# Human-Centered xAI: Towards Overcoming Interpretation Biases in Biomedical Signal Analysis

Luca La Fisca

luca.lafisca@umons.ac.be

Friday 20th October, 2023

A dissertation submitted to the Faculty of Engineering

of the University of Mons, for the degree of Doctor of Philosophy in Engineering Science

Supervisor: Prof. B. Gosselin

Co-supervisor: Prof. L. Lefebvre

This thesis was supported by the Fonds pour la Formation à la Recherche dans l'Industrie et l'Agriculture (FRIA-FNRS)

# Jury members

Prof. **Laurence Ris** - Université de Mons, President
Prof. **Bernard Gosselin** - Université de Mons, Supervisor
Prof. **Laurent Lefebvre** - Université de Mons, Co-supervisor
Prof. **Thierry Dutoit** - Université de Mons
Prof. **Philippe Fortemps** - Université de Mons
Dr. **Cyril Pernet** - University hospital of Copenhagen
Prof. **Robert Oostenveld** - Radboud Universiteit
Prof. **José Antonio Oramas Mogrovejo** - Universiteit Antwerpen

"Learn from the past, set vivid, detailed goals for the future, and live in the only moment of time over which you have any control: now."

Denis Waitley

# Abstract

In the ever-evolving landscape of biomedical signal analysis, the pursuit of accuracy can be hindered by interpretation biases. These biases can arise at different stages of the research process, from study design to publication.

The heart of this thesis lies in Explainable Artificial Intelligence (xAI), striving to illuminate the intricacies of "black-box" decision-making processes that underlie most contemporary deep learning algorithms. Our objective is to mitigate the potential for biased conclusions stemming from these non-transparent models. Therefore, we propose a novel xAI approach, termed *human-centered explainable AI*, which leverages intra and inter-subject similarities to extract pivotal features forming the basis for classification or regression tasks. Inspired by human decision-making processes based on comparisons, this technique is applied to sleep data. It yields a novel severity measure for sleep apnea events and uncovers electroencephalographic (EEG) biomarkers associated with severe sleep apnea events.

To promote the prioritization of clinician-interpretable features by AI models, we investigate the incorporation of "white-box" analyses that provide human-friendly representations of the recorded signals. However, these analyses can

also introduce biases and, although easy to understand, may inadvertently limit data exploration. This limitation could cause us to overlook important factors or effects beyond the scope of the analysis. As a response, we propose techniques to address vulnerabilities in experimental protocol design, sub-optimal recordings, and misrepresentations of target information. Focusing on EEG signal processing, this thesis introduces standardized frameworks covering the evaluation of confounder factors' effects, EEG preprocessing, and the validation of brain source reconstruction.

In summary, our overarching goal is to counteract interpretation biases in biomedical signal analysis, thereby fostering transparency, precision, and ethical integrity. Our future endeavors are aimed at extending the human-centered xAI approach to encompass multimodal data and diverse medical applications. This journey holds the promise of not only technological advancement but also a profound shift towards reliable medical diagnostics and research.

# Acknowledgements

I would like to begin by expressing my gratitude to my advisor, Professor Bernard Gosselin, for agreeing to supervise this thesis. He provided unwavering support throughout the entire journey and kept me motivated, especially during times when doubts arose, especially during Covid pandemic.

I also extend my thanks to my co-advisor, Professor Laurent Lefebvre, who introduced me to the fascinating world of neurophysiology and enabled me to undertake this interdisciplinary thesis.

My heartfelt appreciation goes to my laboratory head, Professor Thierry Dutoit, for providing us with an enviable working environment, allowing us the freedom to thrive in our unique ways, and always being attentive to our needs.

I am deeply grateful to Professor Laurence Ris, without whom this thesis would never have come to fruition. I will always remember our initial meeting, which transformed the past four years of my life.

I would like to acknowledge Dr. Cyril Pernet for initiating the collaboration that, to my great surprise and joy, has endured over time. Your guidance, insights, and questions pushed me beyond my limits, fostering learning and enabling achievements I couldn't have imagined attaining on my own.

I extend my thanks to Professor Robert Oostenveld for hosting me during a brief yet intense experience, during which I gained insights into the world of

industry within my field of expertise (which was still in its early stages of development).

I am grateful to all the members of my thesis committee and jury for your questions, remarks, advice, and, above all, your kindness, which allowed me to view my work with a fresh perspective.

I would like to express my gratitude to the "Fond National de la Recherche Scientifique" (FNRS) for funding this thesis through the "Fond pour la Formation à la Recherche dans l'Industrie et l'Agriculture" (FRIA).

Special thanks to Dr. Marie Bruynel for the wonderful collaboration with the sleep clinic at St. Pierre Hospital, which allowed us to harness the capabilities of the developed models to improve patients' lives.

I acknowledge Céliane for her work alongside me during her internship, which opened up new successful avenues of exploration.

My sincere thanks to my colleagues of the faculty of psychology, Cynthia, Aurélie, Erika, and Isabelle, for the numerous discussions that enhanced my understanding of various aspects of cerebral mechanisms.

I would like to express my appreciation to my colleagues at the ISIA Lab for making these four years as enriching as they were enjoyable. Conducting this thesis with all of you was the best opportunity I could have hoped for. I must especially mention the extraordinary ISIA Babies team, who, despite growing older, remains utterly fantastic: Maio, Nathan, Baba, Victor, and Noé. You are all a bunch of crazy guys, and I cherished every significant moment we shared.

I offer my thanks to my family for always believing in me, supporting me at various stages of my journey, and contributing to shaping the person I am today.

In particular, I extend my gratitude to my parents and my sister, Lisa, for always pushing me to exceed my limits, offering constant guidance, and ac-

companying me in the various choices I made in life. Seeing pride in your eyes is my greatest motivation. I love you.

To my family in law, Carmelo, Sylvie, Gerlando, and Lia, thank you for always considering my choices with kindness and showing me that you are always proud of my achievements.

I want to thank my friends, Loris, Maxime, Mengo, Simon, Baptiste, Charlotte, and Nicolas, who shared many intense moments and have always been there, both in good times and bad. Having you by my side gives real meaning to my life, and you continuously make me a better person.

Lastly, I wanted to reserve a special place of honor for the one who agreed to share her life with me by becoming my wife this year, Ornella. You already know, but without you, I could never have reached this point. Your unwavering support, comforting words, unique sense of humor, and boundless love allow me to keep moving forward tirelessly. I can never thank you enough for all the sacrifices you made for me throughout these four years. Rest assured, the deadlines are over! (Well, for now...) I am immeasurably fortunate to have you in my life, my beloved Namour! I love you deeply.

# Contents

**Introduction** **3**

**1 Fundamentals** **7**

1.1 Introduction to Biomedical Signals . . . . . 8

1.1.1 Electroencephalography . . . . . 10

1.1.2 Polysomnography . . . . . 18

1.2 Biomedical Signal Processing . . . . . 21

1.2.1 EEG Signal Processing . . . . . 23

1.2.2 PSG Signal Processing . . . . . 25

1.2.3 Machine Learning . . . . . 27

1.3 Explainable AI . . . . . 33

1.4 In Brief . . . . . 40

**2 Potential Biases** **41**

2.1 Planning Biases . . . . . 43

2.2 Data Collection Biases . . . . . 44

2.3 Analysis Biases . . . . . 45

2.4 Publication Biases . . . . . 46

2.5 Biases in Focus . . . . . 47

2.6 In Brief . . . . . 49

**3 Datasets** **51**

3.1 Priming Dataset . . . . . 52

3.1.1 Stimuli and Experimental Task . . . . . 52

3.1.2 Participants . . . . . 53
3.1.3 Data Acquisition . . . . . 54
3.2 Obstructive Sleep Apnea Dataset . . . . . 56
3.2.1 Experiment and Participants . . . . . 56
3.2.2 Data Acquisition . . . . . 56
3.2.3 Preprocessing . . . . . 57
3.3 In Brief . . . . . 60

**4 Planning Phase: Confounding Bias Evaluation 61**
4.1 LIMO EEG . . . . . 63
4.2 Method . . . . . 64
4.2.1 Variable Selection . . . . . 66
4.2.2 Linear Modeling . . . . . 68
4.2.3 Statistical Inference . . . . . 72
4.2.4 Effects Separability . . . . . 75
4.3 Results . . . . . 78
4.3.1 Discussion . . . . . 84
4.4 In Brief . . . . . 87

**5 Data Collection Phase: ERP Preprocessing 89**
5.1 Scope . . . . . 90
5.2 Proposed Framework . . . . . 91
5.2.1 FieldTrip . . . . . 91
5.2.2 Visual Inspection . . . . . 92
5.2.3 Ocular Artifacts Reduction . . . . . 93
5.2.4 Detrending and Filtering . . . . . 94
5.2.5 Segmentation and Downsampling . . . . . 95
5.2.6 Line Noise Removal . . . . . 96
5.2.7 Muscle Artifacts Reduction . . . . . 96
5.2.8 Baseline Correction and Re-referencing . . . . . 98
5.3 In Brief . . . . . 101

**6 Analysis Phase: Source Localization Benchmarking** **103**
6.1 Source Reconstruction . . . . . 104
6.1.1 Forward Modeling . . . . . 104
6.1.2 Inverse Modeling . . . . . 107
6.2 Related Work . . . . . 109
6.3 Proposed Framework . . . . . 111
6.3.1 Parameters Selection . . . . . 112
6.3.2 Source Selection . . . . . 113
6.3.3 Pseudo-Source Signal Generation . . . . . 113
6.3.4 Pseudo-EEG Signal Generation . . . . . 117
6.3.5 Performance Evaluation . . . . . 118
6.3.6 Benchmark . . . . . 120
6.4 Discussion . . . . . 122
6.5 In Brief . . . . . 123

**7 Human-Centered Explainable AI** **125**
7.1 Genesis . . . . . 126
7.2 Model Architecture . . . . . 132
7.3 Use Case: Cat/Dog Classification . . . . . 136
7.4 Discussion . . . . . 139
7.5 In Brief . . . . . 142

**8 xAI for Obstructive Sleep Apnea Assessment** **143**
8.1 Obstructive Sleep Apnea Assessment . . . . . 144
8.2 Model Architecture . . . . . 145
8.2.1 xVAEnet . . . . . 146
8.2.2 xAAEnet . . . . . 151
8.3 Biomarkers Identification . . . . . 156
8.4 Obstructive Sleep Apnea Severity Scoring . . . . . 163
8.5 In Brief . . . . . 175

Conclusion 177

Bibliography 181

List of Figures 211

List of Tables 229

# Acronyms

**AAE** Adversarial Auto-Encoder.

**AAL** Anatomical Automatic Labeling.

**AE** Auto-Encoder.

**AHI** Apnea-Hypopnea Index.

**AI** Artificial Intelligence.

**ANN** Artificial Neural Network.

**AOA** age of acquisition.

**BCE** binary cross-entropy.

**BCI** Brain Computer Interface.

**CAE** Convolutional Auto-Encoder.

**CAM** Class Activation Mapping.

**CCA** Canonical Component Analysis.

**CHU S$^{t}$-Pierre** Centre Hospitalier Universitaire Saint-Pierre.

**CNN** Convolutional Neural Network.

**COBIDAS** Committee on Best Practices in Data Analysis and Sharing.

**CSF** Cerebro-Spinal Fluid.

**DA** Desaturation Area.

**DL** Deep Learning.

**ECG** Electrocardiography.

**EDF** European Data Format.

**EEG** Electroencephalography.

**EEMD** Ensemble Empirical Mode Decomposition.

**EEMD-CCA** Ensemble Empirical Mode Decomposition-Canonical Component Analysis.

**EMG** Electromyography.

**EOG** Electrooculography.

**ERP** Event-Related Potential.

**FEM** Finite Element Method.

**fMRI** functional MRI.

**GAN** Generative Adversarial Network.

**GLM** General Linear Model.

**GradCAM** Gradient-weighted Class Activation Mapping.

**HMM** Hidden Markov Model.

**ICA** Independent Component Analysis.

**IoT** Internet of Things.

**LDA** Linear Discriminant Analysis.

**LIME** Local Interpretable Model-Agnostic Explanations.

**LOSO** Leave One Subject Out.

**LPA** Left Pre-Auricular.

**MAE** Mean Absolute Error.

**MCC** Multiple Comparison Correction.

**MEG** magnetoencephalography.

**ML** Machine Learning.

**MLP** Multilayer Perceptron.

**MNE** Minimum-Norm Estimation.

**MNI** Macroscopic Anatomical Parcellation.

**MRI** magnetic resonance imaging.

**MSE** Mean Squared Error.

**MWF** Multi-channel Wiener Filter.

**NAF2P** NAsal AirFlow.

**NAF2P** Nasal airflow.

**NREM** non-REM.

**NYC-Q** New York Cognition Questionnaire.

**ODI** oxygen desaturation index.

**OHBM** Organization for Human Brain Mapping.

**OSA** Obstructive Sleep Apnea-hypopnea.

**PCA** Principal Component Analysis.

**PDPs** Partial Dependence Plots.

**PLMI** periodic limb movement index.

**PR** Pulse Rate.

**PRV** Pulse Rate Variability.

**PSG** Polysomnography.

**Pshift** phase shift.

**PSP** post-synaptic potential.

**ReLU** rectified linear unit.

**REM** rapid eye movement.

**RISE** Randomized Input Sampling for Explanation.

**RMSE** root-mean-square error.

**RNN** Recurrent Neural Network.

**ROI** region of interest.

**RPA** Right Pre-Auricular.

**rs-EEG** resting-state EEG.

**$SAO_2$** Oxygen SAturation.

**SEREEGA** Simulating Event-Related EEG Activity.

**SHAP** SHapley Additive exPlanations.

**SLP** Single-Layer Perceptron.

**SNR** signal-to-noise ratio.

**SOTA** state-of-the-art.

**SpO2** Pulse oximetry.

**STFT** Short-Time Fourier Transform.

**SVD** Singular-Value Decomposition.

**SVM** Support Vector Machine.

**t-SNE** t-distributed stochastic neighbor embedding.

**tanh** hyperbolic tangent.

**TFCE** Threshold Free Cluster Enhancement.

**VAB** ABdominal belt Voltage.

**VAE** Variational Auto-Encoder.

**ViTs** Vision Transformers.

**VP-P** peak-to-peak voltage.

**VTH** THoracic belt Voltage.

**xAAEnet** eXplainable Adversarial Auto-Encoder network.

**xAI** Explainable AI.

**xVAEnet** eXplainable Variational Auto-Encoder network.

# Introduction

## Motivation

The challenges posed by biases in biomedical signal analysis are multifaceted, affecting not only the robustness of research outcomes but also casting shadows on the very foundation of medical science's credibility and advancement. The intricate nature of biomedical signals, such as Electroencephalography (EEG), introduces complexity that, when combined with biases, can obscure accurate interpretation and understanding.

In a landscape where reproducibility struggles to gain foothold, the potential ramifications echo far beyond the realm of academia. As we strive for breakthroughs that can transform patient care, the implications of unreliable findings and misguided conclusions are deeply concerning. Each instance of unreproducible research translates into missed opportunities to enhance medical diagnoses, treatments, and overall well-being.

The driving force behind this research lies in the recognition that biases, arising from an array of analytical choices and statistical methodologies, profoundly impact on the validity of our findings. The adoption of inappropriate methods or the subtle skewing of interpretations can inadvertently distort results and lead us down misguided paths. To tackle these challenges head-on, it is paramount to meticulously investigate, quantify, and mitigate these biases.

However, the complexity of the issue is not limited to one isolated phase of research. Rather, it extends its tendrils across the entire investigative journey, whether during the preliminary stages, active trials, or the critical post-trial analyses. Biases introduced at any juncture have the potential to reverberate throughout, casting doubt on the very conclusions we draw and ultimately undermining the reliability of our research.

In this context, this thesis seeks to navigate the intricate landscape of biases within biomedical signal analysis. Through an exploration of EEG and Polysomnography (PSG) signals, the research endeavors to mitigate specific biases that can distort interpretations and hinder the advancement of medical science. By identifying and addressing biases at different stages, from experimental design to publication, the study aspires to contribute to a more accurate and reproducible understanding of physiological phenomena.

## Objectives

Within the context of addressing challenges associated with biases in biomedical signal analysis, this thesis primarily focuses on EEG and PSG signals. The research objectives encompass a multi-faceted approach, aiming to uncover and mitigate biases throughout various stages of the analytical pipeline:

- Streamline the initial phase of study design by questioning the necessity of balancing specific features across the experimental conditions.
- Mitigate the impact of suboptimal data collection through the implementation of a standardized preprocessing technique.
- Strengthen the robustness and reliability of data analysis by evaluating the quality of a specific derived data representation.
- Alleviate biases in data interpretation by introducing explainability into existing models inherently lacking in transparency.
- Evaluate the potential of Explainable AI (xAI) to unravel underlying physiological mechanisms, providing valuable insights into complex signal interactions.

## Contributions

The original contributions addressing the aforementioned challenges and fulfilling the stated objectives encompass:

- Proposing a framework to explore the impact of covariates (confounders) on Event-Related Potential (ERP) signal interpretation, thereby enhancing the integrity of Brain Computer Interface (BCI) experiments and facilitating their implementation.
- Implementing a robust preprocessing technique for ERP signals, enhancing data quality and analysis outcomes.
- Introducing a comprehensive validation framework to benchmark the accuracy and reliability of various EEG source localization methods.

- Developing a pioneering *"human-centered xAI approach"* inspired by human decision-making processes. Applied to PSG signals, this approach allows: 1) the derivation of a novel objective severity score for Obstructive Sleep Apnea-hypopneas (OSAs), providing an alternative to the highly criticized Apnea-Hypopnea Index (AHI); 2) the identification of biomarkers responsible for severity variations.

## Organization of this Dissertation

**Chapter 1** provides a foundational understanding of biomedical signals, with a specific focus on EEG and PSG. It explores signal processing techniques, machine learning algorithms, and explainable AI, while also describing potential biases affecting these signals.

**Chapter 2** describes the potential biases that may impact the interpretation of results in research involving biomedical signals.

**Chapter 3** delves into the acquisition of experimental data for this thesis.

**Chapter 4** introduces our contribution aimed at mitigating biases arising during the planning phase. This chapter focuses on reducing confounding bias by proposing a framework to quantify the influence of confounders on data interpretation.

**Chapter 5** addresses biases that may emerge during the data collection phase. We propose to mitigate measurement biases by employing a standardized framework for preprocessing ERP signals.

