Analysis and Removal of Artifacts in Electroencephalographic Recordings using Microstate Analysis and Randomization Statistics

Jamil Chowdhury

A Thesis
in
The Department
of
Electrical and Computer Engineering

Presented in Partial Fulfillment of the Requirements
For the Degree of
Master of Applied Science at
Concordia University
Montréal, Québec, Canada

August 2020

CONCORDIA UNIVERSITY
SCHOOL OF GRADUATE STUDIES

This is to certify that the thesis prepared

By: Jamil Chowdhury

Entitled: Analysis and Removal of Artifacts in Electroencephalographic Recordings using Microstate Analysis and Randomization Statistics

and submitted in partial fulfillment of the requirements for the degree of

Master of Applied Science (Electrical and Computer Engineering)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

_____ Chair
Dr. H. Rivaz

_____ External Examiner
Dr. J. Yan (CIISE)

_____ Internal Examiner
Dr. H. Rivaz

_____ Co-Supervisor
Dr. W.-P. Zhu

_____ Supervisor
Dr. Y. Zeng (CIISE)

Approved by: _____
Dr. Y.R. Shayan, Chair
Department of Electrical and Computer Engineering

_____ 20___ _____
Dr. Mourad Debbabi, Interim Dean,
Gina Cody School of Engineering and
Computer Science

Abstract

Analysis and Removal of Artifacts in Electroencephalographic Recordings using Microstate Analysis And Randomization Statistics

Jamil Chowdhury

Electroencephalography (EEG) is a popular method to detect brain-neuron activities because of its high temporal resolution. However, very often, various types of biological and non-biological signals contaminate EEG recordings. These non-neural signals create EEG-artifacts, which cause unintentional control in the brain-computer interface related applications and difficulty in the analysis and interpretation of EEG-data. While these artifacts corrupt and mask the underlying neural activity in general, the contaminated EEG data due to the contraction and expansion of the scalp-muscles are called electromyogram (EMG) artifacts. In particular, the frontalis and temporalis scalp-muscles seriously affect the EEG-signals ranging from 0-200 Hz frequency band. This thesis studies the most common EMG artifacts originating from these two brain regions. Its aim is to analyze and remove the EMG artifacts using microstate analysis and randomization statistics.

The thesis first presents a brief literature review of the EEG-artifacts, followed by the preprocessing and analysis of the EEG recordings using EEG signal-power analysis. The purpose of this analysis is to detect the EMG contaminated EEG data-segments or epochs due to frontalis and temporalis scalp muscles (EMG-artifacts). The preliminary step in this analysis includes the transformation of the EEG epochs into the frequency domain through discrete-Fourier transform. Then the signal-powers of the EEG epochs are calculated and compared to some threshold values. These threshold values are selected based on the mean signal-power amplitudes of the EEG-epochs of the highly contaminated EEG data channels representing the frontalis and temporalis brain regions.

Electric potentials from the frontalis and temporalis region of the brain project a set of spatial patterns on the scalp surface. These spatial patterns can be clustered into a set of representative maps called microstates. Using microstate analysis, the EMG-contaminated and non-contaminated EEG epochs, obtained from signal-power analysis are clustered into an optimal number of microstates. This number best explains the data variance of both

groups of EEG epochs. The difference between these microstate features can be used to distinguish artifactual and pure EEG epochs. To find the significant-differences, we have calculated the feature-differences of these microstates after a random group-wise shuffling of the EEG epochs many times to generate a distribution of the feature-differences. The research hypothesis of this distribution is that the differences in features have occurred by chance. To reject this hypothesis, we compare the probability of this distribution to the difference in features obtained before group-wise random shuffling of the EEG epochs. This technique is called multivariate randomization statistics. It has a higher statistical power compared to classical statistics to find a statistically significant difference.

In this thesis, we analyze the EEG recordings of four subjects to detect the EMG artifacts by EEG signal-power analysis. We propose a method to remove EMG artifact from EEG recordings in two steps. In the first step, we cluster the EMG contaminated and non-contaminated EEG epochs obtained from signal-power analysis into ten optimal microstates and calculate three temporal features. In the second step, through randomization-statistical analysis, we differentiate between the artifactual and pure EEG epochs and reconstruct the EMG-artifact free EEG data. Finally, we validate the proposed method by comparing it with independent component analysis (ICA), a signal processing technique for separating the additive sub-components of a multivariate signal. We have found that our proposed method gives similar results to that of ICA. Our research findings suggest that a combination of microstate analysis with randomization statistics be an effective-method in the removal of EMG-artifacts.

Acknowledgments

I want to express my profound gratitude to the Almighty God for helping me towards finishing the thesis. I am deeply grateful to my supervisors Dr. Wei-Ping Zhu and Dr. Yong Zeng, for their unconditional support and crucial suggestions and comments. With their steady support and guidance, my graduate study and research was always a smooth journey. It would have been impossible to finish this thesis without their meticulous observations, comments, and inspirations. It is a great fortune for me that I had the chance to work with them in the field of academic research. I was always on track by their constant enthusiasm and care. I learned countless things from them that surely would help me in achieving my future goals.

I would also like to thank all my colleagues in the Design Lab. I was lucky to work with them in such a research-oriented environment. It was an excellent experience working with the past and present group members. I want to render special thanks to Wenjun Jia and Lucas House for their valuable suggestions and directions in the EEG research and experiment. I want to thank all my friends at Concordia.

Last but not least, I want to thank my beloved mother, Noor Jahan Begum, for supporting me from the beginning of my higher education, encouraging, and giving unconditional love in my life. I want to thank my sisters, Farhana Akter, Farzana Akter, and Dilruba Akhter, for always believing in my capabilities. I am also very thankful to my brother-in-law, Saad Quader, for his good wishes and suggestions. I am especially grateful to KH. Arif Shahriar, Mohammad Shafayet Islam, Sakib Shuvo and Mosabbir Khan Shiblu,

who have always encouraged me towards success. I would also like to thank my friends Md Imtiaz Uddin Johan, S N Saif, Amir Pirhadi, Cesar Ciepelli, Abrar Alvi Chowdhury, Md Ashikuzzaman, Kazi Mustafizur Rahman, Keyu Pan, Abrar Hussain, Azfar Adib, and Chisty Bhuyian for their help in different stages of my study.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and motivation

The existence of electrical activity of the brain was first discovered by Richard Canton in 1875. It was the foremost attempt in the electrophysiology of the human brain, giving rise to the concept of electroencephalogram (EEG). This concept depicts brain electrical activity in human beings. As such, the word electroencephalogram was coined after this concept [3]. The first recording of human EEG was done by Dr. Hans Berger a German psychiatrist way back in 1924 [4]. His works made the measurement of EEG from the human scalp possible. EEG is nowadays widely utilized in the field of neuroscience, cognitive science, cognitive psychology, neurolinguistics, and psychophysiological research. Apart from its more traditional use in the clinical assessment of consciousness research, EEG is frequently used for the investigation of different brain conditions like determining the type and location of epileptic activity or for the analysis of sleep disorders as well as other neurological dysfunctions like encephalopathies, neurological infections, dementia etc. [5].

However, obtaining pure EEG signals is very challenging, since it is very difficult to get the ground truth data due to the unavoidable presence of artifacts. Artifacts are non-neural

signals originated from different body parts like eyes, scalp muscles, skin, skull contractions, external environment, and experimental error. No matter whatever the application is it is essential to get as many clean recordings as possible. Unfortunately, this is always hindered by the presence of artifacts. The artifacts are a constant problem in the research field of EEG signal processing as they take various forms, mask the underlying neural activity and distort the signals. Thus, due to the presence of artifacts, the EEG signal processing becomes faulty and incorrect and often results in wrong interpretation of EEG activity.

The two most dominant artifacts that highly corrupt EEG signals are electrooculography (EOG) and electromyography (EMG). By placing the reference eye channels in nearby eye locations, it is possible to remove the EOG artifacts using linear regression technique and signal subtraction. Unfortunately for tackling the EMG or muscle artifacts, dedicated reference channels are not available. As such it is difficult to remove the EMG artifacts that have high amplitude, broad frequency range, variable distribution of topography.

So, removal of the artifact is necessary to unmask the masked neural signal. There are many existing methods to handle the artifacts in EEG signal processing. The rudimentary technique is to remove artifacts by band pass filtering. Also, there are other methods like independent component analysis (ICA) and blind source separation (BSS).

The theoretical basis in the blind source separation method is that the neural signal and artefactual signals are not co-activated simultaneously. Besides, the combination of the BSS and canonical correlation analysis (CCA) is another method for muscle artifact removal [6]. The underlying assumption is that the contamination of EMG in EEG signals is not the same for frequencies ranging from 0 Hz to 200 Hz [7]. Moreover, the frequency dependence varies with active muscle and EEG recordings (EEG signals) from different parts of the brain. As such, the frequency analysis of the additive sub-components separated from the multivariate EEG signals can be an approach to detect and remove the contamination. In signal processing the method to separate the additive sub-components

of a multivariate signal is called independent component analysis (ICA). This method of separation is based on the assumption that the sub-components are non-Gaussian signals and are statistically independent from each other [8].

However, independent component analysis (ICA) requires a large amount of experimental data for the classification of the components into pure EEG and artifactual components. This is a bottleneck to this process. Alternative concepts can be used in case of less amount and short duration of data, particularly in complex cognitive experiments. As such the main method of this research is the combination of concepts like EEG microstate analysis and randomization statistics to analyze and remove muscle artifacts from EEG recordings obtained from complex cognitive experiments.

The main motivation of this thesis is to explore concepts like EEG microstate analysis and randomization statistics to tackle the EMG contaminations due to frontalis and temporalis scalp muscles. In EEG microstate analysis [9] the brain activity can be modeled as a time sequence of non-overlapping microstates with variable duration. These microstates are sub-second quasi-stable configurations of the scalp-potential field maps. These maps quickly change to another sub-second quasi-stable configuration. The configuration of these transient microstates is physiologically significant and carry information on how the brain processes information [10]. This concept of the microstates is very much different from waveform analysis.

On the other hand, randomization statistics [11] is a powerful tool for conducting statistical analysis with high accuracy and reliable results. It needs fewer assumptions than classical statistics. This method allows the construction of custom-tailored tests for the specific research question of interest. This method is computationally heavy since at least 1000 random runs are necessary to obtain reliable results [11]. However, this obstacle is becoming less and less due to the rapid growth of computing power which is affordable with personal computers [12, see chapter 8].

This thesis investigates the combinations of the concepts of randomization statistics and microstate analysis to remove the EMG artifacts by analyzing the signal-power of the highly EMG prone EEG data segments in the frequency band 45-70 Hz [13].

1.2 Objective of the thesis

The principal objective of this thesis is to analyze and remove muscle or EMG contaminations in EEG signal recordings due to the contraction of frontalis and temporalis scalp muscles. The research focus is on the muscle artifacts analysis and removal from the EEG recordings. This study targeted the EMG contaminations of EEG signals for the development of a new method to remove such contamination or artifacts. First, the EMG artifacts (contaminated EEG data-segments) are analyzed using EEG signal-power spectrum in the 45-70 Hz frequency band. Next for the removal process, EEG microstate analysis, and randomization statistics are combined as a new method to remove the EMG-artifacts. The second objective is to evaluate the performance of EEG microstate analysis and randomization statistics in detecting and removing the muscle artifacts from the EEG recordings of cognitive tasks in comparison to ICA and multiple artifact rejection algorithm.

1.3 Contribution

The contributions of this thesis are as follows:

- A standard preprocessing of raw experimental EEG data is provided for data analysis.
- An analysis of EEG signal-power in the 45-70 Hz frequency band is given for the extraction of the EMG contaminated EEG segments or epochs due to the contraction of frontalis and temporalis scalp muscles.
- An analysis of EEG microstates on the preprocessed data is conducted.

- A noble approach is proposed to remove the EMG contaminated EEG epochs or segments based on EEG microstate analysis and randomization statistics.

- A comparison of the performance between the proposed method and an independent component analysis (ICA) combined with multiple artifact rejection algorithm [14] is shown.

1.4 Thesis organization

The rest of the thesis consists of five chapters. A short overview of each of the chapters is as follows:

Chapter 2 starts with the primary background of the physiological signal in the human brain particularly electroencephalography (EEG). The related signals are also described along with the artefactual signals in the EEG recordings.

Chapter 3 shows a brief explanation of the experimental setup and design for the acquisition of EEG signals.

Chapter 4 analyzes the EMG contaminated EEG epochs or segments which are produced due to the frontalis and temporalis scalp muscles by calculating the 45-70 Hz frequency band signal-power and subsequently removes the contaminated epochs or segments using EEG microstate analysis and randomization statistics.

Chapter 5 validates the proposed-method, i.e., the combination of EEG microstate analysis and randomization-statistics, by comparing it with the ICA using multiple-artifacts rejection.

Chapter 6 provides the concluding remarks of this study and the results obtained and gives a future direction for further research.

Chapter 2

Background Material

2.1 Electroencephalography (EEG)

Electroencephalography (EEG) is a means of measuring the electrical activity of the brain. The human EEG consists of a complicated set of brain waves or signals [15]. These signals are detectable from the scalp electrode (small metal discs) because of the fortuitous architecture of neurons in the human brain. In the neocortex of humans, there are rigidly packed arrays of columns containing six neurons which are orthogonal to the pia matter right below the skull [15]. For this certain arrangement in the neocortex, electric potentials from neurons transmit to the skull where their potential difference can be measured. These scalp potential differences are extremely faint about one-millionth of a volt only. However, this measurement is hindered and distorted by the insulated layers (e.g., spine fluid, skin, skull etc.) between the brain cortex and the electrodes [15].

If all these obstacles are still manageable for interpretation of EEG data, then the negative and positive brain potentials cancel out each other and the difference in valence is detected. This represents only a part of the electrocellular activity beneath the electrode [15].

While EEG signals record the differences in voltage, how the signal is viewed can be set up in a variety of ways called montages. For instance, in bipolar montage, each waveform

in the EEG signal represents the difference in voltages between two adjacent electrodes. For example, F3-C3 represents the voltage difference between channel F3 and its adjacent channel C3. The EEG signal acquisition device repeats this process to get the montage of the whole scalp through the entire array of electrodes.

However, the shape of the brain is volumetric and irregular. As such it is difficult to interpret EEG data if not impossible. The reliable interpretation of EEG data has been done in many research fields like epilepsy, sleep, psychology etc. under many conditions and contexts. Because of its high temporal resolution (in milliseconds range) and non-invasive nature, EEG is widely used for monitoring neuron communications in the brain [15].

