity and impulsivity. The alpha band with a frequency of 8 to 12 Hz is the PDR in neurotypical adults and is associated with relaxed wakefulness and increased attention. The beta band frequency is in the region of 12 to 30 Hz. Increased beta band activity can be associated with an increased attention state. The gamma band frequency is between 30 to 100 Hz and is the highest considered band in EEG studies. High gamma-band activity correlates to blood oxygenation and is associated with a high functional cortical state during sensory stimulation and high performance of cognitive tasks [19, 26, 40–42].

2.3 Artefacts

2.3.1 Overview of artefacts

EEG signals are extremely weak, and their analysis can be influenced due to contamination from artefacts. These artefacts are activities that do not directly originate from the brain but are still present in the EEG data [5].

Artefacts deliver a significant detrimental effect on EEG due to large variations in temporal and spectral contamination. Some artefacts may contaminate several neighbouring channels, while others contaminate only a single channel. In addition, some artefacts appear as regular periodic events, such as pulse artefacts, while some artefacts are extremely irregular [9].

One can categorise artefacts into non-physiological and physiological artefacts. The three primary physiological artefacts discussed in the literature that most typically occur are ECG, EOG and EMG. ECG artefacts originate from the heart and are present in the form of a pulse or heartbeat when an electrode is placed on or near a blood vessel. EOG artefacts mainly originate from eye movements and blinks. EMG originates from any muscles movements, including muscle groups from the neck and the face such as the cheeks, forehead, jaws, tongue etc. Other less intrusive physiological artefacts include the movement of the head, limbs or other movements and tremors [2, 9, 13, 14].

Non-physiologic artefacts can be categorised into instrumental and interference artefacts. Typical interference includes line noise, powerful signals from nearby alternating current power lines occurring between 50 and 60 Hz, and

interference from nearby electrical equipment. Other less intrusive interference artefacts include magnetic, sound, optical and electromagnetic waves, white noise and pink noise. Instrumental artefacts include dysfunctional electrodes, electrodes movement, electrodes pop, dysfunctional cables, cable movement, impedance mismatch and poor electrical ground [2, 9, 13].

2.3.2 Artefacts focused on in this research

This thesis focuses on participants that are non-compliant with artefact preventive guidelines, such as those who use commercial EEG devices in uncontrolled environments and those with ADHD, increasing the likelihood of physiological artefacts contaminating the EEG data [3, 10, 15]. Therefore, the scope of artefacts was reduced to only physiological artefacts. EOG artefacts have been shown to deliver the largest detrimental effect on EEG [11]. EMG artefacts are generally acknowledged to be more difficult to eliminate than other types of artefacts [12]. Furthermore, the contamination of EEG by ECG artefacts constitutes a serious problem for the automatic interpretation and analysis of EEG recordings during sleep [10]. Therefore, EOG, EMG and ECG artefacts were chosen as the three most critical physiological artefacts for testing the cleaning methods.

2.3.3 Electrooculography

EOG artefacts originate from eye blinks, eye movements (referred to as saccades when not related to blinks) and, less frequently, eye flutter and rapid eye movement (REM) sleep, which can propagate over the scalp and contaminate the EEG data [3, 9, 43].

Eye blink artefacts originate from the potential generated through the eyelid sliding down over the positively charged cornea. These generated potentials propagate across the entire scalp with amplitudes often significantly higher than those of EEG. Although eye blinks originate from the anterior region, their effects are considerable over the entire scalp, decreasing in intensity from anterior to posterior regions. Eye blinks have a large intersubject variability, with natural occurring eye blinks having smaller amplitudes and shorter duration than forced blinks [14, 44–46].

Saccade artefacts originate from changes in orientation of the retina and cornea dipole [9, 15]. Saccades and eye blinks both encapsulate particular frequency characteristics but differs from each other significantly [47]. Saccades usually display a lower average voltage and lower range in voltage than eye blinks [43]. Furthermore, saccades show a similar average frequency but a higher frequency range than eye blinks [43]. Vertical saccades influence midline electrodes more, while lateral saccades influence lateral electrodes more [47].

EOG is a combination of mostly eye blinks and saccades [44]. EOG artefacts are often removed using a reference channel and regression methods. A

limitation of these methods is that the EOG reference data can also be cross-contaminated by the EEG data, causing a possible removal of brain information from the data by these methods [3]. EOG signals, similar to EEG signals, are considered non-stationary signals, meaning that the frequency spectrum changes with time [48]. The amplitude of EOG signals is generally significantly higher than that of the EEG, occurring mainly within the amplitude ranges of 10 to 100 μV in comparison to EEG commonly occurring between the ranges of -50 to 50 μV [2, 3, 48]. Eye blinks have a strong energy presence in the delta and theta bands, sometimes showing energy traces in the higher alpha and beta bands [44]. In contrast, EEG occurs throughout, from the delta to the gamma band, with a peak at the alpha band (8 to 12 Hz) [3, 48]. EOG signals last only a few seconds, as they arise from eye movements and blinks [14, 48, 49].

2.3.4 Electromyography

Contamination of EEG data by muscle activity is a well-recognised complex problem arising from different muscle groups [50, 51]. Any muscle contraction or stretch near an electrode recording site can result in EMG artefacts affecting the 'clean' EEG signal [15]. The degree of muscle contraction and stretch affects the amplitude and waveform of EMG artefacts. Regression methods cannot be applied to EMG as with EOG because they originate from multiple sources [3]. EMG sources include the movement of many muscles, including muscle groups from the neck and face, such as the cheeks, forehead, jaws and tongue, from head movement, chewing, swallowing, clenching, talking, sniffing, and facial contractions [2, 9, 13, 14].

EMG presents a wide spectral distribution contaminating all the standard frequency bands. It is, however, most significant in the higher frequency bands, with most literature assuming that the EMG artefacts only affect the higher frequencies, starting at 15 to 20 Hz and upwards [15, 50, 52–54]. The EMG activities often have a temporal amplitude significantly higher than the EEG data, such as EOG [31, 55, 56]. The amplitude of EMG data has a peak in the 20 to 30 Hz range in the frontalis location [50]. The time series of EMG follows a spontaneous bursting behaviour with a temporal and spectral distribution similar to Gaussian noise [56–58]. Additionally, EMG contamination and EEG have substantial statistical independence both temporally and spatially. This implies that the independent component analysis (ICA) methods could effectively identify and remove EMG artefacts [9].

2.3.5 Electrocardiography

ECG artefacts originate from the heart and occur in the EEG data as a pulse or heartbeat when an electrode is placed on or near a pulsating blood vessel such as a scalp artery [13]. ECG signals display a simple, characteristic and

periodic time-frequency characterization pattern [15, 59, 60]. The amplitude of the ECG artefacts is relatively low compared to the amplitudes of EOG and EMG artefacts. However, the amplitude of the ECG artefacts also greatly depends on the relative position of the electrode to the blood vessel and the anatomy of the participant [15, 61]. A major problem with the repetitive and regular patterns of ECG artefacts is that it may sometimes be mistaken for epileptiform activity when the ECG has low amplitudes relative to the EEG [15].

2.4 Cleaning methods

2.4.1 Overview of methods

The difficulty of removing artefacts is a significant obstacle in EEG signal preprocessing and a prerequisite for reliable signal analysis [2, 9, 13–15]. The time efficiency of cleaning methods is also a significant obstacle for developing real-time EEG applications, such as brain-computer interface (BCI) and neurofeedback (NFB) applications [14].

A review of the most popular and relevant methods identified in EEG artefact removal was conducted. The review included the methodology, advantages and limitations of linear regression, source decomposition, adaptive filtering and BSS techniques and can be found in Appendix A. Linear regression and source decomposition techniques were not further considered in this research due to the limitations identified [62–66].

Methods for removing artefacts are primarily developed and tested for removing only one type of artefact at a time [10, 13–15, 67, 68]. Artefacts are diverse and contaminate EEG data with a large variety of intensities, types, locations, combinations, and durations. Participant variability, e.g. statistically different EEG data for various participants, is another factor to consider when removing artefacts [9]. Therefore, artefact removal methods must be able to handle a large variance in artefacts and EEG characteristics [9].

Current cleaning methods can be categorised into three different groups of techniques. The first is artefact avoidance, which includes strict guidelines for participants to avoid moving or blinking during the experiment and gazing at a central fixation point. This may become challenging when the participant is non-compliant or has severe ADHD. The second technique is artefact rejection, which removes contaminated data trials completely, identified either through visual inspection or by automatic identification methods. The third technique is preprocessing the EEG to separate the artefacts from the 'clean' EEG data [69]. With ADHD research and consumer-grade EEG equipment, the first technique is not always feasible, requiring an increase in the development of artefact identification and removal methods used in the second and third techniques [2, 3, 11, 15, 69].

The techniques discussed in Appendix A.1, A.2 and A.3 are preprocessing-based techniques, with adaptive filtering discussed in Appendix A.4 an artefact rejection technique.

2.4.2 Methods for the combination of artefacts

The majority of research on artefacts and the artefact removal methods focus on the effects of one or a few artefacts and seldom on the combination of these artefacts [10, 13–15, 67]. EEG processing that generalises to multiple types of artefacts remains a significant challenge [11, 32, 69–71]. Furthermore, the combination of artefacts significantly increases when data collection moves from the clinical environment to more relaxed settings associated with commercial products [11].

The most popular methods used for research are component-based methods [3, 15, 72, 73]. A significant disadvantage of these methods is that they can only identify as many components as channels used. Therefore, with an increase of independent artefacts, the likelihood of effectively identifying all the components decreases [15, 32, 74].

Currently, no one method is the most effective and efficient for removing a wide range of artefacts [3, 11]. Therefore, artefact removal algorithms for numerous types of artefacts in multiple scenarios still need to be identified [3, 11, 69].

2.4.3 Methods used in this research

The research scope has been reduced to identifying and removing artefacts from the recordings of non-compliant participants, and therefore to BSS techniques, specialising in biomedical signals and removing physiological artefacts [15, 75]. Currently, BSS methods are also the most popular category for artefact removal in EEG research [3, 15, 72, 73]. Under the BSS methods category, CCA and ICA are popular methods, as seen in Figure 2.4. The two most popular ICA based methods used in research are Extended Infomax and SOBI, as seen in Figure 2.4. CCA is a classic BSS method and the third BSS method to be used. It is frequently used in BCI research, which is associated with commercial products and, therefore, with a less controlled participant, having more artefact inducing behaviours. BCI application requires time efficiency and superior performance in EMG removal, which are characteristics associated with CCA [12, 15, 76].

2.4.4 Blind source separation

BSS methods are called blind since neither the sources nor the mixing process is known in advance [14]. These methods decompose the signal into several estimated sources, in which the sources identified as artefacts are rejected and

the signal reconstructed. By employing BSS techniques, it is possible to separate signals from multiple sources into several components, which become more effective when the number of measured signals increase [64]. The effectiveness of the BSS technique depends on a specific set of assumptions that must first be made; for example, the sources are uncorrelated, independent, non-Gaussian, instantaneously propagated or linear [15]. BSS methods are ultimately limited because artefacts containing components may also include neural activity besides the pure artefact itself [77]. BSS techniques may not be effective on EEG recordings that contain highly non-stationary artefacts such as EOG [15]. When there are limited electrodes used for EEG recording, the information related to brain activity may be lost by excluding sources containing brain activity. There is also the risk of human error in BSS methods when components are mislabeled as artefacts but contain neural activity [64]. Although BSS algorithms have not proven to be the most effective approach, they are most often used with EEG data that is prepossessed [15, 78, 79]. It has been demonstrated that BSS variants can remove EOG artefacts equivalent to the 'gold standard' regression methods that use an EOG reference signal [15, 78, 79]. One of the significant disadvantages of BSS methods is that the automated classification of artefact components is not straightforward [15, 73]. Other disadvantages are that they inaccurately assume the stationarity of the sources and are computationally expensive [15, 73].

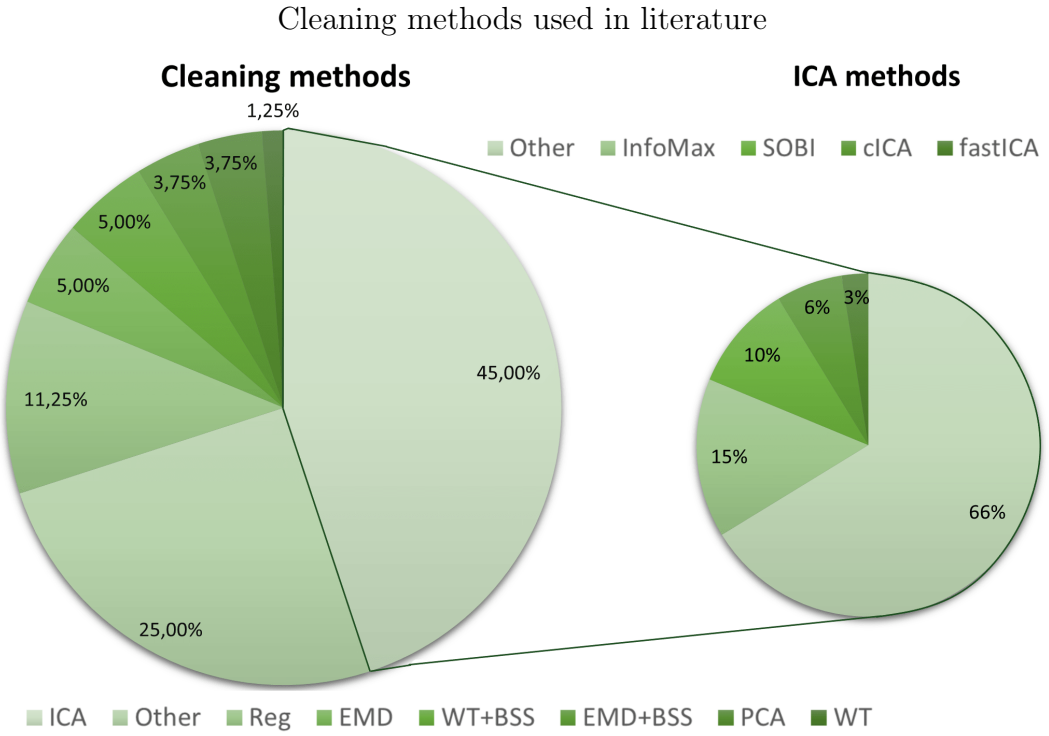


Figure 2.4: The popularity of methods in literature based on data from [15].

2.4.4.1 Independent component analysis

ICA methods comprise several methods that aim to estimate the unmixing matrix, $W = [w_i(j)]_{n\times n}$, which is used to separate the measured data, $X = [x_i(j)]_{n\times N}$ into its estimated sources $S = [x_i(j)]_{n\times N}$ through imposing statistical independence of the sources. In the equations mentioned above, N represents the number of samples and n, the number of channels [15]. ICA methods are based on three main assumptions: the ICA component projections are linearly summed at the scalp electrodes, the sources of the measured EEG are independent, and that the source activities characterise a non-Gaussian distribution. With the assumption of independence, even if artefacts are caused by brain activity and therefore not independent, e.g. if EMG is triggered by motor cortex activity, the timing between the resulting artefacts and the triggering brain events will still vary across trials, allowing them to represent independence. Further discussion on the mathematical implementation and statistical assumptions of ICA methods can be found in the Materials and Methods chapter [5, 45, 75].

ICA algorithms can be branched into those based on exploiting higher-order statistics (HOS) of the signals, such as Extended Infomax, and those based on using second-order statistics (SOS), such as SOBI [80]. HOS-ICA approaches find a linear transformation for the estimated sources to be as independent as possible, from which the approaches can identify the artefact components. SOS-ICA methods are based on decorrelating the data in the time domain [15]. ICA methods are typically the preferred methods when it comes to removing EOG and ECG artefacts [81].

Extended Infomax can adapt to and find both sub- and super-Gaussian sources [82, 83]. Frølich and Dowding [82] compared three of the most widely used ICA methods and two other linear decomposition methods and found that Extended Infomax performed the best in terms of the dipolarity score proposed and the event-related desynchronisation (ERD) peak score. A polarity score is a simplified but useful measure of the physiological plausibility of the identified ICA sources [82]. ERD is calculated as the change in signal power relative to a reference period in a given frequency band [82]. Extended Infomax has also been the most thoroughly justified for removing EOG artefacts in other literature [83]. A more in-depth statistical and mathematical discussion of Extended Infomax can be found in the Materials and Methods chapter.

It has also been stated that SOBI stands out as the best method for removing EOG artefacts [8, 84]. Literature differs in conclusions on whether SOBI or Extended Infomax performs the best in removing EOG artefacts [8, 83, 84]. SOBI has also been reported to generally perform better than other methods in removing ECG artefacts and has been demonstrated to be consistent across a wide range of participants [85, 86]. This makes SOBI a feasible method for routine EEG analysis and interpretation, in addition to removing artefacts [85].

2.4.4.2 Canonical correlation analysis

Along with ICA, CCA is a classic method. The effectiveness of BSS methods depends largely on the different statistical assumptions of the estimated sources [87]. CCA solves the problem by utilising SOS information such as SOBI and forcing the sources to be mutually uncorrelated and maximally correlated with a predefined function [87]. Thus, CCA finds two bases in which the correlation matrix between the variables is diagonal, and the correlations on the diagonal are maximised [87]. Therefore, the CCA method produces the source signals that are uncorrelated with each other, maximally autocorrelated and ordered by decreasing autocorrelation [87]. Since EMG artefacts have a broad frequency spectrum, their autocorrelation is low, while the autocorrelation of EEG rhythms is relatively high [57].

CCA is often proposed as a more reliable method for the removal of muscle artefacts in the scalp in comparison to ICA methods due to the exploitation of the autocorrelation of muscle activity being weaker than that of brain activity [12, 15, 88]. CCA has also been shown to clearly identify EOG artefacts for removal compared to those of popular ICA methods [89]. Further discussion on the statistical and mathematical implementation of the CCA method can be found in the Materials and Methods chapter.

2.4.5 Automated methods overview

The transition from expensive and medical-grade to consumer-grade EEG equipment for research has led to decreased preventative measures, increased data collection, and the increased likelihood of artefact contamination [1, 14, 16, 17]. Therefore, to ensure that the preprocessing of the EEG does not become a bottleneck before applying quality analysis, the automation of the BSS methods is essential. Automated methods are also preferable because they eliminate the subjectivity associated with the non-automated methods [8]. The BSS methods mentioned above require the use of manual input to identify the artefact containing components. If the artefact identification process can be automated, then the BSS methods would be fully automatic. Furthermore, adaptive thresholding techniques also exist, which are simple, efficient and fully automatic.

2.4.5.1 Automation of blind source separation methods

As mentioned above, to automate the semi-automatic BSS methods, the identification of the artefact components has to be automated. Unfortunately, One of the significant disadvantages of BSS methods is that the automated classification of artefact components is not straightforward [15, 73]. The developed methods found in the literature were each successful in certain applications but also had significant limitations. These limitations were due to the com-

plexity of these approaches, including the use of machine learning algorithms, additional reference channels, and additional computationally inefficient algorithms [8, 90–95]. The problem of using machine learning algorithms is that it requires manual input from experts to effectively label the training data into clean EEG and artefacts, respectively [90, 91, 94]. Additional information on the methodology, results and limitations can be found in Appendix C.

2.4.5.2 Thresholding methods

In addition to their simplicity, thresholding methods can be effective since artefacts such as EMG, EOG and ECG often have greater amplitudes, therefore, making them distinguishable from EEG [10, 15, 44, 50].

Due to its simplicity and efficiency, threshold-based methods are often considered an alternative for automated real-time applications [15]. However, applying the same threshold value can result in significant detection errors in some participants' EEG data due to large individual variance in the shapes and amplitudes of, for instance, the eye blink artefacts in EOG. Therefore, the threshold values need to be adapted for each participant's EEG [96]. Mognon et al. [97] adjusted the threshold value of their method using the expectation-maximisation algorithm. Their method, however, was not fit for real-time application and processing due to the time-consuming nature of the maximisation procedure [97]. Geetha and Geethalakshmi [98] utilised the thresholding technique initially proposed by Otsu [99] and Breuer et al. [100] for image binarisation and used the 80th percentile of the individual data distribution as the threshold value. However, both methods were still based on empirical parameterisation and, as a result, time-consuming [96]. Despite the relatively good performance of these adaptive thresholding methods, it was found that individually customised thresholding led to greater accuracy in detecting EOG artefacts due to the high variation between the EOG of participants [96].

2.5 Evaluating cleaning methods

Despite the number of techniques developed for removing artefacts, a method that combines high accuracy and algorithmic efficiency is still lacking [3]. Another limitation is the practical testing and comparison of the developed cleaning methods. The greatest challenge for evaluating the performance of artefact removal methods is that the noiseless signal is not known a priori [11, 15]. Therefore, it is necessary to develop tools that allow objective measurement and comparison of the performance of new and current algorithms to select the optimal one for a specific scenario.

Multiple validation procedures for real-life EEG signals have been proposed in recent years. Researchers, however, do not agree on a single mechanism for evaluating and comparing the performance of artefact removal methods.

Numerous evaluation methods based on the use of real data were discussed in Appendix B, but were not considered in this thesis due to the limitations identified [66, 84, 97, 101–104].

The evaluation methods discussed in Appendix B was based on using real contaminated data, but each method was not straightforward due to the lack of knowledge of the clean EEG signal. Fortunately, to some extent, it is possible to determine the clean EEG through realistic simulations, either from computer-generated EEG signals or from controlled 'clean' EEG recordings. A primary advantage of simulated EEG is that the quality of the signal can be evaluated before and after artefact removal using standard evaluation measurements, such as the SNR, thus enabling comparison with other studies. Therefore, the following section will focus on simulated data [105].

2.6 Simulation methods

2.6.1 Overview

Simulations have historically played a significant role in the development of cleaning methods and can be generated using techniques ranging from very simple to more complex [4, 15]. Simulated contaminated EEG enables the use of SNR, which compares the energy of the frequency domain of the 'clean' EEG signal to that of the artefacts [15].

Semi-synthetic techniques can range from using actual 'clean' EEG contaminated with simulated artefacts to using separately measured artefacts with simulated EEG [45, 67]. Simulating techniques also include simulating the EEG and artefacts to using actual 'clean' EEG contaminated with separately measured artefacts [67, 77, 106].

It is possible to simulate some characteristics of a recorded EEG relatively accurately. However, characteristics such as synchronisation between channels, contamination by different artefacts and the effect of artefacts on physiological sources are more challenging to simulate. Simulations are still considered a preliminary evaluation, and actual contaminated EEG data must be used as the ultimate test for evaluating the true performance, reliability and reproducibility of any artefact removal method. Two main methods of simulating EEG data have been identified: the linear mixture model and the forward model [15, 77, 107]. Due to the limitations identified with the forward model, it was not further considered in this thesis, and the discussion and mathematical implementation can be found in Appendix D [107–109].

2.6.2 Linear mixture model

The linear mixture model is the simplest simulation technique. The method is based on linearly adding simulated EEG and different simulated artefacts.

The accuracy and simplicity of this type of simulation depend on how close the characteristics of the simulated EEG and artefact signals are to actual EEG and artefacts. The mathematical representation from the standard assumption that the contaminated data, $X^{(c)} = [x_{i,j}^{(c)}]_{n\times N}$, is a linear mixture of the original clean EEG, $X^{(s)} = [x_{i,j}^{(s)}]_{n\times N}$, and artefacts, $X^{(a)} = [x_{i,j}^{(a)}]_{n\times N}$, such as described by equation 2.1 where N is the number of sample points and n the number of channels [15, 105]:

$$\mathbf{X^{(c)}} = \mathbf{X^{(s)}} + \mathbf{X^{(a)}} \tag{2.1}$$

The semi-simulated techniques can range from using actual 'clean' EEG, contaminated with simulated artefacts, to using real separately measured artefacts with simulated EEG [45, 67]. Techniques also range from simulating the EEG and artefacts to using actual 'clean' EEG contaminated with real separately measured artefacts [77, 106, 110].

The accuracy of these methods is limited as a result of the lack of incorporation of the influence of artefact generation on the physiological sources. An example is the alpha band's amplitude decreasing due to mental effort exerted when participants contract their temporal or frontalis muscles [15, 50].

2.6.3 Signal to noise ratio for performance evaluation

One advantage of using simulated EEG data is that one can assess the quality of the signal before and after the artefact removal through standard performance measures. The metric most commonly employed to represent the signal's energy, compared to the artefacts' energy, is the SNR. More specifically, the signal to noise ratio can be defined as the ratio of the power spectral density (PSD) of the clean EEG to the PSD of the artefacts [15, 105].

The SNR is based on the linear mixture model stating that the contaminated data, $X^{(c)}$ from equation 2.1, is a linear mixture of the clean EEG, $X^{(s)}$, and artefacts, $X^{(a)}$ [15, 105]. Equation 2.2 describes the SNR for one channel, i, where x represents the sample point at a certain channel for the PSD of the EEG data. For the simulation of the data, the artefact data, $X^{(a)}$ is known, and therefore it is simple to quantify the amount of contamination using the SNR as shown in equation 2.2, with the variables described in Section 2.6.2 [15, 105].

$$SNR_i = 10\log_{10}\left(\frac{\sum_{j=1}^{N} x_{i,j}^{(s)}}{\sum_{j=1}^{N} x_{i,j}^{(a)}}\right) \tag{2.2}$$

When testing the effectiveness of the cleaning methods, we do not directly know the amount of artefacts, $X^{(a)}$, that are still present when the simulated contaminated data is cleaned. $X^{(a*)}$ can however be calculated as shown in equation 2.3, where $X^{(k)} = [x_{i,j}^{(k)}]_{n\times N}$ is the cleaned data. Therefore, to calculate the SNR of the cleaned data and the amount of artefact data removed, one

can use equation 2.4, where the ideal situation would be for the denominator to be zero, meaning the SNR would be infinity, and the cleaned data matches the clean data [15, 105].

$$X^{(a*)} = X^{(k)} - X^{(s)} \tag{2.3}$$

$$SNR_i = 10 \log_{10} \left(\frac{\sum_{j=1}^{N} x_{i,j}^{(s)}}{\sum_{j=1}^{N} (x_{i,j}^{(k)} - x_{i,j}^{(s)})} \right) \tag{2.4}$$

2.7 Main findings in literature

2.7.1 Literature summary

Due to the high density of neurons, the cerebral cortex is considered the most significant region in EEG studies. The 10-20 standard is the most common electrode placement system and has enough electrodes and resolution to accommodate most clinical applications. The 10-20 standard is efficient and balances time, effort, cost, and accuracy. There are five standard frequency bands, namely the delta, theta, alpha, beta, and gamma bands, each having established associations and being widely used in the research and clinical context. The literature found that the most significant physiological artefacts with the greatest detrimental effect on EEG are EOG, EMG, and ECG artefacts. BSS methods are often used to preprocess EEG data despite not found to be the most effective methods. BSS methods, such as Extended Infomax, outperform most methods and have similar performance to the 'gold standard' regression methods in removing EOG artefacts without requiring a reference channel. Researchers do not yet agree on a single mechanism for evaluating and comparing the performance of artefact removal methods. Simulations have historically played a significant role in the development and comparison of cleaning methods. A primary advantage of simulated EEG is that the quality of the signal can be evaluated before and after the artefact removal, using standard evaluation measurements, such as the SNR, enabling comparative analysis with other studies.

2.7.2 Main limitations identified

Although BSS methods are the most commonly used preprocessing methods, they still require manual input to identify the artefact containing components, posing the risk of subjective error and requiring an impractical amount of time and effort on the part of the researcher or practitioner. BSS methods are computationally expensive and therefore not suited for real-time applications. There is no clear answer to which ICA methods perform the best between SOBI and Extended Infomax, with artefacts such as EMG, CCA also shows

similar results to the ICA methods. With the automation of the BSS methods, it was found that the solutions were often impractical due to them being overly complex or requiring manual input to label the training data of the machine learning algorithms. Simulations are still considered a preliminary evaluation, and real contaminated EEG data must be used as the ultimate test for evaluating the true performance, reliability and reproducibility of any artefact removal method. The attempt to develop a more realistic simulation using the forward model has much more room for error than the more straightforward approach of the linear mixture model.

2.7.3 Assumptions and decisions made

EOG, EMG and ECG are the most detrimental physiological contamination in EEG data and were, therefore, the main focus of this thesis. The use of the standard EEG bands and the 10-20 standard allowed this thesis to be validated using other literature.

BSS methods were chosen due to their well-earned popularity amongst researchers and medical practitioners. Extended Infomax and SOBI were used as these are the most popular BSS methods. CCA was also used as it is a very popular method for potential commercial applications. The BSS methods are effective but impractical when cleaning large datasets because they are computationally expensive and require manual input. Therefore, it was decided to focus on fully automating the BSS methods to make them practical for use with large EEG datasets. Furthermore, due to the existing automated BSS methods being complex, it was decided to develop a simpler yet effective approach. Due to the BSS methods being computationally expensive, a time-efficient adaptive thresholding method was also developed, based on the standard deviations of the data of each participant.

The performance measures for real contaminated data are complex and not yet well established. The linear mixture model can also be inaccurate if the characteristics of the simulated data do not match those of the actual data. Due to this limitation, it was decided to use the linear mixture model with real ‘clean’ EEG data and measured artefact data. The artefact data duration and combinations were further adjusted within reasonable and realistic limits. These adjustments were made in order to maximise the variety of contamination to more robustly test the cleaning methods. SNR was chosen as the performance measure for the cleaning methods in this thesis as both the ‘clean’ EEG and artefacts were known. The use of SNR enabled the comparison of the results to other research. The final performance measure used was the amount of time the BSS and auto threshold methods took to identify the components and artefacts.

Chapter 3

Materials and Methods

3.1 Use of reference data

The semi-synthetic data was based on real data for it to be as realistic as possible. The manually cleaned EEG data that was published by Klados and Bamidis [111] in 2019 for research purposes and used as the 'clean' EEG data in their research was also used in this thesis. To ensure that the 'clean' EEG data had no significant contamination by external or physiological artefacts, two experts in the field, namely Klados and Bamidis [111], performed a thorough manual inspection and removed any artefacts found in the data. With regard to the artefacts, actual EOG and ECG data from publicly available data repositories from Klados and Bamidis [111] and Khamis et al. [112] were used. In this thesis, the researcher simulated the EMG data to represent the same spectral distribution of those represented by Goncharova et al. [50] for frontalis and temporalis EMG contractions.

The choice of locations for the semi-synthetic dataset of the current research in this thesis was based on the previous work of the same researcher. The previous investigation was based on measuring the EEG signals of ADHD children who were mostly non-compliant, generating excessive EOG and EMG artefacts [3]. This relates to the current thesis, based on removing physiological artefacts such as EOG, EMG and ECG from general non-compliant or uncontrolled participants. The OpenBCI Ultracortex 'Mark IV' EEG headset was used to measure the EEG signals from ADHD children between the age of 7 to 12. The headset was used with the combination of a Cyton and Daisy amplifier, thus allowing only 16 EEG locations. Only the most relevant locations for ADHD children were chosen with guidance from literature [19], and in consultation with van der Westhuyzen [113], an occupational therapist with international certification for biofeedback and neurofeedback. Due to the current research focusing on physiological artefacts from non-compliant participants and the previous research on ADHD children producing excessive physiological artefacts, using the same locations was assumed to be adequate.

The locations used were Fz, Cz, C3, C4, P7, P8, O1, O2, F3, F4, T7, T8, P3, and P4, which are all shown in Figure 2.3. Even though ADHD is often diagnosed as a disorder in the frontal lobes, Fp1 and Fp2 were excluded because of the high risk of over contamination due to excessive blinking and anxiety-induced movements commonly found in ADHD children [3, 19]. Pz was also excluded because it is less important than Fz and Cz for ADHD and less affected by artefacts such as EMG and EOG than other channels [19, 36].

In this thesis, the researcher preserved the EEG and EOG data at the original sample frequencies of 200 Hz, the ECG was downsampled from 500 Hz to 200 Hz, and the EMG was simulated atn 200 Hz. The sources of the data used to create a semi-synthetic database can be found below:

EEG

- The 'clean' EEG was attained from manually cleaned EEG data, available for research purposes if referenced from Klados and Bamidis [111].

EOG

- The EOG was attained from recorded HEOG and VEOG data, available for research purposes if referenced from Klados and Bamidis [111].

EMG

- The EMG was based on the amplitude spectra of the EEG during temporalis and frontalis contractions from Goncharova et al. [50].

ECG

- The ECG was attained from recorded ECG data, available for research purposes if referenced from Khamis et al. [112].

3.2 Electroencephalography description and validation

3.2.1 Electroencephalography dataset description

The EEG data used as the base for the semi-synthetic data was obtained from Klados and Bamidis [77]. The EEG data used consists of 50 datasets with 19 channels each and a mean time-span of 30.1, varying on average with 3.0

seconds from the mean in both directions. The datasets were sampled from 27 healthy subjects, of which 14 were male with a mean age of 28.2 ± 7.5 years, and 13 female with a mean of age: 27.1 ± 5.2. The EEG data was recorded during an eyes-closed (EC) resting state for each subject [77]. For simplifying the simulation, each dataset will refer to that of a simulated "participant", meaning that the semi-synthetic data consists of 50 "participants".

The EEG was recorded at 19 locations and placed according to the 10-20 standard, with odd indices referenced to the left and even indices to the right mastoid, respectively. In contrast, the central electrodes (Fz, Cz, Pz) were referenced to half the sum of the left and right mastoids. The signals were sampled at 200 Hz, bandpass filtered between 0.5 and 40 Hz, and notch filtered at 50 Hz to remove the line noise. Experts manually inspected each EEG dataset in the field to ensure that there was no significant biological or external artefact contamination, as described in Section 3.1 [77]. The 16 locations used here were Fz, Cz, C3, C4, P7, P8, O1, O2, F3, F4, T7, T8, P3 and P4.

3.2.2 Electroencephalography dataset validation

When analyzing the frequency bands, the following ranges were used with delta, < 4 Hz, theta, 4-7 Hz, alpha, 8-12 Hz, beta, 12-30 Hz, and gamma $>$ 30 Hz, based on the frequencies used for ADHD-EEG research [26, 40, 41].

Figure 3.1 provides a compact summary of the distribution of the different frequency band powers using boxplots. Each subplot is positioned according to the 10-20 standard. The y-axis of each subplot represents the power (μV^2) and the x-axis represents the frequency bands from delta to gamma.