**Chapter 6** presents our approach to reduce modeling bias, which can occur during the analysis phase of experiments. Specifically, we concentrate on benchmarking brain source localization methods used in EEG studies.

**Chapter 7** introduces our innovative *human-centered xAI* approach, carefully crafted to mitigate confounder exploitation bias that typically affects the interpretation of results from "black-box" algorithms.

**Chapter 8** demonstrates the practical application of our *human-centered xAI* approach in evaluating the severity of OSA events, emphasizing the discovery of EEG biomarkers and the implementation of an improved scoring method.

# Chapter 1

# Fundamentals

## Contents


**1.1 Introduction to Biomedical Signals** . . . . . **8**
1.1.1 Electroencephalography . . . . . 10
1.1.2 Polysomnography . . . . . 18
**1.2 Biomedical Signal Processing** . . . . . **21**
1.2.1 EEG Signal Processing . . . . . 23
1.2.2 PSG Signal Processing . . . . . 25
1.2.3 Machine Learning . . . . . 27
**1.3 Explainable AI** . . . . . **33**
**1.4 In Brief** . . . . . **40**


This chapter provides the fundamental knowledge necessary to delve into the themes explored in this thesis. **Section 1.1** presents the electroencephalography, from the fundamental principles of neuron activation to the signals captured by electrodes. Additionally, we provide insights into the realm of polysomnography. **Section 1.2** delves into the traditional procedures used for processing both EEG (Section 1.2.1) and PSG (Section 1.2.2) signals, while **Section 1.2.3** introduces machine learning algorithms that play a significant role in the processing of these signals. **Section 1.3** offers insights into the realm of xAI, including discussions on existing methods for comprehending the decisions made by specific Artificial Neural Networks (ANNs).

## 1.1 Introduction to Biomedical Signals

Biomedical signals, with their ability to unveil intricate physiological patterns, stand as invaluable tools that bridge the realms of clinical diagnosis and scientific exploration. These signals not only provide windows into the dynamic inner workings of the human body but also serve as conduits for unraveling the complexities of health and disease.

Throughout history, the quest to understand the human body's functioning has driven the exploration of various physiological signals. From the early rudimentary pulse measurements to the sophisticated monitoring technologies of today, the journey of biomedical signal analysis has been marked by a relentless pursuit of insight. These signals, originating from within the body's intricate systems, have evolved from being mere indicators of basic vital signs to becoming catalysts for comprehensive diagnostic assessments.

The history of biomedical signals traces an intriguing trajectory through the annals of scientific inquiry, spanning centuries of relentless exploration and technological evolution. Dating back to antiquity, the conceptualization of vital signs provided a rudimentary glimpse into the body's inner workings, with pulse and breath serving as early indicators of life. The Renaissance witnessed pioneering investigations into blood circulation, as William Harvey's groundbreaking work revolutionized our understanding of cardiovascular dynamics [1]. This period heralded the inception of quantitative measurement, as Santorio Santori introduced the concept of quantifying physiological variables through his invention of the medical thermometer [2].

The subsequent centuries witnessed a confluence of disciplines, as physics, engineering, and medicine converged to shape the domain of biomedical signal analysis. The $19^{\text{th}}$ century heralded the discovery of electricity's role in physiology, propelling the understanding of nerve impulses and paving the way for the Electrocardiography (ECG) and EEG. In the early $20^{\text{th}}$ century, the pioneering efforts of Willem Einthoven facilitated the recording and interpretation of ECG signals [3], unraveling the heartbeat's electrical signature. Concurrently, the introduction of imaging techniques such as X-rays [4] and ultrasounds [5] illuminated the anatomical landscape, adding a new dimension to diagnostic capabilities.

The latter half of the 20$^{th}$ century witnessed a surge in technological advancements that redefined the possibilities of biomedical signal analysis. The advent of microelectronics birthed portable monitoring devices, democratizing the accessibility of physiological data. Further breakthroughs in signal processing, combined with the burgeoning realm of computational methods, enabled sophisticated analysis and interpretation of complex signals.

Today, the history of biomedical signals continues to unfold with unprecedented dynamism. Innovations like functional MRI (fMRI) and magnetoencephalography (MEG) allow for non-invasive probing of brain function. Wearable sensors and Internet of Things (IoT) devices bring real-time monitoring into everyday life, revolutionizing personalized healthcare. As technology and scientific understanding forge ahead, the trajectory of biomedical signals continues to intertwine with the ever-evolving quest to decipher the intricate symphony of the human body's physiological processes.

Biomedical signals are a cornerstone in both clinical and research domains. In the clinical sphere, they enable physicians to monitor patient health, diagnose conditions, and tailor treatment plans with precision. From detecting irregular heart rhythms through electrocardiograms to assessing respiratory health using spirometry, these signals furnish critical information that guides medical decisions. Simultaneously, researchers harness these signals as tools for scientific exploration.

However, these signals are not devoid of limitations. Artifacts stemming from external interferences, individual variability, and the inherent complexity of physiological systems can introduce noise and distortion into the recordings. The challenge lies in disentangling these nuances from the genuine signals to extract accurate and meaningful information.

In this thesis, our focus hones in on two powerful biomedical signals: EEG and PSG. These signals, rooted in the cerebral landscape and sleep patterns, respectively, stand as windows to the intricate dynamics of brain function and sleep physiology. Their advantages are manifold. EEG, with its high temporal resolution, enables the tracking of rapid neural changes. PSG, through its comprehensive approach, facilitates the assessment of sleep architecture. Both these signals, due to their non-invasive nature, hold the potential to be recorded repeatedly, offering a wealth of information that can be harnessed to derive insights with clinical and scientific implications.

As we traverse the intricate landscape of EEG and PSG signals, our endeavor is to unveil their underlying intricacies, latent potentials, and inherent constraints. In this pursuit, our objective is to intersect the ongoing discourse of biomedical signal analysis, thereby not only augmenting our comprehension of human physiology but also catalyzing strides in the realm of medical diagnostics and scientific inquiry.

Within this ambit, our foremost aspiration is to critically examine the prevailing landscape of signal interpretation. We intend to unravel and scrutinize the multifaceted biases that permeate the current paradigms of EEG and PSG signal interpretation. By delving into these biases, we aspire to illuminate the latent distortions that can inadvertently taint the extraction of insights from these signals. This endeavor is anchored in the pursuit of not merely uncovering the signal's raw potential, but also in mitigating the inadvertent influences that cloud its veracity.

Furthermore, our pursuit extends beyond a mere elucidation of existing biases. We aim to proffer a multifarious spectrum of methodologies aimed at minimizing the encroachment of these biases. By harnessing a range of analytical approaches, we endeavor to establish a robust framework that can enhance the fidelity of signal interpretation. These approaches span from meticulous preprocessing techniques to sophisticated machine learning algorithms that strive to encapsulate a comprehensive view of signal dynamics, thus paving the way for an unadulterated comprehension.

### 1.1.1 Electroencephalography

#### Biological Basis

Neurons, often described as signal receivers, processors, and transmitters, play a pivotal role in generating the electrical activity observed through EEG. As a neural signal propagates, it creates an electric field at the core of EEG measurements. As shown in Figure 1.1, the intricate interplay of neuronal processes involves post-synaptic potentials (PSPs) within dendrites and action potentials (APs) carried along axons. At synapses, neurotransmitter release triggered by an AP alters membrane permeability, giving rise to PSPs. If multiple PSPs accumulate to reach a threshold, an AP is generated, leading to a neuronal "spike" as voltage-sensitive channels open, allowing ions to flow.

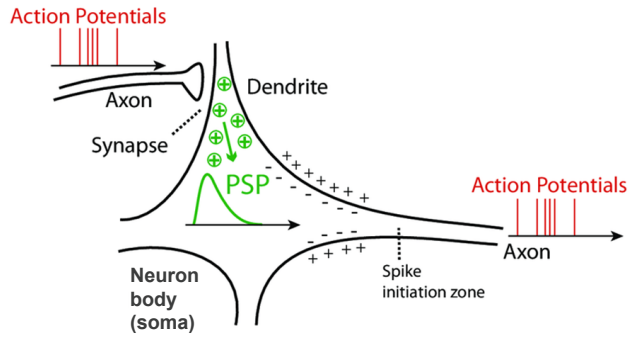



**Figure 1.1.** Action potential (AP) and post-synaptic potential (PSP) in neuron. Action potentials traverse chemical synapses, reaching the neuron's dendrites. These interactions result in the emergence of post-synaptic potentials, whose cumulative effect gives rise to subsequent action potentials, capable of propagating along the neuron's axon. Adapted from [6].

The presence of PSPs and APs creates minute intracellular currents and associated electric fields. While these fields are too small to be measured directly outside the head, they can summate to generate measurable signals. However, the temporal characteristics of PSPs and APs, illutrated in Figure 1.2, impact their summation potential. PSPs with their duration of around 10 ms are better candidates for producing measurable electric fields than the millisecond-duration APs, which are harder to synchronize for summation. Notably, these currents must also have a common direction for successful summation, which is made easier by the monophasic nature of post-synaptic potentials (PSPs).

The pivotal role of pyramidal neurons, shown in Figure 1.3A, in generating detectable electric fields within the cortex cannot be overlooked. These neurons, organized in structured assemblies, constitute a significant portion of the neocortex. Their unique geometry allows the summation of fields generated by PSPs. Particularly, large pyramidal neurons in cortical layer 5 play a significant role. These neurons are organized in parallel, have similarly oriented arrangements, and receive synchronous inputs. As shown in Figure 1.3B, a dipole exists between their soma and apical dendrites, resulting in potential behavior that mimics current flow.

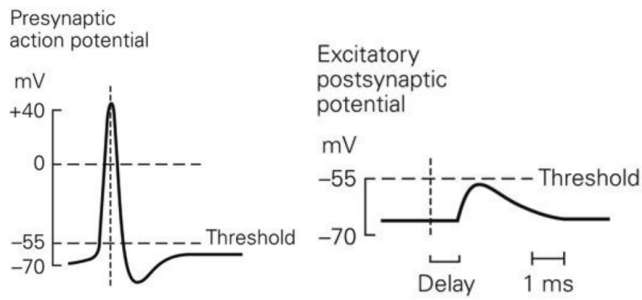



**Figure 1.2.** Temporal Comparison: Action Potential (AP) vs. Post-Synaptic Potential (PSP). The action potential exhibits a biphasic waveform with an initial positive peak (when excitatory), lasting approximately 1 ms. In contrast, the post-synaptic potential features a monophasic waveform (positive when excitatory), extending for about 10 ms. The PSP emerges approximately 1 ms after the peak of the action potential.

## A Brief History

The evolution of EEG unfolds as a remarkable testament to humanity's ceaseless pursuit of understanding the enigmatic realms of brain activity. Emerging from a confluence of pioneering discoveries, EEG's history began with the pioneering endeavors of Hans Berger in the early 20$^{\text{th}}$ century. In 1924, Berger's groundbreaking experiments demonstrated that the brain's electrical activity could be recorded non-invasively from the scalp, inaugurating the era of EEG [8]. The subsequent decades witnessed the refinement of recording techniques, with the introduction of standardized electrode placements and amplification systems.

The mid-20$^{\text{th}}$ century marked a transformative phase for EEG, driven by technological advancements that spurred its clinical utility. The discovery of distinct EEG patterns corresponding to different sleep stages [9] and neurological conditions laid the foundation for diagnostic applications. EEG rapidly found its place in clinical practice, aiding in the diagnosis of epilepsy [10], sleep disorders, and neurological pathologies. Simultaneously, the understanding of EEG's underlying neural generators deepened, catalyzed by advancements in signal processing and source localization techniques [11].

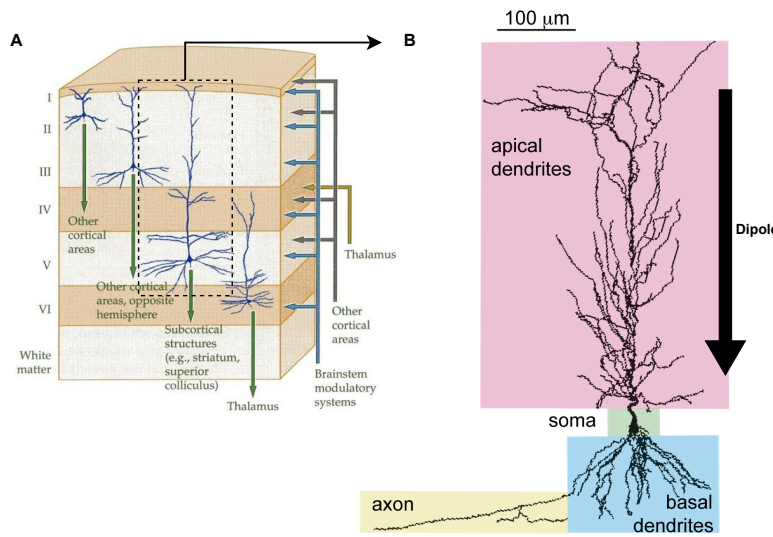



**Figure 1.3.** Pyramidal Neurons and Dipole Generation. (A) Illustration of various cortical layers with their associated brain regions. (B) Emphasis on the unique geometry of a $5^{\text{th}}$ level pyramidal neuron, highlighting the generation of a dipole due to the distinct orientation of apical dendrites relative to the soma. Adapted from [7].

The latter half of the $20^{\text{th}}$ century ushered in a phase of EEG's widespread adoption, fueled by the advent of digital technology. Computerized EEG systems allowed for real-time monitoring, enhancing diagnostic accuracy and enabling long-term monitoring of brain activity. The integration of EEG with other physiological signals, such as fMRI, further enriched its potential in cognitive neuroscience research and clinical investigations.

In the contemporary era, EEG's history intersects with cutting-edge developments in computational neuroscience [12] and Artificial Intelligence (AI) [13, 14]. High-density EEG arrays, coupled with advanced algorithms, unlock new vistas for deciphering complex brain dynamics. Additionally, wearable EEG devices and portable systems extend EEG's reach beyond clinical settings, enabling applications in neurofeedback [15], brain-computer interfaces [16], and cognitive enhancement.

In summation, the history of EEG epitomizes the interplay between scientific curiosity and technological progress. From Berger's pioneering work to the present, EEG has evolved from a nascent experimental technique to a multidimensional tool that unravels the intricacies of brain function. As EEG continues to illuminate the realms of neural activity, its journey mirrors the inexorable march of science, bridging the gap between human cognition and technological innovation.

### Signal Recording

EEG signals are captured by placing multiple electrodes on the scalp, which detect the electrical field generated by the brain's neurons. EEG signals reflect the collective firing of neurons and can reveal important information about brain functions such as cognition, sleep stages, and neurological disorders.

The EEG electrodes placed on the scalp detect the net current flow, whether positive or negative, originating from cortical neurons. The 10/20 system, shown in Figure 1.4A, offers standardized methods for electrode placement, ensuring consistent data collection. For higher density recordings, the 10/10 system, shown in Figure 1.4B, has been proposed by the American Electroencephalographic Society. These electrodes, characterized by low impedance ($5 - 10k\Omega$), can be arranged in bipolar or unipolar montages (cf. Figure 1.5), providing valuable insights into the brain's intricate electrical activity.

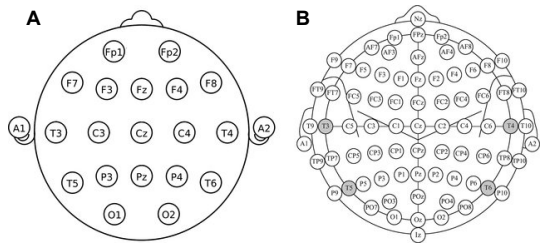



**Figure 1.4.** Electrode Placement Systems. (**A**) In the 10-20 system, electrodes are positioned based on anatomical landmarks using a grid pattern. The electrodes are placed at specific percentages (10 % and 20%) of distances between key landmarks on the scalp, providing consistent and repeatable electrode positions. (**A**) The 10-10 system further refines electrode placement by adding additional positions, allowing for more precise spatial coverage. Reproduced from [17].

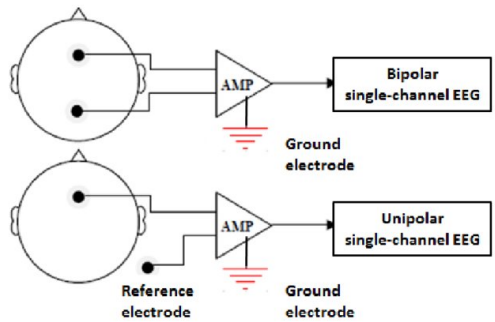



**Figure 1.5.** Comparison of EEG Cap Montages. The figure illustrates two commonly used electrode placement configurations in EEG recordings: Bipolar (A) and Unipolar (B). The bipolar montage (A) involves pairing adjacent electrodes to measure the potential difference between them, facilitating the detection of local electrical activity and providing insights into the scalp voltage gradient. In contrast, the unipolar montage (B) pairs each electrode with a common reference electrode, capturing the individual electrical activity at each electrode site and enabling a comprehensive understanding of neural dynamics across the scalp. Reproduced from [18].

In contemporary EEG setups, electrodes are linked to amplifiers (known as an active setup) responsible for converting neural electrical activity into measurable signals. These signals are then digitized and stored for subsequent analysis using computer systems. Due to the minute amplitude of neural signals, usually within the range of a few microvolts, EEG recordings require the utilization of highly sensitive and noise-resistant recording equipment.

### Signal Features

EEG's non-invasiveness and high temporal resolution make it a favored choice for monitoring rapid changes in brain activity. Different frequency bands, at different magnitude ranges, within the EEG signal offer insights into various brain states [19]:

- Delta (0.5 - 4 Hz): Associated with deep sleep and certain neurological disorders ($5 - 250\mu V$).
- Theta (4 - 8 Hz): Often seen during drowsiness and early sleep stages ($20 - 200\mu V$).
- Alpha (8 - 13 Hz): Dominant during relaxed wakefulness and closed eyes ($5 - 120\mu V$).
- Beta (13 - 30 Hz): Common during active thinking and alertness ($5 - 50\mu V$).
- Gamma (30 - 100 Hz): Associated with cognitive processes and sensory integration (around $10\mu V$).