2.1.1 Brain lobes and rhythms

Brain lobes are the anatomic parts of the brain. Among these anatomic part, the cerebrum is the biggest. According to Terminologia Anatomica (1998) and Terminologia Neuroanatomica (2017) the cerebrum is divided into six lobes. They are:

1. Frontal lobe,
2. Parietal lobe,
3. Occipital lobe,
4. Temporal lobe,
5. Limbic lobe,
6. Insular cortex.

The first four brain lobes of the human brain are shown in Figure 2.1 [1]

In this thesis, the focus is on the frontal and temporal lobes of the brain. In these two lobes, the EMG contamination of EEG data due to the frontalis and temporalis muscle is

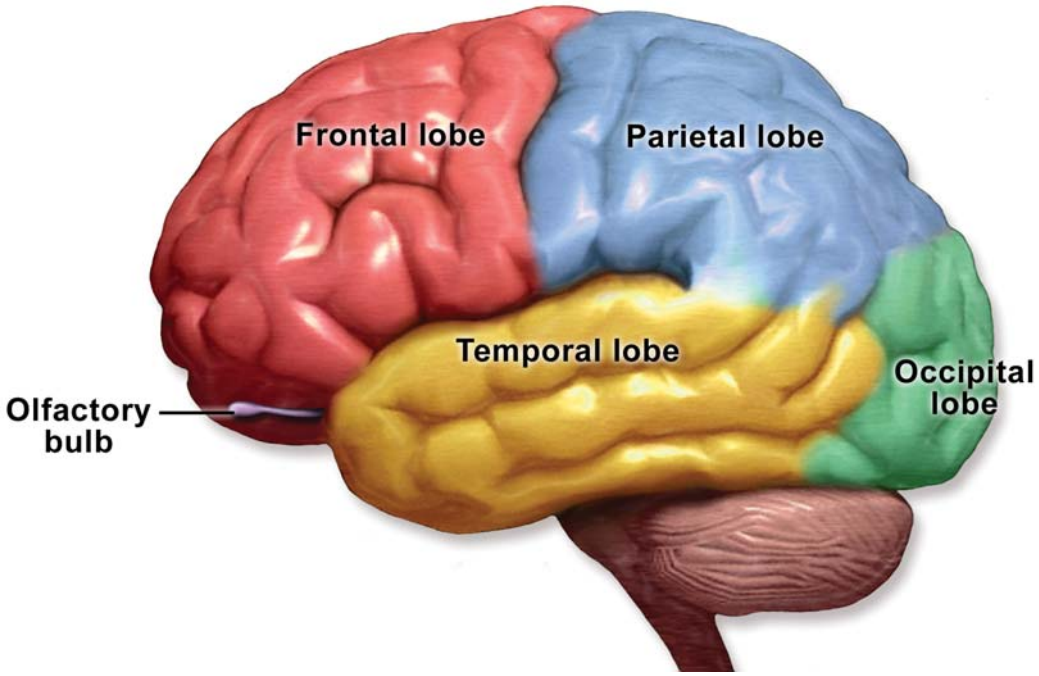

Figure 2.1: Brain lobes from [1]

very high and a proper frequency band to detect is 45-70 Hz [13]. The advantage of choosing this band is that EEG signals in this band have low amplitude and the contaminated EEG signals will have high amplitudes [13].

Frontal lobe: The part of the brain at the very front of the cerebral hemisphere is called the frontal lobe. It has delicate neurons containing dopamine. It exercises the attentional and motivational tasks of the human brain in addition to controlling the significant cognitive skills like problem-solving, emotional expressions, language, judgment, and sexual behavior.

Temporal lobe: The temporal lobe is located on two sides of the cerebral hemisphere below the lateral fissure. This lobe is related to the understanding of a language, memories with vision, and human emotions. The functions of this lobe are auditory processing and auditory memory [16].

Brain rhythms: In EEG signal processing there are five major brain waves classified by different frequency ranges. These waves or rhythms are δ (delta), θ (theta), α (alpha), β (beta) and γ (gamma). Delta wave reflects the EEG activity at a low frequency of 0.5-4 Hz. It is primarily linked to EEG synchronized sleep in the human brain. On the other hand, theta activity is observed in the frequency range of 4-8 Hz. This activity is related to the active and efficient processing of the brain and is dominant during relaxed state and eye open [17].

Next comes the alpha wave which is in the frequency range of 8-13 Hz. The beta wave has frequencies from 13 to 30 Hz and appears when the brain is engaged in visual or cognitive activities [17]. The gamma wave has a higher frequency, ranging from 30 to 70 Hz. In this research, analysis of the gamma waves is conducted to remove the muscle contamination from the EEG signals.

2.1.2 Event related potentials

The event-related potential is a class of EEG related to external events. The EEG signals that are generated due to specific external events like cognitive, motor events [18] are called event-related potentials (ERPs). The ERPs measure how the brain responds to external events. These are typical electro-physiological responses of the human brain due to an external stimulus. These are usually locked in time and divided into two groups: Endogenous and exogenous ERPs. The ERP-studies being noninvasive help researchers to evaluate the functions of the brain.

2.2 EEG artifacts

In clinical neurophysiology, artifacts are any potential difference due to the extra-cerebral source, recorded in the tracing of EEG. These artifacts obscure the EEG signals and lead to misinterpretation and false conclusions. The contamination of EEG signals due to these artifacts is a well-recognized problem in clinical neurophysiology and experimental electroencephalography.

Hence, it is very challenging to handle the artifacts in EEG related studies. The first challenge is to recognize, identify, and determine the sources of artifacts in EEG signals. The second step is to remove those. In some EEG and ERP related studies, people detect and remove the EEG artifacts. However, these artifacts may have the same characteristics: frequency distribution, rhythmicity, and recurrence that exist in the recorded brain potentials. Therefore, the removal of such-artifacts may also remove the useful EEG signal. It thus becomes difficult to differentiate between activities that are of artefactual or cerebral origin. These unwanted artifacts impact the EEG signals found in low amplitudes in the range of microvolts. Hence it becomes complicated to remove EEG artifacts.

2.2.1 Sources of EEG artifacts

There are three major types of sources of artifacts. They are:

1. Environmental artifacts
2. Experimental error
3. Physiological artifacts

Environmental artifacts: This type of artifacts originates from power leads that are present in the surroundings of the body. It can be observed in the form of 50/60Hz noise. This also arises due to electrical interference for the emission of electromagnetic radiation from an external source. It is a principal source of interference in bio-electric measurements because of the capacitive coupling of measurement cables and the main cables of the devices.

Experimental error: This type of artifact occurs due to human error during experimental setup, motion of the subject during data recording, incorrect procedural setup, poor application of electrodes. However, the motion of the subject creates a large amount of error and it is highly detrimental for many physiological signal recordings. The motion also damages the bio-potential measurements in the body, such as ECG and EEG. Subject motion can cause the position of the electrode on the skin to alter. This movement can cause a variation in the distance between the recording electrode and the skin, which results in a corresponding change in the electrical coupling causing signal distortion. Experimental artifacts, relating to motion, in the recorded signals are more difficult to remove as they generally do not have a predetermined narrow frequency band and their spectrum often overlaps with that of the desired signal.

Physiological Artifact: The physiological artifacts are changes in the desired EEG signal due to other physiological processes in the body. Major artifacts are mostly detected in the physiological measurements of eye movement-related artifacts, cardiac signals, and muscle tension signals. Blinking of eyes also causes involuntary movement of the retina as well as muscle movements of the eyelids. As, the eyes are proximity to the brain, when the signal propagates over the scalp, it can appear in the EEG signal as an artifact.

In brief the physiological sources can be one or more of the following:

1. Eye movements, muscle movements (EMG)
2. Muscular artifacts like chewing, swallowing, clenching, sniffing, talking, scalp contraction
3. Cardiogenic movements of the heart, heart beats, ECG artifacts that have QRS complex of poor quality
4. Sweat

2.2.2 Features of EEG artifacts

The EEG artifacts have some striking features which can be used efficiently for the purpose of detection and removal. Some prominent features can be as follows:

1. A relatively large amplitude with respect to that of interested cortical signals like pure EEG signals
2. High potential-difference values for the blinking of eyes or for the vertical eye movements due to the difference between upper and lower EOG reference channels [19]
3. The noise induced by motion artifacts that sometimes masks the neural signal [20]

2.2.3 Typical EEG artifacts

There are three common types of physiological EEG artifacts [21]. These are:

1. Electroculogram(EOG)
2. Electromyogram (EMG)
3. Electrocardiogram (ECG)

Electrooculogram (EOG): The electrooculogram (EOG) is the measurement of electrical activity produced by eye movement, which is normally strong enough to be recorded along with the EEG [5, 22]. This type of signal produces interference. The intensity of this interference depends on the adjacency of the brain-electrodes to eyes. This intensity also depends on the locomotion of the eyes. Blinking eyelids is another prominent reason for the contamination of EEG signals. Moreover, the amplitude of the blinking artifact is generally much larger than that of the original EEG activity [22]. This amplitude is significantly larger in the frontal electrodes than that in other electrodes. In literature, the ocular artifacts are called OAs or EOG-artifacts. In this thesis, we shall adopt the latter for further references [22].

Electromyogram (EMG): Electromyogram or myogenic activity is the tracing of electrical activity generated due to the contraction of the muscle tissues on the body surface. These muscular tissues can be skeletal, smooth, and cardiac muscle tissue. The amplitudes of the interference signals depend on the type of muscle-tissue contraction [5]. As such, it is difficult to stereotype the muscle artifacts in EEG signals. These artifacts are referred to in the literature as MAs or EMG artifacts. We shall use the latter throughout this thesis [5]. The cranial EMG artifacts have several properties that adversely affect the background-EEG signals [23, 24]. EMG artifacts have a wide spectral distribution from 0 to 200+ Hz [7]. It affects all the classic EEG bands like alpha, beta, and delta. Also, the EMG artifacts exhibit less repetition than other biological artifacts. Thus, it is more challenging to characterize the EMG artifacts, since these artifacts arise from the activities of spatially distributed, functionally independent muscle groups, having distinct topographic and spectral signatures [7].

Electrocardiogram (ECG): Electrocardiogram (ECG) is the acquisition of electrical signals arising from the heart. In comparison to the EEG signals of the brain, the amplitude of this type of signal is relatively low. These signals originate from the natural heartbeats that have repetitive characteristics and recurring waveform patterns. These two features greatly help to detect the presence of ECG artifacts in EEG signals. The ECG is routinely measured along with cerebral activity, making this artifact easier to detect and correct since a reference heartbeats-waveform is usually available [5]. These types of artifacts are called cardiac artifacts (CAs) or ECG artifacts in the literature.

2.2.4 Consequences of EMG artifacts

EEG signals become contaminated due to the muscle contraction or expansion, the motion of the subject, electrode movements. As such, data-analysis become difficult that often results in misleading conclusions or findings. In addition, the muscle artifacts distort the original EEG signals and render data analysis more difficult [25]. On top of that, in brain-computer interface (BCI) applications, these EEG artifacts cause unintentional control and decrease the classification accuracy [25].

2.3 State of the art of artifact removal

The most common approaches for EEG artifacts removal can be filtering of EEG data using a band pass filter and subtracting the artifactual signals from the raw signals using a regression technique. Apart from these two methods, decomposition of EEG data into independent components of neural and non-neural components and blind separation of the sources of EEG signals are mostly used in handling the artifacts of EEG data.

As such, the most common techniques for the removal of EEG artifacts may include:

1. Adaptive Filtering
2. Blind Source Separation (BSS)
3. Independent Component Analysis (ICA)
4. Canonical Correlation Analysis (CCA)
5. Empirical Mode Decomposition (EMD)

Adaptive Filtering: An adaptive filter is a linear filtering system. In this system, variable parameters control the system's transfer-function. This type of filter can optimally adjust those parameters by receiving the feedback from the output of the system. It requires a reference channel to make a comparison of the desired output with the derived one [26].

Figure 2.2 shows a simple block diagram of adaptive filtering. Let *s(n)* be the combination of the original EEG signal *x(n)* and additive artifact *r(n)*. Now, if the artifact source *v(n)* is available from an EOG or ECG channel, then adaptive algorithms like Least Mean Square (LMS), Recursive Least Square (RLS) can be used for removing artifacts. The artifact-free signal *x'(n)* will be an estimate of the original signal x(n). The theoretical assumption is that the desired EEG signal and artifact signal are independent [26].

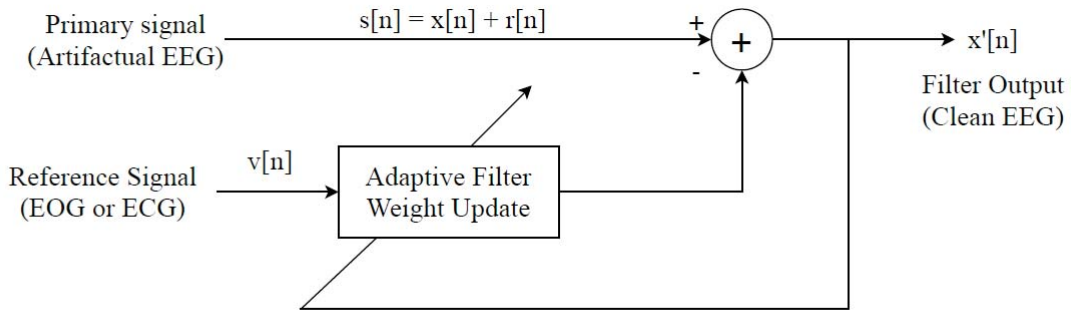

Figure 2.2: Typical use of adaptive filtering in canceling physiological artifacts with available artifact source channel as reference

For handling ocular interference, we can adopt adaptive filtering. For removal, it will depend on the specific application and whether it is online/offline or on the availability of proper reference channels.

Blind Source Separation (BSS): Blind source separation (BSS) technique is one of the most popular techniques for the detection and removal of EEG artifacts. This method extracts the individual unknown source signals from their mixtures. It estimates the unknown mixing channels by using information observed within the mixtures obtained from each channel's output, having very little information about the source signals and the mixing channels. Let *X* be the observed EEG signals from multiple-channel recordings that are assumed to be a linear mixture of the sources *S* plus additive noise vector *N*, that is,

$$X = AS + N \tag{2.1}$$

Here the goal is to estimate the linear mixture matrix *A*. Let *W* be the estimated matrix of *A*. The matrix, *W* is estimated by an iterative process to determine the estimated version of the source signals. Therefore, the estimated version of the source signals can be written as follows:

$$\hat{\mathrm{S}} = \mathrm{WX} \tag{2.2}$$

Here, $\hat{\mathrm{S}}$ is the estimated version of the source signals. Figure 2.3 shows a sample block diagram of the BSS method.

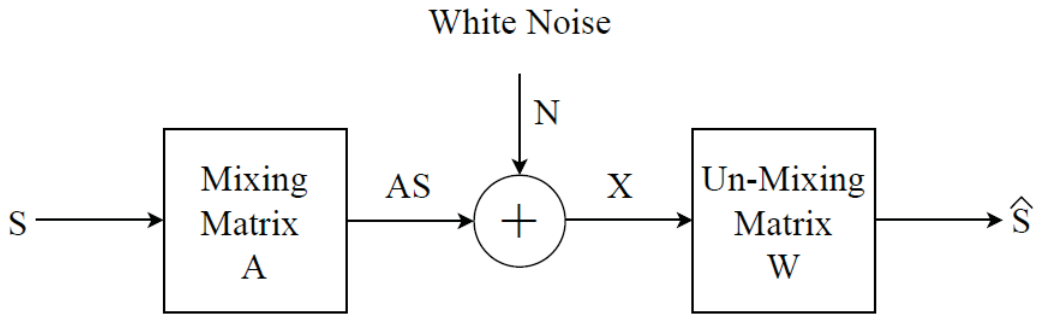

Figure 2.3: Demonstration of the blind source separation method

Independent Component Analysis: Independent component analysis (ICA) is a blind source separation technique based on the assumption that the signal sources are linearly independent [26]. ICA is a method for finding the underlying factors or components from multivariate (multi-dimensional) statistical data [27]. ICA generates a weight coefficient for each factor that measures the linear inter-dependence between the signals and channels. These factors extracted by the ICA are lower than or equal to the number of channels in the EEG data [12, see Chapter 5]. What distinguishes ICA from other methods is that it looks for the components in the EEG data that are both statistically independent, and non-Gaussian [27]. However, the main problem of ICA based artifact detection and removal method is the manual selection of artifactual independent components (ICs). The process can be automatic by labeling the ICs through the calculation of some features. The features quantify the probability of the IC to be an artifact. This process is the combination of ICA with another method, for instance, Empirical Mode Decomposition(EMD) or classifiers like Support Vector Machine(SVM). Even in such cases, the ICs may retain some residual neural signals.