Analysis of the distributions from Figure 3.1, produced from 50 datasets, shows that the delta activity is more prominent in the frontal and midline regions in the EC state, similar to results from Barry et al. [20], Barry and De Blasio [114] and [115]. The theta activity appears to be more midline, central and frontoparietal dominant similar to the results from Barry et al. [20], Barry and De Blasio [114], Michels et al. [116]. In addition, EC alpha activity is more prominent in the parietal and frontoparietal regions, comparable to that found by Barry et al. [20], Barry and De Blasio [114], Matsuura et al. [115]. The alpha band of the datasets is by far the most dominant frequency band similar to that of Chorlian et al. [117]. This dataset shows a slightly left-side dominant distribution for alpha in the frontal, central and parietal regions, similar to the data of Chorlian et al. [117].

Cave and Barry [118] found that females have greater overall amplitudes in delta, alpha, and beta bands as well as enhanced midline activity in theta and parietal and midline activity in the alpha and beta bands. This establishes significant differences between male and female EEG activity, justifying the use of similar numbers of males to females used for the EEG datasets [118].

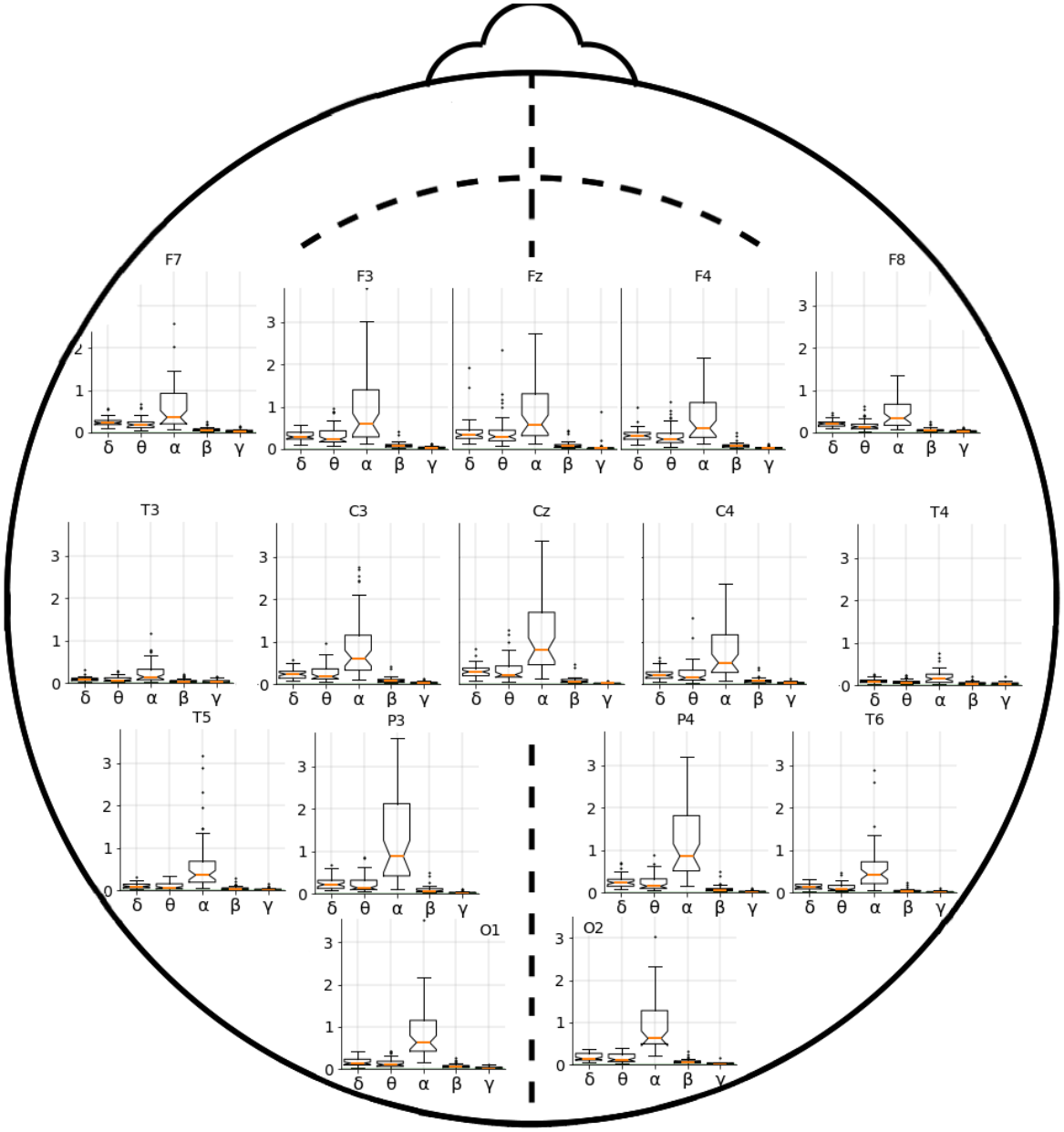


Figure 3.1: The subplots of the power μV^2 on the y-axis and different standard frequency bands, delta, alpha, beta, and gamma, $(\delta, \theta, \alpha, \beta, \gamma)$ on the x-axis. The subplots are positioned according to the 10-20 standard, with the relevant titles above them Klados and Bamidis [111].

3.3 Electrooculography reference and use

3.3.1 Electrooculography reference dataset description

Eye movement-related artefacts have the largest detrimental effect on EEG [11]. The simulated EOG signals are based on the combination of the vertical-EOG (VEOG) signals which are equal to the upper minus the lower EOG electrode recordings, and the horizontal-EOG (HEOG) signals which are equal to the left minus the right EOG electrode recordings Klados and Bamidis [77]. The VEOG and HEOG signals were independently propagated and then combined using their corresponding propagation factors, which describe their distribution in percentages to the relevant locations, acquired from Otavio G. Lins [44].

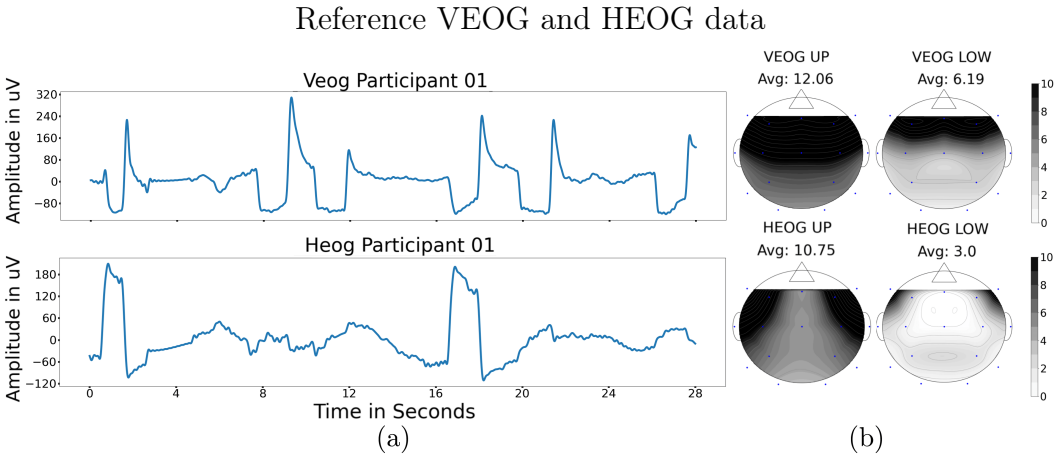


Figure 3.2: (a) Time series of VEOG and HEOG for participant one. (b) The topography from the propagation factors of the VEOG and HEOG with a scale representing the propagation factor values directly. The average propagation values are shown above the topographies. The maximum propagation factors are shown on the left and the minimum on the right topographies Klados and Bamidis [111].

The VEOG and HEOG were obtained by Klados and Bamidis [77] and measured from the same participants used for the EEG by Klados and Bamidis [77] during an eyes-opened (EO) state, using four electrodes placed above and below the left eye and another two on the outer canthi of each eye. The signals were sampled at 200 Hz and bandpass filtered between 0.5 and 5 Hz [77]. Figure 3.2(a), shows the VEOG and HEOG of one of the 50 participants. As seen in Figure 3.2(a), the VEOG, which consists mostly of eye blinks, has a much higher amplitude with a spiking characteristic, while the HEOG mostly consists of horizontal eye movements and has a lower amplitude with smoother characteristics.

The propagation factors are percentages used to determine the distribution of the VEOG and HEOG to the relevant channels, which were acquired from

Otavio G. Lins [44]. 23 Participants, 14 females and 9 males with ages ranging from 13 to 47, with a mean of 32, were used to determine the propagation factors. There were no significant differences between the male and female ages, and all the participants were healthy and had no history of ophthalmologic problems. Otavio G. Lins [44] used linear regression while simultaneously measuring EEG, VEOG, and HEOG to determine the minimum and maximum ranges of the propagation factors of the eye movements to the relevant channels. Figure 3.2(b) shows the lower and upper ranges for both the VEOG and HEOG. The units on the left bar directly correlate to the propagation factors. From Figure 3.2(b), it is clear that the VEOG has a more frontal distribution, with the HEOG having a more front-temporal distribution.

3.3.2 Creating individual electrooculography and EOG-EEG data

As mentioned in Section 3.2.1, each dataset is referred to as a 'participant'. There are 50 participants, with corresponding EEG datasets of 16 locations each. For each participant, a unique EOG contaminated EEG dataset was simulated.

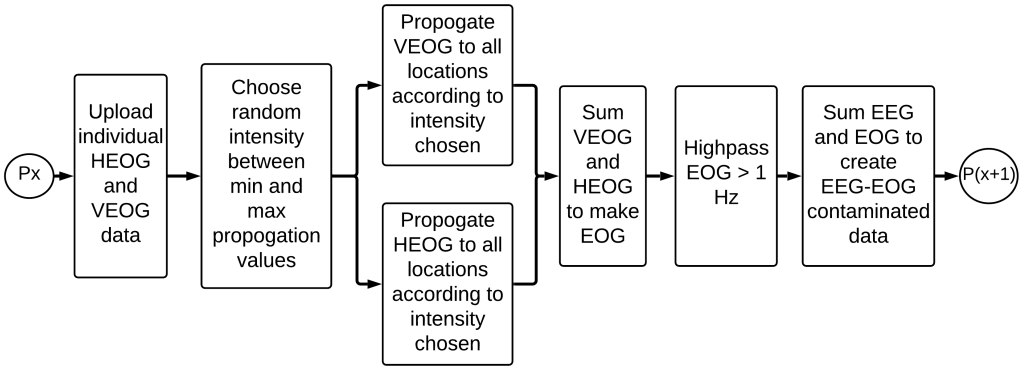


Figure 3.3: Flow diagram of creating the semi-synthetic EOG data.

The flow diagram in Figure 3.3 describes the process followed to create the EOG contaminated EEG data for each participant, as indicated by the 'P' on the left and right side of Figure 3.3. The first step in the simulation was to upload the EEG data of the relevant participant. Then, as seen in the first block of Figure 3.3, the HEOG and VEOG corresponding to the participant was uploaded. For the second block, 10 equally spaced propagation factors were generated between the minimum and maximum propagation factors for each location for VEOG and HEOG. Therefore, creating two, 10 by 16 matrices, summarised as (propagation factors from minimum to maximum for either HEOG or VEOG) by (locations). A random intensity value between

0 and 9 for each participant was chosen, which the algorithm used to fetch the corresponding propagation factors for distributing the VEOG and HEOG across the scalp. The intensity value was randomly chosen using a symmetric probability distribution. After choosing the propagation intensities, the corresponding propagation factor was used to propagate the VEOG and HEOG data, meaning that a 1 by 16 matrix for both the VEOG and the HEOG data was created by duplicating the VEOG, and HEOG signals 16 times each and multiplying them by the corresponding location's propagation factor. After propagating the VEOG and HEOG, they were summed together per channel to create the EOG data. The EOG data was then highpass filtered from 1 Hz to remove low-frequency shifts present in the data. Finally, following the linear mixture model, the EEG and EOG data were summed together per location to create the contaminated EOG-EEG data Urigüen and Garcia-Zapirain [15].

3.3.3 Creating 50 electrooculography and EOG-EEG datasets

One of the main aims of this research is to test the robustness of the cleaning methods. As one EOG contaminated EEG dataset is not enough to facilitate decisive conclusions, 50 EOG contaminated EEG datasets were created and varied as much as possible within realistic limits, validated by other literature using the SNR. Due to the HEOG and VEOG of each participant already being unique, the only independent control variable available to increase the variability of the semi-synthetic data was the intensity of the propagation factors. The propagation factor for each location for the VEOG and HEOG ranged between a certain minimum and maximum value. The propagation intensity was, therefore, randomly varied for each participant using a symmetric probability distribution.

3.4 Electromyography reference and use

3.4.1 Electromyography reference data description

It is generally acknowledged that EMG contamination of EEG is more difficult to eliminate than other types of artefacts [12]. The semi-synthetic EMG data was based on the amplitude spectra of the EEG during temporalis and frontalis at 15% contractions in combination with the amplitude spectra of the frontalis and temporalis data combined at four different percentages from Goncharova et al. [50]. Figure 3.4(a) shows the average amplitude spectra of the EEG data at 15% contraction of the frontalis and temporalis muscles for one channel. Figure 3.4(b) shows the average amplitude spectra of the EEG data at four different EMG contractions respectively for one channel. The data from Figure 3.4(a) and 3.4(b) was available for each of the 32 and 64 locations, respectively,

from which the 16 locations, Fz, Cz, C3, C4, P7, P8, O1, O2, F3, F4, T7, T8, P3, and P4 were used for the semi-synthetic EMG data.

The data in Figure 3.4(a) was obtained and averaged by Goncharova et al. [50] from 25 healthy adults, 13 female and 12 male, ranging between 16 and 53 years, with a mean age of 35 years. None had a history of any neurological or psychiatric disorders or was on chronic medication [50].

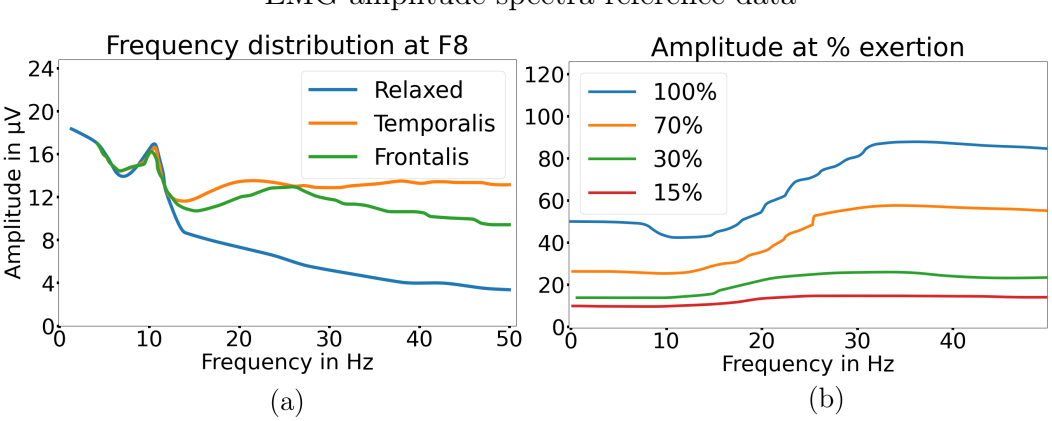


Figure 3.4: (a) The reference data of the amplitude spectra at the F8 location for the relaxed state and the 15% contraction of the temporalis and frontalis muscles. (b) The reference data of the amplitude and frequency distribution at different percentages of EMG contraction measured at the left frontalis position [50].

The EEG signals from which the data on Figure 3.4(a) originate were recorded from 64 standard locations referred to the right earlobe using an electrode cap, along with four bipolar EEG signals (right and left frontalis and anterior temporalis muscles) [50].

The EEG signals, as shown for location F8 in Figure 3.4(a), were recorded at an EO relaxed state and at 15% isometric contraction of the frontalis muscles (produced by raising eyebrows) or the temporalis muscles (produced by jaw clenching). Visual feedback was provided in percentage so that each participant could maintain the target level of muscle contraction. Furthermore, the average time measured for each participant was 153.6 seconds for relaxation, 128.9 seconds for 15% frontalis contraction and 126.0 seconds for 15% temporalis contraction. The signals were sampled at 256 Hz, and bandpass filtered between 0.1 and 100 Hz.

The data of which Figure 3.4(b) represents one channel, was also obtained and averaged by Goncharova et al. [50] from 10 participants, of which three were female and seven male, selected from the same 27 adults for the data from Figure 3.4(a). The data from which Figure 3.4(b) represents one channel, was recorded using 32 locations and four bipolar EMG signals with the same cup electrodes. The EEG and EMG were measured during frontalis or temporalis

muscle contraction using visual feedback for controlling the required contraction levels of relaxation, 15% maximum, 30% maximum, 70% maximum, and maximum voluntary contractions of the frontalis muscles (produced by raising eyebrows) or the temporalis muscles (produced by jaw clenching) [50].

The averaged data from Figure 3.4(b) represents one channel with an average time span of 58.9 seconds for relaxation, 49.1 seconds for 15%, 52.4 seconds for 30%, and 52.8 seconds for 70% contraction. The signals were sampled at 512 Hz and bandpass filtered between 0.1 and 200 Hz. The averaged amplitude spectra of the signals were recorded from four facial locations and the frontal half of the scalp during relaxation and during four levels of frontalis and temporalis muscle contractions (15, 30, 70, and 100%). 32 Standard locations, along with four bipolar EMG signals, were measured [50].

3.4.2 Creating individual electromyography and EMG-EEG data

To describe the process of creating the semi-synthetic EMG contaminated EEG dataset for each participant, it is useful to summarize the two references used and the independent variables available. The first reference used was the amplitude spectra of the EEG at each location during 15% muscle contractions for each location. The second reference was the amplitude spectra of the EEG for four different percentages of muscle contraction for each location.

The independent variables, which one can change to increase the variability of the EMG contamination across all 50 participants, are the duration and the percentage contraction of the frontalis and temporalis muscles. The researcher in this thesis adjusted the duration by changing the number of points used to represent the amplitude spectra of the frontalis and temporalis contractions. The second independent variable is the percentage of contraction. This independent variable was developed using the data for the different percentages of contractions, as seen in Figure 3.4(b) for one channel.

The data of the contractions for each percentage was re-presented as seen in Figure 3.5(a), where the y-axis is the amplitude in μV, and the x-axis is the percentage contraction. For each frequency, where only 15, 25, 35, and 45 Hz are shown in Figure 3.5(a), the algorithm fitted a linear line between the four percentage points by minimizing the squared error. This enabled the calculation of the gradient m_f, relating the change in amplitude (ΔV_f) with the percentage change (ΔP_f) at each frequency (f), as represented in equation 3.1. Figure 3.5(b) shows the results of the calculated gradients for each frequency. The gradient, m_f, from Figure 3.5(b), derived from Figure 3.5(a) for each frequency was then used in conjunction with the data from the frontalis and temporalis contractions at 15%, which is V_{15f} in equation 3.2. To calculate the new amplitude at a certain percentage and frequency V_{xf}, where x represents the new percentage of contraction, equation 3.2 was used. Equation

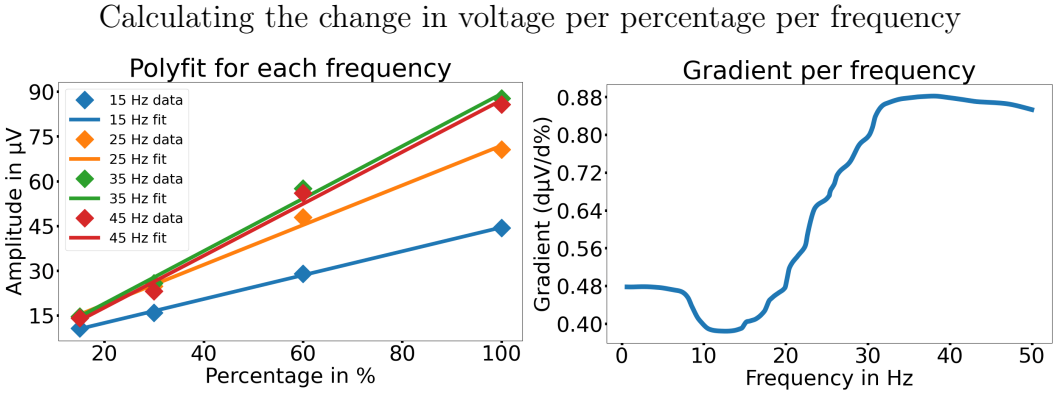


Figure 3.5: (a) Four random frequency segments from Figure 3.4(b), with the percentage as the x-axis and the amplitude on the y-axis, showing the linear line fitted through the points from each frequency to enable the estimation of the gradient for the change in voltage to the change in percentage for each frequency point. (b) The result from Figure 3.5(a) represents the gradient, which is the change in voltage per the change in percentage per frequency [50].

3.2 combined the information of the change in amplitude per percentage per frequency from Figure 3.5(b) and the amplitude of the frontalis and temporalis contraction at 15% from Figure 3.4(a) to calculate the amplitude V_{xf} at a new chosen percentage and frequency.

$$m_f = \frac{\Delta V_f}{\Delta P_f} \tag{3.1}$$

$$V_{xf} = m_f(x - 15) + V_{15f} \tag{3.2}$$

With the above processes understood, the EMG contaminated EEG data simulation can be effectively explained. From the flow diagram in Figure 3.6, the first step for each participant was to upload the amplitude spectra for the resting state EEG and the frontalis and temporalis contractions EEG at 15% for each location. The next step was to subtract the resting state EEG from the frontalis and temporalis data so that the remaining data represented the pure EMG artefacts and could be added later to the separate 'clean' EEG used in this thesis. The EMG was then highpass filtered at 15 Hz for two reasons. The first is that numerous literary works assume that EMG artefacts only affect high frequencies from 15 to 20 Hz and upwards [15, 50, 52–54]. The second reason is that Goncharova et al. [50], from which the researchers obtained the data, stated that due to the experimental setup, when measuring the frontalis and temporalis contractions, the alpha peak, between 8 and 12 Hz, was lower, but not due to EMG contamination. Rather, it was due to the active mental effort required to maintain the contractions at a certain percentage [50]. Due

to the goal of only acquiring pure EMG contamination with the subtraction of the resting state EEG, the researcher in this thesis also avoided the frequencies closer to the alpha band.

The third and fourth step was to upload the data of the amplitude spectra of the EEG at the four different frequencies for each location. From this data, as explained above, the relationship between the amplitude in μV and percentage contractions per frequency was calculated to make the percentage contraction an independent variable.

The fifth and sixth steps were to randomly choose a contraction percentage and time span between the set ranges, using a symmetric probability distribution. The researcher in this thesis then used the contraction percentage and time span to adjust the points and amplitudes at each frequency of the frontalis and temporalis reference data to simulate the separate artefacts in the time domain.

Based on the adjusted references, the seventh and eighth steps were to simulate the time domain frontalis and temporalis data, based on the Inverse Fourier transform, for each location. The temporalis and frontalis data were then separately and randomly added to the first half and the second half of the 'clean' EEG data respectively, to create the EMG contamination. The randomness for the EMG contamination was based on a symmetric probability distribution.

3.4.3 Creating 50 electromyography and EMG-EEG datasets

The 50 EMG contaminated EEG datasets were simulated for the SNR to have a large variety within a reasonable range, validated by literature. The fixed reference data consisted of the frontalis and temporalis amplitude spectra at

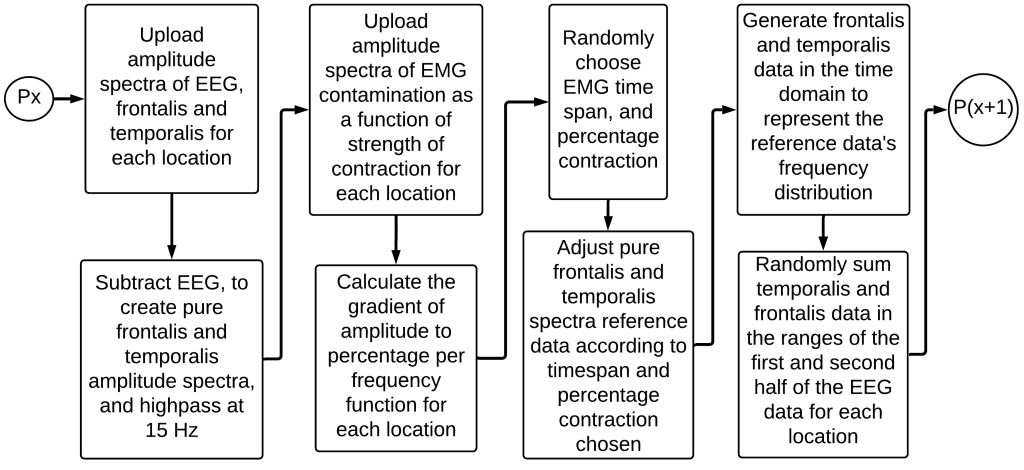


Figure 3.6: Flow diagram of creating the semi-synthetic EMG data.

15% contraction and the amplitude spectra at four different contractions for each location. The second reference data was measured from fewer participants and therefore not used as the basis data, but rather as a basis for a function to adjust the amplitude according to a certain percentage of the 15% contraction data.

The cleaning methods had to be tested on numerous data with a high variation to enable conclusions on their robustness. Therefore, the two independent variables, the time span and percentage contraction had to be varied as much as possible. The results of the SNR for all the channels when varying the contraction percentage and keeping the minimum and maximum time constant were analyzed and compared to literature to ensure that valid ranges were chosen that did not exceed realistic limits. The same was done by varying the time and keeping the minimum and maximum contraction percentage constant.

3.5 Electrocardiographyy reference and use

3.5.1 Electrocardiographyy reference dataset description

The contamination of EEG by ECG artefacts constitutes a serious problem for the automatic interpretation and analysis of EEG recordings during sleep due to its low frequencies [10]. The ECG varies significantly between participants, such as in large inter-individual voltage variations [10]. The ECG signals simulated for the semi-synthetic data are the most closely related to real data compared to the simulation of the EOG and EMG. The ECG data used was sampled from 300 single lead-I pulse recordings obtained by Khamis et al. [112] in a telehealth environment. The signals were sampled at 500 Hz using electrodes that the participants held in each hand. A reference plate was also positioned under the pad of the right hand. The collected raw data was not bandpass filtered previously. From the 300 recordings, the researcher in this thesis chose 10 random ECG recordings. The criteria for the chosen ECG data were minimal contamination and shifts by visual inspection. Other important criteria were that the 10 datasets had to vary from each other in shape, frequency, noise, shift and amplitude as for maximal variance as seen in Figure 3.7. The small number of 10 samples used was justified for the purpose of high variance due to the additional variation in the ECG temporal amplitude, the ECG sample chosen per participant and the channel at which the ECG data was added for each participant.

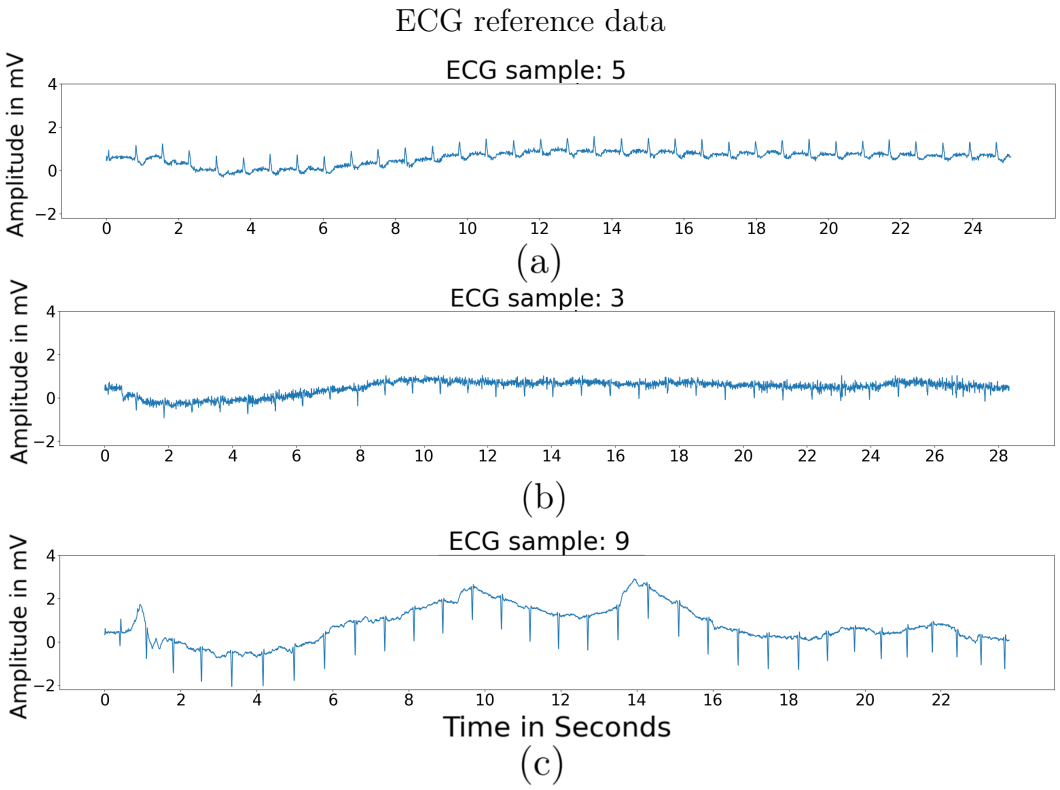


Figure 3.7: (a,b,c) Each Figure represents one of the 10 raw ECG datasets used as the reference data for ECG contamination, with time in seconds in the x-axis and mV on the y-axis [112].

3.5.2 Creating individual electrocardiographyy and ECG-EEG data

To effectively describe the process of creating the semi-synthetic ECG contaminated EEG data, it is helpful to understand what cannot be changed, what can, and why. The 10 obtained ECG datasets are from actual measured data and can only be adjusted to an extent. What can be changed is the range to bandpass the data and the amplitude and channels to which the ECG is added.

The average beats per minute were used as a guide to choose the bandpass filter frequency. The average heart rate of a healthy adult is generally considered to range between 60 and 100 beats per minute [60]. The ECG pulse has a QRS complex, representing depolarization of the ventricles. The QRS complexes are three closely related waves that can be observed on the ECG, namely the Q, R and S waves. To maintain the information of these QRS peaks and reduce the baseline shifts and high-frequency noise present in the obtained raw data, the ECG data was bandpass filtered between 3 and 4 Hz [59, 60, 119].

The amplitude reduction and channels to which the ECG was added, were

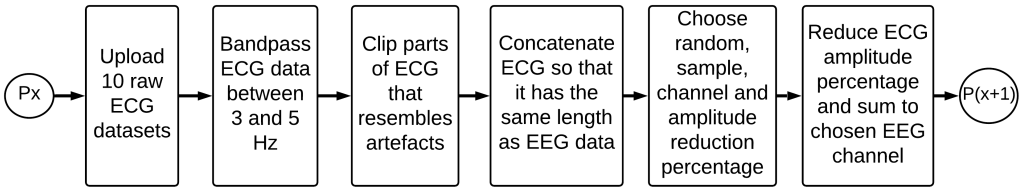


Figure 3.8: Flow diagram of creating the semi-synthetic ECG data.

based on the origin of the ECG. ECG artefacts occur when an electrode is placed over pulsating vessels, such as a scalp artery [15, 120]. Pulse artefacts usually affect only one electrode as they are caused by pulsating scalp arteries lying directly under the electrode and they can be minimized to zero by placing the electrodes correctly [15, 120]. Due to the unlikeliness of more than one electrode lying directly over a scalp artery, but to still contaminate each dataset, one random ECG sample contaminated one channel for each participant [15, 120]. The original amplitude of the ECG would also be reduced by a randomly chosen percentage to increase variation and compensate for the relative position of the electrode to the artery. All the random generations were based on a symmetric probability distribution.

Understanding the more complex parts of the ECG simulation will allow a higher level explanation based on the flow diagram to be better understood. The first step in the flow diagram in Figure 3.8, was to upload the 10 chosen ECG data sets. The second step was to bandpass filter all the ECG data between 3 and 5 Hz to ensure that the activity present is from the source of a pulse and not other sources or artefacts such as EMG. The third part was to clip parts of the ECG signals that showed unusual patterns possibly due to non-physiological artefacts such as a disruption in the signal. Then, given that the clipped parts of the ECG were shorter than the 'clean' EEG data, and the ECG usually occurs throughout the entire time duration because the electrodes remain stationary, and the ECG has a repetitive pattern, the ECG data was concatenated until it represented the same length as the participant's EEG data. A random ECG sample out of the ten samples and the reduction percentage to reduce the amplitude were then chosen. The randomness was based on a symmetric probability distribution. To be compatible with the 'clean' EEG data before finally adding the ECG to the chosen EEG channel, the ECG data was downsampled from its original 500 Hz to 200 Hz.

3.5.3 Creating 50 electrocardiographyy and ECG-EEG datasets

The 50 ECG contaminated EEG datasets were created to have an SNR with distribution as wide as possible within a reasonable and comparable range to other literature, therefore supporting one of the main aims of testing the ro-

bustness of cleaning methods. The fixed data was the 10 ECG sample datasets.

The cleaning methods had to be tested on numerous data, with a high variation, to enable conclusions on their robustness. To account for this, the two independent variables, the choice of samples and percentage reductions, were identified to be varied within limits set by literature. The random EEG channel to which the algorithm added the ECG was not considered a variable because each channel for all 50 participants was contaminated roughly equally in amount. All 10 ECG samples were used for the ECG contamination in order to increase variation. Therefore, the percentage reduction was the only remaining independent variable. The results of the SNR for all the channels when varying the percentage reduction while keeping the minimum and maximum ECG sample (determined by amplitude) constant were analysed and compared to literature. This was done to ensure that valid ranges were chosen that did not exceed realistic limits.

3.6 Combined reference and use

3.6.1 Combined dataset description and methodology

The problem of removing different kinds of artefacts simultaneously is of great significance [68]. The simulation of the EEG dataset contaminated by a combination of EOG, ECG and EMG is only dependant on the simulations of the pure EOG, ECG and EMG artefacts. Due to the effort put into increasing the variability of the EOG, ECG and EMG as much as is reasonable, the results of the combined contamination had a high variation.