EEG is well-suited for examining both resting-state activity and ERPs. In resting-state analysis, EEG captures spontaneous fluctuations in neural activity while individuals are at rest. By analyzing the connectivity patterns of different brain regions, researchers can infer functional networks and gain insights into brain organization and dynamics. ERPs, on the other hand, involve analyzing EEG responses to specific events or stimuli. These brief, transient signals represent cognitive processes and sensory perception. Event-related potentials provide insights into the timing and sequence of cognitive events, helping researchers uncover how the brain processes information and responds to external stimuli.

While EEG has opened doors to understanding cognitive functions, its spatial resolution is limited due to the nature of electrical field propagation in the brain and surrounding tissues. This phenomenon, known as volume conduction, arises from the varying conductive properties of brain tissue, cerebrospinal fluid, and the skull. As electrical currents generated by neural activity spread through these media, the resulting electric potentials observed at the scalp electrodes can be blurred and indistinct, making the precise localization of neural sources challenging. Additionally, the weak depth sensitivity of EEG contributes to its limited ability to distinguish between neural activity occurring at different depths within the brain.

Moreover, the EEG signal is susceptible to various noise artifacts that can obscure meaningful neural information. These artifacts can stem from multiple sources, including muscle movements (electromyographic artifacts), eye movements (ocular artifacts), and external interferences (electromagnetic artifacts). Muscle artifacts, for example, result from muscle contractions, often occurring during head or body movement. Ocular artifacts are caused by the electrical potentials generated by eye movements and blinking. External interferences, such as power line noise or electronic device emissions, can infiltrate the EEG signal during recording. An extensive description of these artifacts is provided in Chapter 5.

Efforts to improve EEG signal quality and accuracy have led to the development of various preprocessing techniques and artifact removal methods. These methods aim to identify and filter out unwanted signals, thereby enhancing the signal-to-noise ratio and preserving the integrity of the neural information. However, despite the existence of such artifact removal algorithms, they are not infallible and can sometimes introduce errors or distortions themselves. Therefore, it is crucial to meticulously record the EEG signal under the cleanest possible conditions, employing strategies such as proper electrode placement, participant immobilization, and shielding against external interferences. This proactive approach to data collection can significantly enhance the quality and reliability of EEG recordings, ultimately leading to more accurate and meaningful insights.

### 1.1.2 Polysomnography

#### A brief history

The narrative of PSG unfolds as a compelling saga in the exploration of human sleep patterns, revealing the mysteries of the nocturnal realm through intricate physiological monitoring. PSG's inception can be traced back to the mid-20$^{th}$ century, when the recognition of distinct sleep stages prompted the quest to capture the dynamics of sleep architecture. The pioneering work of Aserinsky and Kleitman in the 1950s marked a pivotal moment, as they introduced the concept of rapid eye movement (REM) sleep and non-REM (NREM) sleep stages [20], laying the foundation for PSG.

The following decades witnessed the gradual integration of multiple physiological signals into PSG, culminating in a comprehensive view of sleep. EEG, Electrooculography (EOG), and Electromyography (EMG) emerged as the cornerstones of PSG, capturing brain activity, eye movements, and muscle tone, respectively. The simultaneous recording of these signals during sleep unveiled the intricate choreography of sleep cycles, including transitions between REM and NREM stages.

Advancements in technology during the latter half of the 20$^{th}$ century propelled PSG from a niche research tool to a cornerstone of sleep medicine. Innovations in signal processing, amplification, and data storage enhanced the accuracy and fidelity of PSG recordings. These developments transformed PSG into an indispensable diagnostic tool.

The 21$^{st}$ century ushered in a new era for PSG, characterized by portability and data integration. Miniaturization of recording devices enabled ambulatory PSG studies, empowering sleep monitoring beyond clinical settings. Simultaneously, the fusion of PSG with other physiological signals, such as heart rate variability [21], enriched the understanding of sleep-related pathologies. Moreover, the integration of PSG data with computational algorithms facilitated automated sleep stage scoring [22], streamlining analysis and diagnosis.

Today, PSG stands as a testament to the intricate intersection of technology, medical science, and human physiology. Its evolution from rudimentary observations to a comprehensive diagnostic modality epitomizes the relentless pursuit of understanding the nocturnal dimensions of human existence. As PSG

continues to evolve, its narrative weaves a tapestry that unravels the enigmatic terrain of sleep, fostering breakthroughs in clinical practice, research, and the broader exploration of human well-being.

### Signal Recording

The intricate nature of sleep processes requires the utilization of various sensors to capture specific aspects of physiological activity. As shown in Figure 1.6, the PSG sensors are composed of:

- EEG electrodes to monitor brain activity.
- EOG sensors to detect eye movements.
- ECG to record heart activity.
- Nasal airflow (NAF2P) sensor to monitor the passage of air through the nasal passages.
- EMG electrodes to track muscle activity.
- Position sensors to monitor body position during sleep.
- Pulse oximetry (SpO2) sensor to measure blood oxygen saturation levels.
- Thoracic and abdominal belts to assess respiratory effort.

### Signal Features

The combination of physiological signals in PSG offers a comprehensive view of sleep architecture and patterns. Each signal provides unique insights into different aspects of sleep physiology and disorders:

- EEG electrodes offers insights into sleep architecture, stages, and abnormalities.
- EOG enables the identification of REM sleep.
- ECG provides information on cardiac rhythm and its variations.
- NAF2P is crucial for identifying limitations in airflow and diagnosing breathing disorders.

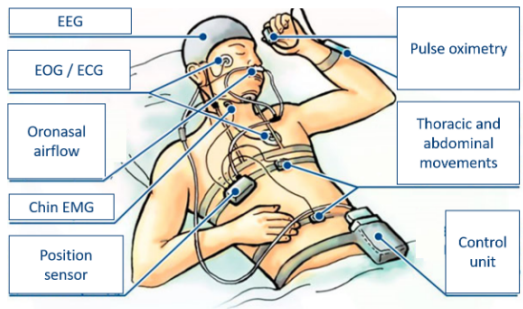



**Figure 1.6.** PSG Sensors Representation. The figure illustrates the different sensors used in PSG recordings to monitor various physiological signals during sleep: EEG electrodes provide insights into brain activity, ECG electrodes capture heart activity, EOG sensors detect eye movements, oronasal airflow sensor monitors the passage of air through the nasal passages, EMG electrodes monitor muscle tone changes, position sensors track body posture, pulse oxymetry sensor measures blood oxygen saturation levels, thoracic and abdominal belts record chest and abdominal wall movements. Adapted from [23].

- EMG discerns muscle tone changes indicative of sleep disorders like sleep apnea.
- Position sensors help identify changes in posture, such as shifts from supine to prone positions or changes in body orientation.
- SpO2 serves as an indicator of respiratory function and can detect conditions such as hypoxia or sleep apnea.
- Thoracic and abdominal belts highlight changes in chest and abdominal wall movements, indicating breathing patterns during sleep.

This multi-modal approach to PSG enables clinicians and researchers to explore the intricate interplay of physiological factors during sleep, contributing to a comprehensive assessment of sleep quality and health.

However, PSG's intricate setup can pose challenges for long-term monitoring and patient comfort. The complex interplay of multiple signals requires sophisticated analysis techniques to derive meaningful interpretations. Despite these complexities, PSG remains a cornerstone in sleep research, enabling the

exploration of sleep patterns' nuances and their implications for overall well-being.

## 1.2 Biomedical Signal Processing

Rooted in the mid-20th century, the inception of signal processing in biomedical contexts was driven by the quest to extract meaningful insights from the cacophony of physiological data. This heralded the birth of signal processing as a potent tool to unravel the intricate dynamics underlying various biomedical signals.

The history of biomedical signal processing is closely intertwined with the evolution of technology. Early endeavors focused on analog noise reduction [24] and visualization techniques [25], facilitating the analysis of signals like ECG, EEG, and EMG. The advent of digital computation spurred a revolution, enabling the development of advanced techniques for filtering [26], feature extraction [27], and signal decomposition [28–30].

The foundational principle of signal processing, Fourier Transform [31], emerged as a transformative catalyst. This mathematical tool bestowed the ability to dissect complex signals into their constituent frequency components, unraveling hidden patterns in physiological phenomena. As signal processing matured, it burgeoned into a multidisciplinary field encompassing diverse techniques like wavelet transforms [28], adaptive filtering [26], and time-frequency analysis [32] represented in Figure 1.7.

Nowadays, deep learning architectures adeptly handle intricate patterns within data, paving the way for automated anomaly detection [34], disease classification [35], and predictive modeling [36]. Moreover, the fusion of signals through data fusion and multimodal analysis enriches the holistic understanding of physiological processes.

In this realm, the synergy of signal processing methods and biomedical understanding continues to flourish. From the depths of historical curiosity to the pinnacle of contemporary computational prowess, the theoretical underpinnings of biomedical signal processing navigate an intricate labyrinth of scientific inquiry, technological innovation, and medical significance.

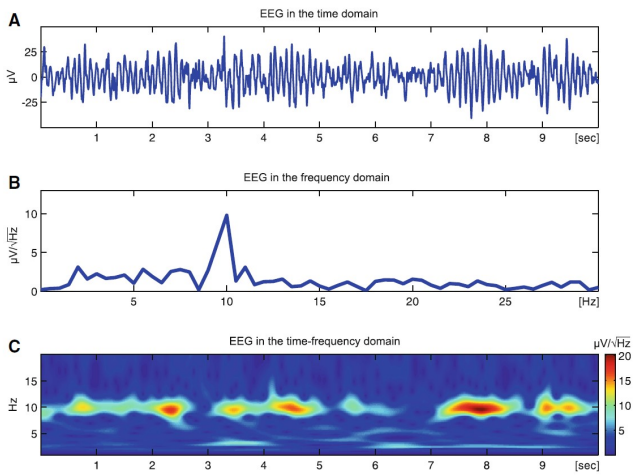



**Figure 1.7.** Time-Frequency Representation of an EEG Signal. **(A)** shows a segment of an EEG signal captured from a single channel. **(B)** illustrates the frequency domain representation of the signal obtained through the Fourier Transform, highlighting the dominant frequency components. **(C)** displays the time-frequency representation of the EEG signal, revealing how its frequency content changes over time. Reproduced from [33].

### 1.2.1 EEG Signal Processing

Processing EEG signals, whether for resting-state analysis or ERPs, involves a series of intricate steps that aim to extract meaningful information from the complex electrical activity of the brain. These methods enable researchers to unravel the intricate neural dynamics underlying cognitive functions and brain processes. The key steps in processing EEG signals are preprocessing, feature extraction, and statistical analysis. For a more comprehensive exploration of the underlying brain processes within specific regions, researchers can optionally incorporate source localization as an additional step.

#### Preprocessing

The initial step in EEG signal processing entails preprocessing the raw EEG data to remove artifacts and enhance the signal quality. This involves techniques such as noise filtering, artifact rejection, and data interpolation. High-pass and low-pass filters are often utilized to eliminate unwanted frequency components, while notch filters help mitigate power line noise and other external interferences. Artifacts caused by eye movements, muscle activity, and electrode drift can be identified and removed through techniques like Independent Component Analysis (ICA) [29] and Principal Component Analysis (PCA) [37]. Proper electrode referencing, either by re-referencing to common average or utilizing more advanced methods like Laplacian referencing [38], also helps in enhancing the signal quality. With the aim of establishing a standardized approach, the Organization for Human Brain Mapping (OHBM) Committee on Best Practices in Data Analysis and Sharing (COBIDAS) proposed best practices for effectively conducting the preprocessing of EEG data, as represented in Figure 1.8 [39].

#### Feature Extraction

For resting-state EEG (rs-EEG), the goal is to characterize the functional connectivity between different brain regions. Measures such as coherence [40], phase synchronization [41], and cross-correlation quantify the interactions between EEG signals recorded at different electrode sites. Extracting features from EEG signals involves transforming the data into a format suitable for

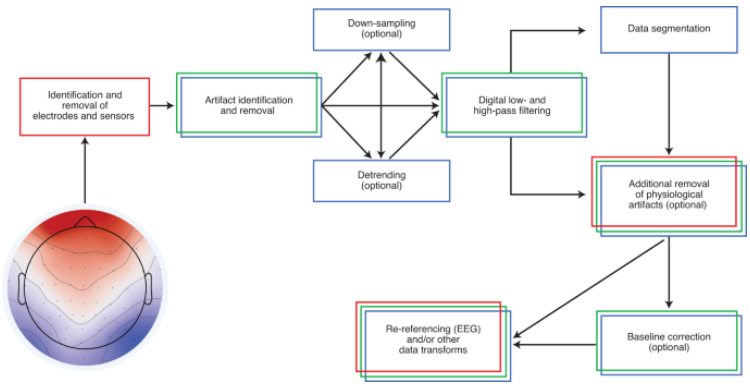



**Figure 1.8.** EEG Preprocessing Steps. The figure illustrates the standard preprocessing workflow for EEG data, as recommended by COBIDAS. Each step impacts the data in the time (blue boxes), space (red boxes), and/or frequency (green boxes) domains. While variations in the order of these steps are permissible based on experimental considerations or specific EEG features under investigation, any deviations should be well-justified. Reproduced from [39].

subsequent analysis. Time-domain features, such as mean amplitude, peak latency, and slope, can provide insights into the temporal characteristics of EEG signals. Frequency-domain features, like power spectral density and spectral entropy [42], reveal information about the underlying neural oscillations. Time-frequency analysis, achieved through techniques such as Short-Time Fourier Transform (STFT) or wavelet analysis [43], uncovers how different frequency bands contribute to neural processing.

### Statistical Analysis

Both resting-state and ERP analyses often require statistical methods to draw meaningful conclusions from EEG data. In rs-EEG, graph theory metrics assess the topology of functional brain networks [44], revealing key nodes and their interactions. Hypothesis testing, permutation testing, and cluster-based methods [45] are employed to identify significant differences between conditions or groups. ERP analysis involves time-locking EEG traces to specific event onsets, followed by averaging to enhance the signal-to-noise ratio. Sta-

tistical tests, such as t-tests or ANOVAs, are used to identify significant differences in ERP waveforms between experimental conditions.

#### Source Localization

Incorporating source localization methods enhances the spatial precision of EEG analysis. Techniques such as Low-Resolution Brain Electromagnetic Tomography (LORETA) [46], Minimum-Norm Estimation (MNE) [47], and beamforming [48] estimate the neural sources responsible for the recorded EEG signals, as further described in Chapter 6. These methods enable researchers to infer the brain regions contributing to observed EEG patterns and gain insights into the underlying neural processes.

In conclusion, processing rs-EEG or ERP requires a comprehensive approach involving preprocessing, feature extraction, statistical analysis, and source localization. These steps collectively contribute to unveiling the intricate neural dynamics and cognitive processes that underlie brain function. By harnessing the power of these techniques, researchers can gain deeper insights into brain activity and cognition, paving the way for a better understanding of neural processes and their implications for human behavior and health.

### 1.2.2 PSG Signal Processing

Signal processing for PSG involves a series of steps aimed at extracting valuable information from the complex data collected during sleep studies. The processing of PSG data encompasses preprocessing, feature extraction, and clinical analysis.

#### Preprocessing

Similarly to EEG, the initial step in PSG signal processing is preprocessing, which plays a vital role in ensuring data quality and reliability. PSG data is prone to various artifacts, including those caused by body movements, electrode detachment, and interference from external sources. The approach to mitigating these artifacts closely mirrors the methods elaborated in Section 1.2.1.

### Feature Extraction

Feature extraction is a crucial step in PSG signal processing, as it transforms raw data into meaningful information for analysis. Various physiological and temporal features are extracted, depending on the research or diagnostic goals:

- Time-domain Derived Features: Features like heart rate variability (HRV) and abdominal-thoracic phase-shift are derived from PSG signals. HRV can be indicative of autonomic nervous system activity during sleep [49], while shifts in abdominal and thoracic movements may indicate respiratory-related disorders or alterations in breathing patterns [50].
- Respiratory Metrics: PSG includes sensors to monitor breathing, and features related to respiratory events, such as AHI [51] and oxygen desaturation index (ODI) [52], are extracted to assess sleep-related breathing disorders like sleep apnea.
- Movement Patterns: Features related to muscle tone and movement, such as periodic limb movement index (PLMI) [53], are derived from the EMG signal to detect movement-related disorders.

### Clinical Analysis

Diverse analyses can be conducted on PSG data:

- Sleep Staging: PSG data is typically used to classify sleep into different stages, including wakefulness (W), NREM sleep stage 1 to 3 (N1, N2, N3), and REM sleep (R). In clinical settings, sleep stages are manually scored by sleep experts.
- Event Detection: PSG data is analyzed to detect and quantify specific events, such as apneas and hypopneas in sleep apnea diagnosis.
- Pattern Recognition: By scrutinizing patterns and trends within PSG data, researchers and clinicians can discern abnormal sleep patterns and disorders, aiding in comprehensive diagnosis and treatment planning.

In conclusion, PSG signal processing empowers clinicians to classify sleep stages, detect respiratory events, and identify symptomatic sleep patterns.

This comprehensive approach is pivotal for advancing our understanding of sleep physiology and diagnosing sleep-related disorders effectively.

### 1.2.3 Machine Learning

Machine Learning (ML), a subset of AI, is a computational paradigm that endows computer systems with the ability to learn and make predictions or decisions based on data, all without explicit programming. This field has revolutionized biomedical signal processing by introducing data-driven approaches that can unveil intricate patterns, relationships, and insights from complex physiological data. In the realm of ML, a fundamental distinction exists between "classical" machine learning algorithms and Deep Learning (DL) models. The classical machine learning approaches necessitate handcrafted features extracted from the signals as input. Examples of these classical methods include logistic regression, decision trees, Hidden Markov Models (HMMs), and Support Vector Machines (SVMs). In contrast, DL models obviate the need for manual feature extraction. These models possess the unique ability to process raw data directly, enabling them to discern and exploit hidden structures that may not be explicitly represented by conventional handcrafted features. Different DL models are designed for various purposes and can be employed with diverse input data for a wide array of applications.

#### Multilayer Perceptron

The cornerstone of historical DL models is the Multilayer Perceptron (MLP), characterized by multiple layers of interconnected artificial neurons. These layers typically include an input layer, one or more hidden layers, and an output layer, as shown on the 3-layer network in Figure 1.9. The introduction of hidden layers injects non-linearity into the model, empowering it to learn intricate patterns and representations from the input data.