As a result, while reconstructing the signals by rejecting the artifactual ICs, distortion is introduced in the neural signals. Moreover, this process can not operate on single-channel EEG data because the number of recording channels must be at least equal to the number

of independent sources [26]. ICA is successful in removing EEG muscular-artifacts. However, the removal of muscle artifacts using the ICA technique is hard. The muscle artifacts are harder to eliminate as the proper reference channels are not available. Even if the reference channels are available, it is still difficult to remove the muscle artifacts. The artifactual signals can generate due to the activation of multiple muscles rather than a single-muscle. As such, there is a disagreement in the literature on whether the ICA-technique is efficient or not. Hence, other methods like canonical correlation analysis (CCA) [6] and empirical mode decomposition(EMD) are also necessary for the efficient removal of EEG muscle artifacts [28].

Canonical Correlation Analysis(CCA): CCA is another algorithm based on the concept of blind source separation. This method uses second-order statistics(SOS), looks for uncorrelated components in the data signal [26]. This process uses a weaker condition than the ICA method. The process seeks statistical independence among the signal components. Unlike ICA, CCA addresses the temporal correlations by finding uncorrelated components. It maps the signal components from maximum to least auto-correlation. The signal component having the least auto-correlation mostly reflects the artifacts because the auto-correlations of neural signals is maximum . The main strong point of this method is being automatic and computationally efficient [26].

Empirical mode decomposition(EMD): This method is a data-driven empirical approach. It is an algorithm that performs on random/stochastic, non-linear, non-stationary processes. As a result, this is ideally suitable for EEG signal analysis and processing. In this approach, the signal *s[n]* is decomposed into the sum of band-limited components or functions *c[n]* called intrinsic mode functions (IMF) with well-defined frequencies. IMFs have an equal no. of extrema and zero crossings. At any point in the curve of the IMFs the

mean value of the maxima and minima must be zero.

Combination of methods for handling multiple artifacts: In the case of two or more artifacts, ICA is a better choice. Others can be visual inspection, correlation analysis, frequency spectrum, iso-potential maps. For processing the ocular and muscular artifacts, ICA can be employed together with wavelet transform (WT). For eliminating the muscle artifacts ICA or CCA or combination of both or combination with other methods can be a better choice. Again, ECG or cardiac artifacts have specific dynamics and are easily separable into different ICs. For this type, we can choose regression with the filtering process and ICA. Alternatively, ICA and wavelet transform can be combined together for handling the cardiac artifacts. A combination of ICA and other methods can be a good option for detecting and removing the eye and muscle artifacts.

Besides, several statistical features are used in machine learning(ML) classifier for threshold calculation in EMD or ICA based methods to improve the overall artifacts-removal process. Moreover, to tackle the most common 3 artifacts: EMG, EOG, and ECG, the following hybrid algorithms can be a good option. These are:

1. Blind source separation (BSS) and Support vector machine (SVM)

2. Adaptive filtering and neural networks

BSS-SVM: This is a hybrid method for removing the EEG artifacts [29]. Here, carefully chosen statistical features are extracted from separated source components after the application of BSS. Next the features are fed into a support vector machine(SVM) classifier to identify and remove the artifactual components. Figure 2.4 shows a sample block diagram of this hybrid process. In this process the second order blind estimation (SOBI) technique has been used. The full process is described in [26, 29].

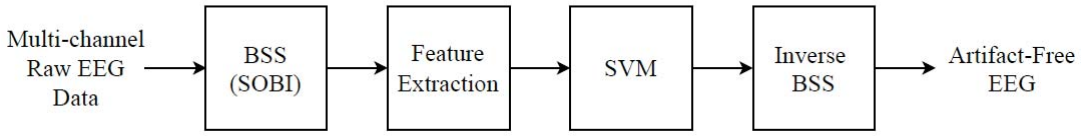

Figure 2.4: A block diagram of the BSS-SVM process

In summary, different methods for the removal of the three most common EEG artifacts: EOG, ECG, and EMG have been discussed. These methods can be suitable for both single-channel and multi-channel EEG data. Table 2.1 shows the types of artifacts, the methods to remove them, and the corresponding number of data channels.

Type of Artifacts	**Methods**	**Channel type**
Muscular	CCA, ICA, combination of ICA and CCA	Multi, Single
ECG	Adaptive Filtering, ICA	Multi
Muscle	CCA	Multi
EOG (Ocular)	ICA, BSS and EMD	Multi
Head movement	ICA	Multi

Table 2.1: Different types of EEG artifacts, method of removing them and corresponding channel type

While these techniques have been proposed to detect and remove different artifacts, numerous studies in the literature have used different measures to validate the algorithms. In general we do not have the optimal choice for the removal of all types of artifacts. On the contrary, for brain-computer interface (BCI) applications, the artifact removal algorithm needs to be efficient enough for online or real-time processing with single or multiple data-channels.

Therefore, this thesis targets BCI applications and investigates the method in [13] for the detection of the frontalis and temporalis scalp muscles contaminations in EEG recordings by conducting the EEG signal-power analysis in the 45-70 Hz frequency range [7, 13]. The thesis also proposes a method to remove the EMG contaminations of EEG signals by using microstate analysis [9] and randomization statistics [30].

Chapter 3

Experimental Design and Signal Acquisition

This chapter focuses on experimental setup and design, signal acquisition, and data preprocessing. The experiment consists of a cognitive task that was designed by Lucas House, a research associate in Design Lab at Concordia University, and approved by the Human Research Ethics Committee at Concordia University. The motivation of this experiment was to find out the relationship among cognitive workload, mental effort, and stress. For estimating the cognitive effort and mental stress, physiological signals like, EEG signals, skin conductance were collected from the research subjects.

3.1 Experimental setup

The BioSemi is one of the top EEG hardware companies that manufacture the EEG signal acquisition system. The Bio-semi machine is a commercial machine for the acquisition of EEG signals ranging from 32 to 256 data channels. With the help of Active II, Bio Semi machine, the EEG signals were acquired.

In this experiment, the total number of EEG signal data channels consists of 64 EEG channels, 8 EOG channels, and one stimulus channel for the trigger. In addition to these channels, there are two extra channels for the skin conductance. Participation in this experiment was voluntary. Four participants having an age range of 24-32 years took part in this experiment. Three participants were right-handed, one having regular eyesight without glasses. One of them had eyeglasses. They belonged to the Faculty of Engineering and Computer Science at Concordia University. The data collected from this experiment was used and analyzed with proper consent from the participants.

EEG signal recording: The most traditional and widely accepted method for recording EEG signals is the International 10-20 system [31] in clinical operations. This system specifies the positions where the EEG electrodes should be placed [16]. It builds a standard comparison among the subjects. Here, the number 10 shows the actual distance of adjacent electrodes of the cap is 10% of the total front-back and 20 indicates that to interval distance is 20% of the right to the left of the skull. Besides, the name of each electrode reflects essential information. The first letter of the name of an electrode represents the area of the brain. The number refers to the displacement from the midline and laterality. The central position electrode in the $10 - 20$ system is at the top of the scalp which is named Cz. For EEG studies in the research laboratories, data are recorded from many channels and pre-processed for subsequent analysis. In this thesis, we used 64 channels of the standard Bio semi 10-20 system in addition to 8 Electrooculogram (EOG) channels and one stimulus or trigger channel. However, today EEG caps with 256 channels are also available for recording processes. The number of channels is selected based on specific research purposes and questions. We used a 64-channel-cap (Bio semi) as per the 10-20 standard system of electrode layout for EEG recordings.

In collecting the EEG, signal to amplifier gain of the bio-semi system was fixed [32].

The Bio semi system comes with the ActiView [33]. The ActiView is an open-source program based on the graphical programming language LabVIEW. In this research-experiment, the LabView run-time engine [33] was necessary to run the ActiView program. The ActiView program handles data acquisition, displays on-screen the recorded EEG signals during the experiment. The program displays ground sensors, EEG-sensors, as well as the other sensors that are used for collecting the skin conductance, eye movements (EOG) of the research-participant. In our case, the "ground" electrodes, unlike the conventionally used ones, are two separate electrodes:

- Common Mode Sense (CMS) active electrode (ideally placed in the center of the measuring electrodes)
- Driven Right Leg (DRL) passive electrode (ideally placed away from the measuring electrodes) [34]

These two electrodes form a feedback loop, which drives the average potential of the subject (the Common Mode voltage) as close as possible to the ADC reference voltage in the AD-box (the ADC reference can be referred to as the amplifier "zero") [34]. The "Biosemi" cap [35] was put on to the experimenter's head in advance to lower the duration of the experiment. There was a small slit for each electrode in the cap where a small amount of gel was placed when the subject first wore the cap. The gel reduced the scalp or skin impedances of the EEG electrodes. The subject had to wear this cap during the whole experiment. The layout of the standard Bio Semi 64 electrodes or channels [2] is illustrated as follows in figure 3.1.

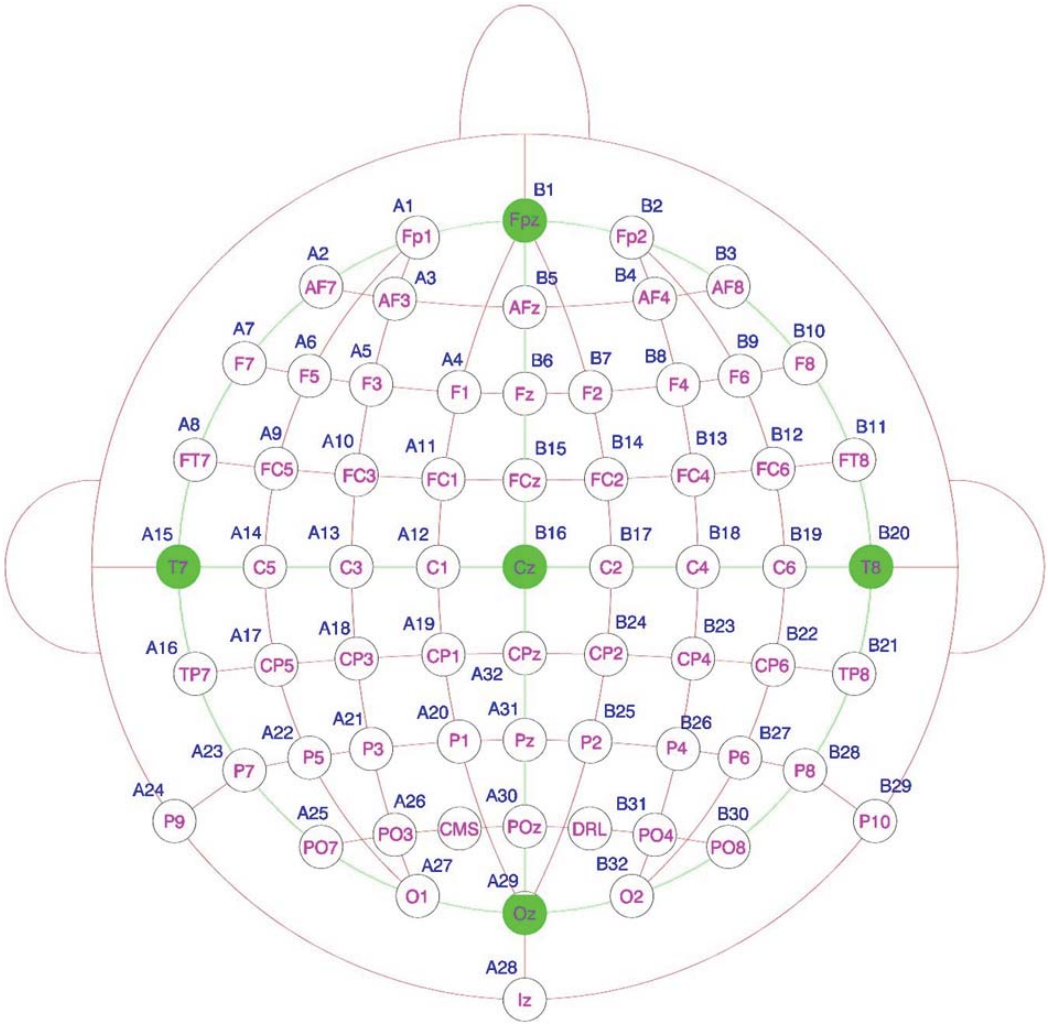

Figure 3.1: Bio Semi 64 channel electrode layout from [2]

3.2 Experimental procedure: N-back task

The main experimental design is based on a working memory task, as described in [36]. In this experimental paradigm, the stimuli consisted of English capital letters randomly drawn and shown on the computer screen. In every 4.5 seconds, the participant saw the stimulus for 200 milliseconds. So, each experimental-trial consisted of 4.5 seconds. In between the two stimuli, a small cross popped up at the center of the computer screen. The identity and location of each letter varied randomly from trial to trial. The subjects performed a continuous English capital letter-matching task. They indicated whether the

current stimulus matched with that presented on the previous one, two, or three trials.

The subjects performed two versions of this matching-task with three levels of difficulty: low (1-back), medium (2-back), and high (3-back). The first version of the task was a verbal, non-spatial task where the participants had to recall the identity of the visual stimulus (letter) presented regardless of the position of the stimulus. In the second version of the matching task, the subjects had to recall the identity and the location of the visual stimulus (letter), shown on the screen. This task is called verbal, spatial-task. In the 1-back verbal, non spatial-task, the participants compared the current stimulus with the stimulus presented one trial before. The participants detected a "match" or "non-match" in the tasks. A "match" between the two stimuli occurred, when the identity of the second stimulus matched with that of the first one, ignoring its position.

If the identity of the two stimuli was not the same, a "non-match" occurred. For the 1-back verbal, spatial-task, "match" between the two stimuli occurred, when the identity and location of the second stimulus matched with that of the first one. If the identity and location of the two stimuli were not similar, a "non-match" occurred. The participants pressed the key "K" on the keyboard when a match was detected. On the other hand, the participant pressed the key "L" on the keyboard when a non-match was detected.

During the medium difficulty-level task (2-back task), the participants compared the current stimulus with that of presented two trials ago. Finally, in the high difficulty level task (3-back task), they compared the current stimulus with that presented three-trial before. The subjects performed all three levels of the matching-task in both versions: spatial and non-spatial. The research-subjects had at most one hour and forty-five minutes for doing the whole experiment. The trials were 4.5 seconds long. The participants performed 24-trials for each of the six task conditions [36]. These were: 1-back spatial, 2-back spatial, 3-back spatial, 1-back non spatial, 2-back-non spatial, and 3-back-non spatial.

3.3 Data preprocessing

The acquisition of the EEG signals was conducted during the completion of the experiment. Each recording had 73 channels with 64 EEG channels, 8 EOG channels, and one trigger channel. The signal sampling rate was 512 Hz. In the pre-processing section, the raw data was imported using the MNE library of Python [37], which was then filtered using a simple bandpass filter with a frequency range of 0.1 to 100 Hz [36]. Filtering is the most traditional preprocessing-technique used for handling the raw EEG data, which tackles the non-neural artifacts like artifacts originating from the equipment and environment. While conducting experiments, the uncontrolled variation arises because of experimental error. Hence, it is impossible to eradicate this type of noise. With a proper frequency-selective filtering approach, the environment-induced noises like the main power leads, white noise etc., can be eliminated to improve the signal to noise ratio. Thus the raw EEG data were filtered with a notch filter at 60 Hz for removing the power line noise. After filtering of raw EEG data, the average EEG-reference was set using the MNE Python package. All these steps has been completed using the MNE package of Python [37] to obtain the preprocessed raw EEG data.