Therefore, following the linear mixture model assumption, the contaminated EEG results from 'clean' EEG summed with the artefacts. The assumption was further extended to the three physiological artefacts being independent and therefore summed together. The results of the final simulation methodology can be described by equation 3.3 and 3.4, where the combined artefacts is CMD_{ij} and the EEG combined with artefacts is CMD_EEG_{ij} [15]. In equation 3.3 the different artefacts, before they were added to the 'clean' EEG data, were summed together to create a combination of artefacts. Finally, in equation 3.4, the combination of the artefacts was added to the clean EEG per participant, i, and for each separate channel j.

$$CMD_{ij} = EOG_{ij} + ECG_{ij} + EMG_{ij} \tag{3.3}$$

$$CMD\_EEG_{ij} = CMD_{ij} + EEG_{ij} \tag{3.4}$$

The purpose of the semi-synthetic data is to test, develop, and quantify the effectiveness and efficiency of the cleaning methods. With the methodology of the semi-synthetic data established, the methodology of the different cleaning methods used can effectively be discussed.

3.7 Blind source separation methods

3.7.1 Blind source separation methods overview

In Figure 3.9, one can see the process used by the BSS methods to clean the contaminated EEG data, for either EOG, EMG, ECG or a combination of these artefacts for each individual participant.

The first step was to upload the contaminated data for a participant. The second step, as indicated in Figure 3.9, was to identify all the components which can be seen as the estimated sources of the measured data, using the statistical restrictions of each method, which were tested separately, be it either Extended Infomax, SOBI or CCA. After the components were estimated by the algorithms, the artefact components were manually identified and marked using the methods described in Section 3.7.3. These artefact components were then manually removed from the mixing and component matrices before calculating the cleaned data, as shown in the last block of the flow diagram and further explained in Section E.1.

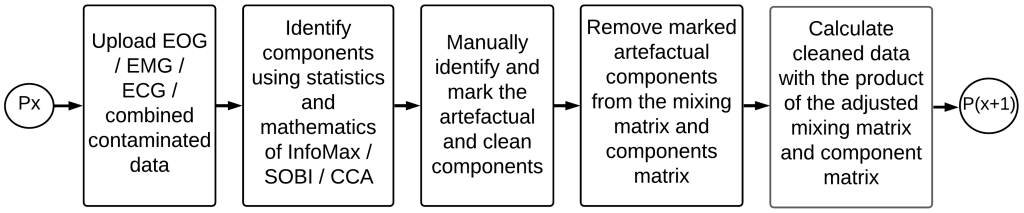


Figure 3.9: BSS process flow diagram.

3.7.2 Blind source separation mathematical methods

The method described below was implemented with Extended Infomax, SOBI and CCA. The only difference is the statistics used to calculate the unmixing matrix W. The components from which the researcher in this thesis can identify the artefact components were calculated as shown in equation 3.5, where $S = [x_i(j)]_{n \times N}$ is the estimated components, with N being the total samples, n the number of channels and M the number of components. BSS methods can only identify as many components as the channels used. Therefore this equation always holds $M \leq n$, and in this case, the BSS algorithms calculated the maximum number of components for each case; therefore, it can be stated that $M = n$. Furthermore, the unmixing matrix, $W = [w_i(j)]_{n \times n}$ is used to separate the measured data into its estimated sources S. Where $X = [x_i(j)]_{n \times N}$ is the data measured from each channel n [13, 15, 121].

$$\mathbf{S} = \mathbf{WX} \tag{3.5}$$

Once the components, S, have been identified using the unmixing matrix, W, the components containing contamination can be identified by closely inspecting the characteristics of the time, frequency and other characteristics. When the contaminated components have been identified, the chosen clean components indexes to keep, $\hat{M}$, are determined, which is used to calculate the cleaned data. The cleaned data is estimated by first computing the mixing matrix, $A = [a_i(j)]_{n\times n}$, which is the inverse of W, as shown in equation 3.6. Then, as shown in equation 3.7, the cleaned data, $\hat{X} = [\hat{x}_i(j)]_{n\times N}$, is calculated by multiplying only the selected clean indexes, $\hat{M}$, of the mixing matrix, $\hat{A} = [\hat{a}_i(j)]_{n\times\hat{M}}$ and components $\hat{S} = [\hat{s}_i(j)]_{\hat{M}\times N}$ [13, 15, 121].

$$\mathbf{A} = \mathbf{W}^{-1} \tag{3.6}$$

$$\hat{\mathbf{X}} = \hat{\mathbf{A}}\hat{\mathbf{S}} \tag{3.7}$$

The mathematical and statistical derivations of the BSS methods can be found in Appendix E. Furthermore, the mathematical and statistical derivations for ICA and CCA can be found in Appendix E.1 and E.2 respectively. Additionally, the mathematical and statistical derivations of Extended Infomax and SOBI, derived from ICA can be found in Appendix E.1.1 and E.1.2 respectively.

3.7.3 Identifying contaminated components

To determine whether a component produced by one of the BSS methods represented an artefact or 'clean' EEG, the researcher in this thesis first analysed five sub-plots, each shown in Figure 3.10 for the four different types of components identified during the cleaning and testing of the methods. Referring to Figure 3.10(a), for EEG, the top subplot represents the time series data of the component, providing valuable information on the amplitude, duration and pattern of the component. EEG components represented the full duration, with a varying pattern throughout the time series, with artefacts being less consistent in duration. The second important subplot is the topography of the identified component, with red indicating a higher amplitude and blue a lower amplitude. An artefact would have a high concentrated amplitude relative to the location from which it originates. The third subplot, the spectrum subplot, shows the frequency distribution of the component, with EEG usually having a peek at alpha and artefacts being much higher at either lower or higher frequencies. The segment image and event-related potential over event-related field (ERP/ERF) subplot consist of two additional subplots. The subplot at the bottom is the component's two-second segments' time series. The top subplot is the amplitude at each separate segment, corresponding at the x-axis with the time series below it. In this case, there are 14 segments, with red indicating high amplitudes and blue low amplitudes at the different time points

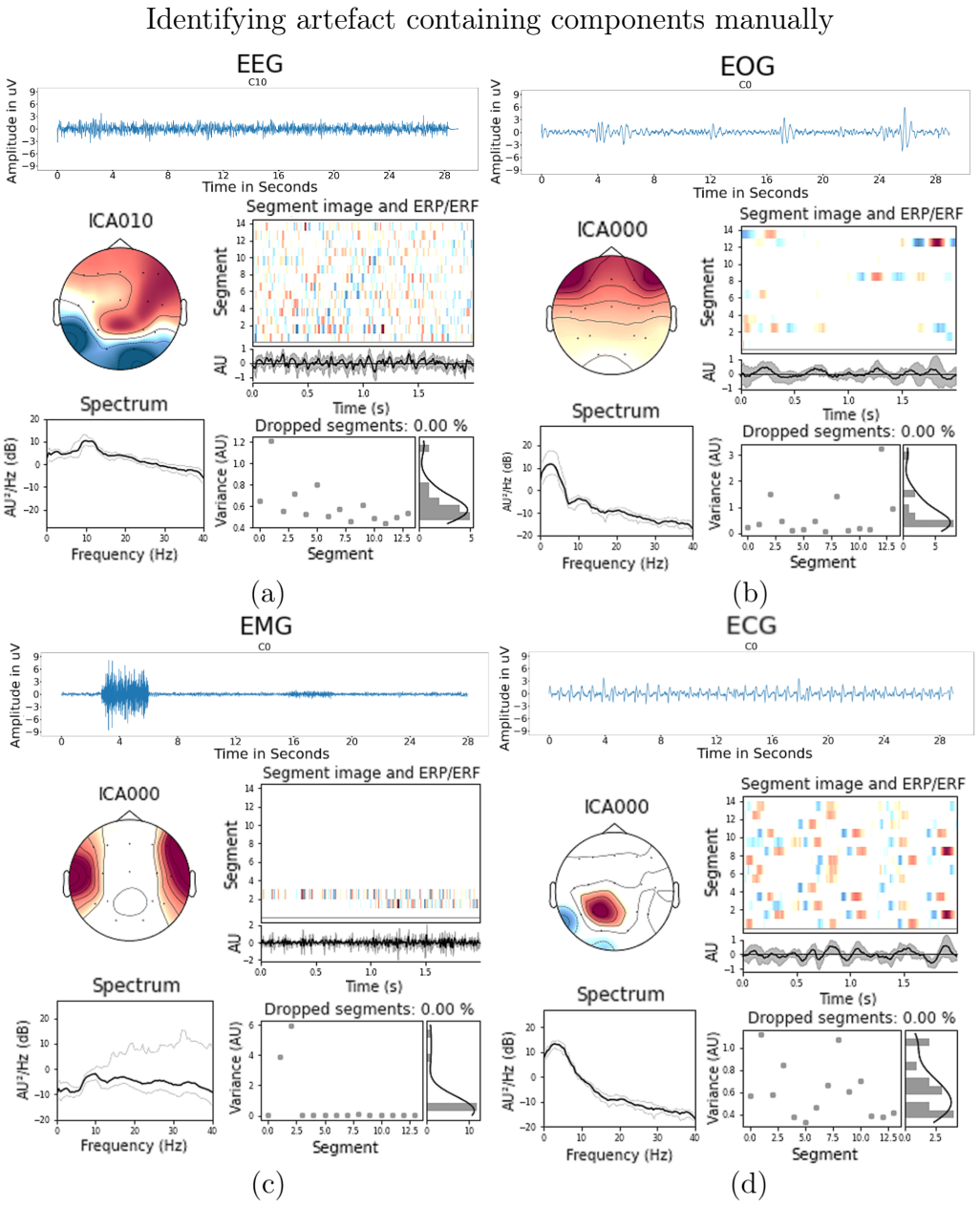


Figure 3.10: There are four parts for each artefact, with five subplots each. The time-series subplot provides information on the amplitude and duration. The topography subplot shows higher amplitude in red and lower in blue. The spectrum subplot shows the frequency distribution. The event-related potential over event-related field (ERP/ERF) subplot has a time series of 2-second segments from the component at the bottom. The top of the ERP subplot represents the amplitude at each separate segment, corresponding at the x-axis of the time of the data below it. The y-axis represents each segment, with the red and blue corresponding to the high and low time series amplitudes respectively. The variance subplot shows the variance for each segment [50, 111, 112].

for that segment, with segments indicated on the y-axis. Artefacts usually show activity at only some segments, while EEG shows continuous activity. The final subplot is the dropped segments subplot, showing the variance for each segment and provides much of the same information as the ERP/ERF subplot, with EEG components varying steadily throughout and artefacts components having less frequent but more significant variance. These five subplots were used in combination with what we know of EEG, EOG, EMG and ECG to identify whether the components were in fact from actual EEG or artefacts; these decisions were based on literature from Sweeney et al. [11], Daly et al. [13], Urigüen and Garcia-Zapirain [15], Barry et al. [20], Otavio G. Lins [44], Goncharova et al. [50], Uhlig et al. [54], Sakai and Wei [59], Jose and Collison [60], Barry and De Blasio [114], Matsuura et al. [115], Chorlian et al. [117].

On identifying pure EEG components, referring to Figure 3.10(a) with the time series, a consistent and wide variance in frequency throughout, with high amplitude alpha activities at times, suggested that the component represented EEG. The topography of the pure EEG showed more widespread activity in comparison with the artefacts. The ERP/ERF segment subplots for the EEG showed activity in all the segments, with more variance in amplitudes for each segment when compared to some artefacts. The spectrum subplot for EEG components represented a pattern following the trend of $\frac{1}{f}$, where f represents the frequency, with a peak at the alpha frequency range. With the variance subplot, the EEG variance resembled a symmetric probability distribution [20, 114, 115, 117].

To identify the EOG components, Figure 3.10(b) shows all the information used to confirm whether a component represented an EOG artefact. With the time series, the EOG components showed high amplitude and low-frequency bursts, resembling the familiar pattern of EOG time series data. The topography of the EOG components showed high amplitude at the anterior regions of the head, as does the topography for the EOG in Figure 3.10(b). For the EOG ERP/ERF segment plot, some segments showed high activity, while others showed very little activity. The reason for this is the inconsistent activity of the EOG. On the ERP/ERF subplot, the bars are also wider because the frequency for the EOG is lower. In the spectrum plots for EOG data, lower frequencies showed higher amplitudes. Just like the ERP segment subplot, the EOG only showed variance at some segments and no variance at others with the dropped segments subplot [11, 15, 44].

Figure 3.10(c) shows all the information used to determine whether the component represented an EMG artefact. With the time series of the EMG components, the EMG only occurred at certain time segments, with no activity at other time segments. The EMG components also had high amplitudes and frequencies. The topography of the EMG originated and was more concentrated at the temporalis or frontalis areas than other artefacts. The EMG component shown in Figure 3.10(c) is contamination from the temporalis area,

originating from jaw clenching and mouth movement. The segment plots for the EMG showed high-frequency activity in only some of the segments, with almost nothing at other segments. The spectrum subplot for EMG showed a high amplitude at higher frequencies, as seen in Figure 3.10(c). The EMG variance showed the same type of distribution as the segment subplot, with the only variance at a few segments [15, 50, 54].

Referring to Figure 3.10(d) on identifying ECG components, the time series of the ECG represented repetitive constant and low-frequency patterns similar to a heart rate. The topography of the ECG components has a high amplitude at one concentrated random location, corresponding to the electrode affected. The segment plots for the ECG showed a repetitive pattern for each segment, with the red and blue bars almost aligning vertically. The bars are also wider because the frequency is lower for the ECG component than the EEG components, as shown on the ERP subplot. The spectrum subplot for ECG data showed a high amplitude at lower frequencies similar to EOG artefacts. With the variance subplot, the ECG, like the ERP segment subplot, showed variance at most segments, closer to the distribution of the EEG but with a higher variance [13, 15, 59, 60].

3.7.4 Identifying contaminated components automatically

For the automation of the semi-automatic BSS methods, the identification of the artefact components had to be automated, being the only part in the BSS methods that relied on manual input. As found in literature, the approaches for automating the BSS methods were often complex, including machine learning algorithms which added significant computational and required manual input from experts to effectively label the training data into clean EEG and artefacts respectively [90–94].

The automation method described below is a novel method developed by the researcher of this thesis. The method developed was a simple but effective method that emphasised the distinguishing amplitude spectra characteristics of the EEG and artefact components for separation using predetermined thresholds. The automation of the artefact components was based on the assumption that the amplitude spectra of the 'clean' EEG components had an alpha peak in the range of 8 to 11 Hz. In comparison, the ECG and EOG artefact components peaked at lower frequencies of 2 to 4 Hz, and the EMG artefacts peaked at higher frequencies in the range of 15 to 49 Hz. The algorithm further emphasised these assumptions by looking at the fourth power of the amplitude spectra, therefore emphasising the peaks and lowering less significant frequency amplitudes. The fourth power was chosen after testing different powers and heuristically identifying it as the most effective power for the algorithm. Figure 3.11 shows the logic of the process with a flow chart,

and Figure 3.12 further explains some of the processes discussed in the flow chart.

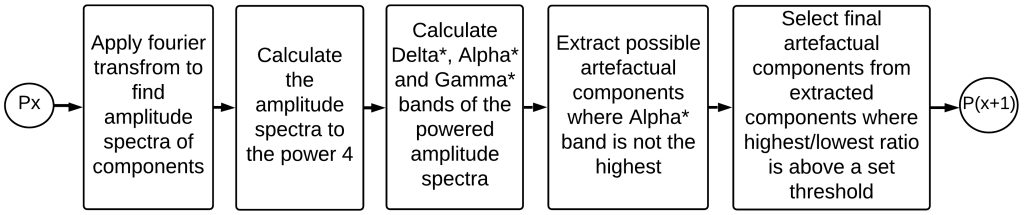


Figure 3.11: Identifying artefact containing components automatically flow diagram.

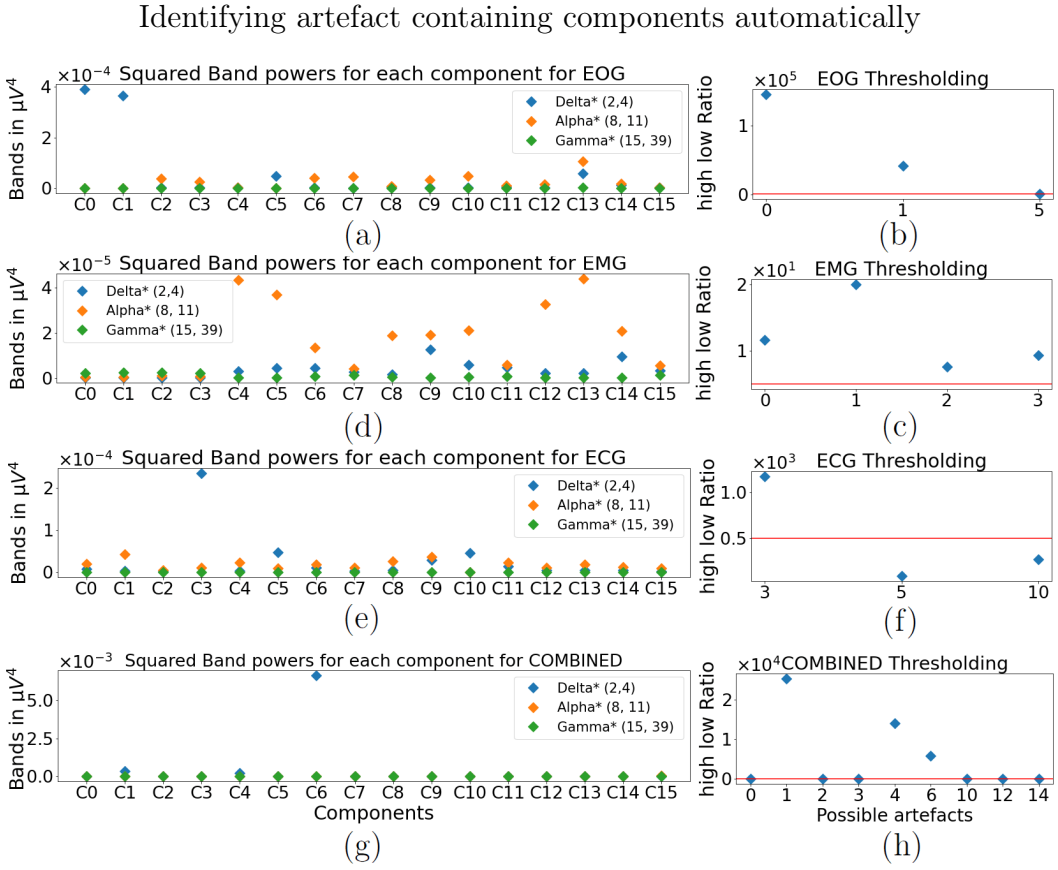


Figure 3.12: The left subfigures represent the amplitude spectra in μV^4 of each frequency band, for each component, for each contamination type. The right subfigures represent the ratios of the highest over the lowest frequency bands of the chosen components, used alongside a threshold to determine whether the suggested components are actual components.

As shown in the flow diagram of Figure 3.11, the first step in the automatic identification of the components was to find the amplitude spectra of each component identified by the BSS methods. The second and third steps, shown in the second and third block Figure 3.11, and as explained previously, were to calculate the average bands of the amplitude spectra to the power of four. In the fourth block, the relationships between the amplitudes of the calculated bands for each component were visualised. Assuming that the 'clean' EEG components have the highest peak, where this was not the case, the components were chosen as suggested artefact components.

Figure 3.12(a) further illustrates this process, showing the fourth power of each bands' amplitude spectra for the EOG component. As seen in Figure 3.12(a), only components zero, one, two and five did not have the modified alpha band as the highest band and were therefore selected for the second stage as possible artefacts components. As described by the last block of the flow chart in Figure 3.11, a threshold was imposed as seen in Figure 3.12(b). Using the ratio of the highest band over the lowest band was based on the hypothesis that it quantified the extent of the component's artefact characteristics. The method, therefore, aimed to emphasise and quantify artefact characteristics with the use of a threshold to determine whether the suggested artefact components were indeed actual artefact components. Referring to Figure 3.12(e) and 3.12(f), one can see that the algorithm selected components three, five and ten as possible artefact components. As shown by Figure 3.12(f), the component that actually represented the ECG component, using the highest band to lowest band ratio, was much higher than the other components. As a result, the component was above the chosen threshold and selected as a definite artefact component to be removed from the data.

3.8 The auto threshold method

The auto threshold method described below, based on the standard deviation of the data of each channel, is a novel method developed by the researcher of this thesis. The auto threshold method is a simple method that was implemented and tested alongside the BSS methods. The justification of the method was the fact that artefacts such as EMG, EOG and ECG often have significantly larger amplitudes than the EEG itself, making a threshold method a possible cleaning alternative [10, 15, 44, 50]. Using a threshold that is not adaptive causes problems since each channel has a different amplitude depending on the region and the quality of electrode contact. Additionally, each channel has different types or combinations of artefacts. In this case, it was decided to make the thresholds individually dependent on the standard deviation of each channel and set the thresholds to be based on a preset number of standard deviations. It was assumed that the data would have a mean of zero microvolts, so the thresholds were equal in magnitude but set above and below

the zero points of the data. By fitting a five-degree polynomial through the data, any form of a shift in the baseline of the data was found and removed by subtraction, further confirming the centre of the distributions to be at zero voltage. The degree of the polynomial was increased until the fifth degree, where it could be heuristically and visually deduced that the slow baseline shifts were the most effectively captured. The number of standard deviations and intersections was then determined from this centred data. As a result, any shift in the baseline of the data did not affect the calculation of the number of standard deviations for finding intersection points of the artefacts. With the deeper calculations in the auto threshold method, the rest of the process can be effectively explained using the flow diagram in Figure 3.13.

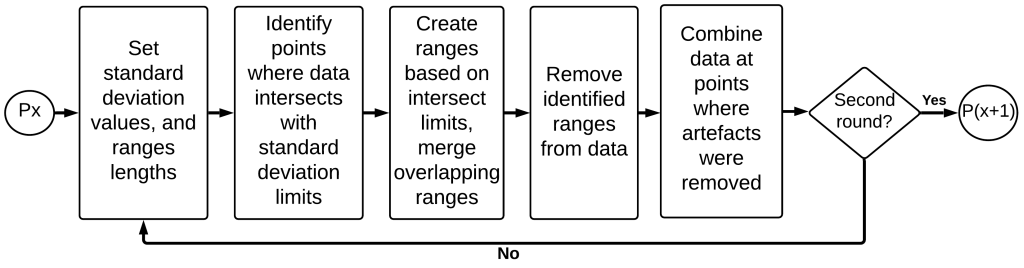


Figure 3.13: Flow diagram of cleaning data with an auto threshold method.

As seen in the first block in Figure 3.13, the initial step in the automated method was to set the parameters for the number of standard deviations and ranges. Highly contaminated data may perform better with a larger number of standard deviations due to the data having larger amplitudes and relatively clean data may perform better with fewer standard deviations. The ranges determined how much data was removed when an intersection was identified. After the values for the number of standard deviations and ranges were set, the data was centred as described above. Each time point that intersected with the thresholds determined for each channel, was marked. Then, as described in the third block of the flow diagram, ranges were determined based on preset range lengths and centred at the intersecting points. The algorithm merged overlapping ranges into one larger range. As shown in the fourth block, the next step was to remove the data between the identified ranges. The fifth step was to restore the continuous state of the data by splicing the points where the algorithm removed the ranges. After this process, it could be expected that the algorithm had removed most of the significant artefacts. The separate removal of the contaminated ranges causes the time series of the channels not to line up. However, due to the Fourier transform providing a summed spectral analysis across the entire signal, the information that would have been gathered by synchronising the time series of the channels is disregarded. Therefore, the synchronising of the channels in the time domain is not necessary in this case.

Due to the large artefacts affecting the standard deviation of the data, the thresholds determined with the first standard deviation may have missed more subtle artefacts. As shown with the decision block in Figure 3.13, the whole process was repeated once again, this time with a smaller number of standard deviations and usually a smaller range.

3.9 Technical implementation

All four methods, namely Extended Infomax, SOBI, CCA and the auto threshold method, were tested and compared to each other and to findings reported in the literature. In this thesis, the researcher tested their performance regarding the increase of the SNR of the contaminated data and the average time it took to identify the components per method.

Regarding the consistency of testing the effectiveness of each method in increasing the SNR of the contaminated data, the researcher in this thesis separately used each method to clean the semi-synthetic EOG, EMG, ECG and combined contaminated data for each of the 50 participants. The results were separately stored for each method, participant and type of semi-synthetic contamination. Using the methods described in Section 3.7.3, the contaminated components were identified and removed, as shown in equation 3.7. Regarding the time that each method took to clean the data, the following can be noted: for the BSS method, only the time before and after the calculation of the components was recorded and saved for each algorithm. For the auto threshold method, only the time it took for the algorithm to identify the ranges of the artefacts were measured. During this process, all other programmes and unnecessary background processes on the computer were closed, except for the ones used to implement the algorithms.

The methods were written on Python 3.8.5, with further information provided in Appendix G.1, while using Jupyter Notebook as the integrated development environment (IDE). The cleaning methods were either written from scratch or partially, using relevant libraries. No toolboxes were used for the implementation of the cleaning methods. Extended Infomax was partially coded, with the use of the MNE-Python 0.23.0 [122] library, combined with NumPy 1.20.1 [123] and Matplotlib 3.3.3 [124] for additional necessary matrix calculations and components visualisations. SOBI was written mainly from scratch, using NumPy 1.20.1 [123] for all the matrix calculations and Matplotlib 3.3.3 [124] for the visualisation of components, topographies, amplitude spectra etc. The CCA algorithm was also implemented from scratch, using mostly the scikit-learn 0.24.2 [125] library, with NumPy 1.20.1 [123] and Matplotlib 3.3.3 [124] for additional matrix calculations and visualisations.

The simulations and cleaning methods were implemented and tested on a HP laptop with a 7,88 GB usable ram and an Intel(R) Core(TM) i7-8550U processor. Therefore the system has a processor base frequency of 1.80 GHz.

The laptop runs on a 64-bit operating system. The operating system used was Windows 10, version 21H1.

3.10 Statistical analysis methods used

The semi-synthetic data and cleaning methods were analysed using boxplots, t-tests and a description of these methods can be found in Appendix F.

The researcher in this thesis analysed the semi-synthetic data by comparing their characteristics to other actual artefacts discussed in the literature. Among the characteristics, the SNR was also used as a performance evaluator to compare the intensity of the different artefact contaminations to each other and the findings in the relevant literature. The SNR for the semi-synthetic data was represented in the form of boxplots. The SNR values for the 16 channels of all 50 participants were grouped. Therefore, the boxplots represented 800 data points per artefact (except for ECG containing 50 data points). The SNR for the semi-synthetic data was also further compared to each other and relevant literature by only grouping certain regions together, for example, only representing the data of the temporal location (T3, T4, T5, and T6). Therefore, four channels for 50 participants, meaning 200 SNR data points, were used for further in-depth comparisons between simulations and literature.

The performances of the cleaning methods were evaluated and compared to each other and to literature. The analysis of the SNR of the cleaned data as well as the difference between the SNR of the cleaned data and the contaminated data was done using boxplots and the t-test. A p-value greater than 0.05 was used as an indicator that there was no significant difference between two datasets. The cleaning methods were also compared to each other and literature using the average SNR of each channel of the cleaned data. The average time that each BSS method took to identify the components was compared to each other and literature using boxplots and the t-test. For the auto threshold method, the average time it took to identify the artefact ranges was used.

Chapter 4

Results and Discussion

4.1 Electrooculography results

4.1.1 Results and discussion of semi-synthetic electrooculography time-series

Figure 4.1(a) shows the results of the temporal propagation and summation of the HEOG and VEOG signals, creating the pure EOG signal at the Cz position. The EOG data oscillates between an amplitude under 100 μV with bursts lasting about one to three seconds. Figure 4.1(b), represents the results of the 'clean' EEG in blue and the contaminated EEG in black, which is the summation of the EOG and EEG signal at location Cz. Figure 4.1(c) is a close-up view of Figure 4.1(b), showing the contaminated data in black and clean data in blue between the zero- and four-second range.

The simulated EOG in Figure 4.1(a) shows similar slow frequency, high amplitude, and brief patterns as the EOG simulated by Zeng et al. [45]. The EOG observed is non-stationary, therefore varying in characteristics such as frequency and amplitude with time. The frequency of a non-stationary wave changes constantly during the process, as observed from real EOG signals by Sanjeeva Reddy et al. [48]. The amplitude of the EOG signal as observed in Figure 4.1(a) also occurs within the amplitude and frequency of 10 to 100 μV and 0 to 10 Hz as observed of EOG signals by Sanjeeva Reddy et al. [48]. The duration of the EOG signals only lasts for a few seconds as observed in EOG by Sanjeeva Reddy et al. [48] and Venkataramanan et al. [49].

As shown in Figure 4.1(b), the 'clean' EEG, having an amplitude typically under 50 μV, was increased through contamination to amplitudes exceeding 100 μV Gratton [47]. Figure 4.1(b) displays how the contaminated EEG shows drastically different patterns and higher amplitudes at certain time segments, completely distorting the parts of the 'clean' EEG time series data when observed from a distance when the EOG signal is present.

When further investigating the contaminated EEG, for instance, between

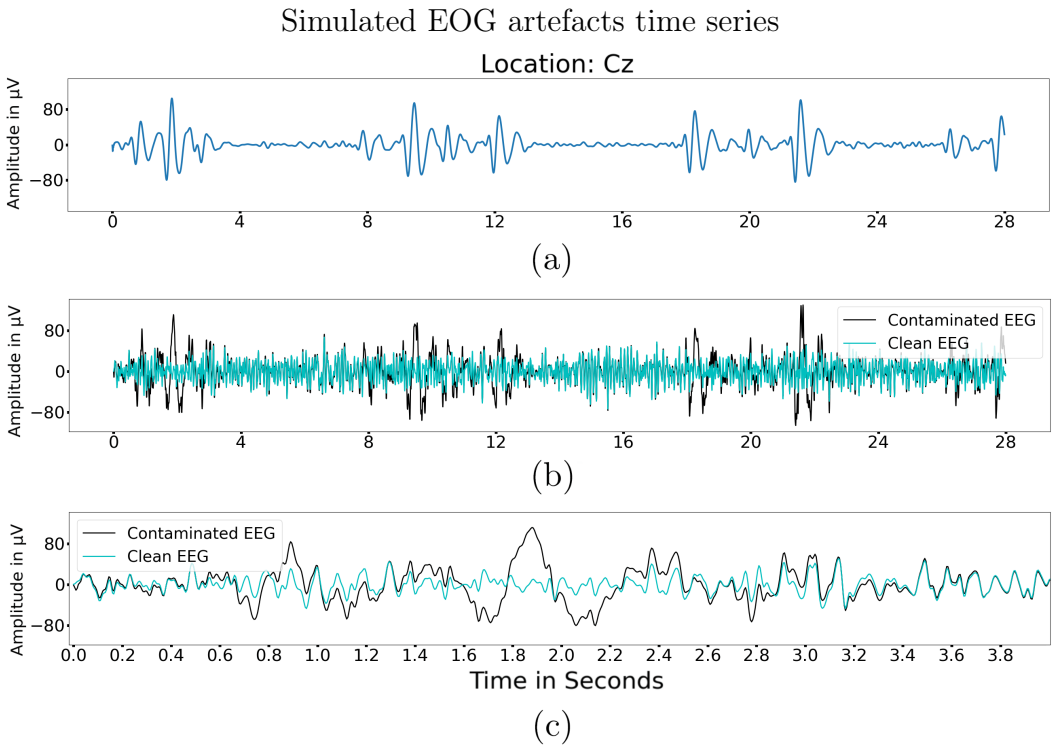


Figure 4.1: (a) Pure EOG data at location Cz after propagating and combining the HEOG and VEOG data. (b) The contaminated and 'clean' EEG. (c) An expanded view of Figure 4.1(b) [111].

zero and four seconds as shown in Figure 4.1(c), one can see that for the most part, the contaminated EEG follows the same frequency patterns as the 'clean' EEG, undergoing mostly a shift rather than a complete change in the pattern. At the peaks of the EOG contamination, such as between 1.8 and 2.0 seconds, the 'clean' EEG patterns are completely obscured.

4.1.2 Results of varying electrooculography and EOG-EEG datasets

Figure 4.2(a) shows the results of the SNR of the semi-synthetic EOG contaminated EEG for all 16 locations when varying the intensity of the VEOG and HEOG propagation values from minimum to maximum. On the x-axis of Figure 4.2(a), the propagation intensity is equally spaced from 10 to 100. The boxplot at each intensity value represents the SNR results of all 50 semi-synthetic EOG contaminated EEG datasets when the corresponding propagation values were kept constant for the simulation. Figure 4.2(b) represents the same results as Figure 4.2(a), with a small difference in representation. In this case, each boxplot only represents one channel for all 50 participants, showing the results of the most affected channel, F8, and the least affected channel, O2, when keeping the propagation intensities constant during the simulation.

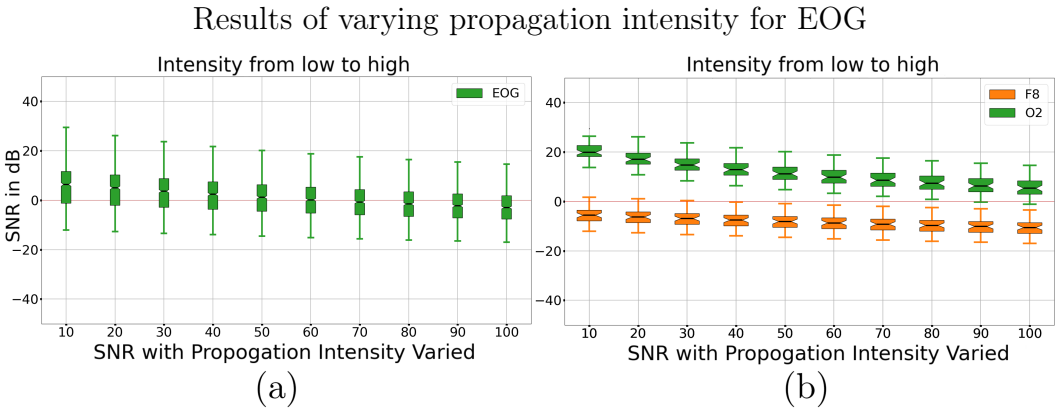


Figure 4.2: (a) The boxplots of the SNR of 50 participants when the intensity of the EOG varied between the minimum and maximum intensities for the whole head. Therefore, each boxplot represents 800 points. (b) The same data representation as Figure 4.3(a), but only for two separate channels. The channels chosen were F8 and O2, representing only 50 data points for each boxplot [111].