Each artificial neuron within these networks processes input data and produces an output, with each input associated with a weight that determines its significance in the computation. These weights enable the calculation of a weighted sum to ascertain the neuron's output. Neurons often incorporate a bias term, an additional parameter that introduces a shift to the weighted sum, allowing

neurons to model relationships that do not necessarily pass through the origin. The adjustment of these weights and biases constitutes a crucial facet of the neural network training process.

Non-linearity is introduced into a neuron's response through an activation function. This function decides whether the neuron should "fire" or become active based on the weighted sum of its inputs. Common activation functions encompass the sigmoid function, rectified linear unit (ReLU) [54], and hyperbolic tangent (tanh), each tailored to different purposes and selected based on the specific problem being tackled.

The training of an ANN entails feeding the network with training data, comparing the network's predictions to actual target values, and iteratively updating its parameters (weights and biases) using optimization techniques such as gradient descent [55] to minimize prediction errors.

Despite the inherent simplicity of its components, MLPs exhibit a remarkable capacity to represent highly complex functions [56]. Consequently, they are capable of achieving exceptional performance on complex tasks.

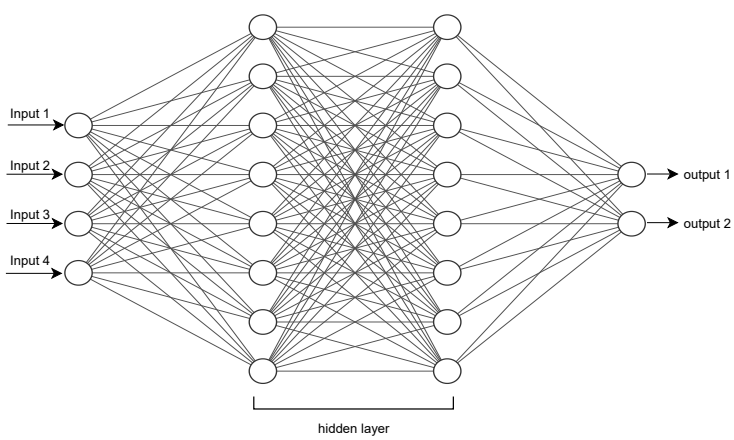



**Figure 1.9.** Three-layer Multilayer Perceptron (MLP) architecture. This MLP comprises an input layer with four input neurons, one hidden layer, and an output layer with two outputs. Each connection between neurons represents a weighted connection, and each neuron incorporates an activation function.

The primary limitation of MLPs pertains to the input format, which is confined to discrete data types, such as patient demographic information. Consequently, when applying these models to time-series data like EEG or PSG signals, a common approach involves transforming the sequential data into numerical feature vectors. In this scheme, specific features should first be extracted from the raw data before being fed to the neural network.

To address these constraints, alternative architectures like Convolutional Neural Network (CNN) or Recurrent Neural Network (RNN) can be employed to directly operate on raw data. Globally emphasizing temporal relationships within the input data, RNNs tend to overlook spatial relationships [57]. Therefore, we have chosen to focus on CNNs.

### Convolutional Neural Network

In contrast to MLPs, CNNs have emerged as a powerful architecture for processing biomedical data, particularly when working with complex and multidimensional inputs like images or time-series signals. CNN are renowned for their ability to capture local patterns and hierarchies within data, making them exceptionally suitable for tasks that require understanding the spatial and temporal relationships in biomedical signals.

As shown in Figure 1.10, CNNs operate by employing a set of learnable filters, also known as convolutional kernels [58]. These filters convolve across the input data, extracting relevant features and patterns at different spatial or temporal scales. This process is akin to how the human visual system recognizes patterns and objects. In the context of biomedical data, CNNs excel in identifying distinctive features within signals, such as specific waveform shapes in ERPs or characteristic patterns in medical images like X-rays and magnetic resonance imagings (MRIs).

One of the most significant advantages of CNNs is their ability to automatically learn and adapt feature detectors directly from the raw input data, obviating the need for manual feature engineering. This feature is particularly advantageous when handling two-dimensional biomedical signals, such as EEG or PSG data, where the various sensors (referred to as channels) are arranged along the first dimension, while the temporal data accumulates along the second dimension.

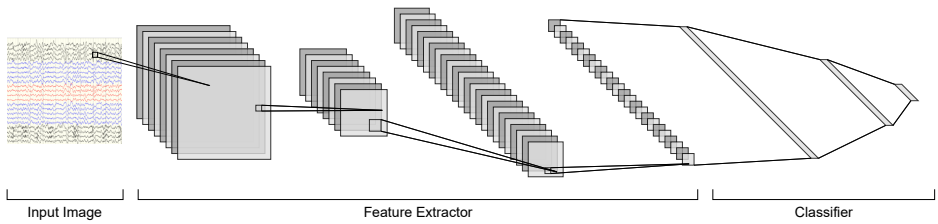



**Figure 1.10.** Convolutional Neural Network (CNN) Architecture for EEG Data Analysis. This figure illustrates the CNN structure for processing EEG data. The CNN comprises three main components: the input image, where raw EEG data is provided as input; the feature extractor, responsible for automatically identifying relevant patterns within the EEG data; and the classifier, which categorizes the EEG signals into distinct classes or states. This end-to-end approach eliminates the need for manual feature extraction and enables comprehensive analysis of EEG signals for various applications in biomedical signal processing.

When employed as an encoding method, as depicted in Figure 1.10, CNNs exhibit considerable prowess in handling classification and regression tasks after being trained in a supervised manner. CNN encoders have found diverse applications in the biomedical field, including:

- Diagnosis: analyzing medical images (e.g., mammograms, histopathology slides) to detect anomalies or pathologies.
- Prediction: forecasting seizures in epileptic patients or predicting sleep stages in PSG data.
- BCI: enabling individuals to control devices or communicate directly through the analysis of brain signals, including EEG and fMRI data.
- Classification: classifying ECGs into different arrhythmia types or rs-EEG signals into cognitive states.
- Automation: segmenting medical images or identifying sleep apnea events in PSG data.

Other purposes such as data denoising and dimensionality reduction can be harnessed by using an architecture that merges both an encoder and a decoder. Such architectures fall under the category of Auto-Encoders (AEs).

### Auto-Encoder

AEs are part of the broader family of unsupervised learning techniques and are particularly adept at capturing informative representations of complex input data.

The fundamental architecture of an AE consists of two main components: an encoder and a decoder. The encoder maps the input data into a lower-dimensional latent space, effectively compressing the input information into a compact representation. This process is akin to dimensionality reduction and feature extraction. The decoder, conversely, reconstructs the input data from the latent space representation [59]. When made of convolutional layers, as illustrated in Figure 1.11, this neural network is called Convolutional Auto-Encoder (CAE).

The training process of AEs primarily aims to minimize the difference between the input data and the output data, effectively encouraging the network to learn a compressed representation of the input.

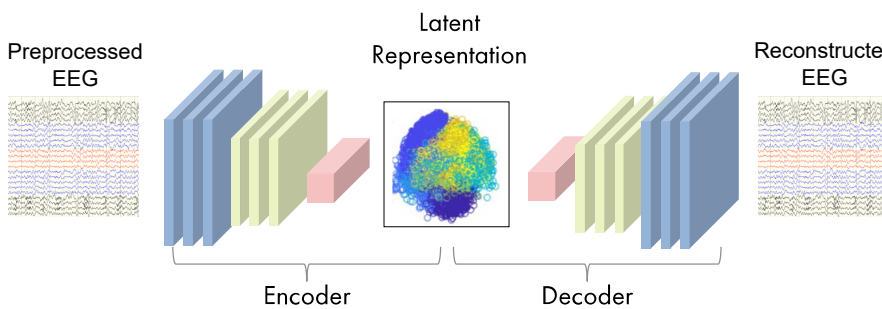



**Figure 1.11.** Convolutional Autoencoder for EEG Data. This figure illustrates a CAE architecture designed for EEG data processing. The CAE consists of an encoder, responsible for capturing salient features from preprocessed EEG input, a decoder for reconstructing EEG data from the learned features, and a latent representation, where the compressed information about the EEG signals is stored. Adapted from [60].

Beyond AEs, Variational Autoencoders (VAEs) introduced probabilistic frameworks that enabled the generation of novel data samples. This innovation unlocked applications like data augmentation, crucial for enhancing the ro-

bustness of deep learning models. The introduction of Generative Adversarial Networks (GANs) marked a paradigm shift, enabling the generation of data samples that resembled real-world distributions. GANs engendered a renaissance in data synthesis and augmentation, empowering data-hungry applications within biomedical signal processing.

The subsequent emergence of transformers reinvented the landscape of sequence data processing. Originally designed for natural language processing, transformers leveraged self-attention mechanisms to model relationships between distant data points. The application of transformers to sequential biomedical data, such as time-series EEG signals, unlocked unparalleled insights into temporal patterns and dynamic phenomena.

### Impact on Signal Processing

DL models have significantly transformed the way we process biomedical signals throughout the workflow:

- Preprocessing: Denoising DL methods, particularly AEs, have gained prominence. AEs are capable of reconstructing signals while effectively removing noise. They achieve this by learning from large datasets of noisy and clean signals, allowing for robust noise reduction [61, 62].
- Feature Extraction: As aforementionned, DL algorithms can automatically extract relevant features for a specific task, such as detecting specific brain states or events, without the need for handcrafted feature engineering. However, this automatic feature extraction makes the decision making process almost impossible to interpret by clinicians.
- rs-EEG Pattern Recognition: DL models can identify functional connectivity patterns among different brain regions. They can reveal hidden relationships within EEG data, shedding light on complex brain network dynamics that were previously challenging to decipher [63].
- ERP Waveforms: ANNs can learn complex temporal patterns within EEG data and can therefore automatically identify and classify ERP components, reducing the manual effort required for ERP waveform identification [64].

- EEG Source Localization: Still in preliminary states, some CNN models have already shown their ability to solve the inverse problem in a distributed dipole model based on simulated EEG data [65].
- Sleep Staging: DL methods are being actively explored for automating the classification of sleep stages in PSG data. They have the potential to enhance the accuracy and efficiency of sleep stage identification, which traditionally relies on manual expert annotation [66].
- Respiratory Event Detection: DL models are increasingly applied to detect respiratory events in sleep studies, such as apneas and hypopneas. They leverage the temporal information in PSG signals to identify these events more accurately and robustly [67].

In summary, ML models have emerged as powerful tools in biomedical signal processing, offering a versatile and data-driven approach to a wide range of applications, from disease diagnosis to brain-computer interfaces. They have greatly simplified and enhanced various processing steps in biomedical signal analysis. However, the choice of the most suitable ML algorithm for a specific application can be a complex task. DL algorithms, while often improving performance, introduce challenges due to their "black box" nature, making it difficult to interpret their decision-making processes. This lack of interpretability can hinder adoption by clinicians who require transparency in decision-making. Consequently, the field of explainable AI has gained significant attention, aiming to bridge the gap between ML's powerful capabilities and the need for interpretable and trustworthy results. This ongoing effort to make ML models more transparent and accountable is crucial for their successful integration into the medical field.

## 1.3 Explainable AI

The advent of xAI stands as a pivotal response to the opaque nature of complex machine learning models, opening new avenues for comprehending their decision-making processes. This paradigm shift marks the confluence of historical precedent and contemporary demand for transparent and accountable AI systems.

The history of xAI is intrinsically tied to the rise of complex models, like deep neural networks, whose inner workings often appear as "black boxes". Initially, we primarily worked with relatively simple models, inherently interpretable ones like decision trees and linear regression. In these models, it was straightforward to trace model decisions back to input features. However, as the field progressed, the allure of highly accurate yet intricate models posed a challenge to comprehensibility. Inherent and post-hoc explainability emerged as two fundamental approaches within the realm of xAI.

Inherent explainability centers on designing models that are intrinsically interpretable. Linear models, decision trees, and rule-based systems fall within this category, as their decision-making processes can be articulated through human-readable rules. However, the trade-off between model complexity and accuracy becomes pronounced in complex tasks, limiting the efficacy of this approach.

In contrast, post-hoc explainability aims to unveil the rationale behind the predictions of complex models after they have been trained. This approach encompasses a range of techniques, including visual explanations, input modification methods, and deconvolution-based methods. Since the latter requires detailed knowledge about the model used to perform the inverse operations [68], which is not always feasible, we will focus here on visual explanation and input modification techniques.

Visual explanation methods such as Class Activation Mapping (Class Activation Mapping (CAM)) and Partial Dependence Plots (Partial Dependence Plots (PDPs)) enhance explainability by providing a visual representation of the explanation. Techniques like CAM directly highlight the regions of input data that contribute most to a particular prediction in the form of a heatmap, as shown in Figure 1.12. On the other hand, PDPs conveys information about the relationship between variables, as exemplified in Figure 1.13.

Input modification methods apply specific changes to the input data to quantify the influence of specific input features on model predictions, shedding light on which aspects drive specific decisions. Techniques like Recursive Feature Elimination [71], SHapley Additive exPlanations (SHAP) [72], and Randomized Input Sampling for Explanation (RISE) [73] fall into this category. The Local Interpretable Model-Agnostic Explanations (LIME) algorithm can also be categorized within this group, although it necessitates an initial step of con-

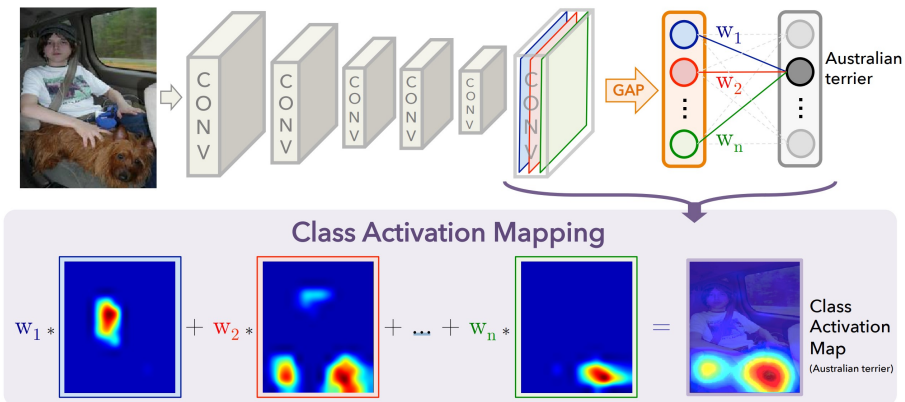



**Figure 1.12.** Class Activation Mapping (CAM) Example. CAM applied to Australian Terrier detection displays the filters from the penultimate layer and the resultant weighted sum of their activations to identify class-specific regions. This technique leverages gradients to identify important regions within images. Reproduced from [69].

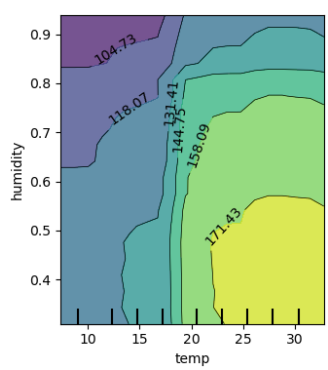



**Figure 1.13.** Example of Partial Dependence Plot (PDP). This figure illustrates the relationship between Temperature (x-axis) and Humidity (y-axis). The PDP visualizes how changes in Temperature and Humidity affect the predicted outcome, with color intensity representing the predicted values (here, the number of bike rentals). Adapted from [70].

structing a linear model that approximates the local boundaries of the initial complex model [74], as illustrated in Figure 1.14.

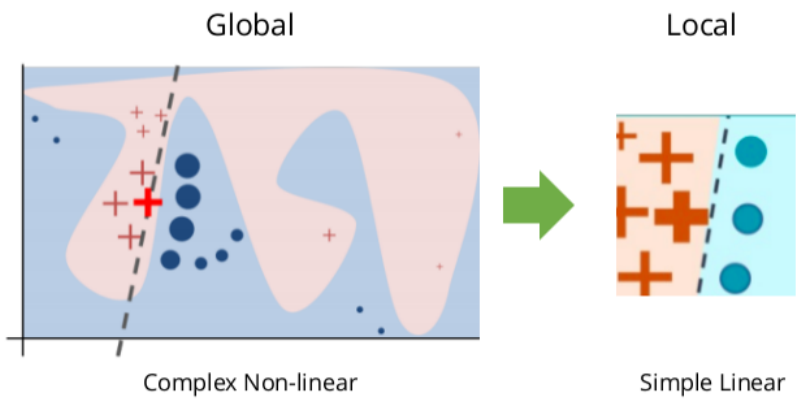



**Figure 1.14.** Intuition for Local Interpretable Model-Agnostic Explanations (LIME). The left side of the figure represents a complex, black-box decision function f (shown in blue/pink) that is unknown to LIME. The bold red cross indicates the instance being explained. LIME constructs a subsample of instances by making slight alterations to the target instance, obtains predictions from f, and weighs them based on their proximity to the instance being explained (indicated by size). On the right side, a linear explanation (represented by the dashed line) is learned to approximate the complex model locally, providing a more interpretable understanding of how the model behaves in the vicinity of the explained instance. Reproduced from [75].

Dhurandhar *et al.* proposed to go beyond highlighting important features for the proper classification of the input by also including relevant negative features whose presence would change the classification of the input [76].

In the realm of DL, certain models offer better explanations than others. For our research, which primarily involves image-like inputs (as discussed in Section 3.2), the two most performing architectures are CNNs [58] and Vision Transformers (ViTs), the architecture of which is depicted in Figure 1.15 [77].

These models predominantly rely on heatmap-based explainability approaches to illuminate their decision-making processes.

CNNs generate feature maps that can be visually interpreted, allowing us to gain insight into the model's behavior by examining learned features or activations in various layers. They excel at capturing local features in images, like edges or textures, yielding visually meaningful and interpretable features that shed light on how the model processes input images.

ViTs, in contrast, are tailored to capture global contextual information, making them potentially more interpretable in tasks that involve long-range dependencies or global context. Their hierarchical structure, featuring self-attention heads, can be visualized and interpreted individually. This provides valuable insights into how different heads attend to distinct features or regions within input images [78]. While ViTs indeed produce attention maps for interpretability, understanding the intricate interactions among self-attention heads remains a challenge. With multiple heads attending to different regions, comprehending the interplay between them and grasping the reasoning behind model predictions can be complex. To gain a more comprehensive understanding, we often need to employ saliency methods in addition to attention maps [79]. However, these approaches may yield results that are not always reliable or intuitive, as shown by Kindermans *et al.* [80].