Chapter 4

EMG Artifact Analysis and Removal

The contamination of EEG signals due to frontalis and temporalis scalp muscle spreads over the entire scalp, which is involved in EEG signal acquisition. It masks the underlying neural activity and distorts the original brain signal; hence it is necessary to get rid of this contamination. As a result, EEG data segments from these two (frontalis and temporalis) brain regions are of particular interest in EMG artifact analysis as they are the most common sources of EEG contaminations or artifacts over the frontal and central head regions [38, 39]. The EMG artifact has both spectral and topographic characteristics as discussed in [7], where it is shown that EMG artifact in the average subject data has a broad frequency range from 0 to 200 Hz and the artifactual signal amplitude is the greatest at 20-30 Hz and 40-80 Hz in the frontal and temporal regions of the brain. In the 20 Hz frequency, the temporal EMG activity shows smaller peaks. Also, the EMG spectra often have beta peaks like EEG data. So a suitable frequency band for detecting EMG contaminations is 45-70 Hz [13]. In this band, the EEG signal amplitudes are much smaller compared to the EEG signal amplitudes at 13-38Hz. The underlying assumption is that the signal peaks that occur between 45-70 Hz will most likely be from EMG activity [13]. According to [40], to detect EMG contaminated EEG data segments, analysis of frequency in the 51-69 Hz of EEG data-channel F7 can be done. This band is also called the muscle band.

However, the EMG contamination is the greatest at the scalp periphery near the active muscles. Hence, it obscures or mimics the EEG alpha, beta, and mu waves over the entire scalp. As a result, the researchers acquire EEG signals from the peripheral scalp locations. To detect the contaminated EEG data-segments (epochs), the researchers have used the EEG data obtained from the frontalis and temporalis scalp positions, since the EEG data in these two scalp positions can be analyzed to remove the EMG artifacts for single-subject data as well as average data of multiple subjects.

In this chapter, we shall discuss the process of analyzing and removing the temporalis and frontalis muscle EMG-contamination (EMG artifact) from the EEG recordings and reconstruct the EMG artifact-free EEG data. The analysis (detection) of EMG contaminations consists of three steps:

1. Data loading and pre-processing
2. Selection of data-channels
3. EEG signal-power analysis in the frequency band of 45-70 Hz [13]

Figure 4.1 gives an overview of the process of EMG artifacts analysis.

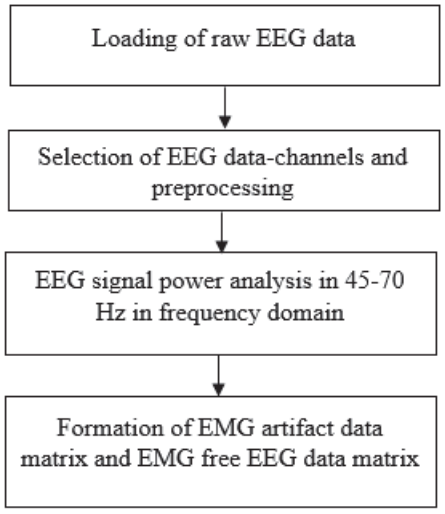

Figure 4.1: The overview of the frontalis and temporalis muscle artifact analysis

The removal process is a combination of the following five steps:

1. EEG microstate analysis [9]

2. Randomization statistics [30]

3. Fit back of EEG microstate maps

4. Interpolation of the data points of the EEG data-channels [37]

5. Data reconstruction

Figure 4.2 provides an overview of the method of removal

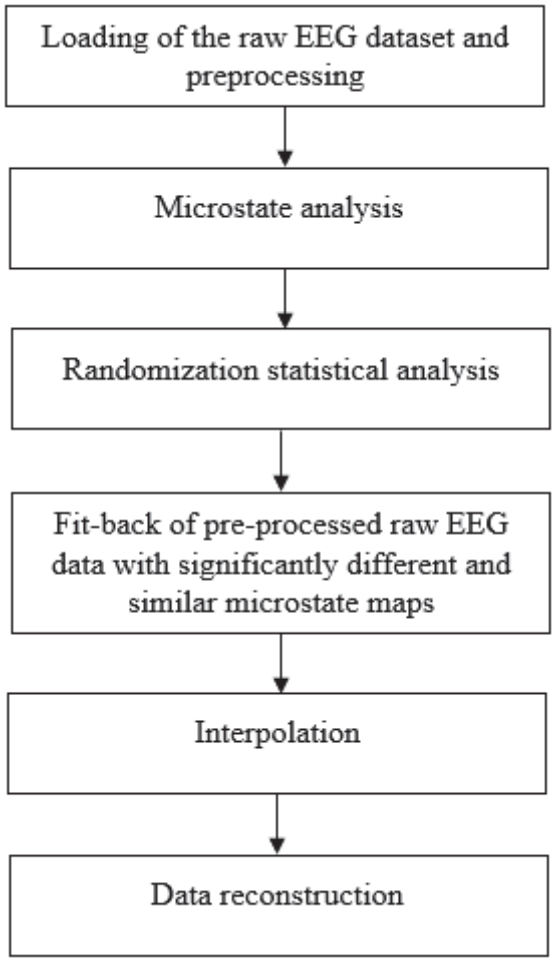

Figure 4.2: The overview of the frontalis and temporalis muscle artifact removal

4.1 EMG artifact analysis

4.1.1 EEG data loading and preprocessing

At first, we load the dataset of the four subjects of the N-back experiment, using the MNE-Python library [37]. In the dataset, there are a total of 73 channels containing 64 EEG channels, 8 EOG channels, and a stimulus channel. The loaded data are bandpass filtered from 0.1 to 100 Hz, followed by notch filtering at 60 Hz. Finally, we set the reference of the EEG-data as "average" EEG-reference using the MNE library of python [37].

4.1.2 Selection of data-channels

After completing the preprocessing steps, we select the EEG data-channels: AF7, AF8, FT7, FT8 as primary data-channels. The frontalis and temporalis scalp muscles contaminate these four EEG data channels. We chose the EEG data of the five adjacent data-channels to each of the primary data channels for the detection and removal of EMG-artifacts. Thus EEG channels Fp1, AF3, F3, F5, F7 are selected for the channel AF7, channels Fp2, AF4, F4, F6, F8 for AF8, channels F5, F7, FC5, C5, T7 for FT7 and channels F8, F6, FC6, C6, T8 for FT8. The frequency band of 45-70 Hz is chosen [13], for EEG signal-power analysis in the frequency domain. One hundred EEG data-segments (epochs), 2 seconds each, are selected for the whole EEG signal power-analysis, in the frequency domain. Each EEG epoch has 1024 time samples because the sampling frequency during the EEG signal acquisition was 512 Hz.

4.1.3 EEG signal-power analysis in frequency Domain

As mentioned above, the 45-70 Hz frequency band is selected to investigate the EMG contamination of EEG data. The power of the EEG signals in this frequency band is analyzed for detecting the EMG contaminated EEG segments or epochs in the pre-processed raw EEG data. The EEG data are scaled into microvolts from volts by dividing each data sample by 10^{-6} to calculate the signals' power in the microvolt range. The equation for calculating the power of the EEG signals in the frequency-domain [13] is as follows:

$$Power_{45-70Hz} = \frac{1}{N^2} \sum_{k=45N/f_s}^{70N/f_s} |X(k)|^2 + |X(N-k)|^2 \tag{4.1}$$

Where, N is the number of time samples in a time interval of interest, f_s is the sampling frequency and $X(k)$ is the k^{th} discrete Fourier transform coefficient as defined below:

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{(-j\frac{2\pi}{N}nk)} (0 \leq k \leq N-1) \tag{4.2}$$

Here, $x(n)$ is a real discrete signal in the time domain. The next step is to determine the threshold value for detecting the EMG contaminated EEG-epochs. The threshold values are set as the mean value of the EEG signals' power obtained from one hundred EEG epochs for each primary channel. Thus, for each of the four primary channels AF7, AF8, FT7, FT8, there are one threshold values. Each primary channel epoch-power is calculated and compared against the corresponding threshold value to detect EMG contaminated EEG epochs. If the amplitude of the EEG epoch's power, obtained from any of the four primary EEG-channel, exceeds the threshold, we also compare the corresponding EEG epoch-power from the nearby five channels. Thus, for each EEG-epoch, we examine the EEG data-channels Fp1, AF3, F3, F5, F7 for AF7 channel, Fp2, AF4, F4, F6, F8 for AF8 channel, F5, F7, FC5, C5, T7 for FT7 channel and lastly F8, F6, FC6, C6, T8 for FT8 channel to detect the EMG contaminated EEG-epochs and store the epoch's data in a data matrix called artifactual data-matrix. This process is repeated for all the hundred epochs to generate the artefactual data having the dimension of the number of channels.

In this sample analysis, the artifactual data matrix has 16 EEG data-channels after removing the duplicate-channels, 100 EEG-epochs, and 1024 data-points for each EEG epoch. If the signal power-amplitude (equation 4.1) of an EEG epoch exceeds the threshold value, we store the data-points of that epoch in the artifactual data-matrix. We mark the data-points of the rest epochs as zeros. The stored artifactual data is formatted to get rid of the zero data-points. This formatting of the artifactual (contaminated) data-matrix is a tricky task to accomplish. The non-zero data-points are extracted from the artefactual (contaminated) data-matrix and formatted into row vectors having the number of elements equal to the number of non-zero data-points. Then the row vectors are reshaped to form the EMG artifact data matrix having 16 channels and the total time-points. We calculate

the total time-points by multiplying the number of EMG contaminated EEG-epochs by 1024 as each epoch contained 1024 time-points. In the analysis of the subject-1 of the N-back dataset, we find 11 EMG-contaminated EEG-epochs. Thus total time-points become 11264. Eventually, the dimension of the EMG artifact data matrix becomes 16 by 11264. We form the EMG free EEG data matrix from the preprocessed raw EEG data (section 3.3) by formatting the preprocessed raw EEG data having the size equal to the EMG artifact data matrix.

Results of EEG signal-power analysis in frequency domain: The first one hundred, two-second preprocessed raw EEG epochs are selected for the signal-power analysis to detect the EMG-contaminated EEG data-segments. The channel-wise data-points are plotted against the time, using the MNE library of Python [37]. The plotting shows high amplitude peaks in the channels over the duration and demonstrates the contaminated EEG data-segments due to frontalis and temporalis muscles. Figure 4.3 shows the sample plot of the EMG-contaminted EEG epochs for a duration of 0.2 seconds.

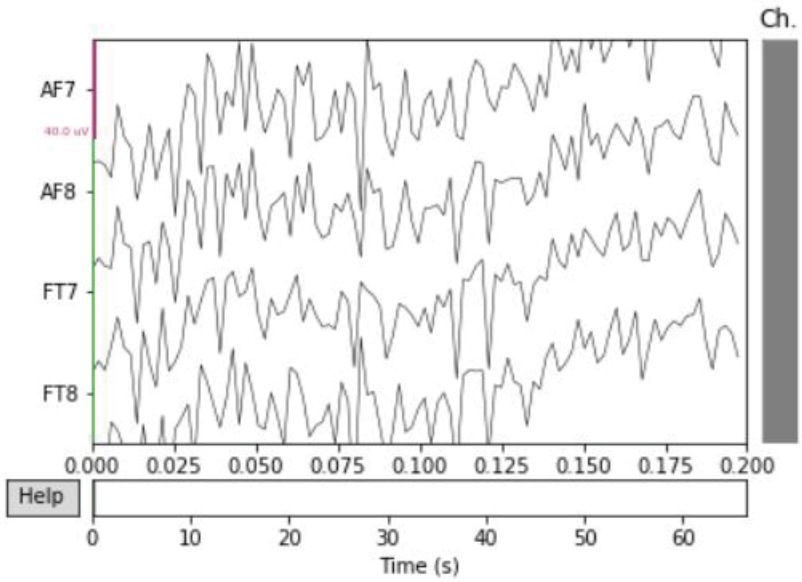

Figure 4.3: EMG-contaminated preprocessed EEG data for a duration of 0.2 seconds obtained from the primary EEG data-channels

We observe from figure 4.3 that there are signal-peaks in the EEG data-channels. This indicates the contamination of the EMG artifacts. In case of the regular EEG data the signal peaks should be low as shown in figure 4.4.

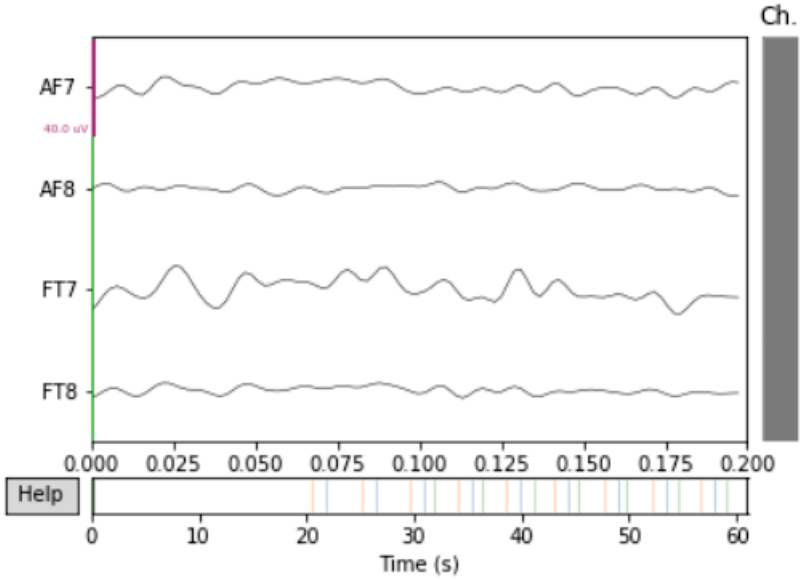

Figure 4.4: Regular EEG data for a duration of 0.2 seconds obtained from the data-channels AF7, AF8, FT7, FT8

We do not observe such low signal-peaks of figure 4.4 in figure 4.3. So figure 4.3 shows the presence of EMG-artifacts, and Algorithm 1 shows the simplified procedure to detect the EMG artifacts.

Algorithm 1: The procedure to detect EMG artifacts

1. Input: 250 seconds EEG data with dimension of no. of channels by no. of time points
2. Keep sampling frequency 512 Hz and band pass filter at 0.1 to 100 Hz
3. Remove line noise by 60 Hz notch filtering and set average-reference
4. Segment EEG data into 2 seconds EEG data-segments (epochs) having 1024 time points
5. Selection of primary EEG data channels (AF7, AF8, FT7, FT8)
6. Calculate threshold i.e mean signal-power of 100 EEG-epochs for each primary channel
7. Calculate signal-power of 100 epochs of 5 channels adjacent to each primary channel
8. Compare each epoch-power against each primary-channel threshold value
9. Store the data points exceeding the threshold values to the EMG-contaminated array
10. Format the array into a MNE-Python raw EEG data object

4.2 EMG artifacts removal

Investigation of the spectral and topographical patterns of the cranial EMG is required to get rid of EMG-artifact in the EEG signal recordings.In the previous section, signal-power analysis has been used to detect the contaminations of the EEG recordings due to frontalis and temporalis scalp-muscles. This analysis can be called the EMG-artifact analysis phase (detection). Once we have detected the contaminated EEG data-segments (epochs), the next step is to remove those. We combine microstate analysis and randomization statistics for the removal of EMG-artifact.

4.2.1 Microstate analysis

Brain microstates are defined as a functional or physiological brain state while the brain performs a neural computation task. These microstates are uniquely characterized by a fixed spatial distribution of the active neuronal generators in the brain, having a time-varying intensity [9]. Electroencephalography (EEG) measures the electric potential of these neuronal generators that project a set of spatial patterns on the scalp surface. This set of spatial patterns can be clustered into a set of representative maps. These maps are called EEG microstates [41].

The purpose of the analysis of these microstates is to compress the EEG recordings (data). There are many data reduction techniques for compressing the EEG data. Among these techniques, the microstate algorithm is very important because of its use in a variety of experiments [9, 42, 43].