To further validate the semi-synthetic EOG contamination, the SNR was compared to the ranges used in related research. The SNR values found in the literature were summarised in Table H.1, Appendix H. The research conducted by Bai et al. [126], Sadasivan and Dutt [127], Merino et al. [128], Naga et al. [129] and Puthusserypady and Ratnarajah [130] involved the contamination of one channel numerous times while varying the SNR from minimum to maximum values, whereas Cheng et al. [131] and Paulson and Alfahad [132] contaminated numerous different channels, but with the same level of SNR, also varied from minimum to maximum for each channel.

From the SNR of the semi-synthetic EOG contaminated EEG, in Figure 4.2(b), F8 has a minimum SNR of about $-18\,\text{dB}$, and O2 has a maximum SNR of about $24\,\text{dB}$, making these the minimum to the maximum range of the SNR of the simulated data. Referring to Table H.1, the ranges are comparable to that of Puthusserypady and Ratnarajah [130], at $-20\,\text{dB}$, and Naga et al. [129], at $23\,\text{dB}$, and well within the limits of those of Paulson and Alfahad [132], at $-40\,\text{dB}$, and Merino et al. [128], at $60\,\text{dB}$. Therefore, it is justified to use the whole range of propagation intensities as it produced SNR values comparable and within limits to those from other studies, maximizing the variation of the semi-synthetic EOG contaminated EEG.

4.1.3 Results electrooculography and EOG-EEG topography

Figure 4.3 shows the results of the average topography of the 50 semi-synthetic simulated EOG contaminations, using a random symmetric distribution to determine the propagation intensity for each participant in combination with the unique VEOG and HEOG data that was recorded for each participant.

The columns represent the different EEG and artefact scenarios in Figure 4.3, and the rows represent the frequency bands. The data represented is the topography of the 'clean' EEG, the EOG contaminated EEG, the pure EOG, and the VEOG and HEOG. The scale used to compare the topography's distribution to itself and the other topographies is a log scale, calculated as $10\log(\mu V^2)$, representing the band powers in dB. The average band powers in μV^2 are shown at the top of each topography map. In Figure 4.3, only the delta and theta bands are shown, as the alpha, beta and gamma bands were not significantly affected by the semi-synthetic EOG data.

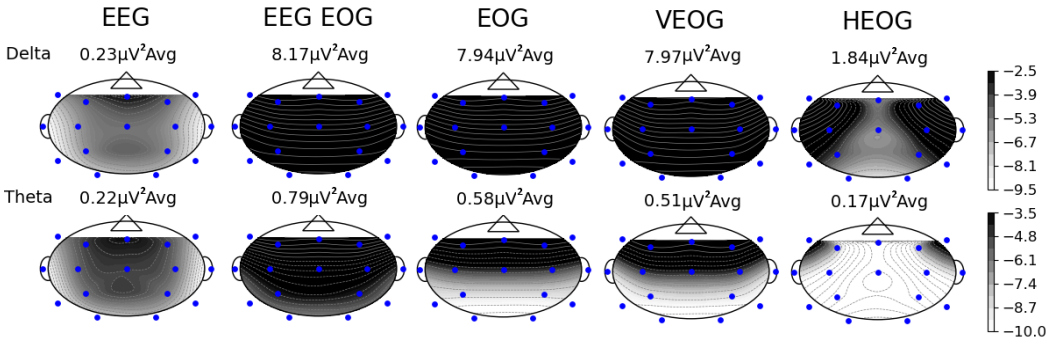


Figure 4.3: The columns represent the different contamination scenarios, and the rows represent the frequency bands. The data represented is the topography of the 'clean' EEG, the EOG contaminated EEG, the pure EOG, and the VEOG and HEOG. The scale used to compare the topography's distribution to itself and the other topographies is a log scale, thus calculating a $10\log(\mu V^2)$, representing the band powers in dB. The average band powers in μV^2 are shown at the top of each topography map [111].

From the topography, one can conclude that the distributions in the lower frequency bands are completely obscured, with the average voltage for the delta band and theta band increased by about 35 and 3.6 times from their original average voltage by the EOG contamination. Therefore, the theta band is the most affected by the EOG data, obscuring any possible EEG topography analysis. Delving into the source of the high voltage, it could be as a result of the VEOG being a stronger contributor to the theta band contamination, with an average voltage of 7.97 μV^2 in comparison to the HEOG with an average voltage of 1.84 μV^2. The HEOG mostly consists of horizontal eye movements. The VEOG also captures most of the activity originating from the vertical eye movements and the eye blinks, which generate potentials through the eyelids sliding down over the positively charged corneas [14, 44–46].

Analysing the HEOG and VEOG from Figure 4.3, it can be seen that the delta band has a higher average power than the theta band and propagates further towards the back of the scalp, causing a larger spatial distortion. The

semi-synthetic EOG can be further validated by comparing the topography of the HEOG and VEOG to other studies, for instance, that from Klados et al. [46], where the HEOG propagation shows the similar frontotemporal concentrated distribution and the VEOG similarly shows a mostly frontal concentrated distribution [46].

4.2 Electromyography

4.2.1 Results and discussion of semi-synthetic electromyography time-series

Figure 4.4 shows the results of the simulation of the temporalis and frontalis time series data, simulated for five seconds, using the temporalis and frontalis frequency from Goncharova et al. [50] as reference. Figure 4.4(a) and 4.4(b), describes the temporalis data simulation and Figure 4.4(c) and 4.4(d) the frontalis. The dashed blue line in Figure 4.4(a) represents temporalis frequency reference data. The solid blue line represents the Fourier Transform of the time-series simulated data which was simulated based on an Inverse Fourier Transform of the reference data. The reference data, namely the dashed lines on Figure 4.4(a) and 4.4(c), were first adjusted according to the chosen percentage at the relevant frequencies by the appropriate amplitude using equation 3.2, as previously discussed.

From Figure 4.4(a) and 4.4(b), one can see that the amplitude spectra of the simulated time-series data follow the amplitude spectra of the reference data, but with a high variance. The limitation of this simulation is the number of data points available in the short time-series simulation. As a result, it is difficult to accurately capture the 0 to 40 Hz range of frequency data when one second only contains two hundred data points. For the best estimation of the reference frequency data, more points are preferable, resulting in simulated EMG with longer durations for more accurate frequency representations.

Figure 4.5 shows the result of simulating the EMG for one channel. Figure 4.5(a) shows the simulated temporalis data in the first half and the simulated frontalis data in the second half of the time series simulation. Figure 4.5(b) shows the 'clean' EEG in blue, and the EMG contaminated EEG in black, therefore the summation of the EMG and the EEG in the time-series domain. Figure 4.5(c) is an expanded view between zero and more or less three seconds, showing more clearly the effect of the EMG contamination on the time-series data.

As seen in Figure 4.5(b), the EMG activities can display a magnitude much higher than the EEG signals, as observed by other studies such as Teng et al. [55], Mucarquer et al. [56] and Lugaresi et al. [31]. When referring to the 4.5(c), one can see that the original rhythm of the 'clean' EEG is completely obscured by the EMG artefacts, making the analysis and interpretation of the

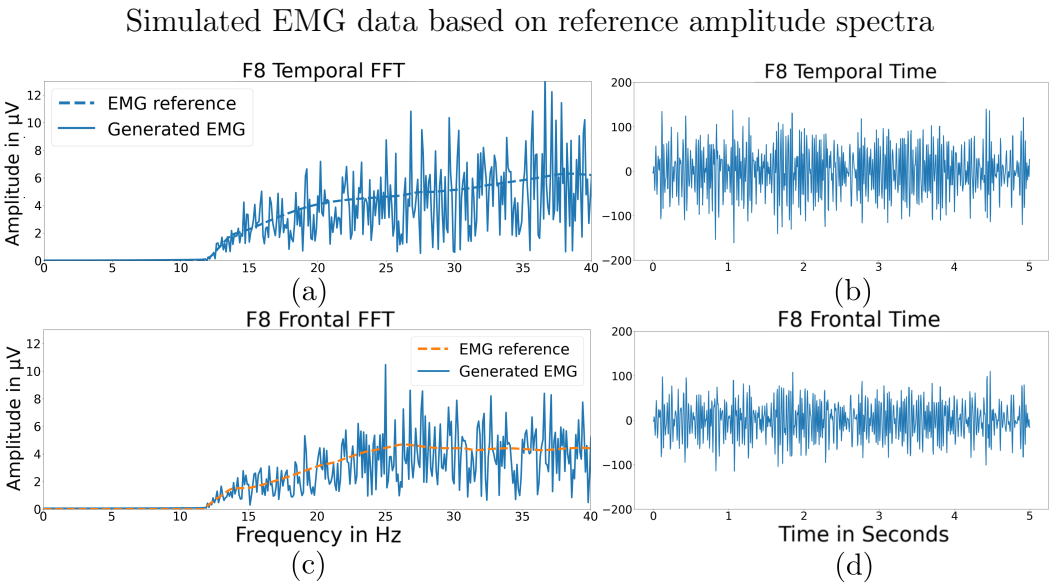


Figure 4.4: (a,c) The solid line represents the amplitude spectra of simulated time-series temporalis and frontalis contamination and the dashed line represents the amplitude spectra from the reference temporalis and frontalis data, adjusted according to the chosen percentage. (b,d) The simulated time-series data of the temporalis and frontalis contamination [50].

EEG signals difficult, as observed by researchers such as Teng et al. [55] and Lugaresi et al. [31]. The EMG follows a spontaneous bursting behaviour of Gaussian noise, also observed of EMG by Mucarquer et al. [56], Liu et al. [57] and Chavez et al. [58]. The distinguishing characteristics that define the time-series of the observed simulated EMG signals remained constant for all the channels and participants, with only the amplitude, frequency distribution and length changing within the valid ranges. This justifies the analysis and validation of the time-series data using only one of the signals.

4.2.2 Results of varying electromyography and EMG-EEG datasets

Figure 4.6 represents the SNR of all 50 participants, with Figure 4.6(a) and 4.6(b) representing the data of all the channels and Figure 4.6(c) and 4.6(d) the channels that were least affected (O2) and most affected (F8). Variations in intensity were carried out between a minimum (10%) and a maximum (100%) while keeping a constant time of a minimum of one second and a maximum of 10 seconds as shown in Figure 4.6(a) and 4.6(c). Figure 4.6(b) and 4.6(d) are the results of varying the time between a minimum and maximum of 1 to 10 seconds while keeping the intensity constant at a minimum of 10% and a maximum of 100%.

To validate and determine the ranges used to simulate the semi-synthetic

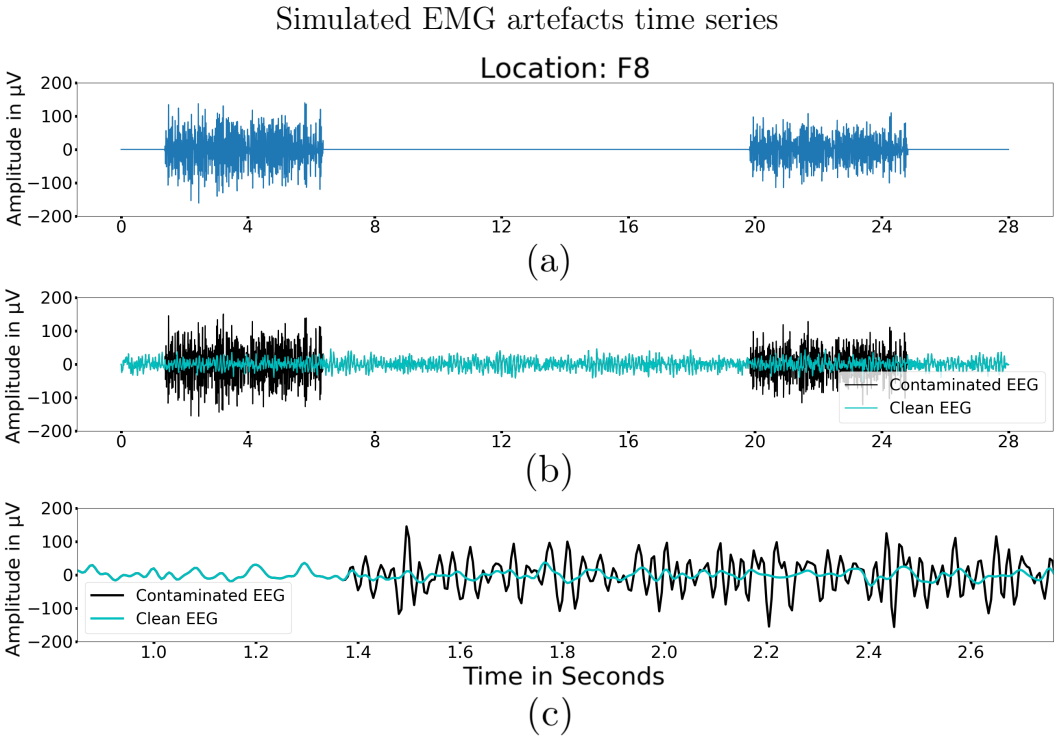


Figure 4.5: (a) Pure EMG data at location F8 after combining temporalis and frontalis data (b) The contaminated and 'clean' EEG. (c) An expanded view of Figure 4.5(b) [50].

EMG contamination, the SNR trends shown in Figure 4.6, were compared to the ranges used in related research. The difference between the simulation of the EMG in comparison to that of the simulated EOG is that the EOG had one independent variable, the propagation factors, and EMG has two independent variables, the contraction percentage and the duration of the EMG contamination. We know that Goncharova et al. [50] measured the original reference data for the EMG at 15% contraction from 27 participants. In contrast, the data used to calculate the function that determines the change in amplitude per percentage per frequency was based on data from 10 participants. This means that the larger the range and further the intensity is from 15%, the less accurate the estimation of the EMG simulation becomes. The researcher has also discussed in Section 4.2.1, that the longer the duration of the simulated EMG, the more points are available to represent the frequency reference data, therefore increasing the simulated accuracy. Therefore, using a larger range of time and a smaller range of percentage contraction is a priority.

An observation of interest in Figure 4.6(c) is that the minimum and maximum affected channels only differ significantly in SNR between the ranges of about 15% to 40%. Thereafter, with the intensity increase, the distribution of the minimum and maximum channels became very similar. This differs from the results of EOG, where the SNR of the most affected and least affected

channels were highly distinguishable from each other. The reason for the similarity in SNR may be due to the source of EMG signals differing from eye blinks originating closer to the surface of the skin. In contrast, EMG originates from under the surface, therefore affecting the propagation, where EMG clearly shows a wider constant propagation than EOG in this case.

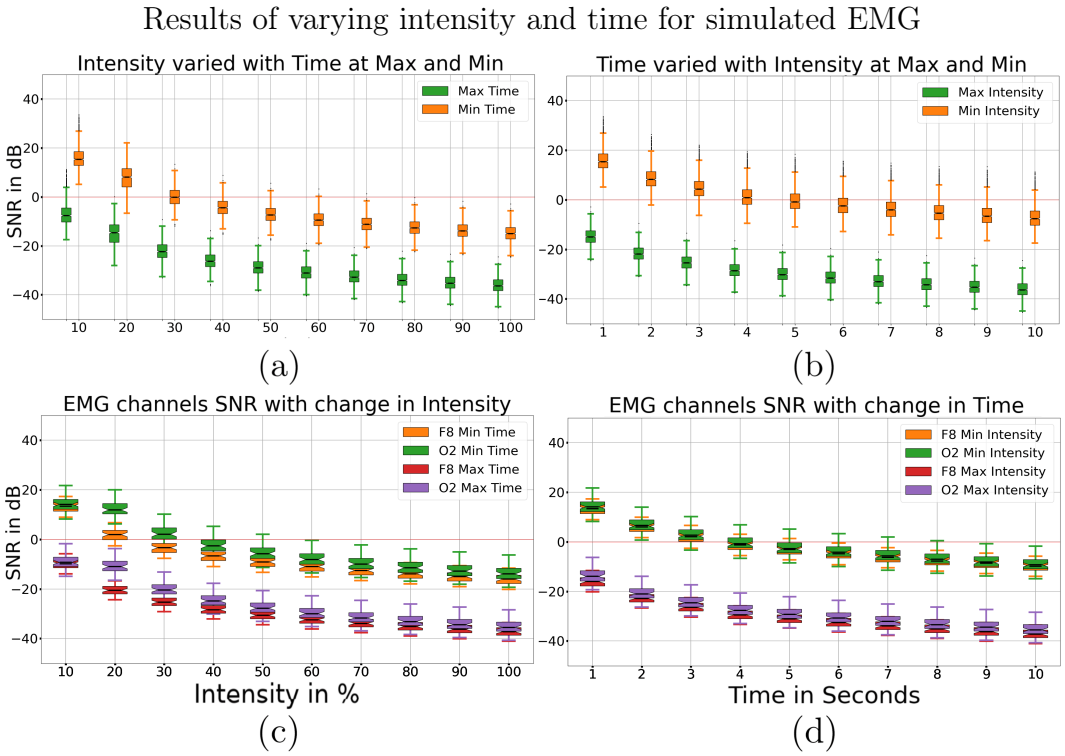


Figure 4.6: (a,c) The boxplots represent the SNR of 50 participants, with Figure 4.6(a) representing the data of all the channels and Figure 4.6(c) the channels least affected (O2) and most affected (F8), while the intensity was varied between a minimum of 10% and a maximum of 100% while keeping the time constant at a minimum of one second and a maximum of 10 seconds. (b,d) The boxplots represent the same as Figure 4.6(a) and 4.6(c), but the time was varied between the minimum and maximum seconds of one and ten seconds while keeping the intensity constant at a minimum of 10% and a maximum of 100% [50].

The SNR found was compared to the ranges used in related research to validate the semi-synthetic EMG contamination further. The SNR values found in the literature were summarised in Table H.2, Appendix H. The research conducted by Chen et al. [12], Liu et al. [57], De Clercq et al. [104], Chen et al. [133], Choudhry et al. [134] and Li et al. [135] involved the contamination of one channel numerous times while varying the SNR from minimum to maximum values. In contrast, Teng et al. [55], Clercq et al. [136] and Magno et al. [137] contaminated numerous different channels, but with the same level of SNR, also varied from minimum to maximum for each channel.

From the SNR of the semi-synthetic EMG contaminated EEG, in Figure 4.6(c), the lowest SNR value of F8 was at −40 dB when the duration and the contraction percentage were set at the maximum values. O2 was found to have the highest SNR at about 23 dB when the duration and the contraction percentage were set at the minimum values. Therefore, −40 dB and 23 dB were the lowest and highest SNR values of the simulated data. Based on the fact that simulations with a longer duration increased the accuracy of the simulation, while simulations further from 15% decreased it, the following decision was made while considering the ranges from Table H.2: the total time range of 1 to 10 seconds and a contraction percentage range of 15% to 35% was chosen, resulting in a total estimated range of −30 dB to 15 dB for the simulated EMG.

Referring to Table H.2, the lowest SNR of −30 dB used was comparable to that of Magno et al. [137], at −30 dB, and within the limits of those of Choudhry et al. [134], at −36.05 dB. The highest SNR of 15 dB used was higher than the SNR data reflected in most of the literature in Table H.2, although those studies involved contaminating one channel. In contrast, this thesis involved contaminating data over the entire scalp, resulting in less affected channels. This was as a result of the locations far from the frontalis and temporalis EMG sources increasing the overall SNR. The highest SNR used for the EMG was also well within the limits of the SNR used by that of Magno et al. [137], at 30 dB. Therefore, using the total time range and a contraction percentage range of 15% to 35% was justified, with an estimated SNR range of −30 dB to 15 dB for the simulated EMG.

4.2.3 Results of EMG and EMG-EEG topography

The columns represent the different contamination scenarios in Figure 4.7, and the rows represent the delta and theta, frequency bands. The data represented is the topography of the EMG contaminated EEG, the pure EMG, and the temporalis and frontalis activity. Unlike the topography maps of the EOG in Figure 4.3, the EEG topography maps were excluded because they were not comparable on the same log scale. However, it is important to note that the original average voltage of the delta and theta bands for the 'clean' EEG is 0.07 and 0.03 μV^2 respectively. Only the beta and gamma frequency topographies were analysed for the EMG, as only these bands were significantly affected and justified by the assumption from the literature that EMG only occurs at frequencies from 15 to 20 Hz and upwards [52–54].

Investigating the propagation of the EMG in Figure 4.7, it is of value to first analyse its sources, which are the frontalis and temporalis EMG. As their names suggest, both the frontalis and temporalis EMG are more prominent at the frontalis and temporalis areas at the beta and gamma frequency bands. With the increase in voltage and frequency, the frontalis and temporalis distribute further to the posterior and central regions, respectively. This may

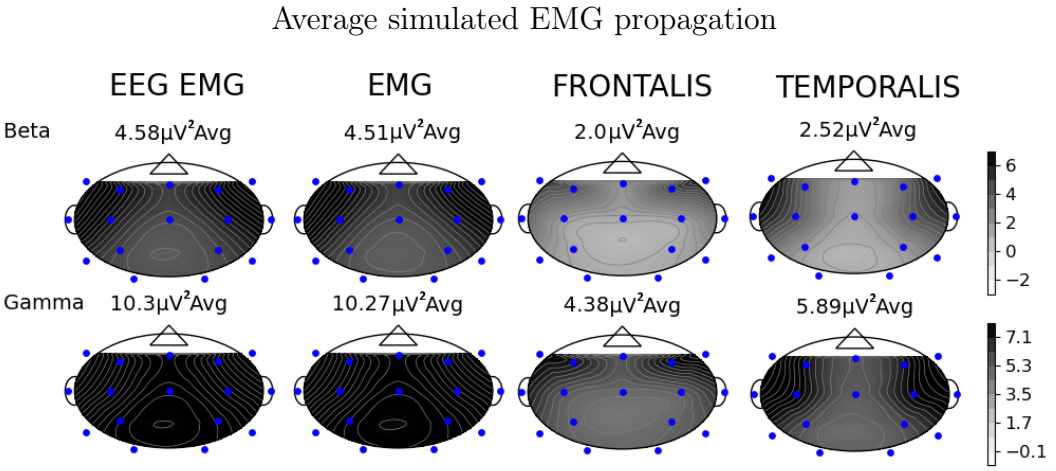


Figure 4.7: The columns represent the different contamination scenarios, and the rows represent the frequency bands. The data represented is the topography of the EMG contaminated EEG, the pure EMG, and the temporalis and frontalis activity. The scale used to compare the distribution of the topography to itself and the other topographies is a log scale, thus calculating a $10\log(\mu V^2)$, representing the band powers in dB. The average band powers in μV^2 are shown at the top of each topography map [50].

be as a result of the higher voltage increasing the distribution capability of the EMG artefact. When combining the frontalis and temporalis EMG, the distribution characteristics are also combined because the frontalis and temporalis EMG have comparable average voltages. Therefore, at the beta frequency band, the EMG has a frontotemporal concentrated distribution. However, at the gamma frequency band, possibly due to the high voltage, the EMG significantly contaminates the entire head. The EMG completely distorts the capability of investigating the topography of the 'clean' EEG, increasing the average voltage at the beta and gamma bands by 65 and 343 times, respectively. The high-frequency activity of the EMG and the propagation of the frontalis and temporalis EMG is similar to the results found from Goncharova et al. [50] and Zeng et al. [45], further validating the simulated EMG.

4.3 Electrooculography results

4.3.1 Results and discussion of semi-synthetic electrooculography time-series

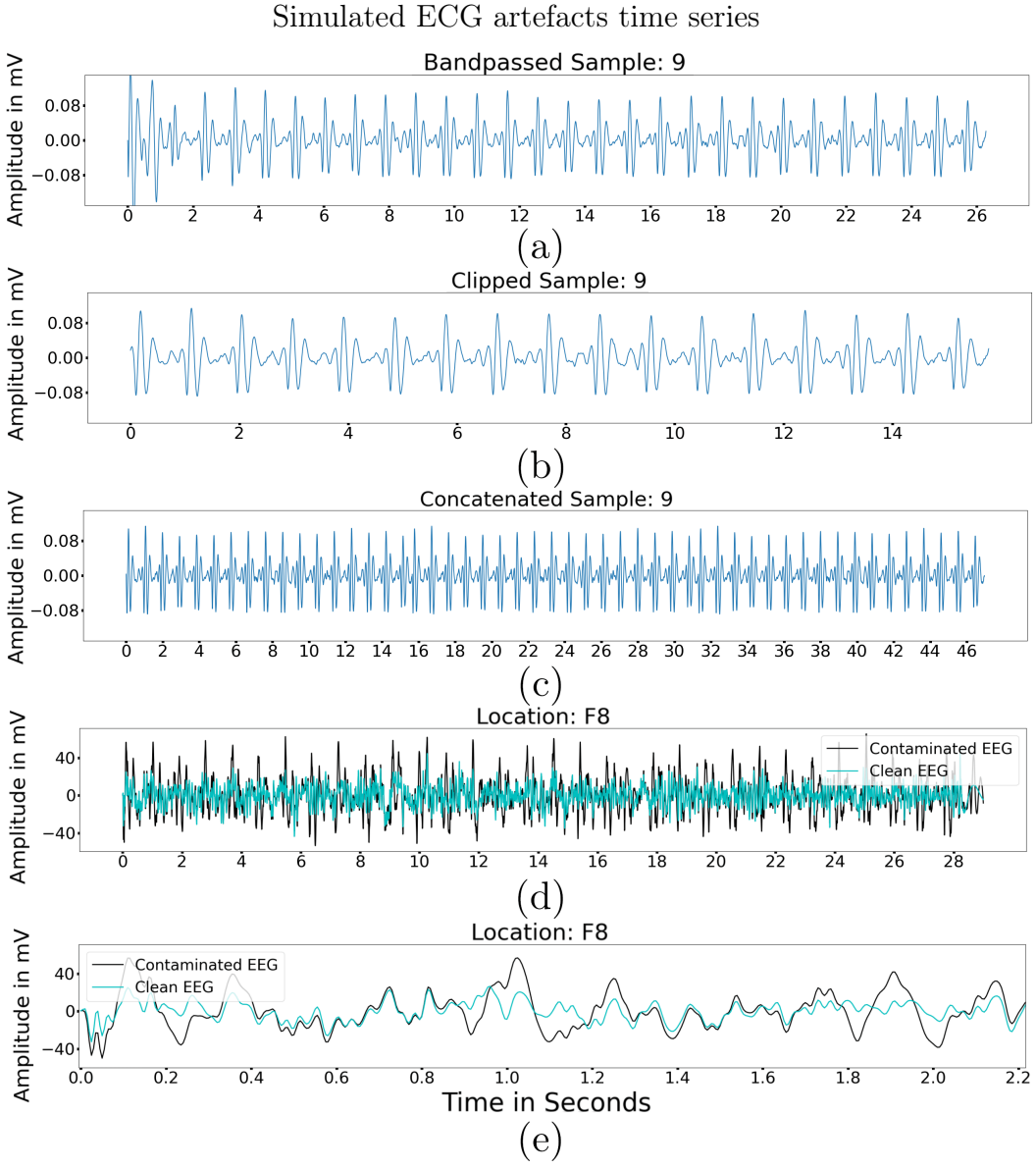


Figure 4.8: (a) The ECG after it has been bandpass filtered. (b) The ECG after possible artefacts has been clipped. (c) The concatenated ECG data to represent a longer time span. (d) The ECG contaminated EEG and EEG after the ECG has been added to a random channel and its amplitude decreased by a random amount. (e) An expanded view of Figure 4.8(d) [112].

Figure 4.8 represents the process of creating the semi-synthetic ECG contaminated EEG. Figure 4.8(a) represents the result of bandpass filtering the ninth raw ECG sample between 3 and 5 Hz. Figure 4.8(b) shows the result of the ECG after possible contaminated sections have been clipped away. Figure 4.8(c) shows the concatenated clipped ECG data to represent a longer time span. Figure 4.8(d) represents the ECG contaminated EEG after the ECG has been added to a random location, in this case, F8, and its amplitude decreased by a random amount. Figure 4.8(e) shows an expanded view of the ECG contamination of Figure 4.8(d) between 0.0 and 2.2 seconds.

Referring to Figure 4.8(b), it is clear that the ECG signal shows a simple, characteristic and periodic pattern, such as also observed by Urigüen and Garcia-Zapirain [15], Sakai and Wei [59], Jose and Collison [60] of ECG time-series data. The amplitude of the ECG in Figure 4.8(b) is relatively low. However, this amplitude greatly depends on the electrode positions and differs for certain participants, such as described by Urigüen and Garcia-Zapirain [15], Dora and Biswal [61]. Referring to Figure 4.8(d), where the blue represents the 'clean' EEG, and the black ECG contaminated EEG, it seems as if the original 'clean' EEG is completely distorted. However, Figure 4.8(e), shows just as was seen with the EOG, that due to the slow rhythms, it contaminates the EEG similar to that of a slow baseline shift. Therefore, the original rhythms of the 'clean' EEG signal are not highly distorted, as noted by [138] ECG contamination.

As further validation of the time-series pattern, Figure I.1, found in Appendix I, shows that the semi-synthetic ECG corresponds to the standard QRS ECG waveforms. As shown in Figure I.1(a), the shape of the actual bandpass filtered data is comparable to the standard QRS shape in Figure I.1(b), further validating the acquired ECG [59, 119].

4.3.2 Results of varying electrooculography and ECG-EEG datasets

Due to the contamination of the ECG only being added to one location per participant, the ranges of the independent variables were determined differently to that of the EMG and EOG simulation. The two independent variables, as discussed before, for ECG, are the percentage reduction of the original signal, relating to the variability in the electrode's position relative to a blood vessel, and the samples, relating to pulse characteristics differences between individuals. When no contamination is added to a signal, the SNR is infinite, making it impossible to calculate the average SNR of the participant if only one channel was affected. Therefore, to decide the percentage reduction and sample ranges, the same ECG contamination was added to each channel for each participant and visualised in Figure 4.9. To reiterate for clarity, in deciding the maximum and minimum SNR ranges for the ECG, all the channels

were contaminated for each participant, including the inter-participant EEG variations in the decision making process. After the ranges of the independent variables, which produced valid SNR values were chosen, only one channel per participant was contaminated by ECG artefacts for the actual semi-synthetic dataset.

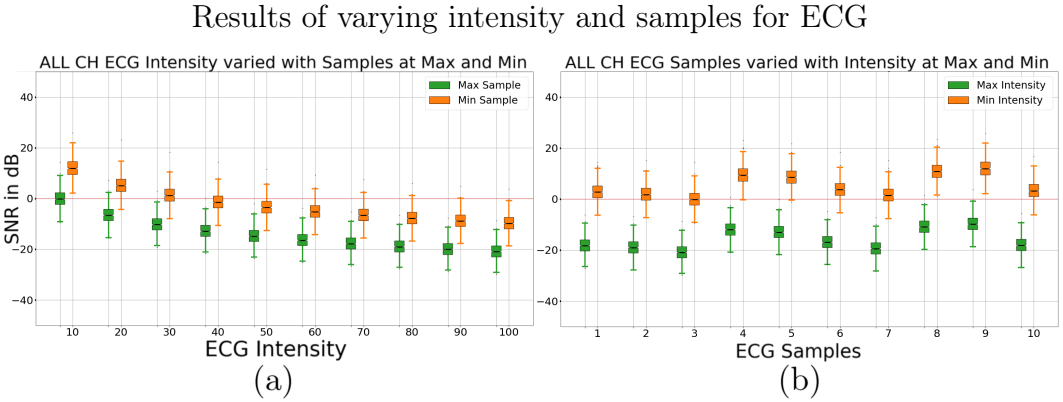


Figure 4.9: (a) The boxplots for the SNR of all 50 participants for all the channels contaminated by the same maximum and minimum sample, three and nine, while changing the percentage reduction from 10% to 100%. (b) The boxplots for the SNR of all 50 participants for all the channels while the maximum percentage reduction of 100% and a minimum of 10% was kept constant while varying from sample 1 to 10 [112].

Figure 4.9(a) shows the results of varying the percentage reduction and Figure 4.9(b) the results of varying the samples. Figure 4.9(a) shows the boxplots for the SNR of all 50 participants for all the channels contaminated by the same maximum and minimum sample, three and nine respectively while changing the percentage reduction from 10% to 100%. To create Figure 4.9(b), the maximum percentage reduction of 100% and a minimum of 10% was kept constant while varying from sample 1 to 10.

Referring to Figure 4.9(b), it can be seen that there is indeed a significant variation between some of the ECG samples relative to others, as for samples three and nine. Some samples do not significantly differ in their SNR values, such as samples one and two. With that stated, combining the different samples, adding them to different locations and changing the reduction percentage, produced a good amount of variation in the SNR of the ECG artefacts.

To further validate the semi-synthetic ECG contamination, the SNR was compared to the ranges used in related research. The SNR values found in the literature was summarised in Table H.3, Appendix H. Sakai and Wei [59], Dora and Biswal [61, 139], Cho et al. [140], Suja Priyadharsini and Edward Rajan [141] and Navarro et al. [142] focused on contaminating a single EEG channel while varying the SNR between a minimum and maximum value. Navarro

et al. [143] and Hou et al. [144] additionally only used one set value for the SNR for one channel.

As can be deduced from the literature and Table H.3, ECG artefacts are not as highly distorting as EMG and EOG and usually have a much higher SNR value. To maximize the variance of inter-individual ECG differences, the researcher in this thesis decided to use all 10 of the raw ECG samples. Therefore, the only independent variable left was that of the percentage reduction. To create viable ECG contamination comparable to that of other literature, the percentage reduction was chosen to range between 10% and 19%, creating a semi-synthetic ECG contaminated EEG dataset with an estimated contamination range between 15 dB and −7 dB SNR. The maximum estimated SNR of 15 dB, is comparable to the 15 dB used by Dora and Biswal [61, 139], Navarro et al. [143] and the minimum estimated SNR of −7 dB is only slightly higher than the −6 dB of Cho et al. [140] and −5 dB of Navarro et al. [143], Hou et al. [144]. Therefore, with the characteristics validated by other literature, the intensity of the contamination is further validated by occurring within and close to the ranges of established literature.

4.4 Combination results

4.4.1 Results and discussion of semi-synthetic combined time-series

Figure 4.10(a) shows the 'clean' EEG results simultaneously at location C3 contaminated by ECG, EMG, and EOG artefacts. In Figure 4.10(a), the original 'clean' EEG is blue and the contaminated EEG is black. Figure 4.10(b) shows an expanded view of between 0.6 and 6 seconds.