In light of these considerations, our choice for our specific application has leaned towards CNNs. Their convolutional filters are easier to interpret compared to the multi-head long-range attention maps of ViTs. While CNNs limit us to local interactions, this limitation aligns with our need for interpretable explanations. The addition of residual connections to a CNN could potentially address the challenge posed by local interactions. However, it's essential to investigate how such additions might impact the model's explainability. Recently, Bohle *et al.* proposed a novel architecture for holistically explainable ViTs [81]. Exploring this avenue could potentially yield the best of both worlds.

While both inherent and post-hoc approaches offer benefits, they also grapple with inherent challenges. On one hand, inherent explainability may restrict model expressiveness, limiting their capacity to capture intricate patterns in data. One the other hand, post-hoc explainability, though valuable for complex models, frequently grapples with the balance between interpretability and

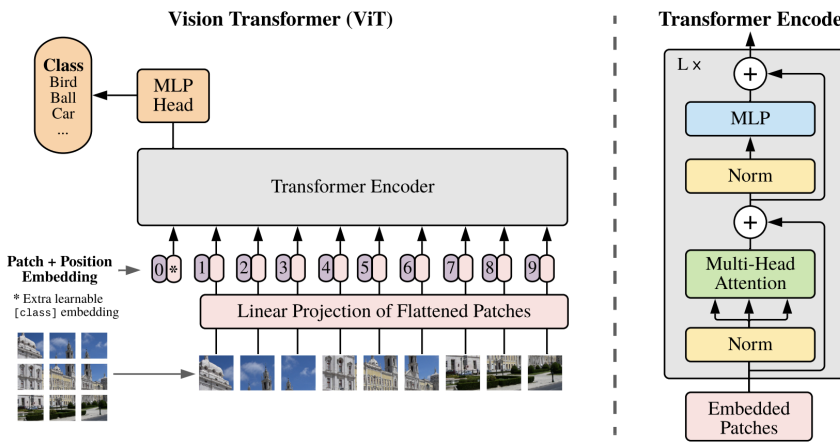



**Figure 1.15.** Vision Transformer (ViT) Architecture. The ViT architecture, depicted on the left, begins by partitioning the input image into consistent patches, which are subsequently linearly transformed. Position embeddings are introduced to these embeddings, creating a sequential representation of vectors. This sequence is then processed through a conventional Transformer encoder, elaborated in detail on the right. To facilitate classification, an extra learnable classification token is incorporated within the sequence. Reproduced from [77].

understandability [82]. Complex algorithms like deep learning models often depend on specific features that elude human perception. Consequently, the explanations they produce may have limited significance. Furthermore, striving for explainability may lead to oversimplification, obscuring subtle nuances and interactions. It can also be influenced by the human inclination to interpret results in a overly positive manner.

To create more human-understandable explanations, Chen *et al.* introduced the concept of prototypical explanations [83]. These explanations involve comparing the input to a prototype of each class and assigning it the label of the closest prototype. In this context, an explanation corresponds to the portion of the input that is most similar to the prototype. While this explanation method aligns with the human decision-making process, which often relies on comparisons, it presents challenges related to how prototypes are constructed and which parts of them are relevant for the given task. Wang *et al.* proposed an approach to encourage models to focus on human-understandable cues during training by removing specific imperceptible features from the inputs [84]. The challenge, therefore, lies in identifying these features before training the deep learning model. Building upon these ideas, this thesis introduces a novel *human-centered explainability* approach that leverages similarities and differences among all items in the training dataset.

Further insight into comprehending AI models includes exploring their interpretability. Instead of being restricted to understanding why a model makes a decision, it delves into how the decision is made by analyzing all the internal operations and unveiling the entire path from the input to the output [85]. Interpretability represents a whole field of study and falls outside the scope of this thesis.

xAI extends beyond technological aspects, encompassing ethical and regulatory dimensions. In Europe, the call for a more ethical AI deployment led to the establishment of the AI Act [86]. This groundbreaking regulation categorizes AI systems based on the level of risk they pose to users. Such regulatory advancements pave the path towards the development of transparent AI models that instill trust, mitigate biases, and facilitate accountability, especially in critical domains like medical diagnostics.

## 1.4 In Brief

Summary of Chapter 1

- EEG is a valuable tool for capturing rapid changes in brain activity and exploring specific frequency bands, though it has inherent spatial limitations. Noise and artifacts often affect EEG signals.
- PSG incorporates a diverse array of sensors distributed throughout the body, providing a comprehensive view of sleep architecture and patterns.
- The advent of machine learning has revolutionized biomedical signal processing, reshaping workflows across various applications.
- Deep learning algorithms, while powerful, pose challenges due to their opacity. Efforts are underway to enhance their transparency and interpretability for clinical use, giving rise to the emerging field of explainable AI.
- Current post-hoc explainability methods face a persistent struggle in achieving a balance between interpretability and human comprehension, a key area of exploration in this evolving field.

# Chapter 2

# Potential Biases

## Contents


2.1 Planning Biases . . . . . 43
2.2 Data Collection Biases . . . . . 44
2.3 Analysis Biases . . . . . 45
2.4 Publication Biases . . . . . 46
2.5 Biases in Focus . . . . . 47
2.6 In Brief . . . . . 49


This chapter delves into the broad landscape of biases in biomedical signal analysis, shedding light on their prevalence and impact. These biases, scattered throughout the research process, pose significant challenges, ultimately compromising the quality and reliability of biomedical studies. Navigating the extensive array of biases in modern research can be daunting, with Chavalarias and Ioannidis documenting a staggering 235 distinct biases potentially affecting biomedical research in their 2010 review [87].

In this thesis, we narrow our focus to the primary biases directly impacting research reliant on biomedical signals, notably EEG and PSG, which are our signals of interest. Inspired by Panucci and Wilkins' methodology [88], we have categorized these biases according to the stages of research where they manifest, from study planning to publication. The identified biases and their classification are visually represented in Figure 2.1.

Through an examination of these biases, this chapter underscores the importance of developing strategies to mitigate their effects, ensuring the integrity of research outcomes and the advancement of medical science as a whole.


**Figure 2.1.** Visualization of potential biases in medical research. This figure classifies the potential biases into four main categories retracing each step of a medical research from the study planning to the publication.

## 2.1 Planning Biases

Certain biases manifest during the planning phase of a study, making it crucial to design research carefully. Flaws in the initial planning can irreparably compromise a study's validity and the generalizability of its findings. In this phase, our biases of interest are:

- **Selection Bias:** This bias impacts the composition of the study population. Inadequate randomization or non-representative sample selection can skew results, making them unrepresentative of the broader patient population. Biases originating from initial patient selection can subsequently influence the interpretation of signal patterns and limit the external validity of the research [89].
- **Classification Bias:** Classification bias emerges when we possess incomplete information about study participants, leading us to categorize some incorrectly into a specific group (control/patient), when they should belong to the other group or even be excluded from the study [90, 91].
- **Confounding Bias:** Confounding bias involves establishing a false association between the desired outcome and a factor not causally related to it. These factors often stem from uncontrolled experimental conditions, referred to as confounders or covariates. Machine learning algorithms can inadvertently exploit this type of bias, making classification tasks easier while overlooking the actual physiological effects under investigation. To mitigate this bias, researchers traditionally strive to balance critical known covariates across experimental conditions [92]. However, identifying all potential covariates is a daunting task, and even if feasible, attempting to balance them all would significantly reduce the data's variability.

Selection and classification biases can only be reduced through a meticulous examination of all participants and are study-specific. In contrast, confounding bias can be mitigated by modeling known covariates and regressing them out before drawing conclusions about related findings [93]. This method has been implemented in the context of ERP data to standardize the assessment of the separability of category-dependent part of the evoked response from the remaining EEG signals, as elaborated in Section 4.

## 2.2 Data Collection Biases

Data collection biases originate during the process of gathering and recording data, potentially introducing inaccuracies into the recorded information. This section examines three prominent data collection biases:

- **Measurement Bias:** Measurement bias arises when data collection instruments or techniques inaccurately assess the variable of interest [94, 95]. In the context of biomedical signal analysis, such as EEG, measurement bias may result from imprecise sensors, suboptimal data acquisition procedures, or the presence of artifacts. Such bias can distort the recorded signals, impacting subsequent analyses.
- **Observer Bias:** Observer bias occurs when individuals involved in data collection, interpretation, or analysis allow their subjective perceptions or expectations to influence the experiment [96]. This bias can lead to changes in participant behavior or alter the experimenter's approach, potentially favoring a specific group of participants and distorting the investigation of physiological effects. Observer bias underscores the importance of standardized protocols and double-blind procedures to minimize the influence of human subjectivity.
- **Performance Bias:** Performance bias refers to inconsistencies in data collection stemming from variations in operator skills, equipment calibration, or other external factors [88]. These inconsistencies can result in unreliable data, undermining the overall quality and trustworthiness of the findings.

Mitigating observer and performance biases requires rigorous training, standardization of data collection procedures, and continuous monitoring to ensure data collection remains objective and free from external variations.

Addressing measurement bias requires careful data collection procedures. Furthermore, a proper preprocessing phase is essential to reduce remaining noise and artifacts. To this end, we propose a preprocessing framework tailored to ERP data. This framework selectively identifies and reduces artifacts within the EEG signals, enhancing the accuracy and reliability of the collected data. Detailed information on this approach is provided in Chapter 5.

## 2.3 Analysis Biases

Analysis biases manifest during the data processing and interpretation phase, impacting the reliability and validity of study outcomes. This section explores three prominent analysis biases:

- **Modeling Bias:** Modeling bias emerges when data representations are employed to model specific features. In EEG analysis, for instance, it is common practice to derive brain source activity, connectivity, or spectrograms from the input time-series data. These transformations rely on specific parameters which inadvertently introduce distortions [97, 98]. Consequently, this bias has the potential to misrepresent the true underlying patterns within the data, ultimately leading to erroneous conclusions.
- **Confounder Exploitation:** Confounder exploitation bias emerges in complex "black-box" models like deep learning algorithms, which obscure their decision-making processes. This opacity may result in the unintended utilization of confounding factors rather than capturing the desired physiological phenomena [99]. Blindly relying on such algorithms can lead to critical errors when implementing them in clinical settings.
- **Expectation Bias:** Expectation bias materializes when researchers hold preconceived notions or expectations that influence data interpretation [100]. These pre-existing beliefs can trigger confirmation bias, causing researchers to unintentionally favor data supporting their expectations and compromising the study's integrity.

Mitigating expectation bias is challenging, as it is rooted in the human tendency to interpret results in a positive light, aligning with initial hypotheses. Addressing this bias involves instilling objectivity and robust statistical techniques in researchers, favoring a rigorous approach over a bias towards positive results.

To counteract modeling bias in EEG studies, we propose implementing a validation framework for brain source localization, detailed in Chapter 6. This framework establishes a standardized benchmark, ensuring the accuracy and transparency of EEG source reconstruction techniques.

Additionally, we present a novel approach, *human-centered xAI*," to mitigate confounder exploitation bias, as described in Chapter 7. This innovative

method enhances the transparency of deep learning models by leveraging intra and inter-subject similarities to extract the most influential features used by the model in its decision-making process. We show the effectiveness of this approach in a sleep apnea severity scoring task.

## 2.4 Publication Biases

Publication biases are intrinsic to the dissemination of research findings and can significantly impact the scientific literature [101]. These biases manifest after the experimental phases and involve the reporting and referencing of study outcomes. In this section, we address three key publication biases:

- **Inflation Bias:** Inflation bias, also known as "p-hacking", refers to the tendency to exaggerate or overemphasize significant findings while downplaying or omitting non-significant. Researchers or journals may prefer to publish studies with statistically significant outcomes, potentially resulting in an incomplete and overly optimistic portrayal of the scientific landscape [102].
- **Reporting Bias:** Reporting bias arises when researchers selectively report specific aspects of their findings, making them less inclined to publish research with negative results [103]. Overall, studies reporting positive are more likely to be published [104]. This selective reporting can distort the overall perception of the research landscape, potentially leading to the adoption of flawed or incomplete conclusions.
- **Citation Bias:** Citation bias occurs when studies confirming existing beliefs or aligning with popular scientific trends receive more citations than those challenging prevailing ideas or introducing novel concepts [105]. This bias can perpetuate scientific paradigms, making it challenging for innovative or dissenting research to gain recognition and influence.

Publication biases pose significant challenges to the integrity and comprehensiveness of scientific knowledge. While not directly addressed in this thesis, an awareness of these biases is crucial for fostering a more transparent and open scientific discourse. Researchers should strive for transparency and honesty in reporting their findings, considering the broader implications of publication biases on the advancement of their respective fields.

Another classification worth noting is the distinction between random and systematic bias. Random biases arise due to sampling variability or measurement precision and are inherent to almost all quantitative studies, being minimizable but not entirely avoidable. In contrast, systematic biases involve reproducible errors leading to a consistently false pattern of differences between observed and true values [106]. This thesis primarily focuses on systematic bias, which relies more on the technical methods used rather than how they are applied to the experiment.

## 2.5 Biases in Focus

Among all the biases described in this chapter, only some of them are addressed in this thesis. These biases can be categorized into "white-box" and "black-box" biases. The former directly impact the recorded data or their related representations, while the latter emerge when non-transparent models are used to extract information from input data, essentially leading to *confounder exploitation* bias.

The selection of the targeted "white-box" biases has been done by evaluating which biases are predominant in the EEG dataset recorded from the priming experiment described in Section 3.1. These biases include:

1. *Confounding bias*: Noticing differences in the shapes of images from different categories, we explore the need to revise the entire experimental protocol to balance the confounders across categories. This is achieved by quantifying the separability between categorical and confounding effects (Chapter 4).
2. *Measurement bias*: Given that the recorded dataset is heavily affected by specific artifacts (eye blinks, head movements, and jaw contractions), we propose a dedicated preprocessing framework tailored to address these artifacts. This approach allows us to selectively retain valuable information from the polluted signals (Chapter 5).
3. *Modeling bias*: Due to the absence of information about the participants' head anatomy and certain imprecisions in the recordings of EEG electrode locations, modeling the data, especially in terms of reconstructing brain source activations, becomes challenging. To tackle this challenge,

we propose a validation framework to assess the accuracy of source reconstruction methods under these conditions (Chapter 6).

To address the various "white-box" biases, we introduce standardized frameworks with the aim of maximizing the reproducibility and adaptability of the proposed solutions.

Our work in reducing *confounder exploitation* bias resides in an innovative approach aimed at enhancing the transparency of deep neural networks. This approach, named *human-centered explainable AI*, is elaborated upon in Chapter 7. We apply this technique to PSG data in Chapter 8, leading to a novel severity measure for sleep apnea events and uncovering EEG biomarkers associated with severe sleep apnea events.

## 2.6 In Brief

Summary of Chapter 2

- Biases in biomedical research can be classified based on the stage of the experiment when they occur: planning, data collection, analysis, or publication.
- Planning biases arise from deficiencies in the experimental protocol design, including issues related to participant selection/classification and uncontrolled confounding factors.
- Data collection biases result from inadequate instruments/techniques, subjective perceptions/expectations, or inconsistencies across trials/participants.
- Analysis biases manifest post-experiment and can be caused by misrepresentation of data, inappropriate choice of analysis algorithms, or preconceived expectations regarding the final results.
- Publication biases stem from selective or biased reporting of study outcomes, influenced by the scientific community's expectations.
- This thesis focuses on specific biases, namely confounding bias, measurement bias, modeling bias, and confounder exploitation bias. These biases serve as the foundation for the proposed solutions.

# Chapter 3

# Datasets

## Contents


**3.1 Priming Dataset** . . . . . **52**
3.1.1 Stimuli and Experimental Task . . . . . 52
3.1.2 Participants . . . . . 53
3.1.3 Data Acquisition . . . . . 54
**3.2 Obstructive Sleep Apnea Dataset** . . . . . **56**
3.2.1 Experiment and Participants . . . . . 56
3.2.2 Data Acquisition . . . . . 56
3.2.3 Preprocessing . . . . . 57
**3.3 In Brief** . . . . . **60**


The accurate analysis of EEG and PSG data requires access to high-quality datasets that exhibit both diversity and representativeness within the target population. This chapter delves into the two datasets utilized in this thesis, each offering a unique perspective on various biases prevalent in biomedical signal research.

The first dataset, detailed in Section 3.1, comprises in-lab recordings obtained from 30 young, healthy subjects, aged between 18 and 35 years. These participants engaged in a visual task centered around a semantic priming paradigm, evoking distinctive ERP responses. The second dataset, as outlined in Section 3.2, encompasses clinical PSG signals gathered from a cohort of 60 patients diagnosed with obstructive sleep apnea (OSA).

Within this chapter, we provide insights into both datasets. This includes an overview of the experimental conditions, a description of the participants, and details on data acquisition procedures.

## 3.1 Priming Dataset

The experimental task designed to collect the EEG data and build our dataset is a semantic priming paradigm, as described in Section 3.1.1. Thirty healthy adults aged between 18 and 35 years old participated in the study (see Section 3.1.2), and EEG data were recorded with a Biosemi Active-Two system, as presented in Section 3.1.3.

### 3.1.1 Stimuli and Experimental Task

Our semantic priming paradigm is based on an intra-modal procedure consisting of processing a target-picture while ignoring a prime-picture [92]. The requested task was to answer the question: "Is the target-picture an existing entity?". The Figure 3.1 shows examples of stimuli for "yes" and "no" correct answer to the task. The participants answered through 2 manual press-buttons (yes or no). Stimuli of the semantic priming paradigm were designed according to three priming conditions: one semantically related condition (taxonomic (e.g. banana-peach) and thematic (e.g. banana-monkey) relation), one semantically unrelated condition (e.g. banana-chair) and one control condition composed of neutral trials (primers being geometrical shapes and targets random existing entity).