A brief overview of this algorithm is as follows. Let us consider an EEG data set having $n_{time\,sample}$ time samples from $n_{channel}$ channels, or electrode locations. So each sample is an array of $n_{channel}$ real numbers, and each element of the sample represents the electric potential at a specific brain-location. This whole array gives a discrete sampling of the continuous electric field. Therefore, we can visualize an EEG dataset as a time series

of changing spatial patterns called maps. This microstate algorithm looks for a small set of spatial patterns to best explain the maximum data-variance by these spatial patterns or maps.

In this study, we have used the commonly employed modified k-means algorithm as introduced in [9] and implemented in [41] by using the programming language Python [44]. The classical k-means algorithm clusters data so that the sum of the squared distances of all the data points to their respective cluster centroids i.e., the arithmetic mean of all the data points assigned to that cluster is minimum. This algorithm progresses in a stochastic manner i.e., in random fashion by using a fixed number of clusters and set the cluster centroids randomly with data samples.

For the EEG recordings, a data sample is an array of electrical potential values at a given time point, having the dimension as the number of EEG data-channels or electrodes. During each iteration, this algorithm sets each data sample to its nearest cluster-centroid updates the clusters and their centroids, taking into account the newly assigned samples.

However, modified k-means (microstate algorithm), as described in detail in [9] does not use this arithmetic mean of data samples for cluster representation. It uses the first principal component of the samples [41]. Due to this the microstate algorithm can ignore the EEG-topographic polarity. Thus, the overall symmetry of the topographic potential remains as the feature to be clustered [9, 43].

We choose this microstate algorithm because of its application to event-related potential (ERP) data sets for a long time. Moreover, our experimental EEG data set is also an ERP dataset. This algorithm transforms the ERP-EEG data set into a sequence of microstate labels with respect to the maximum similarity between the candidate microstates and the actual EEG topography, that is, the configuration of the electric field at the scalp. An extensive description of the primary microstate algorithm is provided in [9] and reviewed in [45]. In this thesis, we have used the EEG topography at the local maxima of the global

field power (GFP) of the EEG channel data as the input for this microstate algorithm.

Global field power: Global field power (GFP) is a parametric assessment of the related-strength of the EEG microstate topographic maps [45]. This map strength is defined as the sum of absolute micro volt values measured at all EEG channels divided by the total number of the EEG channels [46]. EEG researchers compute the GFP as the standard deviation of the momentary potential values. Mathematically,

$$GFP_u = \sqrt{\frac{1}{n}\sum_{i=1}^{n} u_i^2} \tag{4.3}$$

Here, n is the number of EEG data-channels including the reference channel, u_i is the average-referenced potential of the i^{th} electrode. By average reference we refer to the mean of all instantaneous electrode or channels' electric potential values. This average reference potential is calculated by subtracting the mean value (average-reference) from the electric potential value at the i^{th} electrode at time point t. More detailed description is provided in [47–49].

To competitively fit-back the microstate maps into the EEG data set, we calculate both the global explained variance and cross-validation [9, 41] of the EEG microstate maps for each run. A concatenated data-set is formed by averaging the instantaneous EEG data of four subjects obtained from the "N-back experiment" (Chapter 3) across the two main groups: EMG contaminated and non-contaminated. From the concatenated data-set, n data points are randomly selected. Here, the data point is the electric brain potential-difference value obtained from all the scalp electrodes at a given time point (hereafter, template maps) [45]. The number of data points can be from 1 to all the data points or less. Next, the spatial correlation between each of the 'n' template maps and each data point coming for each time point is measured. The spatial correlation between two different time points of

the EEG data from the same group, $C_{u,v}$ is defined as:

$$C_{u,v} = \frac{\sum_{i=1}^{n} u_i \cdot v_i}{\|u\| \cdot \|v\|} \tag{4.4}$$

Where, n is the number of template maps, $\|u\| = \sqrt{\sum_{i=1}^{n} u_i^2}$, $\|v\| = \sqrt{\sum_{i=1}^{n} v_i^2}$, u_i is the average-referenced [47–49] potential of the i^{th} electrode (for a given group, at a given time point t) and v_i is the average-referenced potential of the i^{th} electrode (for the same group but at different time point t') [45].

This process provides a correlation value for each template map as a function of time [45] and for any given time point, one of these 'n' template maps has the highest spatial correlation value. Empirically this process suggests that in the event related potential (ERP) EEG-data a given template map has highest spatial correlation for a sustained period of time. After that another template map is generated having the highest correlation value. This process continues. The global explained variance of the correlation values of these template maps is then calculated [45].

Global explained variance (GEV): Global explained variance (GEV) is a measurement of how well the template maps explain the whole dataset chosen for analysis [45]. Mathematically,

$$GEV = \frac{\sum_{t=1}^{tmax} GFP_u(t).C_{u,T_t}^2}{\sum_{t=1}^{tmax} GFP_u^2(t)} \tag{4.5}$$

Where, $GFP_u(t)$ is the GFP of the data for the condition U at time point t. T_t is the template map assigned by the segmentation for condition U at time point t. The 't' represents the given time point within the data [45].

The average of the template maps, from all the time points, is taken to redefine the template maps when the i^{th} map had the highest spatial correlation. We calculate the spatial-correlation between each of the redefined template maps and time points along with the

resultant GEV. We repeat this sequential process of averaging the time points for the re-generation of each template map, calculating the spatial correlation until we obtain a stable global explained variance (GEV) value. The process of repetition stops when a given set of n template maps do not have a higher GEV value for the given dataset. However, this process creates the possibility of choosing the neighboring time points that might result in a low GEV. Thus to make sure the process yields the highest GEV for a given number of n-template maps, a new set of 'n' template maps is selected. Then, we repeat all the steps as described earlier.

An important point here is that the number of repetition of all these steps is user-dependent. The higher the number is, the higher is the computation time. Now, we retain the highest GEV for the 'n' template maps, and then the same steps are completed for the 'n+1' template maps and can be continued until n equals the total number of data points. These steps mentioned give information about how good the $n, n+1, n+2 ... etc.$ template maps represented the concatenated dataset. A significant factor for this type of analysis is the determination of the optimal number of template maps. These template maps represent the scalp electric potential distribution as topographic maps. The topographic maps are two-dimensional matrices having the dimension of no. of microstate clusters by no. of EEG data-channels or electrodes. As a result, each topographic map is a row-vector having the no. of EEG data-channels as the total number of elements [41].

At this point, the analysis identifies a set of template maps to describe the group-averaged concatenated EEG-data set. However, the issue of how many clusters of the template maps is optimal remains as a bottleneck for this type of analysis. Unfortunately, a definite solution of this issue does not exist. It is because the more the clusters, the higher the value of global explained variance, and the lower the compression of the dataset. If the number of clusters is low, the GEV value will be small and the EEG-dataset will be highly compressed. It is because a small number of template maps will represent the dataset. On

the contrary, with a high number of clusters, the explained variance will be high, but we can not compress the dataset. Therefore, the main target is to determin an optimal number of clusters to achieve a middle-position between these two extreme cases. For this, a method based on the cross-validation criterion is chosen [30].

Determination of optimal number of clusters: After the detection of the contaminated EEG data-segments in NumPy [50] array format through EEG signal-power analysis in the 45-70 Hz band, both the raw non-contaminated and contaminated data are available for subsequent EEG microstate analysis. The EMG contaminated EEG segments are transformed from NumPy array into a raw MNE object using the input-output processing functions of the Python-MNE library [37]. For determining the optimal number of clusters, the dataset of four subjects is shuffled randomly into training and test datasets, each having two-subjects' data .

At first, in the training dataset, the global field power of all the data-points is calculated. Then the peak values are selected for clustering. The number of clusters ranges from 3 to 20. We apply the modified k-means algorithm [9] to the training dataset for each microstate model that is a function of the number of EEG microstate-clusters. The mean spatial correlation between each microstate model and the test dataset is computed and retained. This process is repeated several times (in our case 350 times) for each microstate model. We select the model with the highest average mean correlation as the optimal microstate model [30].

In brief, the entire process can be described as follows:

1. Segmenting the concatenated dataset into the training and test dataset each having 50% of data samples

2. Computing the mean correlation of test data with the microstate clusters generated from the training dataset. Here we have varied the number of microstate clusters

from 3 to 20

3. Averaging the mean correlation for each microstate model ranging from 3 to 20 microstate maps

4. Repeating steps 2 and 3 for a sufficient number of times. For instance, in this case, 350 times

5. Finding the corresponding microstate model giving the maximum average mean-correlation with the test dataset

Once the optimal number of clusters is determined, we again apply the modified k-means algorithm [9] to produce the optimal EEG microstate-maps from the contaminated and non-contaminated EEG data segments.

Results using microstate analysis: It is necessary to find the optimal number of clusters that best explains the EEG dataset of four subjects. So to find the optimal-number of microstate clusters, the process mentioned in 4.2.1 is repeated at first from 7 to 20 times with an interval of 1. Table 4.1 shows the number of clusters for each number of repetitions.

Sr. No.	Number of repetitions	Number of clusters
1	7	10
2	8	9
3	9	8
4	10	9
5	11	6
6	12	8
7	13	10
8	14	9
9	15	10
10	16	10
11	17	9
12	18	10
13	19	6
14	20	8

Table 4.1: The number of repetitions of the process of determining the optimal number of microstate clusters from 7 to 20 times with an interval of 1

It is seen from the table 4.1 that the number of clusters varies from 6 to 10 for number of repetitions less than or equal to 20. So we take the number of repetitions from the range 25 to 350 with an interval of 25. Table 4.2 shows the number of clusters for each number of repetitions.

Sr. No.	Number of repetitions	Number of clusters
1	25	8
2	50	10
3	75	8
4	100	8
5	125	10
6	150	8
7	175	8
8	200	10
9	225	8
10	250	10
11	275	10
11	300	10
12	325	10
13	350	10

Table 4.2: The number of repetitions of the process of determining the optimal number of microstate clusters from 25 to 350 with an interval of 25

From the table 4.2 we see that with the increase in number of repetitions the number of clusters varies from 8 to 10. Figure 4.5 shows the of the number of clusters for number of repetitions in the range of 20 to 350 with an interval of 10.

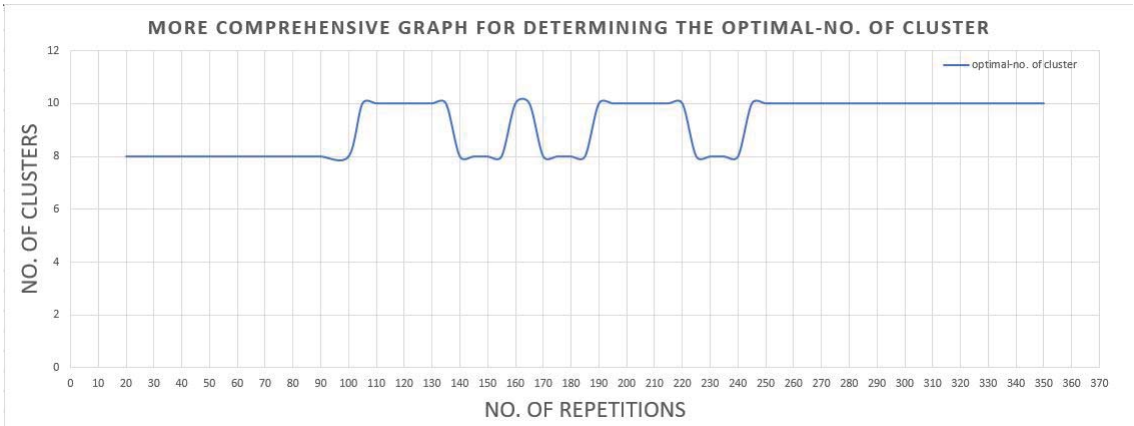

Figure 4.5: Number of clusters vs number of repetitions for 20 to 350 with an interval of 10

It is seen from the figure 4.5 that for number of repetitions ranging from 20 to 250 there is a fluctuation in the number of cluster from 8 to 10. However, this fluctuation does not exist as we increase the number of repetitions from 250 to 350 rather the number of clusters

becomes stable after 250 times. So we can conclude from this pattern that the maximum spatial correlation of the microstate model having 10 clusters occurs with the test dataset of 2 subjects from the "N-back" experiment (Chapter 3). Thus the optimal number of EEG microstate-clusters is ten. These 10 clusters best explain the dataset of four subjects. Figure 4.6 shows these 10 EEG microstate clusters or maps.

Figure 4.6: The 10 Optimal EEG microstate maps or clusters

4.2.2 Randomization statistical analysis

Randomization statistics [11] [12, see Chap 8] test the plausibility of the so-called null-hypothesis which postulates that the variance in the data is unrelated to some assumed structure in the data. Estimating the probability of this null-hypothesis is the goal of a statistical test. The null hypothesis is rejected when this probability is sufficiently low and the alternative hypothesis is taken i.e., accepting that the assumed structure of the data is probably related to the variance of the data and saying the fact that the assumed structure in the data has a significant effect in the data. This procedure requires two steps. The first step is to introduce test-statistics that measure to what extent the variance of the data is related to an assumed data structure [12, see Chap 8] i.e., the test statistics generate some size (magnitude) of the effect called effect-size of the assumed structure in the data. The second step is to estimate how likely, the occurrence of observed-effect size has been noticed due to chance.

In this step, randomization statistics constructs a distribution of the random effect sizes under the null hypothesis by destroying the assumed structure in the data. We break this structure in the data by randomly shuffling the data samples. This randomized data is the first instance of a set of observations made under the null hypothesis. Then the effect size is recomputed. This obtained effect size is one instance of the observed effect size obtained under the null hypothesis. We generate an empirical distribution of the effect sizes that is compatible with the null hypothesis with the repetition of randomizing the data many times and recomputing the effect sizes.

The observed effect size, before random shuffling of the data, is compared with the empirical distribution of the effect sizes formed under the null hypothesis. It provides how likely the observed effect size has occurred while the null hypothesis is valid. If the likelihood is sufficiently low, the alternative hypothesis is accepted. Then, the assumed data structure has a significant effect on data [12].

In this thesis, we randomly shuffle the data 1000 times and use 3 features of the EEG microstate clusters. These are Count of time points, Microstate Onset and Microstate Offset.

Count of time points: Total number of time points in the subject 1 data when a particular microstate was active and assigned to the EEG data [30].

Onset of microstate classes: The first time point of the first assignment of the EEG microstate class or map to the subject 1 EEG data [12, 30].

Offset of microstate classes: The last time point of the assignment of the EEG microstate map to the subject 1 EEG data [12, 30].

In this research, the statistical analysis is conducted in the following steps:

1. A sample population of EEG data is necessary: Sub-population comes from the conditions generated from the signal-power analysis, and it belongs to the group of epochs coming from adjacent channels of the highly EMG prone EEG data-channels

2. We generate the optimal number of EEG microstate maps using the EEG microstate analysis. (section 4.2.1)

3. We calculate three temporal features of the optimal EEG microstate maps. These are the count of time points, microstate onset, offset

4. The difference in microstate clusters with respect to all the features are computed group-wise to find the observed effect sizes

5. We randomly shuffle the data group-wise, and then the effect size is computed again to get the first instance of the random effect size. We repeat this process at least 1000/5000 times to get a distribution of random effect sizes

6. Now, we calculate the probability of how likely the observed effect occurs by chance. We divide the number of random effect-sizes that are equal to or larger than the observed effect-size by the total number of randomization runs

7. To obtain reliable results, having a critical p-value of 5%, 1000 randomization runs are needed. For the results with a p-value of 1%, we need 5000 randomization runs [12, see Chapter 8]

Purpose of randomization statistics: The purpose of randomization statistics is to investigate the statistical properties of the sub-population groups. Randomization statistics highlight the within-subject group error and between-subject group error. This method helps to find out the significant difference between the features of two or more samples from a population group. The primary requirement of this approach is computationally expensive. It requires at least 1000 randomization runs to obtain reliable results. However, this obstacle is becoming less and less due to the rapid growth of computing power. It is affordable with personal computers [12, see Chapter 8]. It also helps to allow the construction of custom-tailored tests for the specific research question of interest. Moreover, this approach is powerful because of fewer assumptions and gives improved performance over classical statistics. Many assumptions in the classical-statistics sometimes bias the research study. Hence, to remove the effect of bias and to obtain much-improved results, the randomization statistics approach has been adopted in this thesis [12, see Chapter 8].