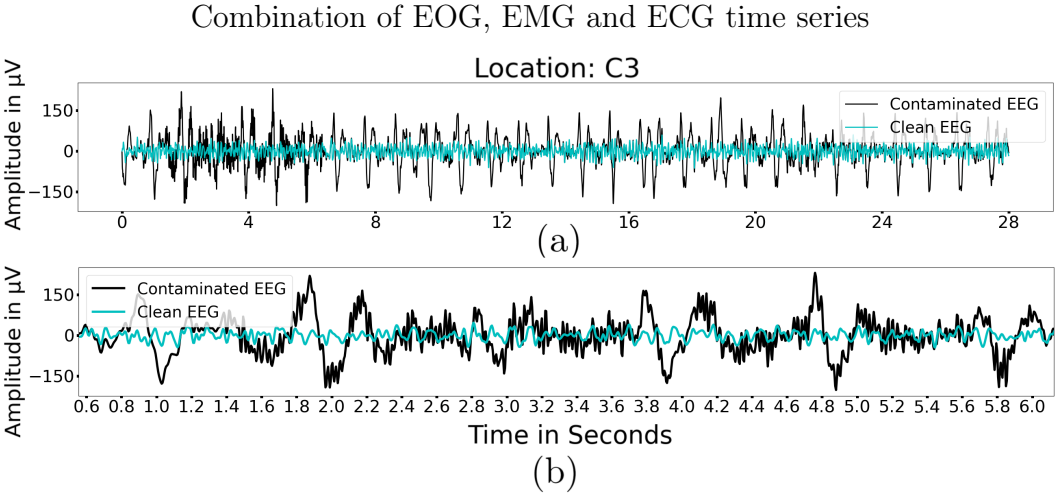


Figure 4.10: (a) Results of combining the ECG, EMG and EOG to the EEG (b) Expanded view of Figure 4.10(a) [50, 111, 112].

The need for robust cleaning methods, and therefore datasets that can test robustness, has been previously concluded from the literature review [9, 15, 68]. In Figure 4.10(b), it is quite clear that the distortion in baseline shift and rhythms of the original signal is high. Also, due to the increase in the number of artefacts, the limitations of the BSS methods capturing the maximum amount of components equal to channels used may decrease the effectiveness of the methods. With the use of the auto threshold method to clean high-intensity combinations of artefacts, not much EEG data will be left if all the estimated artefact ranges are removed.

4.5 Discussion of semi-synthetic contaminated data

Figure 4.11 shows the final SNR results of the simulated data for each contamination for all 50 participants. On the x-axis, the relevant contamination is indicated, and on the y-axis, the SNR value. The boxplots only represent the results of the contaminated channels; therefore, for ECG, the boxplots included only the SNR of the one contaminated channel per participant in the data and for the EMG, EOG and combination of all the artefacts, the boxplots included all the channels in the SNR boxplots.

In Figure 4.11, we can clearly see whether we have succeeded in creating contaminations comparable to each other and those found in the literature. As seen with the literature and Tables H.1,H.2,H.3, as depicted in Figure 4.11, the order of lowest contamination SNR ranges start from EMG to EOG and finally ECG, with the combination of the artefacts logically resulting in the lowest overall SNR.

The EMG contamination ranges between the lowest and highest SNR values of $-30\,\text{dB}$ and $15\,\text{dB}$. This range is validated with its lowest SNR of $-30\,\text{dB}$ being equal to that used by Magno et al. [137] of $-30\,\text{dB}$ and within limits and comparable to the lowest SNR used by Choudhry et al. [134] of $-36\,\text{dB}$. The highest SNR of $15\,\text{dB}$ for the EMG falls between the $30\,\text{dB}$ used by Magno et al. [137] and the lower SNR values used by the other studies. Therefore, the semi-synthetic EMG contaminated data is maximally distributed in terms of its SNR and comparable and within the limits of relevant literature.

The EOG SNR has a lowest and highest SNR of $-18\,\text{dB}$ and $15\,\text{dB}$. The lowest of $-18\,\text{dB}$ is comparable to the $-20\,\text{dB}$ used by Puthusserypady and Ratnarajah [130], and well within limits of the $-40\,\text{dB}$used by Paulson and Alfahad [132]. The highest SNR of $15\,\text{dB}$ is between the highest of $23\,\text{dB}$by Naga et al. [129] and the $10\,\text{dB}$ of Paulson and Alfahad [132] and also well within the limits of the $60\,\text{dB}$ used by Merino et al. [128]. Therefore the SNR of the simulated EOG is comparable to that of other studies and highly distributed while remaining within the limits of the SNR used by other studies.

The ECG contamination ranges between a lowest and highest of −7 dB and 15 dB. The lowest SNR of the ECG is comparable and just a bit lower than the −5 dB used by Navarro et al. [142], Hou et al. [144] and the −6 dB used by Cho et al. [140]. The highest SNR of 15 dB of the simulated ECG is the same as the highest values used by Dora and Biswal [139], Navarro et al. [143] and Dora and Biswal [61]. Therefore, the simulated ECG has an SNR range which is highly comparable to the highest and lowest ranges used by other literature.

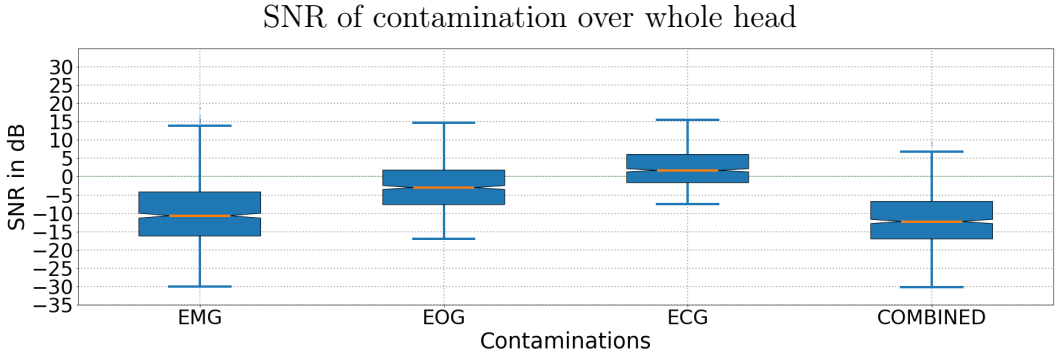


Figure 4.11: The figure shows the final SNR results of the simulated data for each contamination for all 50 participants. On the x-axis, the relevant contamination is indicated, and on the y-axis, the amount of SNR. The boxplots only represent the results of the contaminated channels; therefore, for ECG, the figure includes only the SNR of the one channel per participant in the data. For the EMG, EOG and combined, the boxplots include all the channels in the SNR boxplots [50, 111, 112].

With the ranges of the simulated data validated by comparing the results of the SNR of the simulated artefact contaminations of Figure 4.11 to the literature from the tables H.1,H.2,H.3, we can now start to delve deeper into the validation of the distributions of the SNR. Figure 4.12 consists of two subfigures; Figure 4.12(a) is the SNR distribution of the different regions for the different types of contamination, with Figure 4.12(b) being the legend for Figure 4.12(a), depicting the positions of the Internationally recognised 10-20 standard, with the colours of different locations marked, referring to the locations used to calculate the SNR on the left sub-figure.

Regarding the EMG SNR distribution in Figure 4.12, one can see that the temporal regions have the largest SNR values, followed by the frontal/temporal region, and then the central and finally the occipital/parietal region. This distribution makes sense, primarily because of how close the distributions of the four boxplots are compared to those of the EOG. Furthermore, the boxplots show similar trends to the topography results of the EMG from Figure 4.7. Referring again to Figure 4.7, the researcher in this thesis found that the average temporalis contraction voltages were higher than the average of the frontalis contractions. The boxplots follow the trend that the contaminations are normally distributed, with the temporalis SNR being lower than the frontalis

Comparison of SNR at the different regions for each simulated artefact

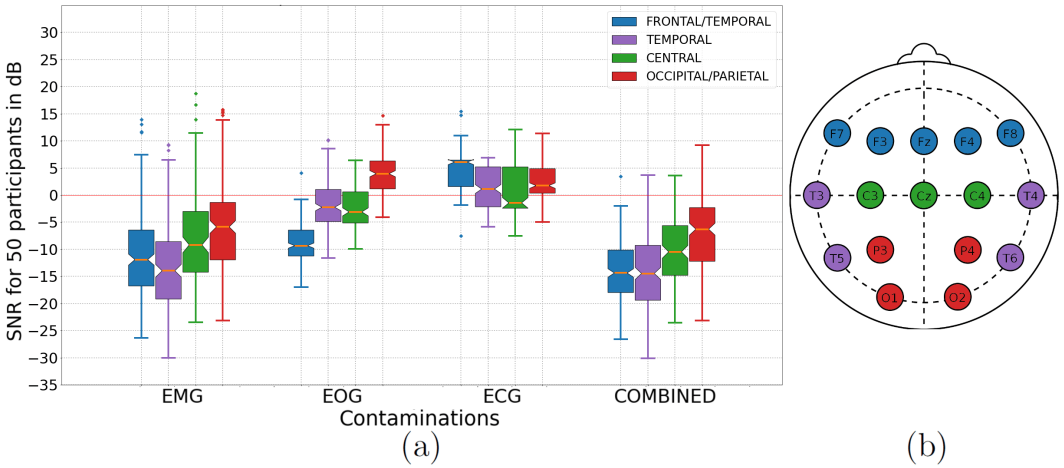


Figure 4.12: (a) The SNR distribution of the different regions for the different types of contamination. (b) The legend depicting the potions of the Internationally recognised 10-20 standard, with the colours of different locations marked, referring to the locations used to calculate the SNR on the left sub-figure [50, 111, 112].

SNR. It makes sense that the central SNR is higher than the occipital/parietal region because it is closer to the frontal/temporal and temporal region, where the EMG originated. Therefore the SNR distribution of the EMG makes sense and is further validated by the topographies from Figure 4.7.

Analysing the EOG SNR distribution in Figure 4.12, it becomes clear that the results are comparable to that of the topography of Figure 4.3, where the SNR of each region differs significantly relative to each other, in comparison to the results of the EMG. The frontal/temporal SNR is also much lower as this is the region most affected by the VEOG, which has the largest voltage, as seen in Figure 4.3. The second most affected region is the temporal area, mostly affected by the HEOG activity. However, the SNR distribution of the central region is close to the SNR of the temporal region, possibly due to the central area being affected by a combination of mostly VEOG activity and some of the HEOG activity, as seen in Figure 4.3. Therefore, the SNR distribution of the EOG follows the trends of the topographies from Figure 4.7 and is further validated.

The ECG data in Figure 4.12 shows similar SNR distributions for each region compared to each other. Ideally, the SNR for each region would have been normally distributed. The randomly chosen ECG samples, reduction percentages, and channels contributed to a symmetric probability distribution. The skewed distribution is due to the limitation of only one channel being contaminated per participant, reducing the total amount of contaminated data.

Finally, analysing the SNR of the combined data in Figure 4.12, one can see that it has the lowest overall SNR. Having the same trend as the EMG, the temporal region of the combined data has the lowest SNR, increasing in the

order of the frontal/temporal, the central, and finally the occipital/parietal regions. Due to the SNR trends of the EOG and EMG being close to each other in order of regions, the ECG trends being the same for all the regions, and the EMG SNR being the lowest and thus having the largest effect on the total SNR, it makes sense that the combination of the three forms of artefacts follows the same trend as the EMG. Interestingly, the lowest values of the SNR for the combination of the artefacts did not change from those observed in the EMG data. The only thing that changed was the highest SNR being significantly reduced. This may be due to the lowest SNR contaminated regions being saturated by contamination and the less contaminated regions being open for contamination by the combinations.

4.6 Cleaning methods results

4.6.1 Threshold method results

Figure 4.13 consists of three sub-figures, showing the process employed by the auto threshold method as described in Section 3.8. EMG and EOG contaminate the 'clean' EEG in Figure 4.13(a). In Figure 4.13(a), the standard deviation is largely due to the EMG contamination raising the average amplitude. Therefore, the standard deviations dependent threshold was set to just above 100 μV, creating a range depending on the amount of threshold intersects at those locations, as shown in Figure 4.13(a). The data within those ranges were removed, resulting in the complete removal of EMG contamination, as seen in Figure 4.13(b). The second threshold was lower and around 60 μV due to the lower overall amplitude. All the data in these ranges determined by the threshold was again removed to create the data in Figure 4.13(c). Figure 4.13(c) shows the cleaned EEG with much less EMG and EOG contamination.

As contamination increases, more data ranges are completely removed, resulting in a significant reduction in the data duration and the loss of brain-related data. With the BSS methods, some real EEG is also removed, but not as completely as with the auto threshold method. In Figure 4.13, the data duration was decreased from about 28 seconds to 14 seconds, resulting in a significant reduction in data points. The question is which loss in data is worse; the loss in complete ranges from the auto threshold method or the continuous real EEG caught and removed in the estimated artefact components from the BSS method.

4.6.2 Results of semi-automatic BSS and auto threshold cleaning

Figure 4.14 shows the results of the average SNR at each location for each artefact. Each sub-figure represents different contamination, with the lines

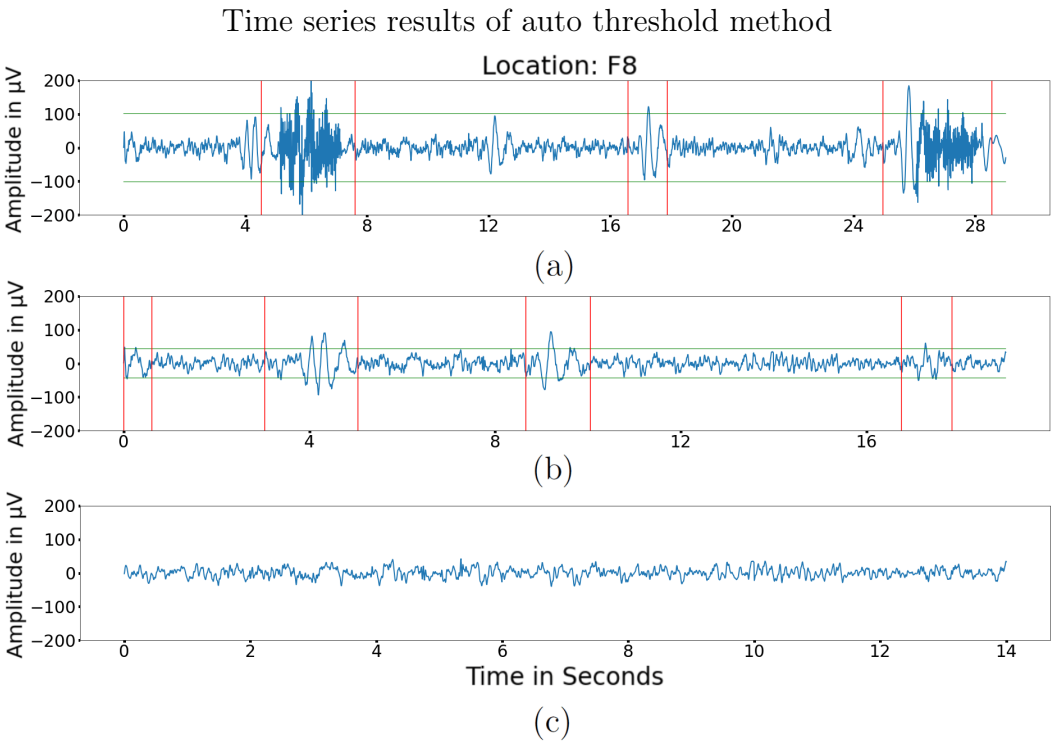


Figure 4.13: This figure shows how the EEG is cleaned in multiple stages using a threshold based on the standard deviation of the data. (a) EMG and EOG contamination time-series. (b) Result from the previous stage where the identified ranges were removed, and only EOG contamination remains. (c) The result after the EOG ranges were removed, and only 'clean' EEG remains [50, 111, 112].

representing the SNR of the cleaned data for each method. The black line represents the average SNR of the original contaminated data.

Figure 4.15 shows the distribution of the difference in SNR between the cleaned data and the original contaminated data for the entire head for all 50 participants. This means that each boxplot shows the SNR difference for 16 locations for each of the 50 participants, therefore representing 800 data points. The x-axis represents the different artefacts used to contaminate the data, and the colours of the boxplots correspond with the methods used, namely the auto threshold method, SOBI, Extended Infomax and CCA.

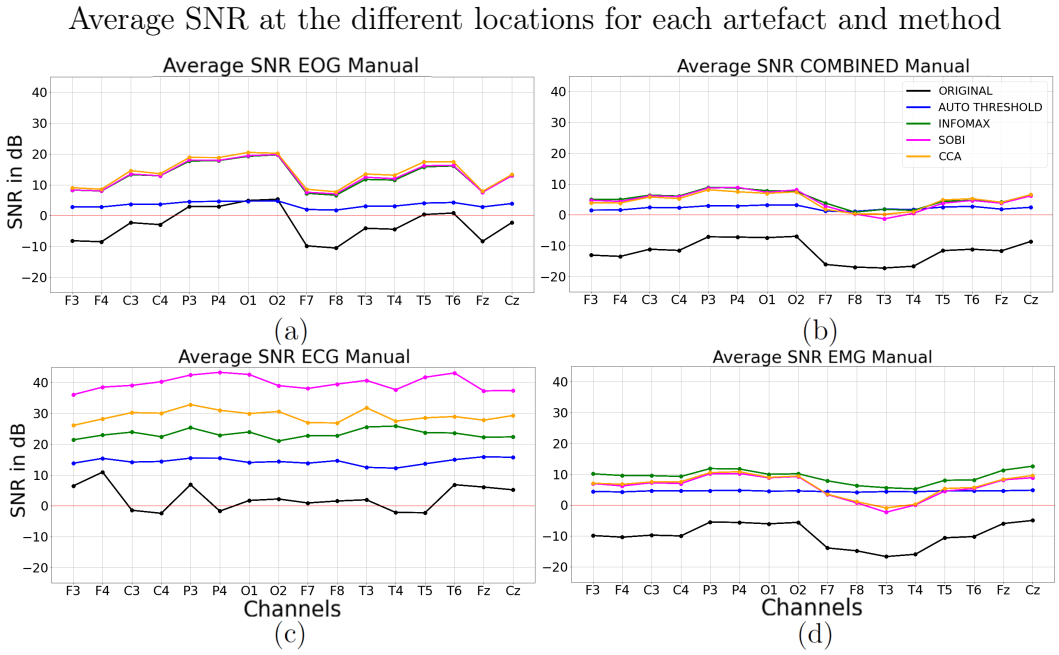


Figure 4.14: The average SNR at the different locations for each artefact type is represented. Each sub-figure shows the new SNR of the cleaned data, cleaned by four different methods, with the black line representing the average SNR of the original contaminated data. (a) EOG (b) COMBINED (c) ECG (d) EMG.

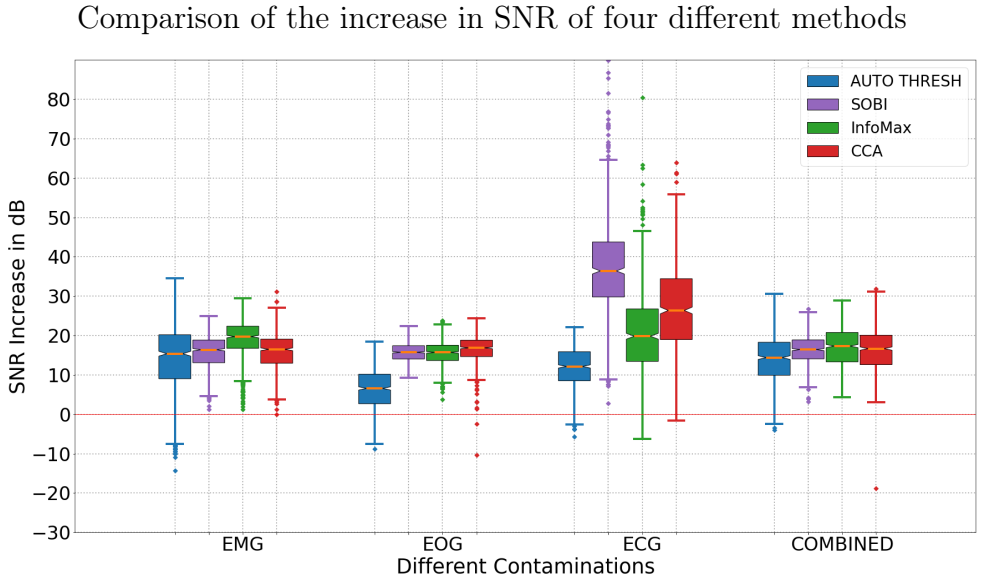


Figure 4.15: The boxplots show the distribution of the difference in SNR between the cleaned data and the contaminated data for the whole head for all 50 participants. The results of the four methods for each type of contamination are shown.

4.6.2.1 Results of removing electromyography artefacts

Both Figures 4.14(d) and 4.15 will first be used to discuss the results of the cleaning methods for EMG contaminated data. Regarding what has been found in literature, CCA is often proposed as a more reliable method for the removal of muscle artefacts in the scalp in comparison to ICA methods, due to the exploitation of the fact that the autocorrelation of muscle activity is weaker than that of brain activity [12, 15, 88]. CCA is better at cleaning largely negative SNR because CCA tends to exaggerate the contribution of the EMG source to the scalp EEG activity. This exaggeration is a possible reason for the impairment of the performance of the CCA method due to the EMG contamination occurring at SNR values as high as 15 dB [76]. ICA methods such as Extended Infomax and SOBI have also been extensively explored with EMG artefact removal. Results from literature show that SOBI and Extended Infomax are as good as CCA at removing EMG from EEG [15, 76]. With some studies claiming that CCA performs better with EMG removal than the ICA methods and others claiming that their performance is the same, it is evident that a dataset with a large variation is necessary to find a clear answer to which method performs better overall in increasing the SNR of EMG contaminated data [12, 15, 57, 76, 88].

The SNR of the original simulated EMG contaminated data ranges between -30 dB and 15 dB, therefore providing a large variation in contamination intensity. Figure 4.15 provides clear overall results for each method. The auto threshold method ranges from the best to the worst performance, increasing the SNR by 35 dB in some cases but decreasing the SNR by 15 dB in other cases, reducing the overall impression of its performance. Therefore, still referring to Figure 4.15, Extended Infomax has performed the best overall with the removal of the EMG artefacts. The t-test was performed on the results of SOBI and CCA in removing the EMG artefacts, with a very close but still significant $P = 0.06 > 0.05$. Thus, we cannot conclude that a significant difference between the performances of the two methods exists. Thus, SOBI and CCA were both second in performance, followed by the auto threshold method. Figure 4.14(d) gives us a better idea of the performance of the methods based on how low the SNR of the EMG is, due to it being varied for each location. Again, we can see how closely the performances for SOBI and CCA were. CCA only performed slightly better at location T3 than SOBI, due to the EMG being at its lowest SNR and relating to the conclusion that SOBI performed better with very low SNR values [15, 76]. It becomes quite clear that in all the locations, the average performance of the Extended Infomax method was the best, with significant differences compared to SOBI and CCA, only when the EMG contamination was at its average lowest values. Finally, it is of interest to note that the auto threshold method has almost the same performance for each channel, whether the EMG contamination was higher or lower. This makes sense due to the simplicity of the auto threshold method, causing it to

remove spikes when it is not EMG, but completely removing EMG when it is pronounced and has a very low SNR. It even outperforms SOBI and CCA at F7, F8, T3, and T4, where the EMG SNR was at its lowest, meaning that when the SNR is so high, it could be better simply to cut the data out, rather than trying to statistically separate it.

4.6.2.2 Results of removing electrooculography artefacts

Both Figures 4.14(a) and 4.15 will also be used to discuss the results of the cleaning methods for EOG contaminated data. From the literature review, it was found that ICA methods such as SOBI and Extended Infomax have become the default choice for removing EOG artefacts from EEG data [15]. Extended Infomax has been the most thoroughly justified for removing EOG artefacts in other literature [15, 83]. It has also been stated in the literature that SOBI stands out as the best artefact removal method for EOG [8, 84]. However, it has also been found that CCA clearly identifies EOG artefacts for removal alongside these popular ICA methods [89]. Clearly, literature produces differences in conclusions on whether SOBI, Extended Infomax and CCA perform the best in removing EOG artefacts. These differences could be due to the subjectivity of these methods, namely the manual identification of the artefact components, with the expertise between researchers not being the same [89].

Referring first to Figure 4.15, one can see that the CCA method performed the best, as it produced the highest SNR increases. At the same time, CCA also has the lowest SNR values, ranging to $-10\,\mathrm{dB}$, meaning that in some cases, the CCA method actually worsened the SNR of the already contaminated data. Still referring to Figure 4.15, the best performance, measured by an increase in SNR and having comparable robustness to Extended Infomax and SOBI, is CCA. A t-test comparing the SNR increased distribution of SOBI and Extended Infomax produced a $P = 0.17 > 0.05$. From this, we cannot conclude that their performance is significantly different. Therefore SOBI and Extended Infomax performed second best with the removal of the EOG artefacts. The auto threshold method performed reasonably well but further decreased the SNR of the data in many cases; this may be due to EOG having been present continuously throughout the dataset, meaning that too much data would have been removed by the threshold method to account for all the EOG artefacts. Referring to Figure 4.14(a), it can again be seen how close the performances of the two ICA methods are, with CCA performing slightly better at each location. Even though Figure 4.14(a) shows that CCA has the average best performance, the lack of robustness that it showed in Figure 4.15 is still somewhat concerning. The results of the auto threshold method in Figure 4.14(a) showed that the SNR of the average results of the cleaned data remained reasonably consistent across all the channels, independently of how low or high the SNR of each channel was. These results show that the auto

threshold method is effective with really low SNR but limited when the SNR is high because the algorithm removes unnecessary data, as observed at channels O1 and O2 in Figure 4.14(a).

4.6.2.3 Results of removing electrooculography artefacts

ECG contamination is periodic and usually has a high SNR. The simulated ECG ranges between −17 dB and 15 dB. Thus, the success of the method also depends on identifying and extracting subtle patterns that sometimes cannot be visually identified in the time series data. The researcher in this thesis noted that with the semi-automatic cleaning of the ECG, the CCA method always identified the artefact components in the same order, unlike the ICA methods, based on stochastic learning methods [12]. ICA methods are typically the preferred methods when it comes to removing ECG artefacts [15, 81], with SOBI reported to generally perform better than other methods in removing ECG artefacts [15, 86].

To get an overall idea of which method performed the best, we again refer to Figure 4.15. We can see that SOBI outperforms the other methods, followed by CCA, Extended Infomax and the auto threshold method. SOBI possibly outperforms the other methods because it is based on second-order statistics, causing a more successful identification of the periodic ECG patterns. The conclusions derived from Figure 4.15, are also clearly repeated in Figure 4.14(c), with SOBI outperforming the other methods, followed by CCA, Extended Infomax and finally, the auto threshold method. Figure 4.14(c) provides valuable information on the effect that the cleaning of the one ECG channel has on the other channels, where the other channels had an original SNR of infinity. When one contaminated channel is cleaned, actual EEG data is also removed from the other channels. Therefore, the higher the data, the more precise the method identifies only the ECG component. An ideal method would leave all the other channels at an infinite SNR. A low SNR in 4.18(c) means that by only removing a contaminated component from one channel, the method reduced the SNR of the whole dataset from infinity to a very low value.

4.6.2.4 Results of removing the combination of artefacts

A major disadvantage of the ICA methods is that they can only identify as many components as channels used; therefore, the likelihood of effectively identifying all the components decreases with an increase of independent artefacts [15, 32, 74]. As a result of this, it can be expected that the superiority of the ICA methods in performance may not be consistent with the combination of contamination. Most research involving cleaning methods usually focuses on only one type of artefact. As observed in the literature, Extended Infomax methods are primarily applied to EOG artefact removal [15, 83], SOBI meth-

ods to removing EOG and ECG [15, 86] and CCA methods to removing EMG [12, 15, 88]. Testing the methods on the combination of these artefacts is relatively unexplored, and therefore the results are considered more valuable. The results will depend heavily on the number of channels affecting the ability of each method to identify the components due to an increase in contaminated components. Finally, the results will also depend on which method has the best balance between effectiveness with each artefact. This balance is slightly skewed due to the combination of the artefacts not being equally contaminated, but contaminated from an overall lower SNR to an overall higher SNR from EMG, EOG and ECG.

As discussed above, the combined datasets consist of lower SNR EMG, then slightly higher SNR EOG and finally the highest SNR ECG. Referring to Figure 4.15, the three BSS methods show very similar results. Evaluating the methods based on the robustness and the overall increase in SNR of the contaminated data, resulted in Extended Infomax performing best for the combination of artefacts. CCA and SOBI are tied at second place, with a t-test showing that it cannot be concluded that there is a significant difference between them, having a $P = 0.57 > 0.05$. It is also noteworthy that the Extended Infomax results are similar to that of CCA and SOBI, with corresponding P values of 0.01 and 0.02 for the t-test when comparing Extended Infomax to these methods. It is also important to note that the auto threshold method is comparable to the other three methods, which may be due to its good performance when removing EMG contamination.

4.6.3 Time results of semi-automatic blind source separation and auto threshold cleaning

The time each method takes to identify the components is an important characteristic in online applications, such as BCI applications and commercial EEG products. Apart from the manual identification of the components, in the algorithm, component identification is the most time-consuming part of the cleaning process.

SOBI was the slowest to identify each component, as shown by Figure 4.16. More specifically, SOBI was the slowest with ECG components and the fastest with all the artefacts combined, thus an un-equivalent relationship was established between the intensity, the number of artefact components and the time necessary to identify the components. Extended Infomax was the second slowest method and showed the same artefact number and intensity to time relationship as SOBI. The expanded view of CCA and auto threshold distribution of time can be found in Figure J.1. Finally, CCA, known for its time efficiency and viability for online use [12, 15, 76], ranges between 0.5 and 2 seconds, as seen in Figure J.1(b), making it the fastest BSS method tested. The CCA algorithm also followed the same relationship between the amount

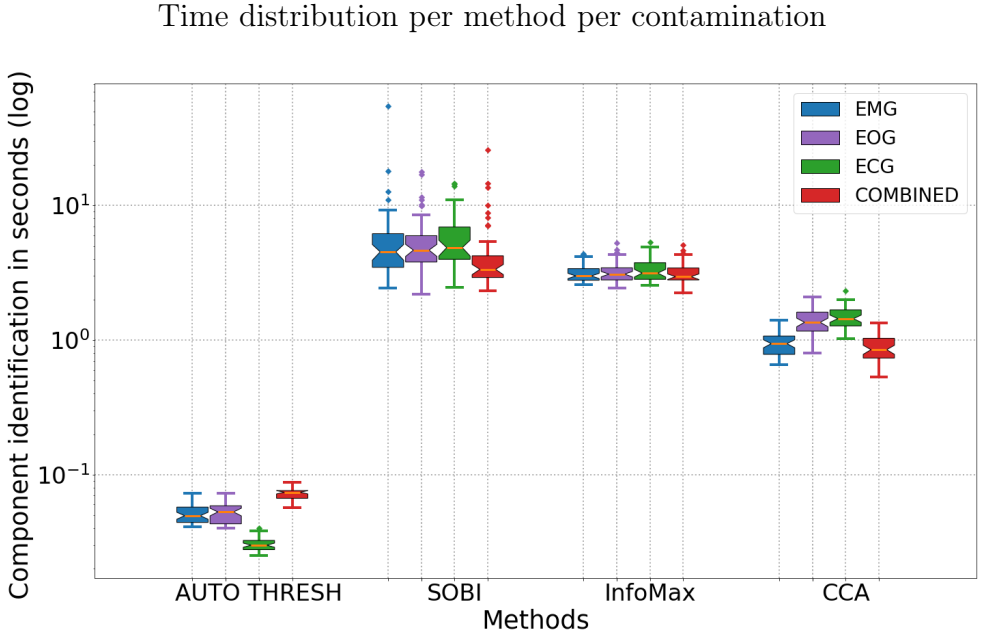


Figure 4.16: The boxplots of the average time in seconds to identify all the components for one participant for all 50 participants. The boxplots are shown for each method and each contamination.

and intensity of artefacts and the time to identify them as Extended Infomax and SOBI. The most significant result is the time taken to identify the possible artefacts of the auto threshold method. As seen from Figure 4.16 and more in-depth with Figure J.1(a), it is clear that the auto threshold method performed about 10, 20 and 100 times faster than CCA, Extended Infomax and SOBI, respectively. Additionally, the auto threshold method achieved comparable results to the BSS methods in cleaning the EMG and combining all three artefacts. Therefore, considering comparable performance results of the auto threshold method and its vast improvement in time, it may be considered the most viable cleaning solution in some contexts.

Thus, with very low SNR EMG contamination, the auto threshold method may be applicable in real-time. Furthermore, it is the only method discussed that can currently be applied in real-time applications as a result of its time efficiency. Due to the thresholds being based on the standard deviation of the data, it requires a few seconds of clean data to start working effectively. With high SNR data, the method is also ineffective. Further developments of this method could be highly effective if there is an additional restriction that the automatic threshold only applies to sudden deviations of the standard deviation, such as bursting artefacts related to EMG.

As another proposed approach, the auto threshold can be used for preprocessing the BSS methods to detect and remove high-intensity EMG artefacts first, since it is superior to BSS at removing EMG contamination with low SNR. This solution would still require that time ranges be removed across all

channels whenever they are removed so that the channels remain time-locked and the BSS remains applicable.

Therefore, it is believed that the auto threshold methodology can be successfully applied to some BCI applications if, as further development is made, it is only applied when there are sudden large changes in the standard deviation. Auto threshold methods would be applicable to BCI applications where participants display excessive muscle movements, such as sport or gaming-related BCI applications or sport-related EEG research requiring immediate analysis.

4.6.4 Results of the fully automated cleaning

Figure 4.17 consists of four sub-figures. The sub-figures compare the results of the BSS methods when the researcher in this thesis chose the components manually and used them automatically. The last sub-figure in Figure 4.17 compared CCA with the auto threshold method. Each sub-figure shows, in the form of boxplots, the increase in SNR from the original contaminated data, for all the participants and all the channels, for each contamination for the automated and semi-automatic methods.