For each category of trials, 38 pictures (19 naturals and 19 manufactured) appeared as target-pictures. Target-pictures appeared twice in the semantic conditions for both taxonomic and thematic trials. The semantically related and unrelated pairs constituted 26.5% of all the experimental items used (38 unrelated + 38 taxonomic + 38 thematic = 114 pairs). The primes in the related condition for manufactured objects were also used as primes in the unrelated condition for natural objects and vice versa. The Figure 3.2 presents examples of the different pairs. The remaining 73.5% of items made-up the control condition with a 2-levels classification (279 prime-target filler pairs (1st level control) + 38 neutral pairs (2nd level control) = 317 pairs). Filler pictures were abstract forms controlling the yes/no response and the neutral pairs controlling the effect of the primer.

| Category of Trials | Number of Pairs | % of Total Experimental Items |
|---|---|---|
| Semantic (Taxonomic) | 38 | 8.8% |
| Semantic (Thematic) | 38 | 8.8% |
| Unrelated | 38 | 8.8% |
| Neutral | 38 | 8.8% |
| Filler | 279 | 64.7% |

**Table 3.1.** Summary of Trial Categories

The trial distribution proposed in this study aims to prevent expectancy strategies [107]. For each stimulus, mean familiarity, age of acquisition (AOA), number of phonemes, lexical and visual frequencies (Lexique380) and visual complexity (five-point Lickert scale) were stored and the visual similarity between each primer-target pair was computed. These features are uncontrolled variables, we consider them as the covariates, or confounders, of our experiment. In total, 431 items were presented per subject, and the order of trials counterbalanced across subjects. In order to familiarize participants with the task, the first 12 pairs presented were training pairs.

### 3.1.2 Participants

Thirty healthy adults, right-handed, French native speakers with normal or corrected-to-normal vision participated in this study. Participants were recruited in the central region of Belgium (15 females out of the thirty participants). They were aged from 18 to 35 years. The mean age was 24.73 years (SD=3.94). Sociocultural level (Poitrenaud scale) was measured according to the highest level of education (1= Elementary School; 2= Middle School; 3= High School; 4= Bachelor Degree; 5= Master Degree; 6= Doctoral Degree). Handedness was assessed using a French version of the Edinburgh Handedness Inventory [108] with all participants being right-handed. Regarding the inclusion criteria, individuals who experienced substance abuse, epilepsy, neurological and/or psychiatric backgrounds were systematically excluded from the study. All subjects gave their informed written consent after the nature and the potential outcomes of the experiment had been explained. This study

**(a)** Existing item
(correct answer = yes)

**(b)** Non-existing item
(correct answer = no)

**Figure 3.1.** examples of stimuli for each answer to the task

(design and protocol) was approved by the Ethical Board Faculty of Psychology and Education of the University of Mons (Belgium) and was conducted in accordance with the Declaration of Helsinki. The participation was on a voluntary basis without financial compensation.

### 3.1.3 Data Acquisition

EEG data were recorded at a sampling rate of 2048 Hz with a Biosemi ActiveTwo system (BioSemi Biomedical Instrumentation, Amsterdam, the Netherlands. AD BOX amplifier) from 64 active Ag/AgCl electrode sites, with a Biosemi headcap arranged in a standard 10-20 layout. The EOG was recorded bipolarily from the outer canthi of both eyes and above and below the left eye. The ground electrode was placed on the forehead between Fp1 and Fp2. Impedance measurements were performed before and after the experiment to ensure the electrode impedance was kept below $10k\Omega$. The preprocessing of these ERP signals, constituting a part of the contributions of this thesis, is detailed in Chapter 5.

**(a)** Semantically related pair (taxonomic)

**(b)** Semantically related pair (thematic)

**(c)** Unrelated pair

**(d)** Neutral pair

**Figure 3.2.** examples of pairs with the primer on the left and the target on the right

## 3.2 Obstructive Sleep Apnea Dataset

Acquiring the OSA dataset played a crucial role in developing the xAI model designed to enhance the transparency of "black-box" algorithms. We established a collaborative effort with the Sleep Laboratory of the Centre Hospitalier Universitaire Saint-Pierre (CHU S$^{t}$-Pierre) to access data from patients with sleep apnea, which included basic EEG signals. This dataset served as a cornerstone in our research, illustrating the effectiveness of our model in uncovering EEG biomarkers. Additionally, it allowed us to devise a novel objective severity score for OSA assessments, providing an alternative to the widely debated AHI.

### 3.2.1 Experiment and Participants

The dataset built for this research comprises PSG data from 72 patients who underwent in-lab PSG sessions (each lasting $\geq$ 8 hours) in 2022 at the Sleep Laboratory, CHU S$^{t}$-Pierre, Brussels, Belgium. Clinicians manually annotated the recordings to identify sleep stages, apnea, hypopnea events, and arousal events following international guidelines. All the selected patients exhibited excessive obstructive respiratory events (apneas or hypopneas) during the night, with at least AHI$\geq$5. The sleep onset was determined when the first epoch of sleep occurs. A preliminary sleep questionnaire was performed and the protocol CE/22-03-03 was approved by the local ethical comittee of the CHU S$^{t}$-Pierre on March 14$^{th}$ 2022.

### 3.2.2 Data Acquisition

The PSG sensors include 2 EOG electrodes (EOG1 under the left eye and EOG2 above the right eye), thoracic and abdominal belt sensors (VTH and VAB) for monitoring respiratory motions, an ECG sensor, a pulse oximetry sensor recording Pulse Rate (PR) and Oxygen SAturation ($SAO_2$), a pressure probe measuring NAsal AirFlow (NAF2P), and 6 EEG electrodes. The reference EEG electrode is placed just above the nasion, and derivations are performed with a right mastoid electrode.

The raw signals were initially recorded at 200 Hz and then downsampled to 50 Hz for storage using the Medatec Brainnet Winrel 5.0 system. Subsequently, the data were converted into Python-friendly files using the MNE-Python package [109], which was employed for signal preprocessing. Since our analysis focuses on differences between apneic events, the dataset exclusively comprises OSA trials, each lasting for 60 seconds. These segments were extracted from manually labeled signals, starting 4 seconds before an OSA event.

### 3.2.3 Preprocessing

The exclusion of PSG trials, based on non-EEG electrodes, was determined by their peak-to-peak voltage (VPP): Trials with $VPP \leq 10^{-7}V$ or $VPP \geq 6 \times 10^{-4}V$ were excluded. Additionally, trials exhibiting statistical outliers in amplitude for ABdominal belt Voltage (VAB), THoracic belt Voltage (VTH), and NAF2P were rejected. A baseline correction was then applied using a 10-second segment preceding each trial.

Two additional signals were computed from the recorded data: 1) the Pulse Rate Variability (PRV), representing the difference between consecutive PR samples, and 2) the phase shift (Pshift), computed as the sample-by-sample phase difference between VAB and VTH phase signals, following the approach suggested by Varady *et al.* [110].

For clarity and simplicity, considering the reduced precision required for neural activity in OSA studies in this thesis, only the 3 left-hand side EEG electrodes were analyzed: FP1 (frontal electrode), C3 (central electrode), and O1 (occipital electrode).

The preprocessing of EEG signals adhered to the COBIDAS MEEG recommendations from the Organization for Human Brain Mapping (OHBM) [39]. Initially, bad trials were rejected through visual inspection. Subsequently, trials significantly affected by ocular artifacts were excluded based on the correlation between the EOG and the FP1 signals. Given that the EEG delta band power exhibits the most variation during OSA occurrences [111], our analysis focused on low-frequency EEG components, achieved by filtering the signals into 2Hz narrow bands: 0-2Hz, 2-4Hz, 4-6Hz, 6-8Hz, and 8-10Hz. Finally, a baseline correction was applied using a 10-second segment preceding each trial.

Being primarily concerned with specific EEG frequency ranges, our preprocessing pipeline focuses on fundamental EEG procedures. To further ensure the robustness of our analysis against potential artifacts, we may consider employing specific techniques, such as those proposed by S. Devuyst in her thesis [112]. These techniques include the removal of cardiac-related components from the EEG using the Independent Component Analysis - Ensemble Averaging (ICA-EA) method [113].

Normalization was performed on a per-channel basis using z-score normalization with values clamped within the [-3; 3] range. After the preprocessing phase, the final dataset consisted of 6992 OSA trials with 23 channels (15 filtered EEG channels and 8 non-EEG PSG channels). These trials were drawn from 60 patients and divided into a training set comprising 4660 trials from 48 patients (referred to as the trainset) and a validation set containing 2332 trials from the remaining 12 patients (referred to as the testset). Figure 3.3 shows an example of the final OSA trials obtained. Each trial can be visualized as a 2D data matrix, resembling an image, where each row represents data from a specific sensor, and the columns represent timestamps.

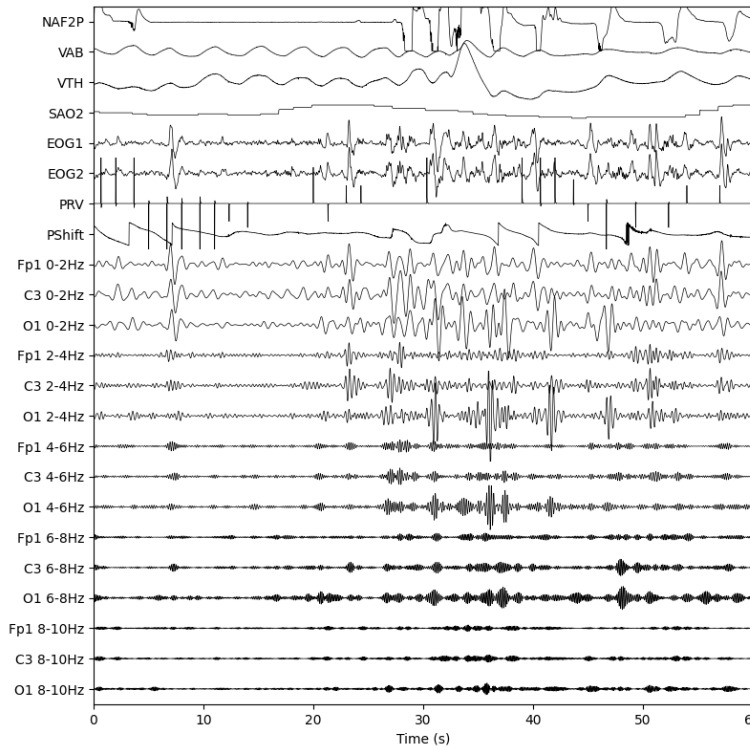



**Figure 3.3.** Example of a 60-second OSA trial showcasing various PSG channels including respiratory (VAB, VTH), oxygen saturation (SAO2), eye movements (EOG1, EOG2), heart rate variability (PRV), phase difference between respiratory signals (Pshift), and filtered EEG signals from three electrodes (FP1, C3, O1) across five 2Hz narrow frequency bands.

## 3.3 In Brief

Summary of Chapter 3

- The Priming Dataset serves as a fundamental cornerstone for the analysis of ERP data. Rooted in a visual priming experiment, this dataset underpins the development of our standardized frameworks. It encompasses investigations into the influence of confounding factors, ERP preprocessing, and brain source reconstruction.
- The Obstructive Sleep Apnea (OSA) Dataset meticulously extracts apnea and hypopnea events from clinical PSG recordings. It provides an invaluable opportunity to examine variations in apnea severity among patients.

# Chapter 4

# Planning Phase: Confounding Bias Evaluation

## Contents


**4.1 LIMO EEG** . . . . . **63**
**4.2 Method** . . . . . **64**
4.2.1 Variable Selection . . . . . 66
4.2.2 Linear Modeling . . . . . 68
4.2.3 Statistical Inference . . . . . 72
4.2.4 Effects Separability . . . . . 75
**4.3 Results** . . . . . **78**
4.3.1 Discussion . . . . . 84
**4.4 In Brief** . . . . . **87**


In the study design phase, we address the issue of confounding bias. Confounding bias occurs when specific features are unevenly distributed between different groups or conditions, potentially leading to biased conclusions.

Traditionally, addressing confounding bias involves attempting to balance confounding factors across experimental groups or conditions. However, achieving perfect balance is nearly impossible, and even if achieved, it may reduce the variability within and between groups, limiting the study's generalizability.

In the context of ERP studies commonly used in BCI experiments, we propose a framework specifically designed to quantify the influence of confounding variables on the studied conditions. Additionally, our framework assesses the separability between these confounder effects and the actual condition-related

effects. If these effects can be distinguished, confounding bias can potentially be mitigated through appropriate modeling of the known confounders. An overview of the framework applied to the priming experiment described in Section 3.1 is given in Figure 4.1.

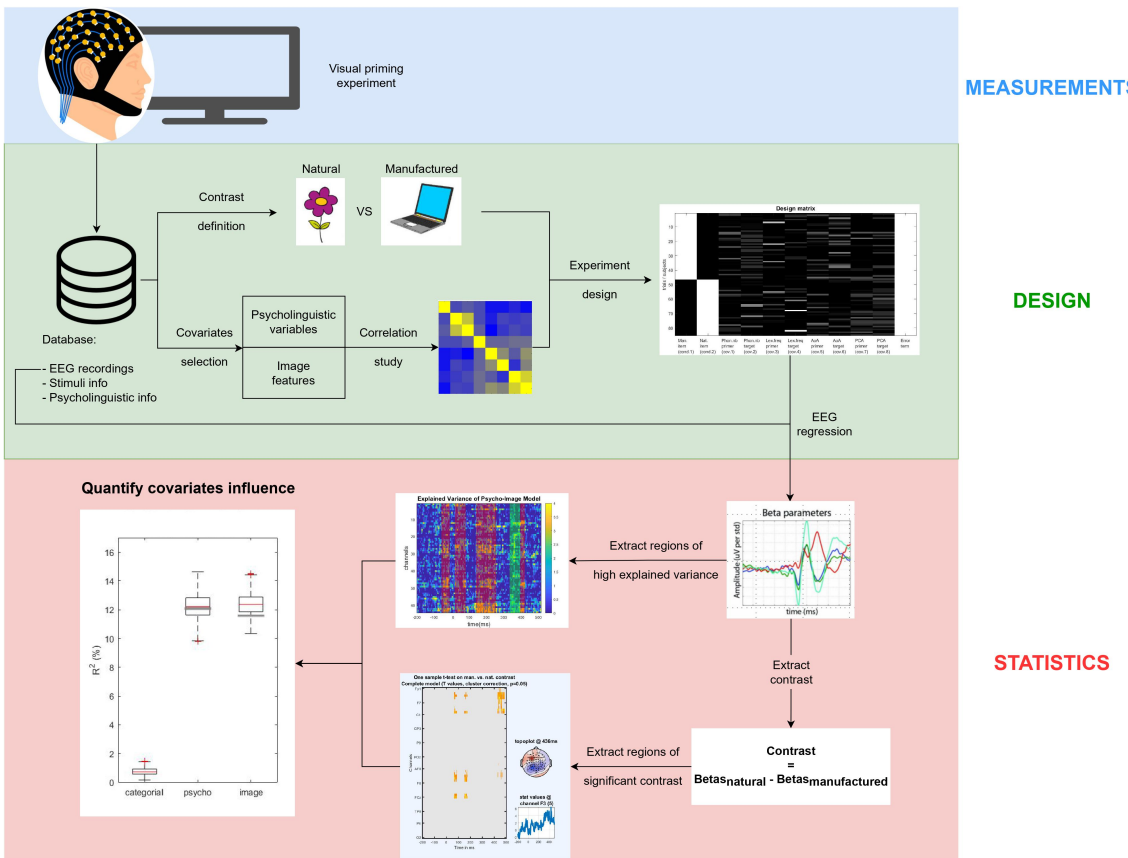



**Figure 4.1.** Framework Overview. The figure illustrates the EEG recordings in blue, the design of the model to include all the needed trials information in a standardized way in green, and the statistical analysis of the regressed ERP to identify covariates influence in red.

This chapter is based on the work led in close collaboration with Dr. Cyril Pernet (University hospital of Copenhagen), especially through the use of the LIMO EEG toolbox [114]:

- "*Biases in BCI experiments: Do we really need to balance stimulus properties across categories*", In Frontiers in Computational Neuroscience, November 22, 2022, DOI: 10.3389/fncom.2022.900571. [115]

First, Section 4.1 describes the LIMO EEG toolbox. Then, Section 4.2 presents the complete framework of the statistical analysis applied on our dataset. Finally, Section 4.3 shows the results of the analysis that consist of highlighting the influence of uncontrolled experimental variables, called covariates, on the statistical contrast between experimental conditions.

## 4.1 LIMO EEG

LIMO EEG is a comprehensive MATLAB-based toolbox designed to facilitate the analysis of electrophysiological data, particularly EEG data. This powerful tool is tailored to the needs of researchers and scientists working with EEG data, offering advanced statistical analysis capabilities and high temporal and spatial resolution. Key features and functionalities of LIMO EEG include:

- General Linear Model (GLM): LIMO EEG leverages the principles of GLM, enabling researchers to model and test various factors and conditions.
- Statistical Analysis: Researchers can perform a wide range of statistical analyses on EEG data, encompassing both univariate and multivariate approaches to identify significant effects and differences.
- Mass Univariate Analysis: The toolbox adopts a "mass univariate" approach, conducting statistical tests independently at each time point and electrode/channel. This approach offers exceptional temporal and spatial precision in analysis.
- Multiple Comparisons Correction: LIMO EEG provides correction methods for handling multiple comparisons, ensuring rigorous control over false positives.

- Visualization: Researchers can create insightful visualizations, including topographical maps of ERPs and visual representations of statistical results.
- Interactive GUI: The toolbox boasts a user-friendly graphical interface within MATLAB, making it accessible to researchers with varying programming backgrounds.
- Open Source: LIMO EEG operates under an open-source framework, granting users the flexibility to access, modify, and tailor the toolbox to their specific research requirements.
- Community Support: An active user community and support forums further enhance its utility, offering solutions to common issues and inquiries.

These features are encompassed in a two-level hierarchical procedure, as illustrated in Figure 4.2. In the first level, subject-specific parameters are estimated by regression for each time point and electrode separately. In the second level, these first-level parameters are aggregated across subjects to compute robust statistics. The inter-subject variance is modeled by the constant term in the first-level regression, while statistical tests (second level) are conducted on the regressed parameters, referred to as beta estimates.

In sum, LIMO EEG stands as a valuable tool for researchers in neuroscience, psychology, and related fields seeking to conduct meticulous statistical analyses on EEG data. Its versatility and precision make it particularly well-suited for research endeavors demanding rigorous control and adaptability within EEG data analysis workflows.