Results of randomization statistical analysis:

In this thesis, we statistically analyze the individual subject 1 of the N-back experiment. This data is formed into two groups based on the data-channels (section 4.1.2) and frequency analysis (section 4.1.3). The two groups are:

1. EMG artifact data matrix

2. EMG free EEG data matrix

The results of the statistical analysis consist of calculating the null hypothesis probability and dividing the EEG microstate clusters into two categories:

1. Significantly different EEG microstate cluster or map

2. Significantly similar EEG microstate cluster or map

Table 4.3 shows the null hypothesis probability for the 10 optimal microstate clusters and the table 4.4 shows the corresponding EEG microstate clusters or map-labels..

Microstate Features	**Null-hypothesis probabilities of 10 microstate clusters**
Count of time points	0.6, 0.795, 0.573, 0.356, 0.191, 0.886, 0.519, **0.001**, 0.718, 0.919
Microstate Onset	**0.001, 0.038, 0.0**, 0.261, **0.046, 0.006,** 0.946, 0.228, **0.018,** 0.579
Microstate Offset	0.126, 0.986, 1.0, 1.0, 0.982, 1.0, 1.0, 1.0, 0.99, 1.0

Table 4.3: The null hypothesis probabilities of microstate clusters with respect to the three microstate quantifiers

Microstate Features	**Significantly Different**	**Significantly Similar**
Count of time points	**7**	0,1,2,3,4,5,6,8,9
Onset of Microstate classes	**0,1,2,4,5,8**	3,6,7,9
Offset of Microstate classes	null	0,1,2,3,4,5,6,7,8,9

Table 4.4: The microstate class or map labels with respect to the three microstate quantifiers

We have determined the category of the labels in table 4.4 by using the probability values mentioned in table 4.3. We reject the null hypothesis and label the EEG microstate class or map as significantly different when the null hypothesis probability of an EEG microstate cluster or map, is equal to or below 0.05 (since the number of randomization runs is 1000 in this case) [30]. On the contrary, if the null hypothesis probability is more than 0.05, we accept the null-hypothesis and label the EEG microstate map as significantly similar. Therefore, the main objective of the statistical analysis is to divide the EEG microstate clusters or maps, into two categories, namely, significantly-different and significantly-similar.

Moreover, this analysis returns the group-wise optimal EEG microstate maps before randomly shuffling the EEG-data. These maps are utilized to fit-back the preprocessed raw EEG data.

4.2.3 Fit-back of raw EEG data using microstate-map labels

From the statistical analysis, we determine significantly different map labels. We also obtain the optimal EEG microstate maps before randomly shuffling the two groups of EEG-data namely, EMG-contaminated and non-contaminated EEG-data (section 4.1.3). In table 4.4, the two categories of map-labels demonstrate the difference between the group-wise microstate clusters. So the non-contaminated EEG epochs are retained by back-fitting the preprocessed raw EEG data (section 3.3) with the significantly-different EEG microstate maps. The purpose of the fit-back is to find the spatial correlation [45](section 4.2.1) between the microstate clusters (maps) obtained from significantly different maps labels and the instantaneous EEG topography of the preprocessed raw EEG data.

It looks for the best spatial-correlation between the optimal EEG microstate maps and the instantaneous EEG data, with at most 3-standard deviations. This process of fit-back also labels each time point in the instantaneous EEG-data with the microstate cluster-label.

In this thesis, we apply this fit-back technique two times. First, we fit-back the EEG microstate maps from group 2 (EMG non-contaminated) to the preprocessed-raw EEG data using the spatial correlation technique. Next, we extract data-segments of the preprocessed raw EEG data fitted with significantly different microstate maps. We track the time-points in the data, labeled with significantly different EEG microstate map labels for the data-extraction process. This process gives us a back-fitted data matrix, that is, the EMG-artifact free EEG data.

However, we could not fit all the time points in the data with the microstate maps. So the corresponding data of those time points are kept in a first-level residual data matrix

called the Rest of back fitted data without the significantly different map labels. Secondly, we fit-back the EEG microstate maps from group 1 (EMG-contaminated) to this residual data matrix using the spatial correlation technique. Then we extract data-segments of the residual data matrix fitted with significantly similar microstate maps to form the EMG-contaminated EEG data matrix. Figure 4.7 shows the flow chart of the process of fit back.

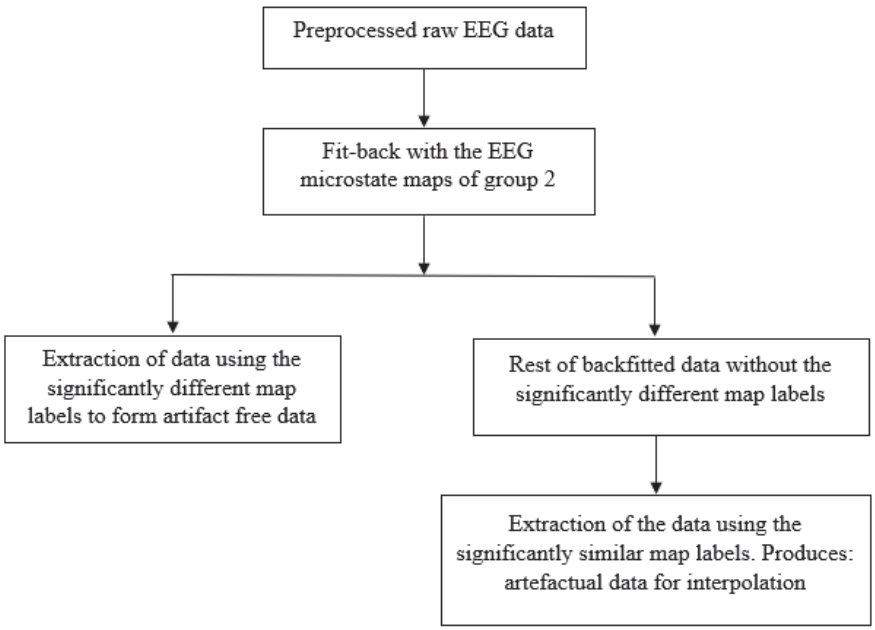

Figure 4.7: The overview of the process of fit-back in EMG-artifacts removal method

Technically, not all time points in the residual data matrix can be fit with the microstate maps. So the corresponding EEG data of those time points are separated and kept in a data matrix called the "residual-error" matrix. This residual-error matrix of EEG data with EMG-artifacts is left behind for analysis. Those data are the errors in the proposed model of the removal of EMG artifacts. In brief, the proposed model can be as follows:

$$Real EEG = Artifacts\, free\, EEG + Artefactual\, EEG + Error \tag{4.6}$$

4.2.4 Interpolation of the EEG data segments

In the randomization statistics analysis, significantly similar map labels are the indicator of the contaminated data segments or epochs (Artefactual EEG of equation 4.6). One solution is to reject those segments. However, this will cause the loss of underlying neural information in those EEG data-segments or epochs. One alternative approach is to use the spherical spline method detailed in [51]. This method projects the EEG-sensor (electrode) locations onto a unit sphere. It then interpolates the EEG-signal at the bad sensor locations based on the signals at the good locations. The process to interpolate EEG data, using spherical spline-interpolation consists of the following steps [37]:

1. To project the good and bad electrodes onto a unit sphere
2. To compute a mapping matrix that maps N good-channels (electrodes) to M bad-channels
3. To use this mapping matrix to compute interpolated data in the bad-channels

In this thesis, the contaminated epochs have a high spatial-correlation with significantly similar microstate clusters or maps. So we have used the MNE-Python method mne.io.raw.interpolate_bads [37] to interpolate those EEG epochs. This method automatically applies the correct-method (spherical splines or field interpolation) to EEG data. In this way, the interpolation technique preserves the underlying neural activity of the EEG signal. This technique also constructs the interpolated EMG-artifact free EEG data matrix.

4.2.5 Data reconstruction

To reconstruct the EMG-artifact free EEG data, we join the two data matrices. These are:

1. The back fitted data matrix that is, the EMG artifact-free EEG data (section 4.2.3)
2. The interpolated EMG-artifact free EEG data matrix (section 4.2.4)

We have used the NumPy library of Python [44,50] to concatenate the two data matrices and to form the EMG-artifact free EEG data.

4.3 Discussion

We first plot the power spectral density (PSD) of the raw EEG data contaminated with EMG-artifacts in the 45-70 Hz frequency band to differentiate between raw EEG data with EMG artifacts and EMG-artifacts free EEG data. The power spectral density (PSD) measure the content of the signal-power against frequency. The spectral density characterizes the frequency components of the signal [52]. We have used the mne.io.raw.plot_psd function of the MNE-Python [37] to plot the PSD of the EEG signals. Figure 4.8 shows a sample PSD plot of 16-channel raw EEG-data with EMG-artifacts without any preprocessing and application of the proposed method.

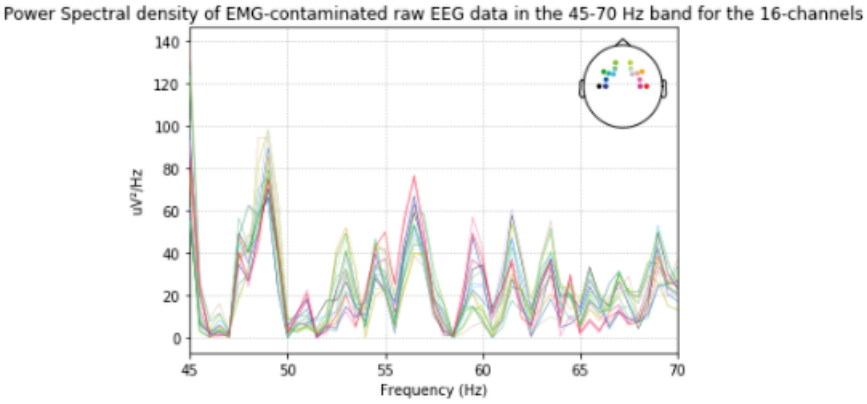

Figure 4.8: A sample PSD plot of the raw EEG data with EMG-artifacts

In figure 4.8, the x-axis represents frequencies from 45 to 70 Hz, and the y-axis is the power-amplitude of the EEG signals in units of $amplitude^2/Hz$. The round-shaped object in figure 4.8 represents the top-view of the brain. The 16 color-dots represent the 16 channels that we have used for EMG-artifacts analysis and removal. We observe the high signal power-amplitude peaks from figure 4.8. The peaks in figure 4.8 show the presence

of EMG-artifacts. The amplitude-peak values range from 20 $(micro\,volt)^2$ to almost 100 $(micro\,volt)^2$ for each frequency in the 45-70 Hz band.

On the contrary, we also plot the PSD of the EMG-artifact free EEG data obtained with the proposed method to observe the power-amplitude peaks. Figure 4.9 similarly shows a sample plot of 16-channel EMG-artifact free EEG-data generated from the proposed method with x-axis as the frequency and y-axis as the signal power-amplitude.

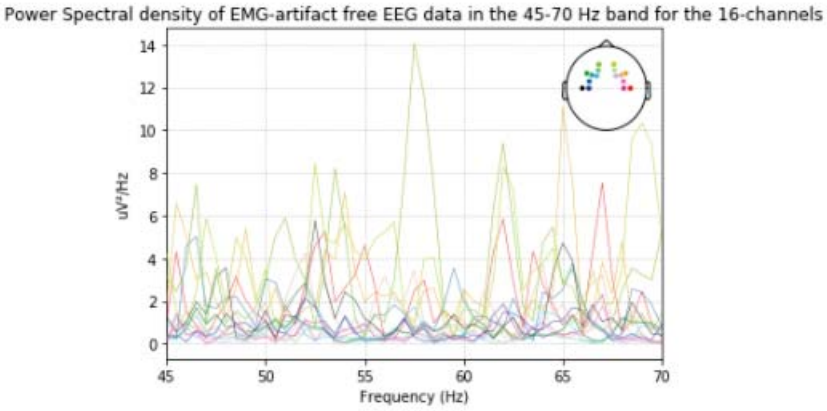

Figure 4.9: A sample PSD plot of the EMG-artifact free EEG data using the proposed method

We observe from figure 4.9 that the power-amplitude peaks are relatively smaller than those of figure 4.8. The peak values range from 2 $(micro\,volt)^2$ to almost 14 $(micro\,volt)^2$ for each frequency in the 45-70 Hz band. Thus the proposed method has removed the EMG artifacts and reconstructed the EEG data and matches the postulates mentioned in [7, 13]. The algorithm 2 summarizes the whole process of the proposed method of EMG-artifact removal and reconstruction of EEG-data.

Algorithm 2: The procedure to remove EMG-artifacts

1. Load raw EEG dataset of n subjects (N-back experimental dataset chapter 3).
2. Complete the steps 1-10 of algorithm 1 to separate the EMG-contaminated and non contaminated EEG-epoch arrays.
3. Repeat step 2 for $n - 1$ times to store n-subjects EMG-contaminated and non contaminated EEG data arrays.
4. For each group-data of n-subjects randomly select n-subjects and split this n-subject data into half to form test and train dataset.
5. Take mean of the test and train dataset to form mean-test data and mean-train data.
6. Input: Number of clusters, n
Run microstate-algorithm on the mean-train data to cluster the data into given n microstates.
7. Calculate spatial correlation of the n microstates with the mean-test data.
8. Repeat the steps 6-8 enough times (18 times for say) for each number of clusters in the range from 3 to 20.
9. For two groups take the average of the two correlation values.
10. Repeat the steps 4-8 for enough times (350 times for say).
11. Calculate mean-average correlation value for each cluster (microstates).
12. Calculate cluster-number having maximum correlation value to find optimal clusters.
13. Load group EEG-epochs of single subject data.
14. Cluster both groups-data into optimal-microstates to get two sets of microstates.
15. Calculate three temporal-features of each microstates for two groups of data.
16. Calculate difference between three temporal features of each group-microstate i.e. the observed effect size.
17. Randomly shuffle group EEG-epochs and repeat steps 13-15 to obtain a random instance of the observed-effect size i.e., rand-effect size.
18. Repeat step 16 many times (1000 times) to form a distribution of the rand-effect size.
19. Calculate how many rand-effect instances are $>=$ to observed effect-size.
20. Probability of null hypothesis = $\frac{number\ of\ random-instances}{total\ no.\ of\ repetitions}$
21. If probability$<=$ 0.05, microstate-label = significant different else, microstate-label = significantly-similar.
22. Fit-back the preprocessed raw EEG data of subject 1 (Step 3 of Algorithm 1) with significantly similar and different microstate-labels.
23. Construct EMG-contaminated EEG data for interpolation and back-fitted EEG data.
24. Interpolate EMG-contaminated EEG data using Legendre Polynomial Expansion (spherical spline interpolation) to form interpolated EEG data.
25. Output: EMG-artifact free EEG data by joining the pure and interpolated EEG data.

Chapter 5

Validation Of the Proposed Method

To validate the method proposed in chapter four, we need to compare the EMG artifacts-free EEG data obtained by this proposed-method and another standard method of EMG-artifacts removal. One approach for making comparisons is to use the simulated EEG data with a controlled amount of noise. Besides, we can compare with real EEG data from verified EEG databases. Moreover, we can make a basic comparison with the method ICA combined with multiple artifact rejection algorithm (MARA).