Comparison between automated and semi-automatic identification of components

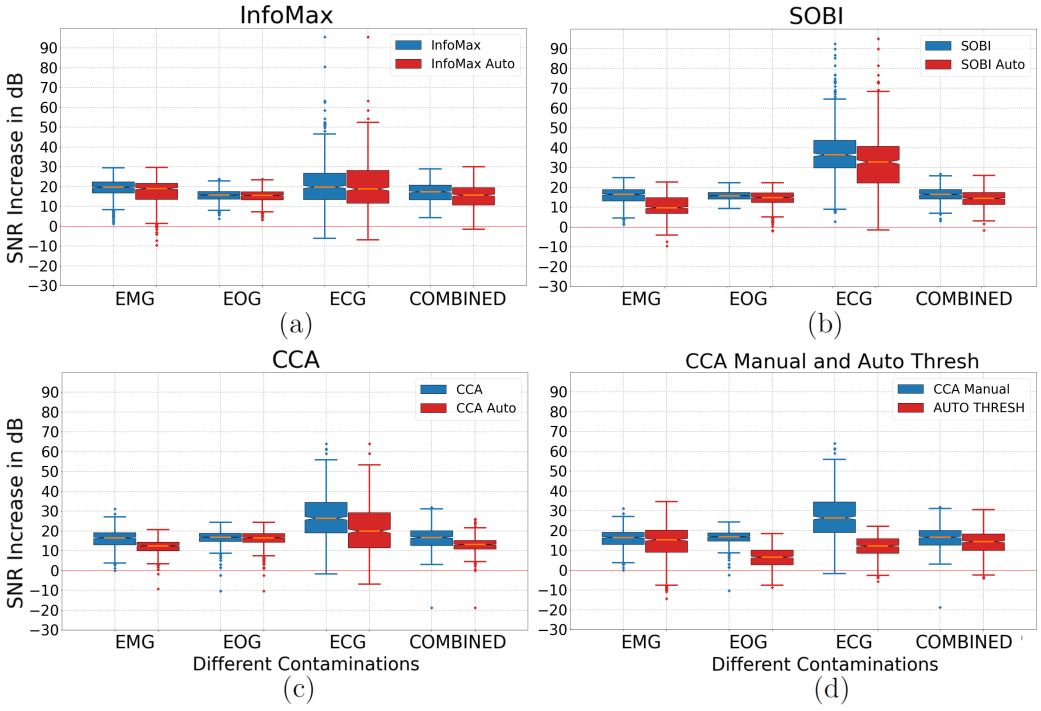


Figure 4.17: The boxplots show the distribution of the difference in SNR between the cleaned data and the contaminated data for the whole head for all 50 participants. The results of four methods for each type of contamination are shown.

Average SNR of the automated and semi-automatic components methods

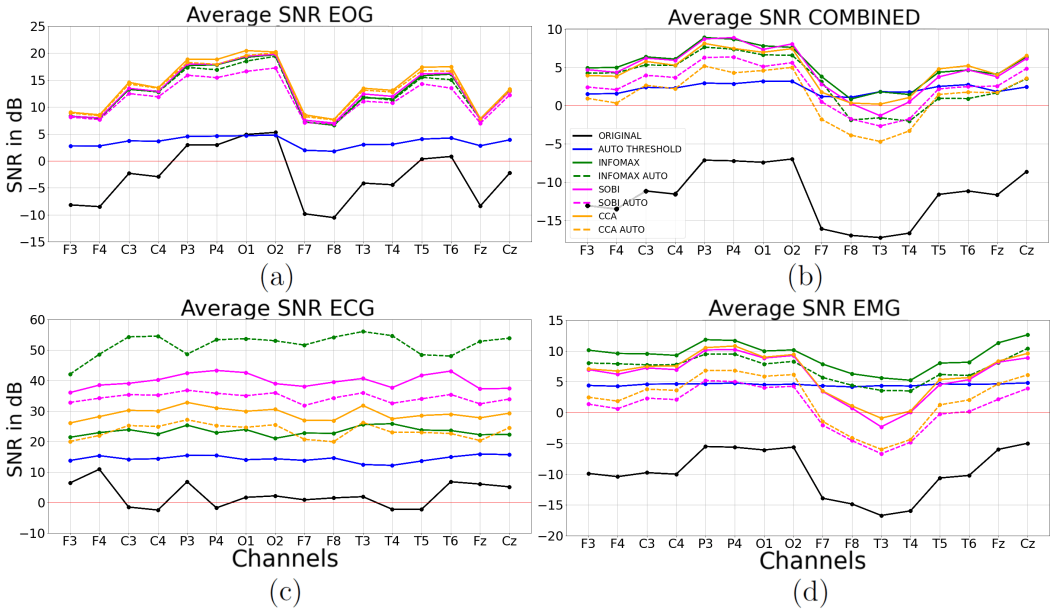


Figure 4.18: The average SNR at the different locations for each artefact type is represented. Each sub-figure shows the new SNR of the cleaned data using four different methods, with the black line representing the average SNR of the original contaminated data (a) EOG (b) COMBINED (c) ECG (d) EMG.

Referring to Figure 4.17, it is clear that the automated and semi-automatic BSS methods produced similar results in most cases. The closest results were for cleaning the EOG when referring to the Figure 4.17(a), 4.17(b) and 4.17(c). More specifically, the automated Extended Infomax, as shown in Figure 4.17(a), was the method that performed the best in removing the EOG artefacts automatically. A t-test comparing the semi-automatic and fully automated Extended Infomax on removing EOG artefacts produced a $P = 0.12 > 0.05$; therefore, we cannot conclude that there is a significant difference between the semi-automatic and automatic results. With the CCA method for cleaning EOG shown in Figure 4.17(c), the results appear to be very similar. However, the two datasets have a $P = 0.03 < 0.05$, therefore we cannot state that there are no significant differences between the two datasets. Overall, one can see that the automated artefact component identification method performed the best with Extended Infomax, producing the closest results across all types of artefacts and the combination thereof. This correlation between Extended Infomax performing the best overall in comparison with the other methods and the components being chosen automatically makes sense because the artefact components identified by the Extended Infomax method contain minimal traces of 'clean' EEG, therefore allowing the automated method to identify the artefact components and to remove them effectively. Another interesting similarity is found in Figure 4.17(d), where the automated threshold

method shows results comparable to that of the semi-automatic CCA method. This could be due to the effectiveness of the BSS methods declining more with an increase in artefact components than the auto threshold method. The greatest differences can be seen with the automatic identification of EMG and ECG artefacts. With the identification of the combination of artefacts, the fully automatic and semi automatic CCA produced similar results with the combination of artefacts, reducing in similarity with EMG and ECG artefacts.

Figure 4.18 shows the SNR of the cleaned data concerning the original contaminated data for the semi-automatic and fully automated methods. Each sub-figure represents a type of contamination. As shown in Figure 4.17, the semi-automatic and automated methods display the closest results for cleaning EOG and the combination of all the artefacts. Regarding Figure 4.18(d), with the cleaning of EMG, CCA and SOBI show big differences between the automated and semi-automatic methods. Still, Extended Infomax has a smaller difference compared to CCA and SOBI. An interesting occurrence is with Figure 4.18(c). It is important to note that the ECG of the cleaned data also represents the average SNR of the channels that had an original SNR of infinity but reduced to a real number as the component removing operation affects those channels. The black line represents the average SNR at those locations, neglecting the data where the channels were not contaminated and therefore having an SNR of infinity. The automated CCA and SOBI both removed more data from the 'clean' EEG than would have been removed with the semi-automatic process. The fact that the automated Extended InfoMax method has a higher average SNR does not mean that the Extended Infomax removed the ECG artefacts more effectively. Rather, the automated Extended InfoMax method, on average, removed much more subtle artefacts that were not ECG artefacts but slight shifts at the start of the signal, therefore not affecting the other channels as much as removing the actual ECG would have.

4.7 Comparison of results to literature

This section compares the findings in this thesis to those found in the literature. As the semi-synthetic datasets have already been developed partially from literature as well as validated and compared with literature, their results are not further compared to the literature in this section.

Regarding the removal of EOG artefacts: Based on the literature, Extended Infomax has been most thoroughly justified in the removal of EOG artefacts [83]. Additionally, other sources have indicated that SOBI is the best method for removing EOG artefacts [8, 84]. Compared to popular ICA methods, CCA has also been shown to identify EOG artefacts for removal clearly [89]. Adaptive thresholding methods have been found to perform accurately due to their effective adaptation to the large variation in EEG and EOG among participants [96]. With the removal of EOG artefacts, no clear method stood out.

Different conclusions were drawn regarding which BSS methods were best for the removal of EOG artefacts. According to the results of this thesis, CCA performed the best, but only slightly better than SOBI and Extended Infomax, which were on par. Despite its adaptive qualities, the auto threshold method performed significantly worse than the other methods in removing EOG artefacts.

Referring to the results of the cleaning methods to remove EMG artefacts: Compared to ICA, CCA is often proposed as a more reliable method for removing muscle artefacts [12, 15, 57, 88]. CCA performs better at cleaning EMG with negative SNR because it is prone to exaggerate EMG contamination influence over scalp EEG activity [76]. The exaggeration may be detrimental to performance when contamination is subtle [76]. Some studies that compared the performance of Extended Infomax, SOBI and CCA in removing EMG artefacts showed similar results [15, 76]. Furthermore, Extended Infomax has been claimed to be the most effective in removing EMG artefacts compared to SOBI [82]. When it came to the performance results of this thesis, Extended Infomax was the best, then SOBI and CCA, followed by the auto threshold method. Although CCA with EMG has gained popularity, the method has not demonstrated robustness to subtler EMG contamination, reducing its overall performance. As expected from the literature, Extended Infomax performed better than SOBI with removing EMG artefacts.

In terms of removing ECG artefacts, SOBI generally outperformed other methods [86]. The results did indeed support this, with ECG outperforming all the other methods in removing ECG artefacts. The second best method was CCA, followed by Extended Infomax and finally the auto threshold method.

There has been limited research comparing these methods for cleaning a combination of all the artefacts, making the results from this thesis more necessary. All the methods performed very similarly with the combination of artefacts, even though they each had their own preferences when used with specific artefacts. In terms of performance, Extended Infomax came in first, followed by SOBI and CCA in second place and auto threshold not so far behind.

Regarding time efficiency: A threshold method is sometimes considered due to its simplicity and time efficiency [15]. In contrast, research has shown that BSS methods are inefficient due to their high computational costs [73]. While BSS methods are computationally arduous, CCA is still a popular choice in BCI research and commercial products, partly because it is time-efficient compared to other BSS methods [12, 76]. As anticipated, the auto threshold method proved to be the most time-efficient, followed by CCA, as expected when compared to the other two BSS methods. The average computation times for Extended Infomax and SOBI were respectively about two and four times longer than CCA.

As mentioned before, the auto threshold method in this thesis was significantly more time-efficient than the other BSS methods. The standard de-

viation of the data determined the thresholds of the method developed. In contrast, Mognon et al. [97] adjusted the threshold values using an expectation-maximisation algorithm which was not fit for real-time application and processing due to the time-consuming nature of the maximisation procedure [97]. Geetha and Geethalakshmi [98] based the thresholding on empirical parameterisation, which was also too time-consuming [96]. The auto threshold method developed performed the best with EMG and the combination of artefacts and was unsuccessful with EOG and ECG. These results differ from Chang et al. [96] who found that individually customised thresholding led to greater accuracy in detecting EOG artefacts due to the high variation between the EOG of participants.

The ability to automate BSS methods are limited since it is not straightforward to classify artefact components automatically [15, 73]. In contrast to those methods found in the literature, the automated approach in this thesis did not require any additional manual input like the labelling of training data, added any significant computational costs by using an additional complex algorithm or require any reference channels [8, 90–95]. Instead, the approach focused on simplicity and was based on the knowledge and emphasis of the amplitude spectra of the artefacts and 'clean' EEG. In this thesis, the semi-automatic and automated BSS methods produced similar results despite the simplicity. In general, the results from the automated EOG component identifications were similar to those obtained from semi-automatic identification, followed by the combination of artefacts and then ECG, with the EMG results being the worst. The researcher in this thesis obtained statistically equivalent results for the semi-automatic and automated identification of the EOG components for Extended Infomax, comparable to the results from Burger and Van Den Heever [90], who used a machine-learning algorithm to automate Infomax. The automation of the ECG identification produced similar but less robust results than the semi-automatic methods. These results were similar to what was found by Hamaneh et al. [92], who automated ICA using a pre-computed template, which was found to not be robust with very low or high SNR ratios. The results of the automated EMG identification in this thesis were found to be the least successful. These results were similar to that from Echtioui et al. [93], who attempted to automate SOBI with the combination of the ADJUST algorithm, but was not successful with EMG artefacts. With the automated identification of the combination of artefacts, the automated Extended Infomax and SOBI were successful for high SNR but performed worse with low SNR values. These results were similar to that found by Daly et al. [91], who combined ICA with a clustering algorithm, which performed best with high SNR but not well with low SNR values with the combination.

Chapter 5

Conclusion

5.1 Review of the project aim

This project aims to identify, automate, and evaluate the robustness of artefact removal methods. This aim was achieved through the accomplishment of the following five objectives:

From a literature review, the most relevant physiological artefacts were identified as EMG, EOG, and ECG artefacts, therefore achieving the first objective. Additionally to the first objective, it was found that literature on the combination of these artefacts was scarce. Therefore, this thesis also provides useful information on how the different artefacts influence each other and the performance of relevant cleaning methods.

An extensive literature review was conducted on EEG cleaning methods related to physiological artefacts such as those mentioned above. As a result, it was determined that only three of the most appropriate and popular methods, namely Extended Infomax, SOBI and CCA, were to be tested. Due to fewer cleaning methods tested, the results of these methods could be analysed more extensively. Therefore, the second objective was accomplished by identifying the most popular and effective methods from the literature. Additionally to the second objective, an auto threshold method was developed, which was time-efficient and comparable in effectiveness to the BSS methods.

The literature found that the SNR was often used as a performance metric for contaminated EEG. Furthermore, the SNR results were often analysed with the use of boxplots and t-tests. Thus, the third objective of identifying the evaluation methods was accomplished.

Simulated data is the most widely used approach for evaluating cleaning methods, but it is still not considered accurate enough for final decisions on the performances of cleaning methods. Due to the popularity but inaccuracy of simulated data, the semi-synthetic dataset was developed to maintain real data accuracy while utilizing the advantages of the SNR metric. Thus, real 'clean' EEG data was contaminated by real EOG and ECG data, with the

simulated EMG data based on real EMG amplitude spectra. Additionally, a combination of these artefacts was simulated. The identified cleaning methods were tested on the developed semi-synthetic dataset. Therefore, the fourth objective was accomplished by developing an accurate semi-synthetic dataset with high variability in measured characteristics for the effective testing of the cleaning methods.

With the semi-synthetic dataset developed, the researcher in this thesis could use the SNR metric to compare and test multiple fully automated approaches to the results of the previous methods and finally develop a successful approach. The semi-synthetic dataset thus enabled the effective development of a fully automated BSS method. The fully automated BSS methods preprocess large contaminated datasets efficiently with comparable effectiveness to the semi-automatic methods, enabling the accomplishment of the fifth objective.

With the development of effective testing procedures, including an accurate and diverse semi-synthetic dataset and simple and standard performance metrics, the researcher could achieve the aim of identifying, automating, and testing the robustness of relevant artefact removal methods.

5.2 Main findings

The effectiveness of each cleaning method was evaluated and compared based on the amount of SNR increase caused by the cleaning method. CCA was the most effective in removing EOG artefacts, with a $P \ll 0.05$ when compared to the other methods. SOBI and Extended InfoMax were less effective but demonstrated similar results to each other with the EOG artefacts with a $P = 0.17 > 0.05$. When it came to removing the EMG artefacts, Extended Infomax was the most effective in increasing the SNR of the contaminated data, with a $P \ll 0.05$ compared to the other methods. SOBI and CCA were less effective than Extended Infomax but did not show a significant difference in removing EMG artefacts when compared to each other, resulting in a $P = 0.06 > 0.05$. All the methods showed distinguishable results with the ECG artefacts, with SOBI outperforming the other methods with a $P \ll 0.05$ when compared to each of the other methods. With the removal of the combination of artefacts, Extended Infomax was the most effective in increasing the SNR of the contaminated data, but not as distinguishable as with its performance for EMG artefacts, having a $P = 0.01$ and 0.02 when compared to CCA and SOBI, respectively. SOBI and CCA demonstrated very similar results with the combination of artefacts, with a $P = 0.57 > 0.05$

It was evident that a particular method worked best with each artefact. However, the BSS and auto threshold methods performed similarly when combining the artefacts. In real-life applications, physiological artefacts often combine. Therefore, even though Extended Infomax was statistically bet-

ter, the BSS methods still showed similar results, indicating that an entirely new approach to cleaning methods is likely essential for further improvements. Moreover, the auto threshold method achieved comparable results to the BSS method in cleaning the EMG and the combination of all three artefacts. The auto threshold method identified the artefacts about 10, 20 and 100 times faster than CCA, Extended Infomax and SOBI, respectively.

In this thesis, the researcher observed similar results when using a simple approach to identify components automatically. Statistically equivalent results were found for both the semi-automatic and automated identification of the EOG components for Extended Infomax with a $P = 0.12 > 0.05$ when comparing the increase of SNR for both approaches. Generally, the automated EOG component identification results were similar to those obtained from semi-automatic identification, followed by the identification from the combination of artefacts and then ECG artefacts. In comparison with semi-automatic identification, the results of the automated EMG identification performance were the least comparable.

In light of the comparable performance of the auto threshold method with the BSS methods and a significant improvement in time, it may be considered a viable cleaning solution in some situations. Therefore, it is believed that the auto threshold methodology can be successfully applied to BCI applications with further development. A few examples of these applications include sport or gaming-related BCIs, or sport-related EEG research requiring immediate analysis.

Since EEG data varies greatly between individuals, the EEG data of numerous participants are required before valid and valuable conclusions may be drawn. A fully automated Extended Infomax, as effective as as the Extended Infomax that required manual intervention for EOG, was developed. Additionally, the automated versions of the BSS methods for each of the artefacts, except for the EMG, provided similar results to the semi-automatic BSS methods. Further development of these methods should enable the effective and efficient preprocessing of large EEG datasets using BSS methods. This method would be especially beneficial for any scenario requiring manual intervention based on BSS methods where it would be impractical or too subjective when a large group of people is involved. By automating BSS methods, raw data from commercial EEG devices, which would normally have been cleaned online for BCI or NFB applications, can be gathered and automatically preprocessed using a slightly less efficient but more effective approach.

The semi-synthetic dataset developed is a simple but useful tool and can be used for testing and comparing the robustness of different cleaning methods in future studies. More development is necessary to automate the BSS methods for EMG, ECG and the combination of artefacts. Still, the fully automated Extended Infomax for EOG can be an effective and efficient tool for the partial preprocessing of large contaminated EEG datasets. The auto threshold method demonstrated comparable results and was up to 100 times

faster than some of the BSS methods. Therefore, in the correct context, despite its simplicity, the auto threshold method is significantly more viable for real-time applications than the BSS methods without compromising too much on effectiveness.

This thesis demonstrates that an accurate and diverse semi-synthetic dataset is an effective tool for comparing conventional and alternative artefact removal techniques, providing a stable datum for comparing the relative performance of each method regarding the most common sources of contamination. Simulations, however, yield preliminary results that can be used as a guide for evaluating and comparing cleaning methods. For future research, one must use recorded EEG data, as a final testbed, to determine whether an artefact removal approach is reliable, reproducible, and performs properly. This thesis further demonstrates that it is possible to automate effective methods, and with further development, the automation could become more robust. With the increase in EEG data acquisition, due to an increase in research interest and commercialisation, fully automatic and effective artefact removal methods are essential. By effectively and practically preprocessing large commercially collected EEG data, we could make accurate analyses and gain new insight into the variations and workings of the human brain.

Appendices

Appendix A

Cleaning methods overview

A.1 Linear regression

Linear regression is one of the first developed preprocessing techniques and is seen as the 'gold standard' [84, 106]. The linear regression method requires a reference channel with which it calculates the proportion of one or various EOG, ECG or EMG references that are present in each EEG channel in the time domain. The artefacts are then separated and removed by subtracting the regression portions [11, 145, 146]. This method is based on the superposition principle and assumption that the signals of each EEG channel are composed of the sum of the clean EEG signal (originating from the brain) and a portion of one or several artefact signals. With regression methods, these artefact signals are measured at their source by reference channels. Thus, regression methods aim to estimate the optimal factor that defines the portion of the artefact signal within each EEG channel. Linear regression has been widely used in removing EOG artefacts due to the locations of the source of the artefact being well defined for the reference channels [69]. Regression methods are currently being replaced by more sophisticated methods due to the disruption of an additional reference channel and their limited effectiveness with artefacts from lesser known or widespread origins, such as ECG and EMG [62, 63].

A.2 Source decomposition

Source decomposition is a method that decomposes every single EEG signal into basic waveforms [69]. It can mainly be categorised into wavelet decomposition, empirical mode decomposition (EMD) and nonlinear mode decomposition (NMD). Wavelet decomposition is an ideal method for biomedical applications because of its versatility [147]. Artefacts are removed by wavelet decomposition in three steps. First, the measured signals are decomposed into levels, then the detail coefficients are filtered, and the signal is reconstructed from the detail coefficients [67, 147]. EMD is a one-dimensional method that

decomposes the measured signal into its basic functions [76, 146]. The key to the success of EMD is that one or more basic functions can represent the signal and the artefacts [76]. EMD has been proven a successful artefact removal method, on its own and in combination with BSS methods [76, 146]. EMD is superior to other signal decomposition methods, such as Fourier or wavelet transformations, since the basis for decomposition is derived adaptively from the data. The ability of EMD to process non-stationary signals is based on the local characteristic time scale of the data [64, 65]. Although EMD adaptively derives the components of decomposition from the data dynamics, it has been suggested that by directly analysing the raw EEG recordings, it may not be able to correct the artefacts [64, 65]. Additionally, the EMD method cannot deal with multidimensional signals and therefore cannot incorporate information from other channels [64, 65]. Applying EMD to the already separated artefact containing components is better than directly to the raw recording. This is one of the main limitations of EMD [64, 65]. NMD breaks down a signal into its nonlinear modes, fully oscillating components and harmonics. NMD consists of four steps: The adaptive extraction of the first harmonic from the synchrosqueezed wavelet transform (SWT), then the determination of possible harmonics, followed by identifying the true harmonics and, finally, the reconstruction of the nonlinear modes [148].

A.3 Blind source separation

Blind source separation (BSS) methods are the most popular artefact removal methods in the research context [15]. BSS methods are component-based methods consisting mainly of principal component analysis (PCA) and independent component analysis (ICA). PCA uses an orthogonal transformation to convert the observations of possibly correlated variables into values of linearly uncorrelated variables called principal components. The objective of the transformation is to produce principal components that have the largest possible variances while being orthogonal to each other [78, 83]. ICA methods unmix linearly mixed signals by imposing the assumption of statistical independence of the sources [83, 149].

A.4 Simple filtering

Simple filtering usually is not an option for removing artefacts from EEG recordings, except for narrowband artefacts like environmental line noise. Thus, the filtering methods adapt the filter parameters to minimise the mean square error between the cleaned EEG and the desired original clean EEG. These filtering methods mainly consist of adaptive, Wiener and Bayes filtering. Adaptive filtering assumes that the ‘clean’ EEG and artefacts are uncorrelated. The

filter generates a signal correlated with the artefact using a reference signal, and then this generated signal is subtracted from the acquired EEG [11]. There are several adaptive filtering algorithms, but adaptive filtering by recursive least squares has shown the best stability, efficiency and fast convergence [66]. It is effective with EOG artefact reduction; however, it has been shown that adaptive filtering is subject to partial removal of neural signals [66]. Wiener filtering is a parametric technique that reduces the mean square error using a statistical approach between the cleaned signal and the desired signal [11]. Wiener filtering overcomes the problem of using additional sensors and extra wiring as required by adaptive filtering [150]. Wiener filtering also shows a significantly greater improvement in increasing the SNR of contaminated data tha than adaptive filtering [150]. A disadvantage of this method is that it only works offline because it needs the whole data set to be applied [150]. Bayesian filtering determines the state of a dynamic system recursively by assuming it is a Markov chain [151]. Bayesian filtering, however, has a high computational complexity, making it an inefficient solution [7].

Appendix B

Evaluation methods overview

Numerous methods to validate and assess cleaning methods exist, where the more straightforward processes use simulated data and the more accurate but complex processes use actual data.

Validation by regression methodologies has been suggested for cleaning methods tested on actual data. Numerous artefact removal methods are evaluated by comparing the correlation of the resulting cleaned data to the reference channel. This method is inconvenient because of the requirement of a reference channel. Furthermore, reference channels are limited with artefacts having less known or widespread origins, such as ECG and EMG. Another method is the standard deviation validation, which compares event-related potential (ERP) consistency associated with eye movements with the EOG reference channels [101].

An alternative approach developed from those proposed by Croft et al. [101], uses regions of interest (ROIs) that are employed to evaluate the specificity and sensitivity of a removal process. Specificity refers to the preservation of neurogenic signals, while sensitivity refers to the attenuation of artefacts. Although this is a promising approach, it is not without challenges. Sensitivity and specificity can only be established with data in which the presence and absence of artefacts are definitive. This method may be deployed using scripted data, such as participants blinking their eyes slowly (to create EOG) or tensing and relaxing muscles in response to instructions (to create EMG). Furthermore, defining ROIs according to myogenic and neurogenic activation peaks is a subjective and challenging process [97, 102, 103].

Sweeney et al. [152] proposed a novel scenario where the researcher controls the EEG recording completely. Their study presents a method for creating two highly correlated signals. One is a reference devoid of artefacts, while the other intentionally contains artefacts. As a result of this controlled scenario, it is possible to apply artefact removal methods to the noisy EEG and compare the resultant signal with the actual clean EEG [152].

Alternatively, some researchers use visual inspection to evaluate the effectiveness of artefact removal methods. By comparing the time, frequency and

spatial characteristics of the cleaned signal to the measured signal, it is possible to evaluate the performance of specific methods visually. Although it is a subjective approach and relies on expert review, it can still reveal whether an algorithm improved signal quality or distorted intervals or frequency bands [66, 84, 104].

Appendix C

Automation of blind source separation methods

C.1 Literature approaches for the automation of BSS methods

Joyce et al. [8] combined SOBI with two EOG reference channels to identify the EOG components automatically. The technique had a high degree of accuracy but was limited due to the reference channels being disruptive and having limited effectiveness with artefacts from lesser-known or widespread origins, such as ECG and EMG [8, 62, 63].

Burger and Van Den Heever [90] combined Infomax and wavelet neural networking (WNN) to identify EOG components automatically. The method was also successful but required a manually labelled training set, consisting of clean EEG and artefacts and used a computationally expensive machine learning algorithm [90].

Hamaneh et al. [92] combined ICA with a pre-computed template and continuous wavelet transformation to identify ECG components automatically. Their method was not very robust with very low or high SNR ratios and required a pre-computed template to be successful [92].

Echtioui et al. [93] combined SOBI with the developed ADJUST [97] technique to identify EOG, ECG and EMG artefacts automatically. The ADJUST method could not detect EMG artefacts and was mainly successful with EOG artefacts [93].

Winkler and Stefan Haufe and Michael Tangermann [94] combined temporal decorrelation source separation (TDSEP) with a linear programming machine (LPM) model to identify EOG and EMG artefacts automatically. The machine learning model had to be trained on 640 components manually labelled by experts. Their method performed similar to the semi-automatic TDSEP but required an inefficient amount of labelled data before application [94].

Daly et al. [91] combined an ICA method, namely TDSEP with lagged auto-mutual information clustering (LAMICA) to identify ECG, EMG and EOG artefact components automatically. LAMIC used the auto mutual information (AMI) for each component to estimate which were artefact components. The technique performed best with high SNR in comparison to other automated methods such as wavelet transformation and multivariate singular spectrum analysis, but not well with low SNRs [91].

Frølich et al. [95] combined Extended Infomax and a developed multinomial regression classifier to automatically identify EOG, ECG and EMG artefacts. This method was thoroughly validated and showed a high classification performance. However, their approach had poor generalisability [95].

Appendix D

Forward model

Researchers have developed 3D models of the brain, skull and scalp to develop more realistic EEG simulations. The brain stimulation is related to source modelling, which is a model of the physiological sources. As part of the simulation of the scalp, the volume conduction of sources throughout the head and skull is modelled. The scalp is the final step in the simulation process, during which the observed electrochemical events that the electrodes record are modelled. These three steps combined are known as the forward model of the brain [107, 108].

The attempts to better quality simulation include developing 3D models of the brain, skull and scalp. Measured EEG is generated by considering dipolar sources and solving the electromagnetic forward problem [15]. Forward models can provide realistic head models and can be readily generated for the analysis of EEG. Equation D.1 describes the mathematical expression of the forward model [107]:

$$\mathbf{x}(\mathbf{t}) = \mathbf{L}\mathbf{j}(\mathbf{t}) + \epsilon \tag{D.1}$$

In equation D.1, $\mathbf{x}(\mathbf{t})$ represents the observed signals from the scalp. Furthermore, the time-dependent $3R$-dimensional vector $\mathbf{j}(\mathbf{t})$ represents the physiological sources at R distinct locations on the cortical surface. The $M \times 3R$ lead field matrix $\mathbf{L}$ describes the relationship between the physiological source and the observable scalp signals at M sensors, therefore the mixing and conduction of the sources. Finally, ϵ is a M-dimensional noise vector [107].

For the successful simulation of realistic models, the models must use a high-resolution average anatomy template. In addition, a realistic volume conductor model must be used. The interacting sources must exert a time-delayed impact on each other. Communication between sources must only occur within a specific frequency range. A realistic source location should have electrical currents that propagate perpendicular to the scalp surface and should be constrained to the cortical manifold. A detailed model must also include a wide variety of locations, spatial extents and depths of sources. Independent

background brain processes, pink noise spectrums, white sensor noises and realistic SNR ranges must also be present [107–109].

Although the forward model attempts to be the most accurate, there are limitations to the EEG-based estimation of functional or effective brain connectivity. Limitations include the disregard of, or insufficient modelling of the source mixing caused by head tissue conductivity, disregard of correlated noise sources and generally an overestimation of the signal-to-noise ratio (SNR) [107–109].

Appendix E

Blind source separation cleaning methods

E.1 Independent component analysis mathematical and statistical methods

The ICA method explanation in this section is based on the explanation from Hyvärinen and Oja [153]. ICA is used to estimate the unmixing matrix W for the calculation sources S, based on two main assumptions: that the sources are statistically independent and non-Gaussian. Therefore, the unknown W and S are calculated by maximising these two assumed attributes.

To simplify the explanation of the ICA method, it has been assumed that each mixture x_j as well as each independent component s_j are random variables, instead of proper time signals. Equation E.1 shows the relationship between the measured data, x, and the sources, s, along with the mixing coefficients, a, where n represents the total number of components, and j represents the individual component. In this application, the total number of components, n, is equal to the total number of sources or channels. As can be seen in equation E.1, ICA is a generative model, meaning that it describes how the observed data x_j is generated by the process of mixing the components s_j.

$$x_j = a_{j1}s_1 + a_{j2}s_2 + ... + a_{jn}s_n, \text{for all j} \tag{E.1}$$

By assuming that each mixture x_j as well as each independent component s_j are random variables, we are effectively now working with $\mathbf{x}$ and $\mathbf{s}$ as vectors, with $\mathbf{A}$ remaining a n by n matrix as seen in equation E.2. Then after estimating the matrix $\mathbf{A}$, the inverse of $\mathbf{A}$, being $\mathbf{W}$, can be used to obtain the independent and non-Gaussian sources, $\mathbf{s}$ as shown in equation E.3

$$\mathbf{x} = \mathbf{As} \tag{E.2}$$

$$\mathbf{s} = \mathbf{W}\mathbf{x} \tag{E.3}$$

The first important principle utilised in the ICA method is that of variables being statistically independent. Statistically independence means that if you have two random scalar variables, y_1 and y_2, they do not provide any information of one another.

To define the technical definition of these variables' statistical independence, we must first define $p(y_1, y_2)$ as the joint probability density function (pdf), and the marginal pdf, $p_1(y_1)$, when considered alone as seen in equation E.4:

$$p_1(y_1) = \int p(y_1, y_2) dy_2 \tag{E.4}$$

With the marginal and joint probability density functions defined, we can finally state mathematically that y_1 and y_2 are statistically independent if the joint pdf can be factorised as shown in equation E.5:

$$p(y_1, y_2) = p_1(y_1)p_2(y_2) \tag{E.5}$$

The definition is then used to derive the most important property of independent random variables, to be used to quantify non-Gaussian later on. Given two functions h_1 and h_2, the statistical independence also means that the relationship in equation E.6 holds:

$$E\{h_1(y_1), h_2(y_2)\} = E\{h_1(y_1)\}E\{h_2(y_2)\} \tag{E.6}$$

Where the definition of E can be seen in equation E.7:

$$E\{h_1(y_1)\} = \int h_1(y_1)p_1(y_1)dy_1 \tag{E.7}$$

The real key to the ICA method's success is the non-Gaussian nature of the sources. Due to the assumption that the sources are independent, the central limit theorem (CLT) can be deployed in the process of estimating the mixing matrix, $\mathbf{W}$. The CLT states that the distribution of the sum of random independent variables tends towards a more Gaussian distribution. Therefore, if we can quantify the Gaussianity, we can use it in combination with the fact that the sum of the independent variables must be more Gaussian than the original variables to estimate the mixing matrix $\mathbf{W}$. To effectively apply the CLT, some rearrangements must first be made to the variables. To estimate one of the independent components, we first consider a linear combination of the x_i, denoted by y in equation E.8. We further define $\mathbf{z}$ in equation E.9, and show the relationship between these variables in equation E.10:

$$y = \mathbf{w}^{\mathbf{T}}x \tag{E.8}$$

$$\mathbf{z} = \mathbf{A}^{\mathbf{T}}\mathbf{w} \tag{E.9}$$

$$y = \mathbf{w}^{\mathbf{T}}x = \mathbf{w}^{\mathbf{T}}\mathbf{A}\mathbf{s} = \mathbf{z}^{\mathbf{T}}\mathbf{s} \tag{E.10}$$

With equation E.10, y can be seen as a linear combination of s_i with weights given by z_i. The w vectors can then be estimated using a chosen measurement of Gaussianity and equation E.10. With the CLT, it is known that the sum of independent variables is more Gaussian than the original individual variables. Therefore, $\mathbf{z}^{\mathbf{T}}\mathbf{s}$ is more Gaussian than any of the s_i sources. Furthermore, $\mathbf{z}^{\mathbf{T}}\mathbf{s}$ is therefore the least Gaussian when $\mathbf{z}^{\mathbf{T}}\mathbf{s} = s_i$, meaning that only one of the z_i elements are non-zero. Thus, the goal is to maximize the non-Gaussian nature of the sources, with the first component already being known as $y = \mathbf{w}^{\mathbf{T}}x = \mathbf{z}^{\mathbf{T}}\mathbf{s}$ where only one of the z_i elements are non-zero.