## 4.2 Method

In this section, we provide a detailed overview of the proposed framework designed to assess the impact of uncontrolled variables on data interpretation. This framework is applied to the priming experiment discussed in Section 3.1. Section 4.2.1 outlines the criteria and methodology for selecting the confounding variables of interest. In Section 4.2.2, we present the development of the linear model employed for regressing the EEG data. This approach builds upon the $1^{st}$-level analysis proposed by LIMO EEG, as illustrated in Fig-

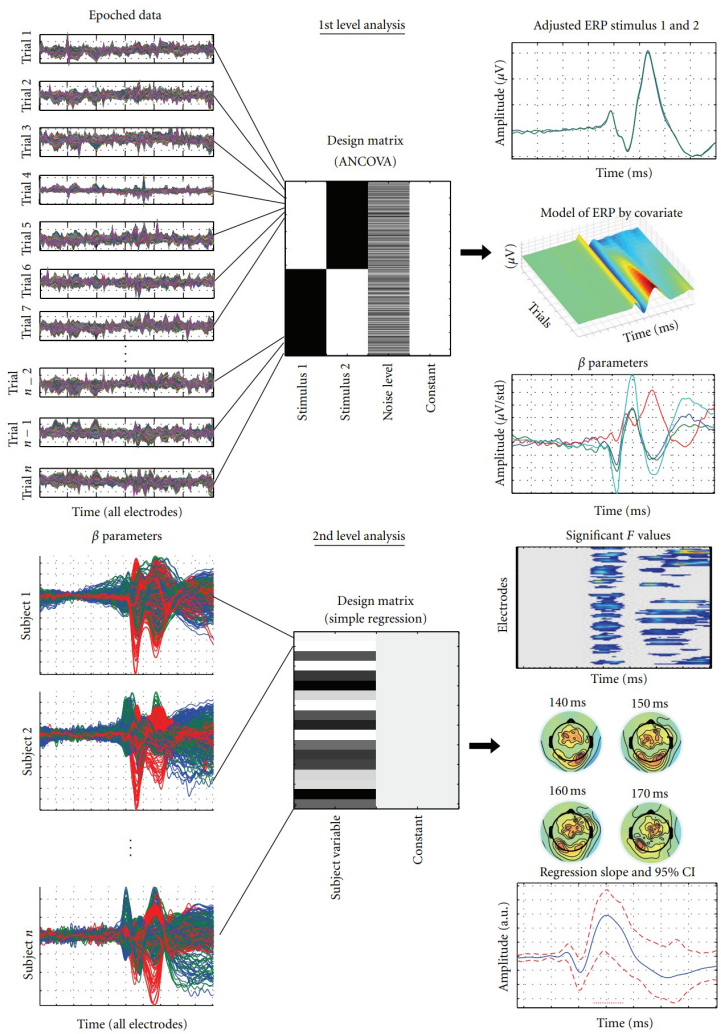



**Figure 4.2.** Hierarchical Analysis Procedure of the LIMO EEG toolbox. At the 1st level (top), individual subject data, comprising all trials, undergo analysis to compute estimated beta parameters. These beta parameters capture the effects of various experimental conditions as specified within the design matrix. At the 2nd level of analysis (bottom), the obtained beta parameters are scrutinized concerning the experimental conditions outlined in the 1st level. This involves testing for statistical significance across all subjects. Reproduced from [114].

ure 4.4. Section 4.2.3 delves into the statistical inference techniques used to extract regions of interest pertaining to either condition-specific effects or confounder effects. This analytical process is based on the $2^{nd}$-level analysis proposed by LIMO EEG, as depicted in Figure 4.7. In Section 4.2.4, we outline the procedure for evaluating the separability between condition-related effects and confounder effects. It's important to note that this study's scope is confined to the selected covariates, and no extrapolation beyond this scope is intended.

### 4.2.1 Variable Selection

The variable selection was done among the psycho-linguistic variables proposed by [116] as well as image properties[1] considering primer and target items separately (cf. Section 3.1.1). One additional covariate measured was the visual similarity between primer and target items. The value of this similarity is defined by a Likert scale [117] during the pre-test of the experiment (1 = primer has a totally different shape than target, 5 = the shape of the primer is the same as the target) [118]. Table 4.1 provides a short description of each covariate.

Having considered many covariates, we first performed a correlation analysis to select the most useful features in order to minimize the model dimensionality while retaining relevant information. This analysis was performed on psycho-linguistic and image variables independently. In Figure 4.3A, we can observe that lexical and movie frequencies are highly correlated (correlation factor (called c) = 0.878), we therefore performed a PCA and kept the first component (explained variance = 93.92%) to represent the common effect. For simplicity, we will call this new variable "psycho frequency" for the rest of the paper. As visual complexity and familiarity were anti-correlated (c = -0.560), a PCA was applied and the first component (explained variance = 87.01%) was defined as "familiarity". Phoneme number and AOA were weakly correlated with other covariates and were thus kept as such. This analysis allowed us to go from 7 to 5 psycho-linguistic dimensions. In Figure 4.3B, we see that entropy, contrast, energy and homogeneity are highly correlated (lowest

[1] The complete description and computation method of each image property are provided in the following repository: https://github.com/numediart/Covariates_Analysis.

| Covariate name | Description |
|---|---|
| Phoneme number | Number of phonemes in the French name of the item |
| Lexical frequency | How often the item appear in the literature |
| Movie frequency | How often the item appear in movies |
| Age of acquisition | At what age we learn the meaning of the item |
| Visual complexity | Level of detail or intricacy contained within the image |
| Familiarity | How often we meet the item in our daily life |
| Imageability | How easily the item will evoke a clear mental image |
| Entropy | Minimal number of bits required to encode the image |
| Contrast | Difference in luminance of the image |
| Correlation | How correlated neighboring pixels are |
| Homogeneity | How close pixel values are to the mean pixel value |
| Energy | Measure of the localized change of the image |
| Compactness | How closely packed the pixels of the item are |
| Ratio | length-width ratio of the item |
| Number of spectral clusters | The variety of frequencies in the image |
| High frequency energy | Energy of spectral cluster with the highest frequency |
| Highest frequency | Centroid of the spectral cluster of highest frequency |
| Maximum spectral distance | Distance between spectral clusters of lowest and highest frequency |
| Visual similarity | How similar the primer and the target picture shapes are |

**Table 4.1.** Description of selected covariates with a separation between psycholinguistic variables and image properties[1].

[1]The complete description and computation method of each image property are provided in the following repository: https://github.com/numediart/Covariates_Analysis.

c = 0.593). The first component of a PCA was chosen to summarize them, reflecting "contrast" (explained variance = 63.76%). The number of spectral cluster, the maximum frequency and the maximum distance between spectral clusters are highly correlated (lowest c = 0.629) and similarly, we used PCA and kept the first component reflecting "image frequency" (explained variance = 78.62%). Correlation, compactness and length-width ratio were considered as independent covariates regarding their low correlation score with other

features, reducing dimensionality from 9 to 7. Figure 4.3C summarizes the correlation between the 12 selected covariates after applying dimensionality reduction.

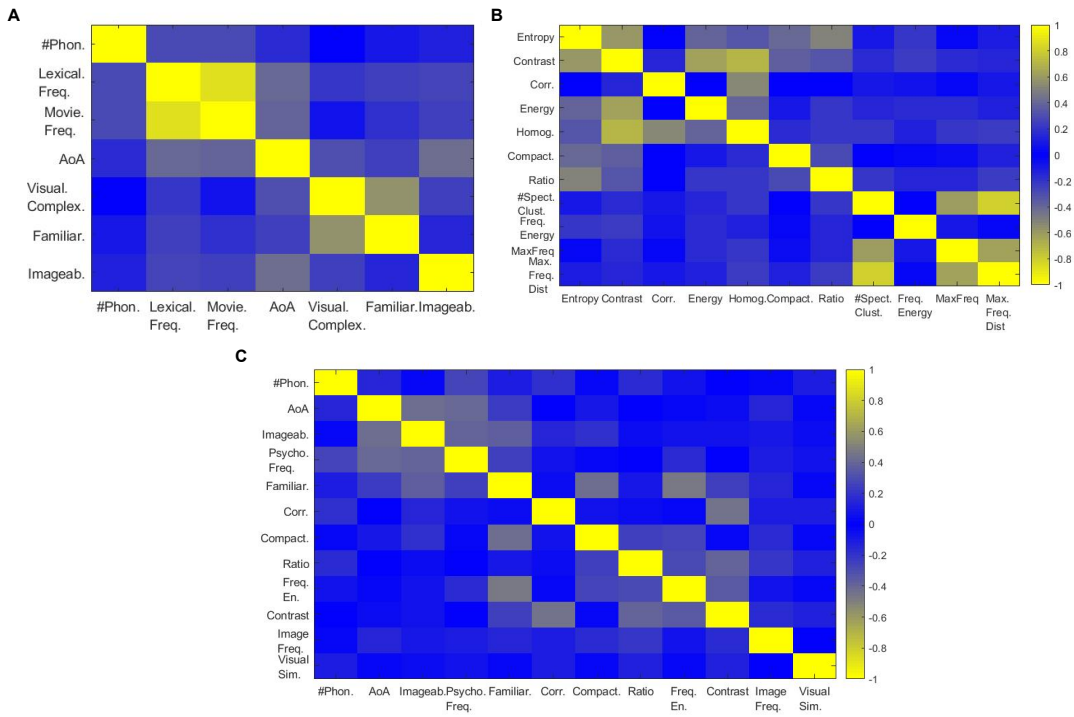



**Figure 4.3.** Correlation analysis of covariates. **(A)** correlation between psycholinguistic variables, **(B)** correlation between image properties, **(C)** correlation between selected covariates.

### 4.2.2 Linear Modeling

To conduct the linear modeling of the ERP data, we adhere to the procedure outlined in Figure 4.4. After selecting the relevant covariates and preprocessing the data as detailed in the framework presented in Chapter 5, we consolidate all the necessary information into a design matrix. This matrix

facilitates the subsequent regression analysis, represented by the beta parameters for each subject, sensor, and variable (categories and covariates).

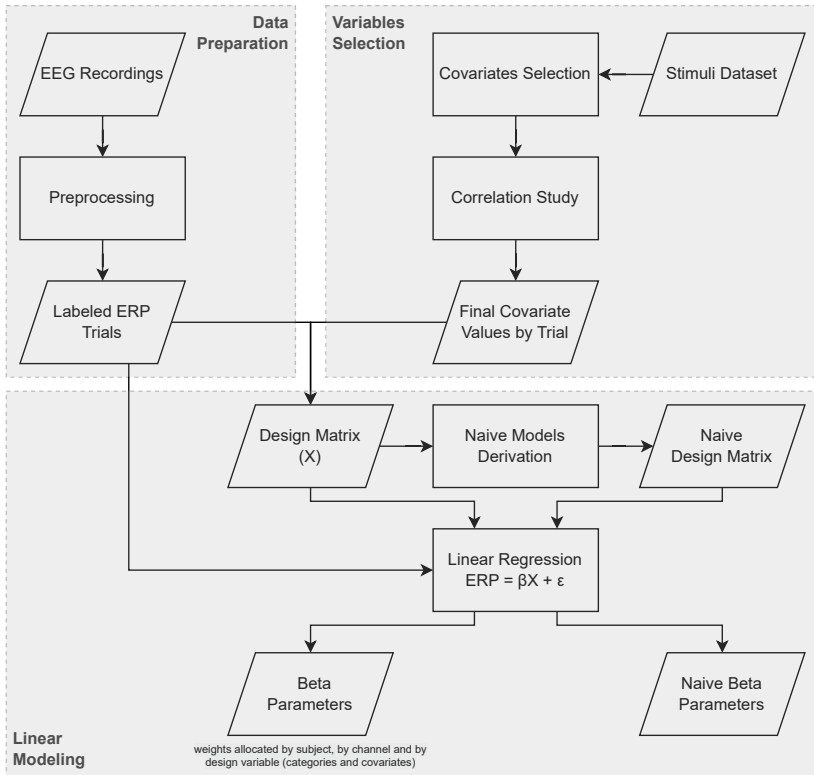



**Figure 4.4.** First-level Analysis. The figure illustrates the three-step process encompassing Data Preparation, Variable Selection, and Linear Modeling, as adapted from the LIMO EEG toolbox to our framework.

In this research, we analyzed the effect of psycho-linguistic and image feature covariates on the ERP independently to identify the most critical one, if there is any difference. To do so, we performed the analysis using four different models. As shown in Figure 4.5 through the corresponding design matrices, the first model (called "categorical model", 4.5A) only considers the categorical variables, the second model (called "psycho model", 4.5B) focuses on psycho-linguistic variables, the third one (called "image model", 4.5C) only takes into account image features and the last one (called "psycho-image model", 4.5D)

encompasses all the covariates. Note that we included the visual similarity in the image model for completeness. The design matrix links each trial with the corresponding category through binary values, the first column representing manufactured items and second column being related to natural items. The covariates, as continuous variables, are represented by their z-score computed throughout all trials. The regression process aims to obtain an optimal representation of the recorded ERPs for each subject. Depending on the designed model, the beta estimates give the linear combination of categorical variable and covariates that best fits the recorded EEG trials using a GLM. The regression is done in a parallel way for each subject using LIMO EEG toolbox through the *limo_batch* function. The computation of beta estimates is presented in Equations 4.1 and 4.2 where ERP represents the recorded EEG trials, $\beta$ the searched parameters, X the design matrix and $\epsilon$ the error term. This operation is performed on one channel at a time, fitting all trials simultaneously.

$$ERP = \beta X + \epsilon, \tag{4.1}$$

$$\beta = diag((X^T X)^{-1} X^T ERP) \tag{4.2}$$

Figure 4.6 represents the trimmed mean (20% of trimming) of the beta estimates across subjects on one electrode (F6) using the psycho model. This example shows that categorical variables have a larger weight on the regression (higher amplitude of the corresponding beta estimates) than covariates and that the constant term encompasses the general ERP behavior following the appearance of two sequential images.

By computing the difference between beta parameters belonging to each of the categorical variable, we can obtain the categorical contrast effect highlighting the ERP variations that are mainly due to the origin of the presented item, i.e. the effect we want to study. Equation 4.3 shows the computation of the contrast signal from beta estimates, where c is the contrast. This operation is performed on each subject and each electrode separately.

$$c = \beta_{manufactured} - \beta_{natural}, \tag{4.3}$$

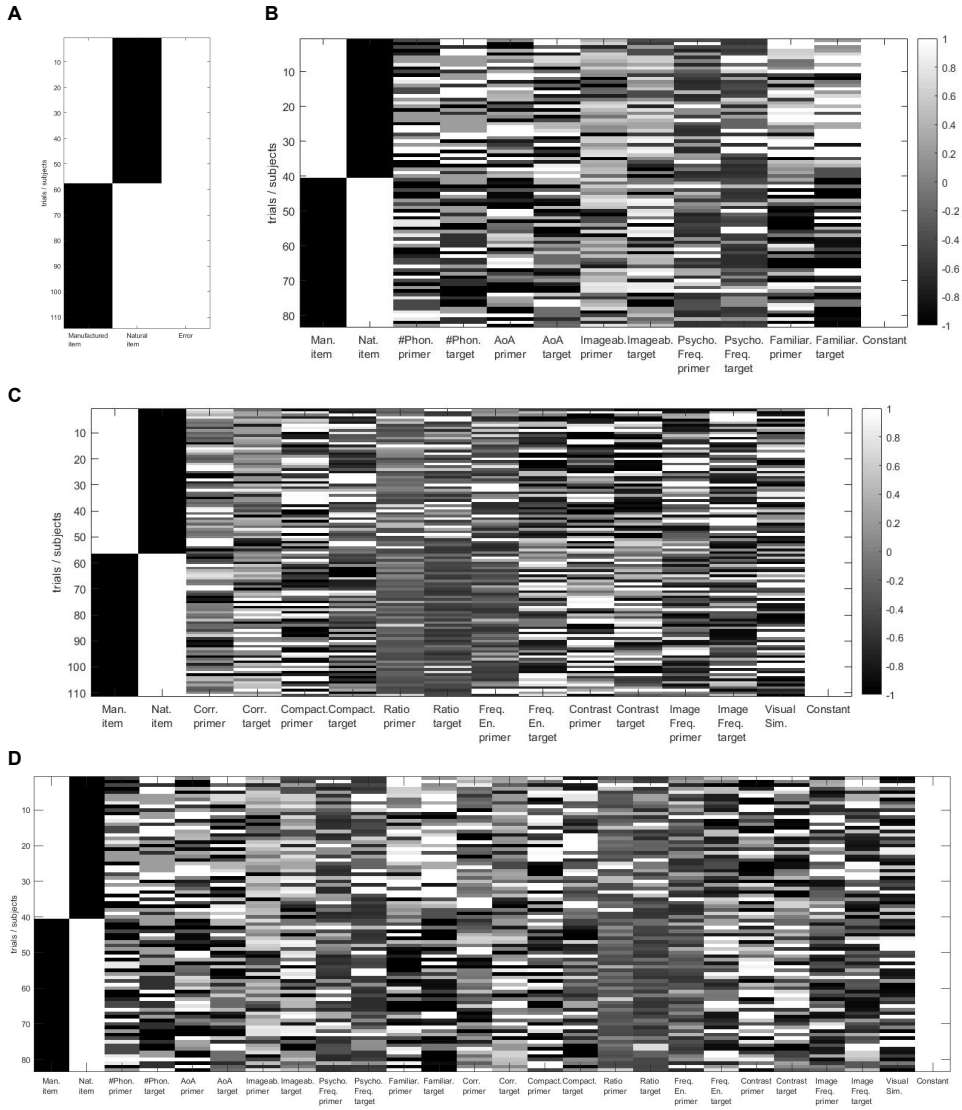



**Figure 4.5.** Design matrices. **(A)** control model (only categories and error terms, 3 dimensions), **(B)** psycho model (13 dimensions), **(C)** image model (16 dimensions), **(D)** psycho-image model (26 dimensions). The two first columns representing the categories are coded as binary values (-1 or 1), while columns corresponding to covariates have continuous values representing the z-score computed thorough all the trials.