There are many other methods for EEG artifacts removal. However, in this thesis, we chose the ICA-MARA method for its well-established results. Also, the ICA method is classical and well recognized in the field of EEG artifacts removal. The ICA with the MARA method is available as a software package in MATLAB. The software-package is the MARA toolbox developed from EEGLAB [53]. This toolbox is well known and accepted in the EEG signal processing domain. Thus it becomes an ideal baseline-candidate for making a comparison of EEG signals despite having no ground truth data.

This comparison with the baseline demonstrates the relative performance of the proposed method in the detection and removal of frontalis and temporalis scalp muscle contaminations (EMG-artifacts) from the EEG recordings. So this chapter validates the proposed method by comparing its performance-results with that of independent component

analysis ICA combined with MARA [14] in removing muscle artifacts from the N-back experimental subject 1 data (Chapter 3).

5.1 ICA with multiple artifact rejection algorithm

For the EEG signals, we can model the brain's electrical activity as a mixture of underlying electric potential source components. In the ICA with the MARA method, the EEG research-experts train classifier using large datasets. The theoretical basis of this method, is that the EEG signal-sources are statistically independent. As such, one can separate these sources from the mixture of EEG potentials. This type of problem is often called a blind source separation (BSS) problem. The researchers separate the source signals statistically to solve the problem.

In the ICA model, the underlying assumption is that the raw EEG data, $X(t)$ obtained in the time domain can be expressed as

$$X(t) = [x_1(t), x_2(t), ..., x_R(t)]^T (t = 1, ..., N), \tag{5.1}$$

where N is the number of sample points and R is the number of channels. The inherent independent components are $S(t) = [s_1(t), s_2(t), ..., s_R(t)]^T$. Thus the mixing model can be as follows:

$$X(t) = WS(t) \tag{5.2}$$

Here, W is the mixing matrix. In general, both the independent components and the mixing-matrix are unknown. We estimate the independent-components and the mixing-matrix, from the mixture $X(t)$ with general assumptions. These assumptions are spatial statistical independence among the independent components and non-Gaussian distribution of the independent components of the EEG signal mixture matrix $X(t)$.

Besides, this method always assumes the mixing-matrix to be square. After estimating the mixing-matrix, its inverse, W^{-1} is also computed. This can be shown by:

$$S(t) = W^{-1}X(t) \tag{5.3}$$

For the removal of EEG artifacts, we can model the EEG data as a summation of neural and artifactual sources. The surrounding noise sources can be assumed to be independent of the underlying brain sources.

In this model we transform the input vectors of the mixture signal $X(t)$ into a signal space where the signals are statistically independent. After this transformation of the mixed EEG signals, the neural sources (independent components) are selected and reconstructed without the artifactual components to create the artifacts free data. The selection process of independent components(ICs) is tedious and not automatic. It differs from expert to expert.

To resolve this user-dependent issue, we selected the automatic-method named Multiple Artifact Rejection Algorithm (MARA) for comparison with the method proposed in chapter 4. The ICA with the MARA method is described in detail in [14]. It is an efficient and reliable process for the detection of all classes of artifacts, such as muscle artifacts (EMG-artifacts). This method constructs six features and incorporates the temporal, spectral, and spatial domain information of ICA components. MARA [14] is an open-source plug-in for a vastly used graphic user interface named EEGLAB [53] for automatically hand labeling the independent-components (ICs) of the EEG signals. However, ICA is sensitive to the slow-drifts. So we filter the data from 0.1 Hz to 100 Hz offline before segmentation and artifact removal process.

Moreover, to analyze the data with a stable, effective and robust version of ICA that avoids over learning the formula for choosing the number of data points in a dataset should be at least 30 * (no. of channels)^2. For example, in our case the EEG data of N-back experimental tasks having 64 data-channels, the number of samples will be $30 * (64)^2 = 122,880$

i.e., 240 seconds of the EEG recordings with a sampling frequency of 512 Hz to reliably decompose the EEG data with ICA.

Substantial information is necessary for the extraction of features in this method. At first, the ICA is applied to the data to generate the independent components. Then the MARA method constructs an initial feature set of 38 features as candidates based on the characteristics of the independent components obtained from the EEG signals. Among these features, 13 come from the component's time series, 9 from the spectrum, and 16 from the component's pattern. A feature selection process as detailed in [14] chose the following six features out of these 38 features.

- k_1, λ, k_2 and fit error

  The k_1, λ, k_2 and fit error are parameters to explain the deviation of a component's spectrum from a prototypical 1/frequency curve mentioned in [14], which is:

$$f \Rightarrow \frac{k_1}{f^{\lambda}} - k_2 \tag{5.4}$$

  These parameters k_1, λ, k_2 can be calculated using three points of the log spectrum of the curve [14]. These are:

  1. The log power at 2 Hz i.e., k_1
  2. Local minimum in the 5-13 Hz i.e., λ
  3. Local minimum in the 33-39 Hz i.e.,k_2

- Alpha and gamma band feature at 8-13 Hz and 31-45 Hz

  These features indicate the average log band power of alpha-band from 8 to 13 Hz and gamma-band from 31-45 Hz band. The alpha peak of the independent component having neural origin can be detected using the feature 8-13Hz [14].

- Range within pattern

  ICA with the MARA method defines the range within a pattern in the scalp maps of the independent components of the EEG signals. This pattern is the logarithm of the difference between the minimal and maximal activation [14]. A broad range within this pattern indicates that the loose electrodes or muscle artifacts generate spatially located scalp maps.

- Current density norm

  ICA can not provide information about the locations of the inherent-neural sources. So it is estimated by the EEG potentials in ICA scalp map.

- Mean local skewness

  Mean local skewness is a time series feature of the components. It is calculated by the mean absolute local skewness of time intervals with a duration of 15 seconds, to detect outliers in the time domain.

After classifying the artefactual components by MARA, we use the graphical user interface(GUI) of the EEGLAB plugin of MATLAB [53] to visualize all the independent-components. It also provides a calibration for artefactual component rejection. In this thesis, we use 64 channels-dataset. Thus, we have 64 independent components for 64 channels. We classify the components by the MARA algorithm to detect and remove the artifacts using the EEGLAB plugin of MATLAB [53].

Comparison of results using PREP preprocessing pipeline: We have analyzed the EMG artifacts free EEG-data obtained from both the methods to find bad-channels. We have used a standardized early-stage preprocessing pipeline (PREP) as detailed in [54]. The bad-channels are EEG data channels having a low signal to noise ratio and very low or no-signal throughout a considerable time of the EEG recording [55]. We use this pipeline to

find the bad-channel ratio. The bad channel-ratio is the ratio of bad data-channels identified in the PREP analysis to the total number of data-channels in the data. This pipeline uses four complex algorithms [54, 55], and four criteria to identify bad EEG data-channels. These are:

1. The extreme amplitude of EEG data (the deviation criterion)

2. Lack of correlation of an EEG data-channel with any other channel (the correlation criterion)

3. Lack of predictability of an EEG data-channel by other channels, that is, the predictability criterion using RANSAC (random sample consensus) [54]

4. unusual high-frequency noise (the hf noise criterion) [54]

In addition to these criteria the algorithm of the PREP pipeline detects bad channels by the EEG data-channels having any NaN (not a number) data called the NaN criterion and significant time periods with constant values or very small values called the flat criterion. The analysis results obtained from the proposed method and the ICA with MARA method are shown in table 5.1 and 5.2.

Bad-channel criterion	**Number of channels**	**Channel name**
NaN	0	none
flat	0	none
deviation	0	none
hf noise	0	none
correlation	0	none
RANSAC	0	none

Table 5.1: PREP analysis of the EMG free data obtained from the proposed method.

Bad-channel criterion	Number of channels	Channel names
NaN	0	none
flat	**4**	**C5, C6, FC5, FC6**
deviation	**2**	**Fp1, Fp2**
hf noise	0	none
correlation	0	none
RANSAC	0	none

Table 5.2: PREP analysis of the EMG free data obtained from the ICA with MARA method.

We see the proposed method performed much better than the method ICA with MARA in the PREP analysis results. The ratio of the bad-channels in the ICA-MARA method is $6/16 = 0.375$. Here 16 is the total number of channels used for the overall analysis. On the contrary, this ratio is 0 for the proposed method. Also, the RANSAC-analysis of [54] fails in the data obtained from the method ICA with MARA. The PREP-analysis [54] of this data shows that a few channels are available to perform the RANSAC method reliably. This analysis also indicates that too many channels failed the quality tests described in [54]. These quality tests determine whether the EEG data-channels are good or bad. In the case of the proposed method, all the EEG data-channels successfully passed the quality tests described in [54].

5.2 Quality metrics of the EMG-artifacts free EEG data

One way of evaluating the quality of the EEG data after the removal of EMG artifacts is to identify the bad EEG data-channels by measuring some parameters. As such, it is necessary to define the criteria for the evaluation of the parameters of the bad channels. We can define the bad channels as those channels that we need to interpolate after the application of the preprocessing steps. The bad channels can have no variation for longer than 5 seconds. These have a small signal to noise ratio or even no signal-amplitude for a considerable time. These are detected based on parameters like deviation, correlation, predictability, and noisiness to the other channels [55, 56]. The main preprocessing steps for the detection of bad channels are filtering with 1Hz high pass filter, removal of power line noise, and application of PREP pipeline [54]. Despite these steps, the EEG data gets affected when we high-pass filter the EEG signals. So researchers have proposed many metrics to serve the purpose. However, we can divide the evaluation metrics broadly into two groups:

1. To measure how well the artifact removal method (algorithm) eliminates the artifact-interference: Degree of artifact removal

2. To quantify how well the algorithm preserved EEG data: Degree of signal preservation

In this thesis, we consider the second metric that tries to quantify the degree of signal preservation and signal quality. For measuring the quality of the data, we calculate the following three metrics [55] from the EMG free data for assessing the performance of the proposed method in comparison with the ICA-MARA method. These are [55]:

Ratio of data-points with overall high amplitude (OHV): The quality measure of the overall high amplitude (OHA) is calculated by the ratio of data-points, d (that is, channels c multiplied by time points t) that have a higher absolute voltage magnitude v of a threshold

value, x μV [55].

$$OHA(x) = \frac{1}{N}\sum_{d}^{N} |v|_d > |x| \tag{5.5}$$

Where, d is data points (no. of channels c multiplied by time points t) and $|x|$ reflects a vector of voltage magnitude thresholds e.g., x = 10 μ V, 20 μ V, 30μ V, 40 μ V, 50 μ V, 60 μ V, 70 μ V, 80μ V, 90 μ V and N reflects the number of data points. Thus, each $OHA(x)$ threshold results in a quality measure that differs in its sensitivity [55].

Ratio of time points with high variance (THV): Similarly, we identify the ratio between time-points t and total time points T, in which the standard deviation σ of the voltage measures v across all channels c exceeds x μV [55].

$$THV(x) = \frac{1}{T}\sum_{t}^{T} \sigma_t\left(v_c\right) > |x| \tag{5.6}$$

Where, time point is t, in which the standard deviation σ of the voltage measures v across all channels c is more than $|x|$ μV, where $|x|$ reflects a vector of standard deviation thresholds and T is the number of time points.

Ratio of channels with high variance (CHV): The same logic applies to the ratio of EEG data-channels, for which the standard deviation σ of the voltage v measures across all time points t exceeds x μV. The channels of high variance (CHV) criterion reflects this ratio. [55]

$$CHV(x) = \frac{1}{c}\sum_{c}^{C} \sigma_c\left(v_t\right) > |x| \tag{5.7}$$

Where, C is total no. of channels, $\sigma_c(v(t)$ is the standard deviation of the voltage v measures of across all time points t greater than $|x|$ μV.

Comparison results: We compare the results obtained from the ICA with the MARA method and the method proposed in chapter 4. The threshold values are $x = 10$ μV, 20 μV, 30 μV, 40 μV, 50 μV, 60 μV, 70 μV, 80μV, 90 μV. Tables 5.3 and 5.4 show the quality metric values for the data obtained from the proposed method in chapter 4 and the method ICA with MARA.

Threshold values (μV)	**OHV**	**THV**	**CHV**
10	14.992058	16.974569	26.463448
20	11.198804	11.307124	20.680853
30	8.357876	8.063291	10.906444
40	6.484651	6.166481	8.729885
50	5.244097	4.322459	3.151252
60	4.356705	2.627016	**0.0**
70	3.654187	1.484150	**0.0**
80	3.080895	0.932752	**0.0**
90	2.642685	0.633356	**0.0**

Table 5.3: The quality metric values obtained from the proposed method

Threshold values (μV)	**OHV**	**THV**	**CHV**
10	0.181556	0.040067	0.0
20	0.049516	0.001875	0.0
30	0.010762	0.0	0.0
40	0.003042	0.0	0.0
50	0.001310	0.0	0.0
60	0.000279	0.0	0.0
70	4.369731e-05	0.0	0.0
80	0.0	0.0	0.0
90	0.0	0.0	0.0

Table 5.4: The quality metric values obtained from the method ICA with MARA

Decently in terms of data quality, the method ICA with MARA performed very well in comparison to the proposed method. However, for the more relaxing threshold values like 30 or more than 30 μV, the two-method give almost the same results. We observe from table 5.3 that for a threshold value of 40 or 50μV, the quality-metric values ranged from 3 to 9. As the threshold value increased, the ratios decreased for the proposed method. The decreasing-pattern indicates that the error-reduction of the proposed EMG-artifacts removal method will generate more similar results to that of the method ICA with MARA. The proposed method is as effective as ICA-MARA method for CHV threshold values in the range of 60 to 90. Hence, for relaxing threshold [55] values like 40 μV or 50 μV or higher than that the two-methods render similar results in terms of data quality.

5.3 Discussion

The comparison of the proposed method with ICA combined with multiple-artifact rejection method is necessary to prove the effectiveness of it. The comparison results demonstrate the effectiveness of the proposed process. As such, we have validated the proposed muscle artifact removal method. Besides, to show the effectiveness of the proposed method, we have done a quantitative assessment of the EEG recordings after the removal of muscle artifacts. We have evaluated the quality of the EEG recordings after removing the EMG-artifacts using three data quality-measurement metrics [55]. However, a criterion should be set to decide whether the EEG recordings after the removal of muscle artifacts can be acceptable or not. It is necessary to check the recording length to determine such criteria. On top of that, such decision highly relies on the number of trials, sample size, effect sizes of interest, and to some extent, on specific research questions. The users or researchers can classify the data-quality into three categories [55]. The categories are:

1. Good, meaning the EEG-data is very clean

2. Ok, meaning the EEG-data is relatively clean

3. Bad, meaning the EEG data has residual-noisy channels even after correction of artifacts [55]

From our comparison, it is clear that the proposed method has performed very well in terms of the bad-channel ratio. However, the ICA with the MARA method has not performed well. Moreover, we can classify the performance of the proposed method with the ICA-MARA into three categories: Good, Ok, and Bad for data quality measures like THV, OHV, and CHV [55]. The performance of the proposed method is "Bad" for strict threshold values like 10-30μV. This performance is "Ok" for relaxing threshold values like 40-60μV and is "Good" for more relaxing threshold values like 70-90μV.

It is clear from tables 5.3 and 5.4 that in the case of experiments with stricter threshold values, our proposed technique will not be optimal. On the other hand, since ICA-MARA requires a certain number of data-points, for EEG-experiments with limited-data, it may not be as effective as our technique. For the EEG case-studies, where quick response is necessary with limited EEG data points, the proposed method can potentially be more effective over the ICA-MARA method to remove EMG artifacts from EEG recordings.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

The muscle artifacts contaminate the EEG signals over a broad range of frequencies. This contamination distorts the original neural signals. While signal acquisition during any complex-cognitive experiment, there is the contraction and expansion in the frontalis and temporalis scalp-muscles of the subject. This contraction and expansion of scalp-muscles can be with or without intention. As a result, the underlying EEG recordings get contaminated with muscle artifacts. Hence, to analyze the EEG signals, it is essential to get rid of these contaminants. Thus one solution to this problem is to conduct an analysis of the raw EEG data-segments and to remove the EMG artifacts.