Maximizing the non-Gaussianity of $\mathbf{w}^{\mathbf{T}}x$ consists of an n-dimensional landscape of vectors $\mathbf{w}$, with 2n local maxima, two maxima's per component, corresponding to s_i and $-s_i$. Therefore to find the components, we need to find the locations of the local maxima's. To create this landscape of $\mathbf{w}$ vectors, we need a quantitative measurement of Gaussianity.

The solving of the ICA problem is based on minimizing or maximizing certain contrast functions, thus transforming the ICA problem into a numerical optimization problem. Due to its computational and theoretical simplicity, the classical contrast function and measure of Gaussianity for ICA is kurtosis. For a normalised version of y, the formula for kurtosis can be seen in equation E.11, with the important properties of the kurtosis equation shown in equation E.12 and E.13 with α being a constant:

$$kurt(y) = E\{y^4\} - 3 \tag{E.11}$$

$$kurt(y_1 + y_2) = kurt(y_1) + kurt(y_2) \tag{E.12}$$

$$kurt(\alpha y_1) = \alpha^4 kurt(y_1) \tag{E.13}$$

To understand how the landscape for kurtosis would look like, we will consider a two dimensional model of $\mathbf{x} = \mathbf{A}\mathbf{s}$, where we assume that s_1 and s_2 has $kurt(s_1)$ and $kurt(s_2)$ values. As previously stated, for the maximization of non-Gaussianity, we already know that one of the components is $y = \mathbf{w}^{\mathbf{T}}x$, where only one of the z_i values are non-zero. Furthermore, the estimation of the W vectors becomes an optimization problem with the goal of maximization the contrast function $|kurt(y)|$, which is described in equation E.15 for two dimensions, developed using equation E.9, E.14 and the kurtosis properties shown in equation E.12 and E.13.

$$y = \mathbf{w}^{\mathbf{T}}x = \mathbf{w}^{\mathbf{T}}\mathbf{A}\mathbf{s} = \mathbf{z}^{\mathbf{T}}\mathbf{s} = z_1 s_1 + z_2 s_2 \tag{E.14}$$

$$kurt(y) = kurt(z_1 s_1) + kurt(z_2 s_2) = z_1^4 kurt(s_1) + z_2^4 kurt(s_2) \tag{E.15}$$

In practice, a random $\mathbf{w}$ vector is initialised, and then the direction of which $|kurt(y)|$, the contrast function, is growing and decreasing the most strongly, is determined based on the available observed data $\mathbf{x}(\mathbf{1}), ..., \mathbf{x}(\mathbf{n})$. Then a gradient method is deployed to find the final values of the $\mathbf{w}$ vectors to find the final unmixing matrix $\mathbf{W}$.

E.1.1 Extended Infomax mathematical and statistical methods

Other methods based on the ICA assumptions of non-Gaussianity and independence of sources have been developed, such as InfoMax. Infomax is a related contrast function derived from a neural network viewpoint [153]. Infomax is based on maximizing the output entropy or information flow of a neural network with non-linear outputs and is thus named InfoMax [153]. Infomax has since been further developed so that the function used to estimate $\mathbf{w}$ adjusts according to the Gaussian nature of the unmixing sources. This method is called Extended Infomax [154]. Extended Infomax consists of three simple steps. The first step is to initialize a random unmixing matrix $\mathbf{W}$; the second step is to repeat the calculations of equation E.16 until it converges; the third step is included in the second step, which is to determine whether the sources are super or sub-Gaussian, determined by a kurtosis, upon which the function $f(\mathbf{f}S)$ is changed to its relevant form, shown in equation E.17 and E.18 [155, 156]:

$$\mathbf{W}(\mathbf{t}+\mathbf{1}) = \mathbf{W}(\mathbf{t}+\mathbf{1}) + \eta(t)(\mathbf{I} - f(\mathbf{S})\mathbf{S}^{\mathbf{T}})\mathbf{W}(\mathbf{t}) \tag{E.16}$$

$$f(\mathbf{S}) = \tanh(\mathbf{S}) \text{ (super Gaussian)} \tag{E.17}$$

$$f(\mathbf{S}) = \mathbf{S} - \tanh(\mathbf{S}) \text{ (sub Gaussian)} \tag{E.18}$$

In equation E.16, $\eta(t)$ is the learning rate function, which specifies the steps for the unmixing matrix updates, and is usually either an exponential function or a constant. $\mathbf{I}$ is an identity matrix of dimensions n by n. The sources can then be found with $\mathbf{S} = \mathbf{WX}$ [155, 156].

E.1.2 Second order blind identification mathematical and statistical methods

Another method based on ICA assumptions is the second order blind identification (SOBI) method. This method consists of five steps as described by Belouchrani et al. [157]. The first step of this method is to estimate the initial

sample covariance, $\hat{\mathbf{R}}(0)$ from the total samples M. With $\lambda_1, ..., \lambda_n$ being the n largest eigenvalues with $h_1, ..., h_n$ being the corresponding eigenvectors of $\hat{\mathbf{R}}(0)$. n is denoted as the the smaller dimension and m the larger dimension of the matrices. The sample covariance matrix is calculated as described by equation E.19, where $\tau \in \{\tau_j | j = 1, ..., K\}$ for a fixed set of time lags K, and H denotes the complex conjugate transpose:

$$\hat{\mathbf{R}}_\tau(\mathbf{x}) = \frac{1}{M-\tau} \sum_{t=1}^{M-\tau} \mathbf{x_t} \mathbf{x_{t-\tau}^H} \tag{E.19}$$

The second step is to estimate the noise variance, $\hat{\sigma}^2$, which is under the white noise assumption, assumed to be the average of the k smallest eigenvalues of $\hat{\mathbf{R}}(0)$, where $k = m - n$. The whitened signals $\mathbf{z}(t)$ in equation E.20 is then calculated through the calculation of each $z_i(t)$, where $1 \leq i \leq n$, showed in equation E.21, to ultimately form the whitening matrix in equation E.22:

$$\mathbf{z}(t) = [z_1(t), ..., z_n(t)]^M \tag{E.20}$$

$$z_i(t) = (\lambda_i - \hat{\sigma}^2)^{-1/2} \mathbf{h}_i^* \mathbf{x}(t) \tag{E.21}$$

$$\mathbf{W} = [(\lambda_1 - \hat{\sigma}^2)^{-1/2} \mathbf{h}_1, ...(\lambda_n - \hat{\sigma}^2)^{-1/2} \mathbf{h}_n]^H \tag{E.22}$$

In equation E.21, $*$ denotes the conjugate transpose of a vector. The third step is to form sample estimates, $\underline{\hat{\mathbf{R}}}(\tau)$ by calculating $\mathbf{z}(t)$ for a fixed set of time lags K. The fourth step is then to obtain a unitary matrix $\hat{\mathbf{U}}$ as a joint diagonalizer of the set $\{\underline{\hat{\mathbf{R}}}(\tau_j) | j = 1, ..., K\}$. The final step is then to estimate the source signals $\hat{s}(t)$, where the method differs slightly from the original BSS method as in equation E.23 from which the mixing matrix $\mathbf{A}$ can be estimated such as in equation E.24, where the # is Moore Penrose pseudoinverse.

$$\hat{\mathbf{s}}(t) = \hat{\mathbf{U}}^H \mathbf{W} \mathbf{x}(t) \tag{E.23}$$

$$\mathbf{A} = \mathbf{W}^{\#} \hat{\mathbf{U}} \tag{E.24}$$

E.2 Canonical correlation analysis mathematical and statistical methods

Canonical correlation analysis (CCA) solves the BSS problem by forcing the sources to be maximally autocorrelated and mutually uncorrelated [121]. Autocorrelation is the degree of similarity between a given time series and a lagged version of itself over successive time intervals. When two variables are uncorrelated, it means that the correlation coefficients concerning each other

are close or equal to zero. The CCA method described below is based on the explanation of the CCA method on the application of EEG signals from Lin et al. [121] and Zhuang et al. [158].

In the CCA method, $\mathbf{Y}(t)$ is the fractionally delayed version of the measured EEG signals $\mathbf{X}(t)$. Therefore, $\mathbf{Y}(t) = \mathbf{X}(t) - 1$, where the number 1 represents a sample shift; therefore, in this research, $\mathbf{Y}(t)$ is delayed by a 200th of a second due to the sampling rate being 200 Hz. Additionally, the mean of each row of the $\mathbf{X}(t)$ and $\mathbf{Y}(t)$ matrices are removed to derive to sets of basis vectors for $\mathbf{X}$ and $\mathbf{Y}$.

Suppose we have two canonical variables, U and V, that comprises of the linear combinations of $\mathbf{X}$ and $\mathbf{Y}$ respectively, as seen in equation E.25 and E.26:

$$U(t) = \mathbf{w_x^T X}(t) \tag{E.25}$$

$$V(t) = \mathbf{w_y^T Y}(t) \tag{E.26}$$

CCA is then used to find the matrices $\mathbf{w_x} = [w_{x_1}, ..., w_{x_n}]$ and $\mathbf{w_y} = [w_{y_1}, ..., w_{y_n}]$ that maximizes the correlation ρ between U and V. In order to maximize the correlation between U and V, the objective function E.27 must be solved, where $\mathbf{C_{xx}}$ and $\mathbf{C_{yy}}$ are the autocovariance matrices of $\mathbf{X}$ and $\mathbf{X}$, and $\mathbf{C_{xy}} = \mathbf{C_{yx}^T}$ are the cross-covariance matrices of X and Y.

$$\underset{W_X, W_Y}{max}\ \rho(U, V) = \frac{\mathbf{w_x^T C_{xy} w_y}}{\sqrt{(\mathbf{w_x^T C_{xx} w_x})(\mathbf{w_y^T C_{yy} w_y})}} \tag{E.27}$$

The canonical coefficients w_x and w_y can be found by setting the partial derivative of the objective function E.27 with respect w_x and w_y to zero, respectively, leading to equation E.28:

$$\mathbf{C_{xy} w_y} = \rho \mathbf{C_{xx} w_x} \text{ and } \mathbf{C_{yx} w_x} = \rho \mathbf{C_{yy} w_y} \tag{E.28}$$

Equation E.28 can be further reduced to a classical eigenvalue problem, if C_{xx} and C_{yy} is invertible and $\rho^2 \in [0\ 1]$ is the eigenvalue, into equations E.29 and E.30:

$$\mathbf{C_{xx}^{-1} C_{xy} C_{yy}^{-1} C_{yx} w_x} = \rho^2 \mathbf{w_x} \tag{E.29}$$

$$\mathbf{C_{yy}^{-1} C_{yx} C_{xx}^{-1} C_{xy} w_y} = \rho^2 \mathbf{w_y} \tag{E.30}$$

$u_i(t)$ and $v_i(t)$ represents the ith pair of canonical variates, with the correlation between these variable represented by ρ_i. Finally, $U(t)$ is related to the estimated components $S(t)$ as shown in equation E.31:

$$S(t) = \mathbf{W X}(t) = U(t) = \mathbf{w_x^T X}(t) \tag{E.31}$$

The unknown mixing matrix $\mathbf{A}$ can be then be calculated from the inverse of the unmixing matrix $\mathbf{W}$, where $\mathbf{W} = \mathbf{w_x}$

Appendix F

Boxplots and t-tests

F.1 Statistical analysis used for simulations and methods

Boxplots are simple and standardized ways to visualize and compare the distribution and other statistical information of the datasets. With a boxplot, one can easily determine whether the data is symmetrical, if it is tightly grouped data, and whether or not the data is skewed. The data points of the boxplot consist of the minimum data point, the lower quartile ($Q1$), the median, the upper quartile ($Q3$), the maximum data point and the outliers. The median is the midpoint in data and is usually represented by a line dividing the boxplot box into two halves. Therefore, half of the data is greater and half is smaller than the median. 25% of the total data falls below $Q1$, and 75% falls below $Q3$. The interquartile range (IQR) ranges between Q1 and Q3 and represents 50% of the data. A whisker connects the minimum and maximum data points to the IQR. Minimum data points are equal to $Q1 - 1.5(IQR)$ and maximum data points are equal to $Q3 + 1.5(IQR)$. We consider values that fall below or above the minimum and maximum scores as numerical outliers.

As one of several statistical tests used to test hypotheses, the t-test is widely used in statistics. The t-test is an inferential statistical method that tests the reliable difference in means between two datasets. Thus, the t-test indicates how reliable it is when considering whether two datasets differ. A t-test measures the variance between two datasets and the variance within each dataset and compares these variances. There is a corresponding p-value for each t-value. A p-value is a measure of the probability that an observed difference occurred y only chance. When the p-value is greater than 0.05, we can be more than 95% confident that the two datasets are the same. A higher p-value indicates that the two datasets are more likely to be the same. A p-value greater than 0.05 indicates that no significant difference exists between the two datasets.

Appendix G

Python and libraries information

G.1 Python reference

Python is developed under an OSI-approved open source license, making it freely usable and distributable, even for commercial use. Developed by the Python Software Foundation. The Python Language Reference is version 3.8.5. Available at `https://www.python.org/downloads/release/python-385/`

Appendix H

Signal to noise tables

H.1 Electrooculography signal to noise Table

Table H.1: Highest and lowest EOG SNR values from literature.

Study	**[126, 127]**	**[131]**	**[128]**	**[129]**	**[132]**	**[130]**
Highest (dB)	5	2	60	23	10	0
Lowest (dB)	-5	-4	1	2	-40	-20

The results from the SNR values in Table H.1 show that the lowest and highest values used for EOG were -40 and 60 dB, which vary highly compared to the values used by other literature.

H.2 Electromyography signal to noise Table

Table H.2: Highest and lowest EMG SNR values from literature.

Study	**[104, 133]**	**[12, 57]**	**[134]**	**[55]**	**[56]**	**[135]**	**[136]**	**[137]**
Highest (dB)	4.8	4.5	-19.4	0	1	4	3	30
Lowest (dB)	-6	0.5	-36.1	-15	-10	-4	-6.8	-30

The results from the SNR values in Table H.2show that the lowest and highest values used for EMG were -36.1 and 30 dB, which vary highly compared to the values used by other literature.

H.3 Electroencephalography signal to noise Table

Table H.3: Highest and lowest ECG SNR values from literature.

Study	**[143]**	**[139]**	**[61]**	**[140]**	**[59]**	**[141]**	**[142]**	**[144]**
Highest (dB)	15	15	15	10	5	1	10	-5
Lowest (dB)	15	8	5	-6	2	0	-5	-5

The results from the SNR values in Table H.3 show that the lowest and highest values used for ECG were -6 to 15 dB, similar to that found in other literature.

Appendix I

Electroencephalography temporal validation

I.1 Comparison of semi-synthetic electroencephalography to standard QRS waves

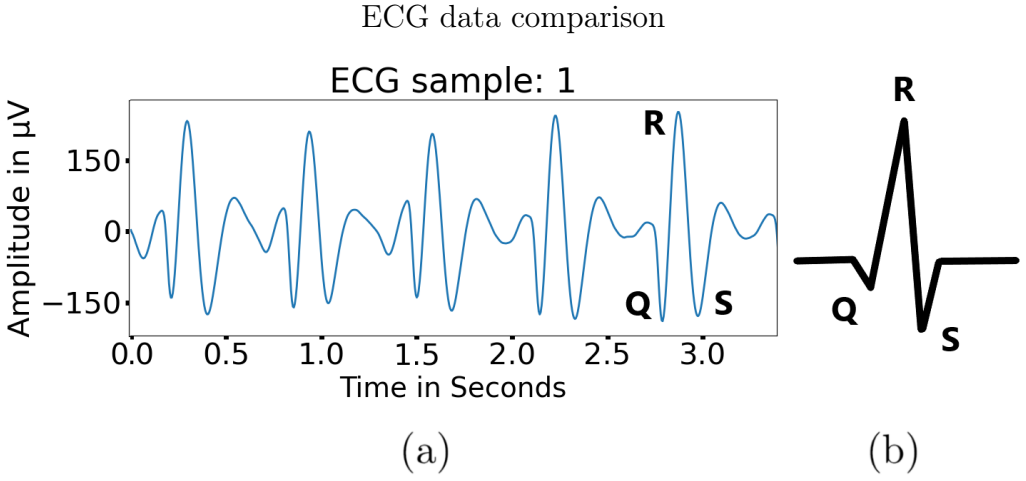


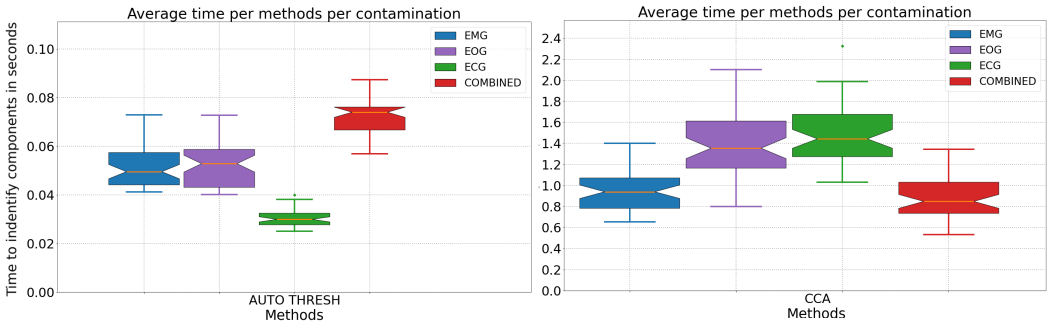
Figure I.1: Expanded view of the shape of actual ECG data used on the left and an image of a standard shape of ECG beat.

Appendix J

Cleaning Time

J.1 Expanded view of time distribution

Expanded view of auto threshold and CCA time distribution



(a) (b)

Figure J.1: (a) auto threshold method time per method per contamination. (b) CCA method time per method per contamination.

List of References

[1] Rytis Maskeliunas, Robertas Damasevicius, Ignas Martisius, and Mindaugas Vasiljevas. Consumer-grade EEG devices: Are they usable for control tasks? *PeerJ*, 2016(3):1–27, 2016. ISSN 21678359. doi: 10.7717/peerj.1746.

[2] Suhas S. Patil and Minal K. Pawar. Quality advancement of EEG by wavelet denoising for biomedical analysis. *Proceedings - 2012 International Conference on Communication, Information and Computing Technology, ICCICT 2012*, pages 10–15, 2012. doi: 10.1109/ICCICT.2012.6398151.

[3] Xiao Jiang, Gui Bin Bian, and Zean Tian. Removal of artifacts from EEG signals: A review. *Sensors (Switzerland)*, 19(5):1–18, 2019. ISSN 14248220. doi: 10.3390/s19050987.

[4] Elham Barzegaran, Sebastian Bosse, Peter J. Kohler, and Anthony M. Norcia. EEGSourceSim: A framework for realistic simulation of EEG scalp data using MRI-based forward models and biologically plausible signals and noise. *Journal of Neuroscience Methods*, 328(August):108377, 2019. ISSN 1872678X. doi: 10.1016/j.jneumeth.2019.108377. URL `https://doi.org/10.1016/j.jneumeth.2019.108377`.

[5] H. N. Suresh and C. Puttamadappa. Removal of EMG and ECG artifacts from EEG based on real time recurrent learning algorithm. *International Journal of Physical Sciences*, 3(5):120–125, 2008. ISSN 19921950.

[6] Molly McVoy, Sarah Lytle, Erin Fulchiero, Michelle E. Aebi, Olunfunke Adeleye, and Martha Sajatovic. A systematic review of quantitative EEG as a possible biomarker in child psychiatric disorders. *Psychiatry Research*, 279 (April):331–344, 2019. ISSN 18727123. doi: 10.1016/j.psychres.2019.07.004. URL `https://doi.org/10.1016/j.psychres.2019.07.004`.

[7] He Chen, Wenqing Chen, Yan Song, Li Sun, and Xiaoli Li. EEG characteristics of children with attention-deficit/hyperactivity disorder. *Neuroscience*, 406: 444–456, 2019. ISSN 18737544. doi: 10.1016/j.neuroscience.2019.03.048. URL `https://doi.org/10.1016/j.neuroscience.2019.03.048`.

[8] Carrie A. Joyce, Irina F. Gorodnitsky, and Marta Kutas. Automatic removal of eye movement and blink artifacts from EEG data using blind component separation. *Psychophysiology*, 41(2):313–325, 2004. ISSN 00485772. doi: 10.1111/j.1469-8986.2003.00141.x.

[9] Md Kafiul Islam, Amir Rastegarnia, and Zhi Yang. Les méthodes de détection et de rejet d'artefact de l'EEG de scalp : revue de littérature. *Neurophysiologie Clinique*, 46(4-5):287–305, 2016. ISSN 17697131. doi: 10.1016/j.neucli.2016.07.002. URL `http://dx.doi.org/10.1016/j.neucli.2016.07.002`.

[10] Stéphanie Devuyst, Thierry Dutoit, Patricia Stenuit, Myriam Kerkhofs, and Etienne Stanus. Cancelling ECG artifacts in EEG using a modified independent component analysis approach. *Eurasip Journal on Advances in Signal Processing*, 2008, 2008. ISSN 16876172. doi: 10.1155/2008/747325.

[11] Kevin T. Sweeney, Hasan Ayaz, Tomás E. Ward, Meltem Izzetoglu, Seán F. McLoone, and Banu Onaral. A methodology for validating artifact removal techniques for physiological signals. *IEEE Transactions on Information Technology in Biomedicine*, 16(5):918–926, 2012. ISSN 10897771. doi: 10.1109/TITB.2012.2207400.

[12] Xun Chen, Aiping Liu, Joyce Chiang, Z. Jane Wang, Martin J. McKeown, and Rabab K. Ward. Removing Muscle Artifacts from EEG Data: Multichannel or Single-Channel Techniques? *IEEE Sensors Journal*, 16(7):1986–1997, 2016. ISSN 1530437X. doi: 10.1109/JSEN.2015.2506982.

[13] Ian Daly, Reinhold Scherer, Martin Billinger, and Gernot Müller-Putz. FORCe: Fully online and automated artifact removal for brain-computer interfacing. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 23(5):725–736, 2015. ISSN 15344320. doi: 10.1109/TNSRE.2014.2346621.

[14] Quentin Barthelemy, Louis Mayaud, David Ojeda, and Marco Congedo. The Riemannian Potato Field: A Tool for Online Signal Quality Index of EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(2):244–255, 2019. ISSN 15344320. doi: 10.1109/TNSRE.2019.2893113.

[15] Jose Antonio Urigüen and Begoña Garcia-Zapirain. EEG artifact removal - State-of-the-art and guidelines. *Journal of Neural Engineering*, 12(3), 2015. ISSN 17412552. doi: 10.1088/1741-2560/12/3/031001.

[16] Phattarapong Sawangjai, Supanida Hompoonsup, Pitshaporn Leelaarporn, Supavit Kongwudhikunakorn, and Theerawit Wilaiprasitporn. Consumer Grade EEG Measuring Sensors as Research Tools: A Review. *IEEE Sensors Journal*, 20(8):3996–4024, 2020. ISSN 15581748. doi: 10.1109/JSEN.2019.2962874.

[17] Mahsa Soufineyestani, Dale Dowling, and Arshia Khan. Electroencephalography (EEG) technology applications and available devices. *Applied Sciences (Switzerland)*, 10(21):1–23, 2020. ISSN 20763417. doi: 10.3390/app10217453.

[18] Reza Yaghoobi Karimui, Sassan Azadi, and Parviz Keshavarzi. The ADHD effect on the high-dimensional phase space trajectories of EEG signals. *Chaos, Solitons and Fractals*, 121:39–49, 2019. ISSN 09600779. doi: 10.1016/j.chaos.2019.02.004. URL `https://doi.org/10.1016/j.chaos.2019.02.004`.

[19] John N. Demos. *Getting Started with EEG Neurofeedback, Second Edition*. W. W. Norton & Company, 2018.

[20] Robert J. Barry, Adam R. Clarke, Rory McCarthy, Mark Selikowitz, Jacqueline A. Rushby, and Elizabeta Ploskova. EEG differences in children as a function of resting-state arousal level. *Clinical Neurophysiology*, 115(2):402–408, 2004. ISSN 13882457. doi: 10.1016/S1388-2457(03)00343-2.

[21] Adam R. Clarke, Robert J. Barry, Rory McCarthy, and Mark Selikowitz. Electroencephalogram differences in two subtypes of Attention-Deficit/Hyperactivity Disorder. *Psychophysiology*, 38(2):212–221, 2001. ISSN 00485772. doi: 10.1111/1469-8986.3820212.

[22] Saeid Sanei and J.A. Chambers. *EEG SIGNAL PROCESSING*. 2007. ISBN 9780470025819.

[23] Stuart Ira Fox. Human physiology—Textbook. 2011.

[24] Paul L. Nunez and Ramesh Srinivasan. *Electric Fields of the Brain The Neurophysics of EEG*, volume 10. 2006. ISBN 9780195050387. doi: 10.3390/brainsci10060379.

[25] C. Stephani, G. Fernandez-Baca Vaca, R. MacIunas, M. Koubeissi, and H. O. Lüders. Functional neuroanatomy of the insular lobe. *Brain Structure and Function*, 216(2):137–149, 2011. ISSN 18632653. doi: 10.1007/s00429-010-0296-3.

[26] Galvan, Federico R., Violeta Barranco, Juan C. Galvan, Santiago Batlle, Sebastian FeliuFajardo, and García. We are IntechOpen , the world ' s leading publisher of Open Access books Built by scientists , for scientists TOP 1 %. *Intech*, i(tourism):13, 2016. doi: http://dx.doi.org/10.5772/57353. URL `https://www.intechopen.com/books/advanced-biometric-technologies/liveness-detection-in-biometrics`.

[27] Marzbani MarH., Marateb H.R., and Mansourian M. Methodological note: Neurofeedback: A comprehensive review on system design, methodology and clinical applications. *Basic and Clinical Neuroscience*, 7(2):143–158, 2016. ISSN 2008-126X. URL `http://www.embase.com/search/results?subaction=viewrecord&from=export&id=L610031835%0Ahttp://sfx.library.uu.nl/utrecht?sid=EMBASE&issn=2008126X&id=doi:&atitle=Methodological+note%3A+Neurofeedback%3A+A+comprehensive+review+on+system+design%2C+methodology+a`.

[28] Stefanie Enriquez-Geppert, Diede Smit, Miguel Garcia Pimenta, and Martijn Arns. Neurofeedback as a Treatment Intervention in ADHD: Current Evidence and Practice. *Current Psychiatry Reports*, 21(6), 2019. ISSN 15351645. doi: 10.1007/s11920-019-1021-4.

[29] Muthuraman Muthuraman, Vera Moliadze, Lena Boecher, Julia Siemann, Christine M. Freitag, Sergiu Groppa, and Michael Siniatchkin. Multimodal alterations of directed connectivity profiles in patients with attention-deficit/hyperactivity disorders. *Scientific Reports*, 9(1):1–13, 2019. ISSN 20452322. doi: 10.1038/s41598-019-56398-8.

[30] Martijn Arns, Hartmut Heinrich, and Ute Strehl. Evaluation of neurofeedback in ADHD: The long and winding road. *Biological Psychology*, 95(1):108–115, 2014. ISSN 03010511. doi: 10.1016/j.biopsycho.2013.11.013. URL `http://dx.doi.org/10.1016/j.biopsycho.2013.11.013`.

[31] E. Lugaresi, G. Coccagna, M. Mantovani, and R. Lebrun. Some periodic phenomena arising during drowsiness and sleep in man. *Electroencephalography and Clinical Neurophysiology*, 32(6):701–705, 1972. ISSN 00134694. doi: 10.1016/0013-4694(72)90106-X.

[32] Yuan Zou, Viswam Nathan, and Roozbeh Jafari. Automatic identification of artifact-Related independent components for artifact removal in EEG recordings. *IEEE Journal of Biomedical and Health Informatics*, 20(1):73–81, 2016. ISSN 21682194. doi: 10.1109/JBHI.2014.2370646.

[33] Christian W. Rempis and Frank Pasemann. Search space restriction of neuroevolution through constrained modularization of neural networks. *Proceedings of the 6th International Workshop on Artificial Neural Networks and Intelligent Information Processing - Workshop, ANNIIP 2010, in Conjunction with ICINCO 2010*, (January):13–22, 2010. doi: 10.5220/0003026200130022.

[34] J. G. Gehricke, Frithjof Kruggel, Tanyaporn Thampipop, Sharina Dyan Alejo, Erik Tatos, James Fallon, and L. Tugan Muftuler. The brain anatomy of attention-deficit/hyperactivity disorder in young adults - A magnetic resonance imaging study. *PLoS ONE*, 12(4):1–21, 2017. ISSN 19326203. doi: 10.1371/journal.pone.0175433.

[35] Bonnie J. Nagel, Deepti Bathula, Megan Herting, Colleen Schmitt, Christopher D. Kroenke, Damien Fair, and Joel T. Nigg. Altered white matter microstructure in children with attention-deficit/ hyperactivity disorder. *Journal of the American Academy of Child and Adolescent Psychiatry*, 50(3):283–292, 2011. ISSN 08908567. doi: 10.1016/j.jaac.2010.12.003.

[36] Deborah R. Simkin, Robert W. Thatcher, and Joel Lubar. Quantitative EEG and Neurofeedback in Children and Adolescents. Anxiety Disorders, Depressive Disorders, Comorbid Addiction and Attention-deficit/Hyperactivity Disorder, and Brain Injury. *Child and Adolescent Psychiatric Clinics of North America*, 23(3):427–464, 2014. ISSN 15580490. doi: 10.1016/j.chc.2014.03.001. URL `http://dx.doi.org/10.1016/j.chc.2014.03.001`.

[37] Jayant N. Acharya, Abeer Hani, Janna Cheek, Partha Thirumala, and Tammy N. Tsuchida. American Clinical Neurophysiology Society Guideline 2: Guidelines for Standard Electrode Position Nomenclature. *Jour-*

nal of Clinical Neurophysiology, 33(4):308–311, 2016. ISSN 15371603. doi: 10.1097/WNP.0000000000000316.

[38] Valer Jurcak, Daisuke Tsuzuki, and Ippeita Dan. 10/20, 10/10, and 10/5 systems revisited: Their validity as relative head-surface-based positioning systems. *NeuroImage*, 34(4):1600–1611, 2007. ISSN 10538119. doi: 10.1016/j.neuroimage.2006.09.024.

[39] Gregor Strobbe. *Geavanceerde voorwaartse modellen voor EEG-bronanalyse Advanced Forward Models for EEG Source Imaging.* 2014. ISBN 9789085787853.

[40] Sandra K. Loo and Russell A. Barkley. Clinical utility of EEG in attention deficit hyperactivity disorder. *Applied Neuropsychology*, 12(2):64–76, 2005. ISSN 09084282. doi: 10.1207/s15324826an1202_2.

[41] Jennifer J. Newson and Tara C. Thiagarajan. EEG Frequency Bands in Psychiatric Disorders: A Review of Resting State Studies. *Frontiers in Human Neuroscience*, 12(January):1–24, 2019. ISSN 16625161. doi: 10.3389/fnhum.2018.00521.

[42] Uri Alyagon, Hamutal Shahar, Aviad Hadar, Noam Barnea-Ygael, Avi Lazarovits, Hadar Shalev, and Abraham Zangen. Alleviation of ADHD symptoms by non-invasive right prefrontal stimulation is correlated with EEG activity. *NeuroImage: Clinical*, 26(July 2019):102206, 2019. ISSN 22131582. doi: 10.1016/j.nicl.2020.102206. URL `https://doi.org/10.1016/j.nicl.2020.102206`.

[43] Rodney J. Croft and Robert J. Barry. EOG correction: A new perspective. *Electroencephalography and Clinical Neurophysiology*, 107(6):387–394, 1998. ISSN 00134694. doi: 10.1016/S0013-4694(98)00086-8.

[44] Patrick Berg Michale Scherg Otavio G. Lins, Terence W. Picton. Ocular Artifacts in EEG and Event-Related Potentials I: Scalp Topography.pdf, 1993.

[45] Ke Zeng, Dan Chen, Gaoxiang Ouyang, Lizhe Wang, Xianzeng Liu, and Xiaoli Li. An EEMD-ICA Approach to Enhancing Artifact Rejection for Noisy Multivariate Neural Data. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 24(6):630–638, 2016. ISSN 15344320. doi: 10.1109/TNSRE.2015.2496334.

[46] Manousos A. Klados, Christos Papadelis, Christoph Braun, and Panagiotis D. Bamidis. REG-ICA: A hybrid methodology combining Blind Source Separation and regression techniques for the rejection of ocular artifacts. *Biomedical Signal Processing and Control*, 6(3):291–300, 2011. ISSN 17468094. doi: 10.1016/j.bspc.2011.02.001. URL `http://dx.doi.org/10.1016/j.bspc.2011.02.001`.

[47] Gabriele Gratton. Dealing with artifacts: The EOG contamination of the event-related brain potential. *Behavior Research Methods, Instruments, and Computers*, 30(1):44–53, 1998. ISSN 07433808. doi: 10.3758/BF03209415.

[48] M. Sanjeeva Reddy, B. Narasimha, E. Suresh, and K. Subba Rao. Analysis of EOG signals using wavelet transform for detecting eye blinks. *2010 International Conference on Wireless Communications and Signal Processing, WCSP 2010*, pages 6–9, 2010. doi: 10.1109/WCSP.2010.5633797.

[49] S Venkataramanan, P Prabhat, S R Choudhury, H B Nemade, and J S Sahambi. Biomedical instrumentation based on electrooculogram (EOG) signal processing and application to a hospital alarm system. In *Proceedings of 2005 International Conference on Intelligent Sensing and Information Processing, 2005.*, pages 535–540, 2005. doi: 10.1109/ICISIP.2005.1529512.