In this research, we have developed a simple graphical user interface to process the raw EEG data. This raw EEG data can be in the "European Data Format (EDF)" or "Biosemi Data Format (BDF)" or "Brainvision " data format. After loading the raw EEG-data, we conduct the analysis (detection) of the EMG contamination on the frontalis and temporalis brain regions. We use the concept of EEG signal-power analysis in the frequency domain for this purpose.

In the preliminary step of this analysis (detection), we calculate the EEG signal-power in the 45-70 Hz band [13]. We chose this band as EEG signals have low amplitudes in this band [13]. If the signal-power amplitude is high in this band, it is most likely a muscle activity [13]. We select a threshold-value of signal-power amplitude to detect the muscle-artifacts. The threshold-value is the mean of the signal-power amplitudes of one-hundred EEG data-segments (epochs) obtained from the primary EEG data-channels AF7, AF8, FT7, FT8 respectively of the real experimental EEG data. Each EEG epoch is two seconds long, having 1024 samples. We examine the signal-power amplitude of each EEG epoch when the signal power-amplitude of that epoch in any of the primary channels, exceeds the threshold. If the value exceeds the threshold value, we detect the EEG-epoch as EMG-artifacts contaminated epochs. In this frequency-domain analysis, high signal-power amplitudes in the 45-70 Hz band of the EEG signals in our obtained results prove the existence of the EEG data segments contaminated with muscle artifacts.

For the process of removal of the EMG-artifacts, we have used the EEG microstate analysis. This analysis determines the optimal number of EEG microstate clusters or maps in the given EEG dataset. We cluster the two groups of EMG-contaminated and non-contaminated EEG epochs and calculate three temporal microstate features. Secondly, we have used the concept of randomization statistics to find the significantly different EEG microstate-maps between these two groups.

In this analysis, we calculate the null-hypothesis probability of each EEG microstate map. We determine this probability based on the difference in the temporal features of the microstates. We reconstruct the EMG artifact-free EEG data back-fitting the preprocessed raw EEG-data with the EEG microstate maps from both the groups "EMG non-contaminated" and "EMG-contaminated" having the significantly "different" and "similar" EEG microstate-maps and interpolating the contaminated EEG data, using the Legendre polynomial expansion (spherical spline interpolation) technique.

In this study we have tried to detect the EMG-contaminated EEG data by applying signal-frequency analysis and removed those by extracting the characteristics of the EMG-free EEG data generated from signal-frequency analysis. We have compared our technique with ICA combined with multiple artifact rejection (MARA). We find that our proposed procedure is more effective than ICA with MARA in terms of bad-channel ratio. It is as effective as ICA with MARA when the data-quality quality measures like THV, OHV, and CHV for the threshold values in the range of 40 to 90 μV [14].

6.2 Future work

The removal of muscle artifacts from the EEG recordings is a classical research problem in the field of EEG signal processing. For many years different researchers have adopted many approaches for the removal of muscle artifacts in EEG signals. Unfortunately, it is hard to find a single best method that is 100% efficient in removing the muscle artifacts from EEG data. Apart from the traditional signal processing techniques like filtering, blind source separation, researchers have combined two or more methods to detect and remove the muscle artifacts. For example, the merging of independent component analysis and surface Laplacian [57].

However, research subjects can produce scalp-muscle movements from different parts of the brain. On top of that, the EEG data varies from person to person. For future research, one can apply the proposed method to analyze a large EEG-database. The database should have more research subjects and variable experimental conditions to detect and remove the frontalis and temporal scalp-muscles contamination of EEG data (EMG-artifacts). More microstate features can be computed, in addition to the proposed ones in this thesis, to understand and investigate the characteristics of EMG-artifacts.

In this thesis, the implementation of the proposed method is a prototype, which can be improved to do an extensive-analysis of average-subject EEG data. Moreover, we have

analyzed 16 EEG data-channels and set the mean value of the signal-power amplitude of the EEG data-segments (epochs) as the threshold. Instead of using this threshold-value other statistical thresholds, for example, the median, the standard-deviation of epoch-powers can be applied.

Bibliography

[1] B. Blaus, "Medical gallery of blausen medical 2014," *Wiki J. Med*, vol. 1, no. 10, 2014.

[2] O. W. Biosemi, "Headcaps," *https://biosemi.com/headcap.htm*.

[3] L. Hu and Z. Zhang, *Introduction: EEG Signal Processing and Feature Extraction*, pp. 1–5. Singapore: Springer, 2019.

[4] J. W. Britton, L. C. Frey, J. L. Hopp, P. Korb, M. Z. Koubeissi, W. E. Lievens, E. M. Pestana-Knight, and E. K. St. Louis, "Electroencephalography (EEG): An introductory text and atlas of normal and abnormal findings in adults, children, and infants," 2016.

[5] J. A. Urigüen and B. Garcia-Zapirain, "EEG artifact removal—state-of-the-art and guidelines," *Journal of Neural Engineering*, vol. 12, p. 031001, apr 2015.

[6] W. De Clercq, A. Vergult, B. Vanrumste, W. Van Paesschen, and S. Van Huffel, "Canonical correlation analysis applied to remove muscle artifacts from the electroencephalogram," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 12, pp. 2583–2587, 2006.

[7] I. Goncharova, D. McFarland, T. Vaughan, and J. Wolpaw, "EMG contamination of EEG: spectral and topographical characteristics," *Clinical Neurophysiology*, vol. 114, no. 9, pp. 1580 – 1593, 2003.

[8] A. Hyvärinen, "Independent component analysis: recent advances," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 371, no. 1984, p. 20110534, 2013.

[9] R. D. Pascual-Marqui, C. M. Michel, and D. Lehmann, "Segmentation of brain electrical activity into microstates: model estimation and validation," *IEEE Transactions on Biomedical Engineering*, vol. 42, no. 7, pp. 658–665, 1995.

[10] P. Nguyen, T. A. Nguyen, and Y. Zeng, "Quantitative analysis of the effort-fatigue tradeoff in the conceptual design process: a multistate EEG approach," in *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, vol. 50190, p. V007T06A020, American Society of Mechanical Engineers, 2016.

[11] B. Manly, "Chapman & hall/crc; boca raton, fl: 2007," *Randomization, bootstrap, and Monte Carlo methods in biology*, 2007.

[12] C. M. Michel, T. Koenig, D. Brandeis, L. R. R. Gianotti, and J. Wackermann, *Electrical Neuroimaging*. Cambridge University Press, 2009.

[13] M. J. Fu, J. J. Daly, and M. C. Cavusoglu, "A detection scheme for frontalis and temporalis muscle EMG contamination of EEG data," in *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 4514–4518, Aug 2006.

[14] I. Winkler, S. Haufe, and M. Tangermann, "Automatic classification of artifactual ICA-components for artifact removal in EEG signals.," *Behavioral and Brain Functions*, vol. 7, no. 1, pp. 30 – 44, 2011.

[15] D. A. Kaiser, "Basic principles of quantitative EEG," *Journal of Adult Development*, vol. 12, no. 2-3, pp. 99–104, 2005.

[16] T. A. Nguyen and Y. Zeng, "Analysis of design activities using EEG signals," vol. 44137, pp. 277–286, 2010.

[17] T. A. Nguyen and Y. Zeng, "Clustering designers mental activities based on EEG power," *Tools and methods of competitive engineering, Karlsruhe, Germany*, 2012.

[18] L. SA, "An introduction to the event-related potential technique," *The MIT Press*, pp. 7–21, 2005.

[19] W. BARRY and G. M. JONES, "Influence of eye lid movement upon electro-oculographic recording of vertical eye movements," *Aerospace medicine*, vol. 36, pp. 855–858, 1965.

[20] E.-R. Symeonidou, A. Nordin, W. Hairston, and D. Ferris, "Effects of cable sway, electrode surface area, and electrode mass on electroencephalography signal quality during motion," *Sensors*, vol. 18, p. 1073, Apr 2018.

[21] C. Brown, "EEG ARTEFACTS," *https://www.slideshare.net/ManchesterEEG/EEG-artefacts*, dec 2011.

[22] R. J. Croft and R. J. Barry, "Removal of ocular artifact from the EEG: a review," *Neurophysiologie Clinique/Clinical Neurophysiology*, vol. 30, no. 1, pp. 5–19, 2000.

[23] B. W. McMenamin, A. J. Shackman, J. S. Maxwell, D. R. Bachhuber, A. M. Koppenhaver, L. L. Greischar, and R. J. Davidson, "Validation of ICA-based myogenic artifact correction for scalp and source-localized EEG," *NeuroImage*, vol. 49, no. 3, pp. 2416 – 2432, 2010.

[24] B. W. McMenamin, A. J. Shackman, L. L. Greischar, and R. J. Davidson, "Electromyogenic artifacts and electroencephalographic inferences revisited," *NeuroImage*, vol. 54, no. 1, pp. 4 – 9, 2011.

[25] M. M. N. Mannan, M. A. Kamran, and M. Y. Jeong, "Identification and removal of physiological artifacts from electroencephalogram signals: A review," *IEEE Access*, vol. 6, pp. 30630–30652, 2018.

[26] M. K. Islam, A. Rastegarnia, and Z. Yang, "Methods for artifact detection and removal from scalp EEG: A review," *Neurophysiologie Clinique/Clinical Neurophysiology*, vol. 46, no. 4, pp. 287 – 305, 2016.

[27] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[28] D. Safieddine, A. Kachenoura, L. Albera, G. Birot, A. Karfoul, A. Pasnicu, A. Biraben, F. Wendling, L. Senhadji, and I. 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.

[29] L. Shoker, S. Sanei, and J. Chambers, "Artifact removal from electroencephalograms using a hybrid BSS-SVM algorithm," *IEEE Signal Processing Letters*, vol. 12, no. 10, pp. 721–724, 2005.

[30] T. Koenig, M. Stein, M. Grieder, and M. Kottlow, "A tutorial on data-driven methods for statistically assessing ERP topographies," *Brain topography*, vol. 27, pp. 72–83, 01 2014.

[31] R. W. Homan, J. Herman, and P. Purdy, "Cerebral location of international 10–20 system electrode placement," *Electroencephalography and clinical neurophysiology*, vol. 66, no. 4, pp. 376–382, 1987.

[32] O. W. BioSemi, "Frequently asked questions (FAQ) amplifier gain," *https://www.biosemi.com/faq/adjust_gain.htm*.

[33] O. W. BioSemi, "Actiview," *https://biosemi.com/download.htm*.

[34] O. W. Biosemi, "Frequently asked questions (FAQ) CMS and DRL," *https://biosemi.com/faq/cms&drl.htm*.

[35] R. Oostenveld and P. Praamstra, "The five percent electrode system for high-resolution EEG and ERP measurements," *Clinical neurophysiology*, vol. 112, no. 4, pp. 713–719, 2001.

[36] A. Gevins, M. E. Smith, H. Leong, L. McEvoy, S. Whitfield, R. Du, and G. Rush, "Monitoring working memory load during computer-based tasks with EEG pattern recognition methods," *Human Factors*, vol. 40, no. 1, pp. 79–91, 1998. PMID: 9579105.

[37] A. Gramfort, M. Luessi, E. Larson, D. A. Engemann, D. Strohmeier, C. Brodbeck, R. Goj, M. Jas, T. Brooks, L. Parkkonen, *et al.*, "MEG and EEG data analysis with mne-python," *Frontiers in neuroscience*, vol. 7, p. 267, 2013.

[38] J. S. Barlow, "Artifact processing (rejection and minimization) in EEG data processing," *Handbook of electroencephalography and clinical neurophysiology. Revised series*, vol. 2, pp. 15–62, 1986.

[39] D. W. Klass, "The continuing challenge of artifacts in the EEG," *American Journal of EEG Technology*, vol. 35, no. 4, pp. 239–269, 1995.

[40] T. Gasser, J. C. Schuller, and U. S. Gasser, "Correction of muscle artefacts in the EEG power spectrum," *Clinical Neurophysiology*.

[41] F. von Wegner and H. Laufs, "Information-theoretical analysis of EEG microstate sequences in python," *Frontiers in Neuroinformatics*, vol. 12, p. 30, 2018.

[42] D. Lehmann, H. Ozaki, and I. Pal, "EEG alpha map series: brain micro-states by space-oriented adaptive segmentation," *Electroencephalography and clinical neurophysiology*, vol. 67, no. 3, pp. 271–288, 1987.

[43] J. Wackermann, D. Lehmann, C. Michel, and W. Strik, "Adaptive segmentation of spontaneous EEG map series into spatially defined microstates," *International Journal of Psychophysiology*, vol. 14, no. 3, pp. 269–283, 1993.

[44] G. Rossum, "Python reference manual," 1995.

[45] M. M. Murray, D. Brunet, and C. M. Michel, "Topographic ERP analyses: a step-by-step tutorial review," *Brain topography*, vol. 20, no. 4, pp. 249–264, 2008.

[46] D. Lehmann, R. D. Pascual-Marqui, and C. Michel, "EEG microstates," *Scholarpedia*, vol. 4, no. 3, p. 7632, 2009. revision #88985.

[47] F. F. Offner, "The EEG as potential mapping: the value of the average monopolar reference.," *Electroencephalography and clinical neurophysiology*, vol. 2, no. 2, p. 213, 1950.

[48] D. Lehmann and W. Skrandies, "Reference-free identification of components of checkerboard-evoked multichannel potential fields," *Electroencephalography and clinical neurophysiology*, vol. 48, no. 6, pp. 609–621, 1980.

[49] R. Pascual-Marqui, "Topographic maps, source localization inference, and the reference electrode: comments on a paper by desmedt et al," *Electroenceph Clin Neurophysiol*, vol. 88, pp. 532–533, 1993.

[50] T. E. Oliphant, *A guide to NumPy*, vol. 1. Trelgol Publishing USA, 2006.

[51] F. Perrin, J. Pernier, O. Bertrand, and J. Echallier, "Spherical splines for scalp potential and current density mapping," *Electroencephalography and clinical neurophysiology*, vol. 72, no. 2, pp. 184–187, 1989.

[52] P. Stoica, R. L. Moses, *et al.*, "Spectral analysis of signals," 2005.

[53] A. Delorme and S. Makeig, "EEGLAB: an open source toolbox for analysis of single-trial eeg dynamics including independent component analysis," *Journal of Neuroscience Methods*, vol. 134, no. 1, pp. 9 – 21, 2004.

[54] N. Bigdely-Shamlo, T. Mullen, C. Kothe, K.-M. Su, and K. A. Robbins, "The prep pipeline: standardized preprocessing for large-scale EEG analysis," *Frontiers in Neuroinformatics*, vol. 9, p. 16, 2015.

[55] A. Pedroni, A. Bahreini, and N. Langer, "Automagic: Standardized preprocessing of big EEG data.," *NEUROIMAGE*, vol. 200, pp. 460 – 473, n.d.

[56] L. J. Gabard-Durnam, A. S. Mendez Leal, C. L. Wilkinson, and A. R. Levin, "The harvard automated processing pipeline for electroencephalography (HAPPE): Standardized processing software for developmental and high-artifact data," *Frontiers in Neuroscience*, vol. 12, p. 97, 2018.

[57] S. Fitzgibbon, D. DeLosAngeles, T. Lewis, D. Powers, E. Whitham, J. Willoughby, and K. Pope, "Surface laplacian of scalp electrical signals and independent component analysis resolve EMG contamination of electroencephalogram," *International Journal of Psychophysiology*, vol. 97, no. 3, pp. 277 – 284, 2015.