[50] I I Goncharova, D J Mcfarland, T M Vaughan, and J R Wolpaw. EMG contamination of EEG : spectral and topographical characteristics. 114:1580–1593, 2003. doi: 10.1016/S1388-2457(03)00093-2.

[51] A Qayoom and W Abdul. Artifact Processing of Epileptic EEG Signals: An Overview of Different Types of Artifacts. *2013 International Conference on Advanced Computer Science Applications and Technologies*, pages 358–361, 2013.

[52] S Lee and M S Buchsbaum. Topographic mapping of EEG artifacts. *Clinical EEG (electroencephalography)*, 18(2):61–67, apr 1987. ISSN 0009-9155 (Print).

[53] P K Sadasivan and D Narayana Dutt. Use of finite wordlength FIR digital filter structures with improved magnitude and phase characteristics for reduction of muscle noise in EEG signals. *Medical and Biological Engineering and Computing*, 33(3):306–312, 1995. ISSN 1741-0444. doi: 10.1007/BF02510504. URL `https://doi.org/10.1007/BF02510504`.

[54] T. Uhlig, A. Merkenschlager, R. Brandmaier, and J. Egger. Topographic mapping of brain electrical activity in children with food- induced attention deficit hyperkinetic disorder. *European Journal of Pediatrics*, 156(7):557–561, 1997. ISSN 03406199. doi: 10.1007/s004310050662.

[55] Chaolin Teng, Yanyan Zhang, and Gang Wang. The removal of EMG artifact from EEG signals by the multivariate empirical mode decomposition. *2014 IEEE International Conference on Signal Processing, Communications and Computing, ICSPCC 2014*, pages 873–876, 2014. doi: 10.1109/ICSPCC.2014.6986322.

[56] Juan Andres Mucarquer, Pavel Prado, Maria Jose Escobar, Wael El-Deredy, and Matias Zanartu. Improving EEG Muscle Artifact Removal with an EMG Array. *IEEE Transactions on Instrumentation and Measurement*, 69(3):815–824, 2020. ISSN 15579662. doi: 10.1109/TIM.2019.2906967.

[57] Qingze Liu, Aiping Liu, Xu Zhang, Xiang Chen, Ruobing Qian, and Xun Chen. Removal of EMG Artifacts from Multichannel EEG Signals Using Combined Singular Spectrum Analysis and Canonical Correlation Analysis. *Journal of Healthcare Engineering*, 2019, 2019. ISSN 20402309. doi: 10.1155/2019/4159676.

[58] M. Chavez, F. Grosselin, A. Bussalb, F. De Vico Fallani, and X. Navarro-Sune. Surrogate-based artifact removal from single-channel EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(3):540–550, 2018. ISSN 15344320. doi: 10.1109/TNSRE.2018.2794184.

[59] Motoki Sakai and Daming Wei. Detection of Electrocardiogram Mixed in Electroencephalogram by Stationarization. 9(2):61–62, 2007.

[60] Anthony D Jose and D Collison. The normal range and determinants of the intrinsic heart rate in man1. *Cardiovascular Research*, 4(2):160–167, 1970. ISSN 0008-6363. doi: 10.1093/cvr/4.2.160. URL `https://doi.org/10.1093/cvr/4.2.160`.

[61] Chinmayee Dora and Pradyut Kumar Biswal. Efficient detection and correction of variable strength ECG artifact from single channel EEG. *Biomedical Signal Processing and Control*, 50:168–177, 2019. ISSN 17468108. doi: 10.1016/j.bspc.2019.01.023. URL `https://doi.org/10.1016/j.bspc.2019.01.023`.

[62] Maarten De Vos, Katharina Gandras, and Stefan Debener. Towards a truly mobile auditory brain-computer interface: exploring the P300 to take away. *International journal of psychophysiology : official journal of the International Organization of Psychophysiology*, 91(1):46–53, jan 2014. ISSN 1872-7697 (Electronic). doi: 10.1016/j.ijpsycho.2013.08.010.

[63] P Berg and M Scherg. Dipole modelling of eye activity and its application to the removal of eye artefacts from the EEG and MEG. *Clinical physics and physiological measurement : an official journal of the Hospital Physicists' Association, Deutsche Gesellschaft fur Medizinische Physik and the European Federation of Organisations for Medical Physics*, 12 Suppl A:49–54, 1991. ISSN 0143-0815 (Print). doi: 10.1088/0143-0815/12/a/010.

[64] Hong Zeng, Aiguo Song, Ruqiang Yan, and Hongyun Qin. EOG artifact correction from EEG recording using stationary subspace analysis and empirical mode decomposition. *Sensors (Switzerland)*, 13(11):14839–14859, 2013. ISSN 14248220. doi: 10.3390/s131114839.

[65] Gang Wang, Chaolin Teng, Kuo Li, Zhonglin Zhang, and Xiangguo Yan. The Removal of EOG Artifacts from EEG Signals Using Independent Component Analysis and Multivariate Empirical Mode Decomposition. *IEEE Journal of Biomedical and Health Informatics*, 20(5):1301–1308, 2016. ISSN 21682194. doi: 10.1109/JBHI.2015.2450196.

[66] S Romero, M A Mañanas, and M J Barbanoj. Ocular reduction in EEG signals based on adaptive filtering, regression and blind source separation. *Annals of biomedical engineering*, 37(1):176–191, jan 2009. ISSN 1573-9686 (Electronic). doi: 10.1007/s10439-008-9589-6.

[67] Christiaan Burger and David Jacobus Van Den Heever. Removal of EOG artefacts by combining wavelet neural network and independent component

analysis. *Biomedical Signal Processing and Control*, 15:67–79, 2015. ISSN 17468108. doi: 10.1016/j.bspc.2014.09.009. URL `http://dx.doi.org/10.1016/j.bspc.2014.09.009`.

[68] Jing Hu, Chun sheng Wang, Min Wu, Yu xiao Du, Yong He, and Jinhua She. Removal of EOG and EMG artifacts from EEG using combination of functional link neural network and adaptive neural fuzzy inference system. *Neurocomputing*, 151(P1):278–287, 2015. ISSN 18728286. doi: 10.1016/j.neucom.2014.09.040. URL `http://dx.doi.org/10.1016/j.neucom.2014.09.040`.

[69] Jesus Minguillon, M. Angel Lopez-Gordo, and Francisco Pelayo. Trends in EEG-BCI for daily-life: Requirements for artifact removal. *Biomedical Signal Processing and Control*, 31:407–418, 2017. ISSN 17468108. doi: 10.1016/j.bspc.2016.09.005. URL `http://dx.doi.org/10.1016/j.bspc.2016.09.005`.

[70] Sim Kuan Goh, Hussein A. Abbass, Kay Chen Tan, Abdullah Al-Mamun, Chuanchu Wang, and Cuntai Guan. Automatic EEG Artifact Removal Techniques by Detecting Influential Independent Components. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 1(4):270–279, 2017. ISSN 2471285X. doi: 10.1109/TETCI.2017.2690913.

[71] Chong Yeh Sai, Norrima Mokhtar, Hamzah Arof, Paul Cumming, and Masahiro Iwahashi. Automated classification and removal of EEG artifacts with SVM and wavelet-ICA. *IEEE Journal of Biomedical and Health Informatics*, 22(3):664–670, 2018. ISSN 21682194. doi: 10.1109/JBHI.2017.2723420.

[72] Junfeng Gao, Chongxun Zheng, and Pei Wang. Online removal of muscle artifact from electroencephalogram signals based on canonical correlation analysis. *Clinical EEG and Neuroscience*, 41(1):53–59, 2010. ISSN 15500594. doi: 10.1177/155005941004100111.

[73] Roberto Guarnieri, Marco Marino, Federico Barban, Marco Ganzetti, and Dante Mantini. Online EEG artifact removal for BCI applications by adaptive spatial filtering. *Journal of Neural Engineering*, 15(5), 2018. ISSN 17412552. doi: 10.1088/1741-2552/aacfdf.

[74] D. Mantini, M. G. Perrucci, S. Cugini, A. Ferretti, G. L. Romani, and C. Del Gratta. Complete artifact removal for EEG recorded during continuous fMRI using independent component analysis. *NeuroImage*, 34(2):598–607, 2007. ISSN 10538119. doi: 10.1016/j.neuroimage.2006.09.037.

[75] Arnaud Delorme, Terrence Sejnowski, and Scott Makeig. Enhanced detection of artifacts in EEG data using higher-order statistics and independent component analysis. 34:1443–1449, 2007. doi: 10.1016/j.neuroimage.2006.11.004.

[76] Doha Safieddine, Amar Kachenoura, Laurent Albera, Gwénaël Birot, Ahmad Karfoul, Anca Pasnicu, Arnaud Biraben, Fabrice Wendling, Lotfi Senhadji, and Isabelle Merlet. Removal of muscle artifact from EEG data: Comparison between stochastic (ICA and CCA) and deterministic (EMD and wavelet-based)

approaches. *Eurasip Journal on Advances in Signal Processing*, 2012(1), 2012. ISSN 16876172. doi: 10.1186/1687-6180-2012-127.

[77] Manousos A. Klados and Panagiotis D. Bamidis. A semi-simulated EEG/EOG dataset for the comparison of EOG artifact rejection techniques. *Data in Brief*, 8:1004–1006, 2016. ISSN 23523409. doi: 10.1016/j.dib.2016.06.032. URL `http://dx.doi.org/10.1016/j.dib.2016.06.032`.

[78] Christopher J James and Christian W Hesse. Independent component analysis for biomedical signals. *Physiological measurement*, 26(1):R15–39, feb 2005. ISSN 0967-3334 (Print). doi: 10.1088/0967-3334/26/1/r02.

[79] Carlos G. Puntonet and Elmar W. Lang. Blind source separation and independent component analysis. *Neurocomputing*, 69(13-15):1413, 2006. ISSN 09252312. doi: 10.1016/j.neucom.2005.12.018.

[80] Gundars Korats, Steven Le Cam, Radu Ranta, and Mohamed Hamid. Applying ICA in EEG: Choice of the Window Length and of the Decorrelation Method. *Communications in Computer and Information Science*, 357 CCIS:269–286, 2013. ISSN 18650929. doi: 10.1007/978-3-642-38256-7_18.

[81] Jorge Iriarte, Elena Urrestarazu, Miguel Valencia, Manuel Alegre, Armando Malanda, César Viteri, and Julio Artieda. Independent component analysis as a tool to eliminate artifacts in EEG: A quantitative study. *Journal of Clinical Neurophysiology*, 20(4):249–257, 2003. ISSN 07360258. doi: 10.1097/00004691-200307000-00004.

[82] Laura Frølich and Irene Dowding. Removal of muscular artifacts in EEG signals: a comparison of linear decomposition methods. *Brain Informatics*, 5(1):13–22, 2018. ISSN 21984026. doi: 10.1007/s40708-017-0074-6. URL `https://doi.org/10.1007/s40708-017-0074-6`.

[83] Tzyy Ping Jung, Scott Makeig, Marissa Westerfield, Jeanne Townsend, Eric Courchesne, and Terrence J. Sejnowski. Removal of eye activity artifacts from visual event-related potentials in normal and clinical subjects. *Clinical Neurophysiology*, 111(10):1745–1758, 2000. ISSN 13882457. doi: 10.1016/S1388-2457(00)00386-2.

[84] Sergio Romero, Miguel A. Mañanas, and Manel J. Barbanoj. A comparative study of automatic techniques for ocular artifact reduction in spontaneous EEG signals based on clinical target variables: A simulation case. *Computers in Biology and Medicine*, 38(3):348–360, 2008. ISSN 00104825. doi: 10.1016/j.compbiomed.2007.12.001.

[85] Akaysha C. Tang, Matthew T. Sutherland, and Christopher J. McKinney. Validation of SOBI components from high-density EEG. *NeuroImage*, 25(2):539–553, 2005. ISSN 10538119. doi: 10.1016/j.neuroimage.2004.11.027.

[86] Ian Daly, Martin Billinger, Reinhold Scherer, and Gernot Müller-Putz. On the automated removal of artifacts related to head movement from the EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 21(3): 427–434, 2013. ISSN 15344320. doi: 10.1109/TNSRE.2013.2254724.

[87] B S Raghavendra and D Narayana Dutt. Wavelet Enhanced CCA for Minimization of Ocular and Muscle Artifacts in EEG. *International Journal of Medical, Health, Biomedical, Bioengineering and Pharmaceutical Engineering*, 5(9):1027–1032, 2011.

[88] Kevin T. Sweeney, Tomás E. Ward, and Seán F. McLoone. Artifact removal in physiological signals-practices and possibilities. *IEEE Transactions on Information Technology in Biomedicine*, 16(3):488–500, 2012. ISSN 10897771. doi: 10.1109/TITB.2012.2188536.

[89] Fadia Noori Hummadi Al-Nuaimy. A new eliminating EOG artifacts technique using combined decomposition methods with CCA and H.P.F. techniques. *Telkomnika (Telecommunication Computing Electronics and Control)*, 18(5): 2580–2586, 2020. ISSN 23029293. doi: 10.12928/TELKOMNIKA.V18I5.14143.

[90] Christiaan Burger and David Jacobus Van Den Heever. Removal of EOG artefacts by combining wavelet neural network and independent component analysis. *Biomedical Signal Processing and Control*, 15:67–79, 2015. ISSN 17468108. doi: 10.1016/j.bspc.2014.09.009. URL `http://dx.doi.org/10.1016/j.bspc.2014.09.009`.

[91] Ian Daly, Nicoletta Nicolaou, Slawomir Jaroslaw Nasuto, and Kevin Warwick. Automated artifact removal from the electroencephalogram: A comparative study. *Clinical EEG and Neuroscience*, 44(4):291–306, 2013. ISSN 15500594. doi: 10.1177/1550059413476485.

[92] Mehdi Bagheri Hamaneh, Numthip Chitravas, Kitti Kaiboriboon, Samden D. Lhatoo, and Kenneth A. Loparo. Automated removal of EKG artifact from EEG data using independent component analysis and continuous wavelet transformation. *IEEE Transactions on Biomedical Engineering*, 61(6):1634–1641, 2014. ISSN 15582531. doi: 10.1109/TBME.2013.2295173.

[93] Amira Echtioui, Wassim Zouch, Mohamed Ghorbel, Mohamed Ben Slima, Ahmed Ben Hamida, and Chokri Mhiri. Automated EEG Artifact Detection Using Independent Component Analysis. *2020 International Conference on Advanced Technologies for Signal and Image Processing, ATSIP 2020*, 2020. doi: 10.1109/ATSIP49331.2020.9231574.

[94] Irene Winkler and Stefan Haufe and Michael Tangermann. Automatic Classification of Artifactual ICAComponents for Artifact Removal in EEG Signals. *Behavioral and Brain Functions*, 32(6):701–705, 2011. ISSN 00134694. doi: 10.1016/0013-4694(72)90106-X.

[95] Laura Frølich, Tobias S. Andersen, and Morten Mørup. Classification of independent components of EEG into multiple artifact classes. *Psychophysiology*, 52(1):32–45, 2015. ISSN 14698986. doi: 10.1111/psyp.12290.

[96] Won Du Chang, Jeong Hwan Lim, and Chang Hwan Im. An unsupervised eye blink artifact detection method for real-time electroencephalogram processing. *Physiological Measurement*, 37(3):401–417, 2016. ISSN 13616579. doi: 10.1088/0967-3334/37/3/401.

[97] Andrea Mognon, Jorge Jovicich, Lorenzo Bruzzone, and Marco Buiatti. ADJUST: An automatic EEG artifact detector based on the joint use of spatial and temporal features. *Psychophysiology*, 48(2):229–240, feb 2011. ISSN 1540-5958 (Electronic). doi: 10.1111/j.1469-8986.2010.01061.x.

[98] G Geetha and S N Geethalakshmi. Artifact Removal from EEG using Spatially Constrained Independent Component Analysis and Wavelet Denoising with Otsu's Thresholding Technique. *Procedia Engineering*, 30:1064–1071, 2012. ISSN 1877-7058. doi: https://doi.org/10.1016/j.proeng.2012.01.964. URL `https://www.sciencedirect.com/science/article/pii/S1877705812009745`.

[99] Nobuyuki Otsu. A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979. doi: 10.1109/TSMC.1979.4310076.

[100] Lukas Breuer, Jürgen Dammers, Timothy P L Roberts, and N Jon Shah. Ocular and cardiac artifact rejection for real-time analysis in MEG. *Journal of neuroscience methods*, 233:105–114, aug 2014. ISSN 1872-678X (Electronic). doi: 10.1016/j.jneumeth.2014.06.016.

[101] Rodney J Croft, Jody S Chandler, Robert J Barry, Nicholas R Cooper, and Adam R Clarke. EOG correction: a comparison of four methods. *Psychophysiology*, 42(1):16–24, jan 2005. ISSN 0048-5772 (Print). doi: 10.1111/j.1468-8986.2005.00264.x.

[102] Brenton W McMenamin, Alexander J Shackman, Jeffrey S Maxwell, David R W Bachhuber, Adam M Koppenhaver, Lawrence L Greischar, and Richard J Davidson. Validation of ICA-based myogenic artifact correction for scalp and source-localized EEG. *NeuroImage*, 49(3):2416–2432, feb 2010. ISSN 1095-9572 (Electronic). doi: 10.1016/j.neuroimage.2009.10.010.

[103] Brenton W. McMenamin, Alexander J. Shackmanb, Lawrence L. Greischarc and Richard J. Davidson. Electromyogenic Artifacts and Electroencephalographic Inferences Revisited Brenton. *NeuroImage*, 23(1):1–7, 2008. ISSN 15378276. doi: 10.1016/j.neuroimage.2010.07.057.Electromyogenic.

[104] Wim De Clercq, Anneleen Vergult, Bart Vanrumste, Wim Van Paesschen, and Sabine Van Huffel. Canonical correlation analysis applied to remove muscle artifacts from the electroencephalogram. *IEEE Transactions on Biomedical Engineering*, 53(12):2583–2587, 2006. doi: 10.1109/TBME.2006.879459.

[105] Joep J M Kierkels, Geert J M van Boxtel, and Leo L M Vogten. A model-based objective evaluation of eye movement correction in EEG recordings. *IEEE transactions on bio-medical engineering*, 53(2):246–253, feb 2006. ISSN 0018-9294 (Print). doi: 10.1109/TBME.2005.862533.

[106] Garrick L. Wallstrom, Robert E. Kass, Anita Miller, Jeffrey F. Cohn, and Nathan A. Fox. Automatic correction of ocular artifacts in the EEG: A comparison of regression-based and component-based methods. *International Journal of Psychophysiology*, 53(2):105–119, 2004. ISSN 01678760. doi: 10.1016/j.ijpsycho.2004.03.007.

[107] Stefan Haufe and Arne Ewald. A Simulation Framework for Benchmarking EEG-Based Brain Connectivity Estimation Methodologies. *Brain Topography*, 32(4):625–642, 2019. ISSN 15736792. doi: 10.1007/s10548-016-0498-y.

[108] John J. Ermer, John C. Mosher, Sylvain Baillet, and Richard M. Leahy. Rapidly recomputable EEG forward models for realistic head shapes. *Physics in Medicine and Biology*, 46(4):1265–1281, 2001. ISSN 00319155. doi: 10.1088/0031-9155/46/4/324.

[109] Zeynep Akalin Acar and Scott Makeig. Effects of forward model errors on EEG source localization. *Brain Topography*, 26(3):378–396, 2013. ISSN 08960267. doi: 10.1007/s10548-012-0274-6.

[110] Nadia Mammone, Fabio La Foresta, and Francesco Carlo Morabito. Automatic artifact rejection from multichannel scalp EEG by wavelet ICA. *IEEE Sensors Journal*, 12(3):533–542, 2012. ISSN 1530437X. doi: 10.1109/JSEN.2011.2115236.

[111] Manousos A. Klados and Panagiotis D. Bamidis. Research domain, Software and database, Mendeley Data: A semi-simulated EEG/EOG dataset for the comparison of EOG artifact rejection techniques, 2019. URL `https://data.mendeley.com/datasets/wb6yvr725d/4`.

[112] Heba Khamis, Robert Weiss, Yang Xie, Chan-Wei Chang, Nigel H. Lovell, and Stephen J. Redmond. Research domain, Software and database, TELE ECG Database: 250 telehealth ECG records (collected using dry metal electrodes) with annotated QRS and artifact masks, and MATLAB code for the UNSW artifact detection and UNSW QRS detection algorithms, 2016. URL `https://doi.org/10.7910/DVN/QTG0EP`.

[113] Louise van der Westhuyzen. Personal Interview, Jun 2020, Bergvliet. URL `https://www.braindynamics.co.za/`.

[114] Robert J. Barry and Frances M. De Blasio. EEG differences between eyes-closed and eyes-open resting remain in healthy ageing. *Biological Psychology*, 129(August):293–304, 2017. ISSN 18736246. doi: 10.1016/j.biopsycho.2017.09.010. URL `http://dx.doi.org/10.1016/j.biopsycho.2017.09.010`.

[115] M. Matsuura, K. Yamamoto, H. Fukuzawa, Y. Okubo, H. Uesugi, M. Moriiwa, T. Kojima, and Y. Shimazono. Age development and sex differences of various EEG elements in healthy children and adults - Quantification by a computerized wave form recognition method. *Electroencephalography and Clinical Neurophysiology*, 60(5):394–406, 1985. ISSN 00134694. doi: 10.1016/0013-4694(85)91013-2.

[116] Lars Michels, Muthuraman Muthuraman, Rafael Lüchinger, Ernst Martin, Abdul Rauf Anwar, Jan Raethjen, Daniel Brandeis, and Michael Siniatchkin. Developmental changes of functional and directed resting-state connectivities associated with neuronal oscillations in EEG. *NeuroImage*, 81:231–242, 2013. ISSN 10538119. doi: 10.1016/j.neuroimage.2013.04.030. URL `http://dx.doi.org/10.1016/j.neuroimage.2013.04.030`.

[117] David Balin Chorlian, Madhavi Rangaswamy, and Bernice Porjesz. EEG coherence: Topography and frequency structure. *Experimental Brain Research*, 198(1):59–83, 2009. ISSN 00144819. doi: 10.1007/s00221-009-1936-9.

[118] Adele E. Cave and Robert J. Barry. Sex differences in resting EEG in healthy young adults. *International Journal of Psychophysiology*, 161(January):35–43, 2021. ISSN 18727697. doi: 10.1016/j.ijpsycho.2021.01.008. URL `https://doi.org/10.1016/j.ijpsycho.2021.01.008`.

[119] Yukinori Suzuki. Self-Organizing QRS-Wave Recognition in ECG Using Neural Networks. *IEEE Transactions on Neural Networks*, 6(6):1469–1477, 1995. ISSN 19410093. doi: 10.1109/72.471381.

[120] Peter Anderer, Stephen Roberts, Alois Schlögl, Georg Gruber, Gerhard Klösch, Werner Herrmann, Peter Rappelsberger, Oliver Filz, Manel J. Barbanoj, Georg Dorffner, and Bernd Saletu. Artifact processing in computerized analysis of sleep EEG - A review. *Neuropsychobiology*, 40(3):150–157, 1999. ISSN 0302282X. doi: 10.1159/000026613.

[121] Chin Teng Lin, Chih Sheng Huang, Wen Yu Yang, Avinash Kumar Singh, Chun Hsiang Chuang, and Yu Kai Wang. Real-Time EEG Signal Enhancement Using Canonical Correlation Analysis and Gaussian Mixture Clustering. *Journal of Healthcare Engineering*, 2018, 2018. ISSN 20402309. doi: 10.1155/2018/5081258.

[122] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A. Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, and Matti S. Hämäläinen. MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, 7(267):1–13, 2013. doi: 10.3389/fnins.2013.00267. URL `https://mne.tools/stable/index.html`.

[123] Charles R. Harris, K. Jarrod Millman, Stéfan J van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del

Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585:357–362, 2020. doi: 10.1038/s41586-020-2649-2. URL `https://pypi.org/project/numpy/1.20.1/`.

[124] John D Hunter. Matplotlib: A 2d graphics environment. *Computing in science & engineering*, 9(3):90–95, 2007. URL `https://zenodo.org/record/4268928#.YSyr5o4zaHs`.

[125] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. URL `https://scikit-learn.org/stable/`.

[126] Yang Bai, Xiaohong Wan, Ke Zeng, Yinmei Ni, Lirong Qiu, and Xiaoli Li. Reduction hybrid artifacts of EMG-EOG in electroencephalography evoked by prefrontal transcranial magnetic stimulation. *Journal of Neural Engineering*, 13(6), 2016. ISSN 17412552. doi: 10.1088/1741-2560/13/6/066016.

[127] P. K. Sadasivan and D. Narayana Dutt. Development of Newton-type adaptive algorithm for minimization of EOG artefacts from noisy EEG signals. *Signal Processing*, 62(2):173–186, 1997. ISSN 01651684. doi: 10.1016/s0165-1684(97)00123-0.

[128] Manuel Merino, María I. Gómez, and Alberto J. Molina. Envelope filter sequence to delete blinks and overshoots. *BioMedical Engineering Online*, 14(1):1–24, 2015. ISSN 1475925X. doi: 10.1186/s12938-015-0046-0.

[129] Rajesh Naga, S. Chandralingam, T. Anjaneyulu, and K. Satyanarayana. Denoising EOG signal using stationary wavelet transform. *Measurement Science Review*, 12(2):46–51, 2012. ISSN 13358871. doi: 10.2478/v10048-012-0010-0.

[130] S. Puthusserypady and T. Ratnarajah. Robust adaptive techniques for minimization of EOG artefacts from EEG signals. *Signal Processing*, 86(9):2351–2363, 2006. ISSN 01651684. doi: 10.1016/j.sigpro.2005.10.018.

[131] Juan Cheng, Luchang Li, Chang Li, Yu Liu, Aiping Liu, Ruobing Qian, and Xun Chen. Remove diverse artifacts simultaneously from a single-channel EEG based on ssa and ica: A semi-simulated study. *IEEE Access*, 7:60276–60289, 2019. ISSN 21693536. doi: 10.1109/ACCESS.2019.2915564.

[132] Kevin Paulson and Othman Alfahad. Identification of multi-channel simulated auditory event-related potentials using a combination of principal component analysis and Kalman filtering. *ACM International Conference Proceeding Series*, pages 18–22, 2018. doi: 10.1145/3288200.3288211.

[133] Xun Chen, Aiping Liu, Hu Peng, and Rabab K. Ward. A preliminary study of muscular artifact cancellation in single-channel EEG. *Sensors (Switzerland)*, 14(10):18370–18389, 2014. ISSN 14248220. doi: 10.3390/s141018370.

[134] Mahipal Singh Choudhry, Rajiv Kapoor, Abhishek, Anuj Gupta, and Bhaskar Bharat. A survey on different discrete wavelet transforms and thresholding techniques for EEG denoising. *Proceeding - IEEE International Conference on Computing, Communication and Automation, ICCCA 2016*, pages 1048–1053, 2017. doi: 10.1109/CCAA.2016.7813897.

[135] Yongcheng Li, Po T. Wang, Mukta P. Vaidya, Robert D. Flint, Charles Y. Liu, Marc W. Slutzky, and An H. Do. Refinement of High-Gamma EEG Features From TBI Patients With Hemicraniectomy Using an ICA Informed by Simulated Myoelectric Artifacts. *Frontiers in Neuroscience*, 14(November): 1–10, 2020. ISSN 1662453X. doi: 10.3389/fnins.2020.599010.

[136] Wim De Clercq, Anneleen Vergult, Bart Vanrumste, Wim Van Paesschen, and Sabine Van Huffel. Canonical Correlation Analysis Applied to Remove Muscle Artifacts From the Electroencephalogram. 44(5):2583–2587, 2006.

[137] Carlos Magno, Medeiros Queiroz, Steffen Walter, Luciano Brink Peres, Maire David Luiz, Samila Carolina Costa, Adriano Alves Pereira, Oliveira Andrade, and Marcus Fraga Vieira. Single channel approach for filtering EEG signals strongly contaminated with facial EMG. pages 1–28.

[138] Luay Yassin Taha and Esam Abdel-Raheem. EEG signal Extraction Utilizing Null Space Approach. *2019 IEEE 19th International Symposium on Signal Processing and Information Technology, ISSPIT 2019*, (1), 2019. doi: 10.1109/ISSPIT47144.2019.9001818.

[139] Chinmayee Dora and Pradyut Kumar Biswal. Computer Methods and Programs in Biomedicine Correlation-based ECG Artifact Correction from Single Channel EEG using Modified Variational Mode Decomposition. *Computer Methods and Programs in Biomedicine*, 183:105092, 2020. ISSN 0169-2607. doi: 10.1016/j.cmpb.2019.105092. URL `https://doi.org/10.1016/j.cmpb.2019.105092`.

[140] Sung Pil Cho, Mi Hye Song, Young Cheol Park, Ho Seon Choi, and Kyoung Joung Lee. Adaptive noise canceling of electrocardiogram artifacts in single channel electroencephalogram. *Annual International Conference of the IEEE Engineering in Medicine and Biology - Proceedings*, pages 3278–3281, 2007. ISSN 05891019. doi: 10.1109/IEMBS.2007.4353029.

[141] S. Suja Priyadharsini and S. Edward Rajan. An Ecient method for the removal of ECG artifact from measured EEG signal using PSO algorithm. *International Journal of Advances in Soft Computing and its Applications*, 6(1):1–19, 2014. ISSN 20748523.

[142] X. Navarro, F. Porée, A. Beuchée, and G. Carrault. Denoising preterm EEG by signal decomposition and adaptive filtering: A comparative study. *Medical Engineering and Physics*, 37(3):315–320, 2015. ISSN 18734030. doi: 10.1016/j.medengphy.2015.01.006.

[143] Xavier Navarro, Fabienne Porée, and Guy Carrault. ECG removal in preterm EEG combining empirical mode decomposition and adaptive filtering. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pages 661–664, 2012. ISSN 15206149. doi: 10.1109/ICASSP.2012.6287970.

[144] Zhongjie Hou, Yonggui Dong, and Xu Wu. A Template Addition Method for Eigentriple Rearrangement in Singular Spectrum Analysis for Processing Biopotential Signals with Extremely Lower SNRs. *IEEE Sensors Journal*, 20 (6):3142–3150, 2020. ISSN 15581748. doi: 10.1109/JSEN.2019.2957864.

[145] Gabriella Tamburro, David B Stone, and Silvia Comani. Automatic Removal of Cardiac Interference (ARCI): A New Approach for EEG Data. *Frontiers in Neuroscience*, 13:441, 2019. ISSN 1662-453X. doi: 10.3389/fnins.2019.00441. URL `https://www.frontiersin.org/article/10.3389/fnins.2019.00441`.

[146] R J Croft and R J Barry. Removal of ocular artifact from the EEG: a review. *Neurophysiologie clinique = Clinical neurophysiology*, 30(1):5–19, feb 2000. ISSN 0987-7053 (Print). doi: 10.1016/S0987-7053(00)00055-1.

[147] M Unser and A Aldroubi. A review of wavelets in biomedical applications. *Proceedings of the IEEE*, 84(4):626–638, 1996. doi: 10.1109/5.488704.

[148] Markus Waser and Heinrich Garn. Removing cardiac interference from the electroencephalogram using a modified Pan-Tompkins algorithm and linear regression. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, 2013:2028–2031, 2013. ISSN 2694-0604 (Electronic). doi: 10.1109/EMBC.2013.6609929.

[149] Ricardo Nuno Vigário. Extraction of ocular artefacts from EEG using independent component analysis. *Electroencephalography and Clinical Neurophysiology*, 103(3):395–404, 1997. ISSN 0013-4694. doi: https://doi.org/10.1016/S0013-4694(97)00042-8. URL `https://www.sciencedirect.com/science/article/pii/S0013469497000428`.

[150] Meltem Izzetoglu, Ajit Devaraj, Scott Bunce and Banu Onaral. Motion Artifact Cancellation in NIR Spectroscopy Using Wiener Filtering. *IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING*, 52(5):934–938, 2005.

[151] Greg Welch and Gary Bishop. An Introduction to the Kalman Filter. *In Practice*, 7(1):1–16, 2006. ISSN 10069313. doi: 10.1.1.117.6808. URL `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.79.6578&rep=rep1&type=pdf`.

[152] Kevin T. Sweeney, Hasan Ayaz, Tomás E. Ward, Meltem Izzetoglu, Seán F. McLoone, and Banu Onaral. A methodology for validating artifact removal techniques for physiological signals. *IEEE Transactions on Information Technology in Biomedicine*, 16(5):918–926, 2012. ISSN 10897771. doi: 10.1109/TITB.2012.2207400.

[153] A Hyvärinen and E Oja. Independent component analysis: A tutorial. *Notes for International Joint Conference on Neural Networks (IJCNN'99), Washington DC*, 1:1–30, 1999. ISSN 0035-8711. URL `http://scholar.google.com/scholar?q=related:qM2Iwk0laFQJ:scholar.google.com/&hl=en&num=20&as_sdt=0,5%5Cnfile:///Users/tsmoon/Documents/PapersBak/Articles/1999/Hyv?rinen/NotesforInternationalJointConferenceonNeuralNetworks(IJCNN?99)WashingtonDC1`.

[154] Te Won Lee, Mark Girolami, and Terrence J. Sejnowski. Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources. *Neural Computation*, 11(2):417–441, 1999. ISSN 08997667. doi: 10.1162/089976699300016719.

[155] Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. An introduction to independent component analysis. *Journal of Chemometrics*, 14(3): 123–149, 2000. ISSN 08869383. doi: 10.1002/1099-128X(200005/06)14:3<123::AID-CEM589>3.0.CO;2-1.

[156] Guillermo Sahonero-Alvarez and Humberto Calderon. A comparison of SOBI, FastICA, JADE and infomax algorithms. *IMCIC 2017 - 8th International Multi-Conference on Complexity, Informatics and Cybernetics, Proceedings*, 2017-March(March 2017):17–22, 2017.

[157] Adel Belouchrani, Karim Abed-Meraim, Jean François Cardoso, and Eric Moulines. A blind source separation technique using second-order statistics. *IEEE Transactions on Signal Processing*, 45(2):434–444, 1997. ISSN 1053587X. doi: 10.1109/78.554307.

[158] Xiaowei Zhuang, Zhengshi Yang, and Dietmar Cordes. A technical review of canonical correlation analysis for neuroscience applications. *Human Brain Mapping*, 41(13):3807–3833, 2020. ISSN 10970193. doi: 10.1002/hbm.25090.