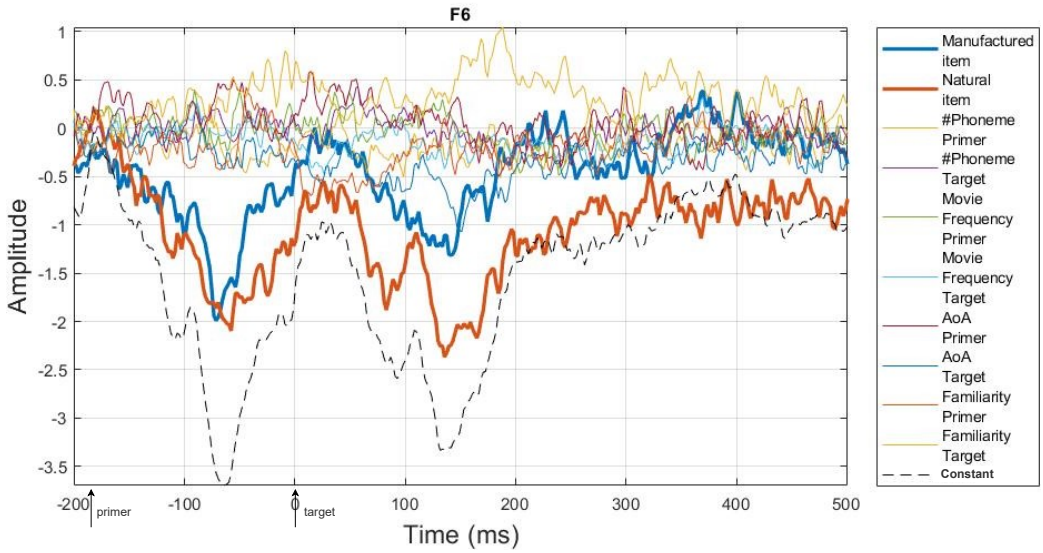



**Figure 4.6.** Trimmed mean (20% of trimming) of beta estimates across subjects on F6 electrode using "psycho" model. The two bold lines represent the categorical variables (manufactured and natural items) while the black dashed line belongs to the constant term. All other signals are related to the covariates (cf. legend). The arrows on the x axis show the appearance of the primer and the target images.

From the contrast parameter, a statistical analysis across subjects can highlight spatio-temporal regions of significant difference between both categories, as described in Section 4.2.3.

### 4.2.3 Statistical Inference

The statistical inference on the regressed signals enables the identification of the spatio-temporal regions of the ERP that support reliable classification between categories and reveals regions susceptible to covariate bias. The process is summarized in Figure 4.7.

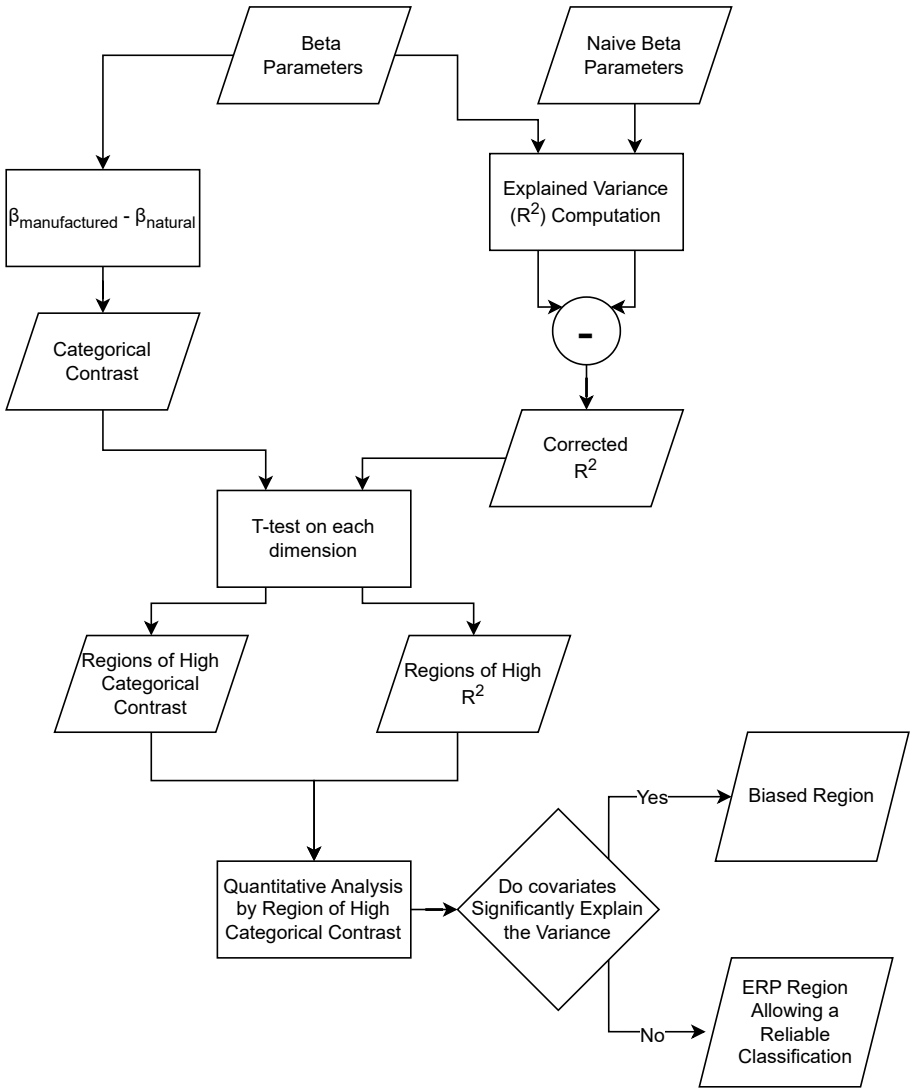



**Figure 4.7.** Analytical framework for identifying regions of interest supporting reliable classification and detecting covariate bias using the 2$^{nd}$ level analysis of LIMO EEG.

LIMO EEG proposes tools to perform robust statistics on regressed parameters, such as the Yuen t-test (i.e. t-test on trimmed mean) alongside bootstrap to account for multiple tests (spatio-temporal clustering and Threshold Free

Cluster Enhancement (TFCE) - [119]) Based on these methods, the second level analysis allows us to identify clusters of significant effect. Highlighting spatio-temporal areas of high categorical contrast is essential to know the regions a BCI algorithm will target to perform the classification task. One-sample t-tests were therefore run across subjects on the contrast parameters obtained from the categorical model as well as from the psycho-image model, followed by a Multiple Comparison Correction (MCC) using spatio-temporal clustering to identify regions that can be targeted by the classifier. Then, a study of the percentage of the ERP variance that is explained by a model is necessary to reveal the regions where the model properly fits the data. To establish a fair comparison between the explained variances ($R^2$) of the different models, the effect of the increase of dimensionality must be controlled. For this purpose, we introduced new models (called "naive" models) whose aim is to simulate the effect of changing the model dimensionality. To build a naive model, we use a design matrix on which the two first columns replicate the initial model (to keep the same category for each trial) and the covariate columns are generated as random vectors from multivariate normal distribution whose mean is zero for every column (as we used the z-score in the initial models) and the covariance matrix has the same rank as the corresponding model. This design matrix is then used to perform the ERP regression as previously described. This process is repeated 30 times with same categorical design but different random values for each naive model type. The beta estimates corresponding to the categories are averaged over the 30 repetitions to allow the study of the effect of the increase in dimensions on the categorical effect, and the $R^2$ values are averaged over repetitions to quantify the increase of explained variance that is due to the dimensionality effect. We therefore created three naive models corresponding to psycho, image and psycho-image models used in the study. The way the explained variances of the different models are combined in the statistical analysis is summarized in Figure 4.9 considering the example of the study of the effect of psycho-linguistic variables on the explained variance. The explained variance belonging to a specific model is computed as the difference between the explained variance of the model and that of the corresponding naive model. By applying a one-sample t-test to the explained variance across subjects followed by an MCC using spatio-temporal clustering, regions where the covariates influence the way a model fits the ERP can be identified. In fact, as the categorical effect is modeled

identically in both the actual and the naive models, the only remaining effect is the covariates influence.

### 4.2.4 Effects Separability

To evaluate how separable the biasing covariate effect is from the desired categorical effect, we have to quantify how their potential correlation affects the variance that can be explained by the regressed signals. Figures 4.8A and B provide a graphical representation of how the different models are combined to extract the contribution of the partial effects required to compute the statistical effects of interest. On Figures 4.8C and D, a representation of the different explained variances, as segments in the associated directions, is given. The correlation effect is totally absent in the ideal case of orthogonality, i.e. zero correlation, between the categorical effect, the psycho covariates effect and the image covariates effect as shown in Figures 4.8A and C. However, this correlation effect is responsible for a loss in explained variance when considering non-orthogonality between the different effects as Figures 4.8B and D illustrate. For sake of clarity, we intentionally omitted the dimensionality effect from Figure 4.8 as adding it would lead to a 4-dimensional problem and would therefore require an additional computation step to obtain the loss in explained variance, as shown in 4.9. When considering all the dimensions, this loss in explained variance due to the correlation between categorical and covariate effects ($R^2$ loss) is computed as the difference between the total categorical effect (identified using the categorical model) and the computed categorical effect. The block diagram presented in Figure 4.9 illustrates that the $R^2$ loss can be computed using Equations 4.4a and b where the "computed" psycho effect is the one used to derive the $R^2$ distribution (cf. Figure 4.12B).

$$\begin{aligned} R^2_{computed\_categorical\_effect} &= R^2_{psycho\_model} - (R^2_{psycho\_image\_model} - R^2_{image\_model}) \\ &= R^2_{psycho\_model} - R^2_{computed\_psycho\_effect} \end{aligned} \tag{4.4a}$$

$$R^2_{loss} = R^2_{categorical\_model} - R^2_{computed\_categorical\_effect} \tag{4.4b}$$

As illustrated in Figures 4.8A and B, the "computed" psycho effect is obtained by subtracting the image model effect from the psycho-image model effect.

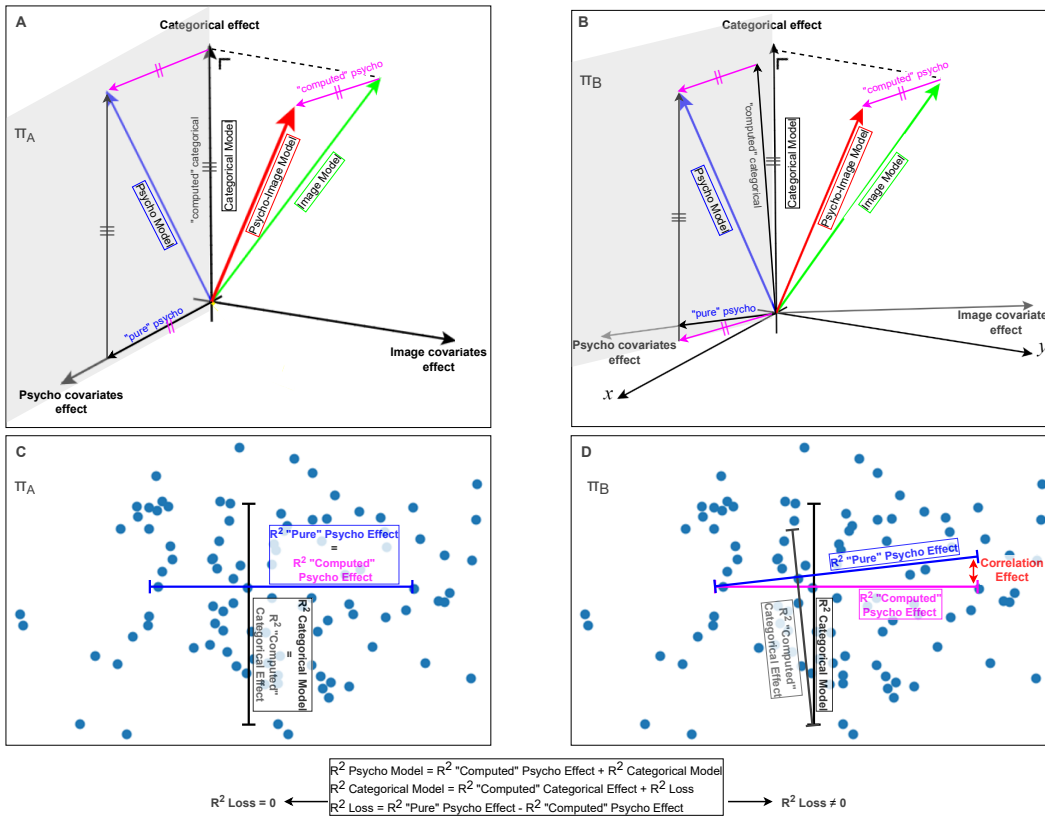



**Figure 4.8.** Geometrical representation of the combination of the different models. Left part relates to the ideal situation where the categorical effect, the psycho covariates effect and the image covariates effect are orthogonal to each other, while right part represents the real-life case of non-orthogonality. **(A)** and **(B)**: Vectorial representations of the categorical, psycho, image and psycho-image models and of the different effects resulting from their combination, with a focus on the psycho covariates effect. **(C)** and **(D)**: The corresponding projections on the $\pi$ planes where the $R^2$ values are computed for a given data set and represented as segments in their corresponding directions. The correlation effect causing the loss in explained variance is represented in red in **D**.

This "computed" psycho effect can be considered as the part of the psycho covariates effect that is not correlated to the categorical effect. Therefore, when

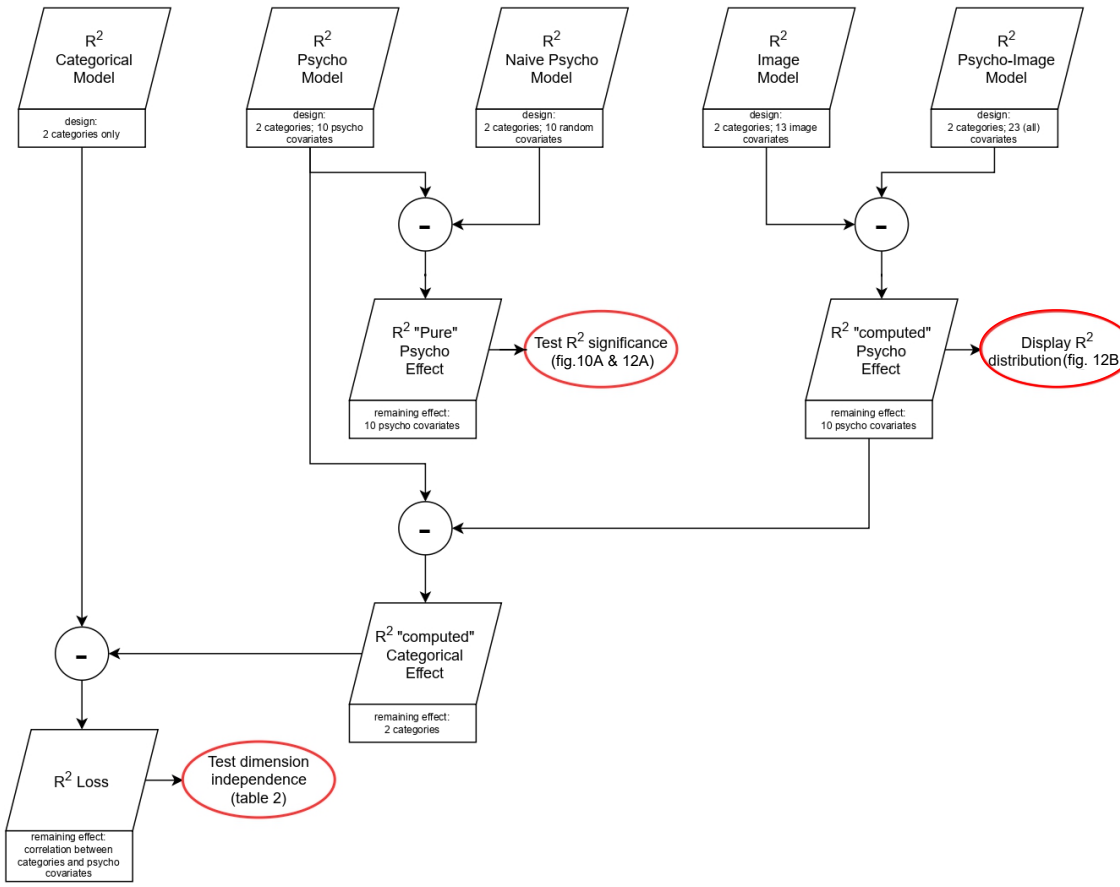



**Figure 4.9.** $R^2$ combination for statistical inference. Example of the study of the effect of psycho-linguistic variables on the explained variance.

subtracting this "computed" psycho effect from the psycho model effect, only the categorical effect remains. This "computed" categorical effect is composed of the actual categorical effect and the part of the psycho covariate effect that is correlated to the categories. The latter component is responsible for the deviation between the categorical model effect and "computed" categorical effect and can be obtained as the vectorial difference between both. These operations can directly be done on the $R^2$ values as the explained variance of a joint effect, e.g. psycho model effect that combines categorical and psycho

covariate effects, is equal to the sum of the explained variances from each of these effects in the ideal case of orthogonality. However when correlated, this sum is affected by the non-orthogonal part of the considered effects and a loss in $R^2$ starts to be propagated across the computations.
The separation between the categorical effect and a specific covariate effect is possible if the $R^2$ loss is significantly lower than the variance explained by the categorical model. This comparison is done within a spatio-temporal cluster of interest using a specific covariates model.
Having proved the separability between categories and covariates, we identify the spatio-temporal regions in the ERP where the categorical effect does not overlap with regions of significant covariate effects. These regions can therefore be used to perform an unbiased classification between the studied categories whatever the balance in the covariate values across those categories.

## 4.3 Results

As the objective of this work is to reveal the influence of the experimental covariates on the distinguishability of the categorical effect on the EEG, we first ran the statistical analysis described in Section 4.2.3 on the psycho-image model to extract both categorical and covariate effects when considering all the selected variables. Using this model, we ensure that the identification of significant categorical contrast was not biased by the experimental covariates. Figure 4.10 shows the explained variance (4.10A) and the categorical contrast (4.10B) of the psycho-image model along with the 20% trimmed mean ERP. The one-sample t-test followed by an MCC using spatio-temporal clustering reveals a cluster of significant categorical contrast from 326ms to 371ms (max T value 5.16 at 334ms on channel F5, corrected p-value = 0.01) and four clusters of significant $R^2$: cluster 1 starts at -62ms and ends at -14ms (max T value 4.78 at -30.42ms on channel C2, corrected p-value = 0.03), cluster 2 starts at 14ms and ends at 75ms (max T value 4.57 at 68.12ms on channel PO4, corrected p-value = 0.03), cluster 3 starts at 133ms and ends at 247ms (max T value 6.29 at 190.73ms on channel PO8, corrected p-value = 0.01) and cluster 4 starts at 383ms and ends at 408ms (max T value 5.61 at 391.21ms on channel C1, corrected p-value = 0.02). As the regions where the variance is significantly explained by the covariates values do not overlap the cluster of significant categorical contrast, we could assume that the identified categorical

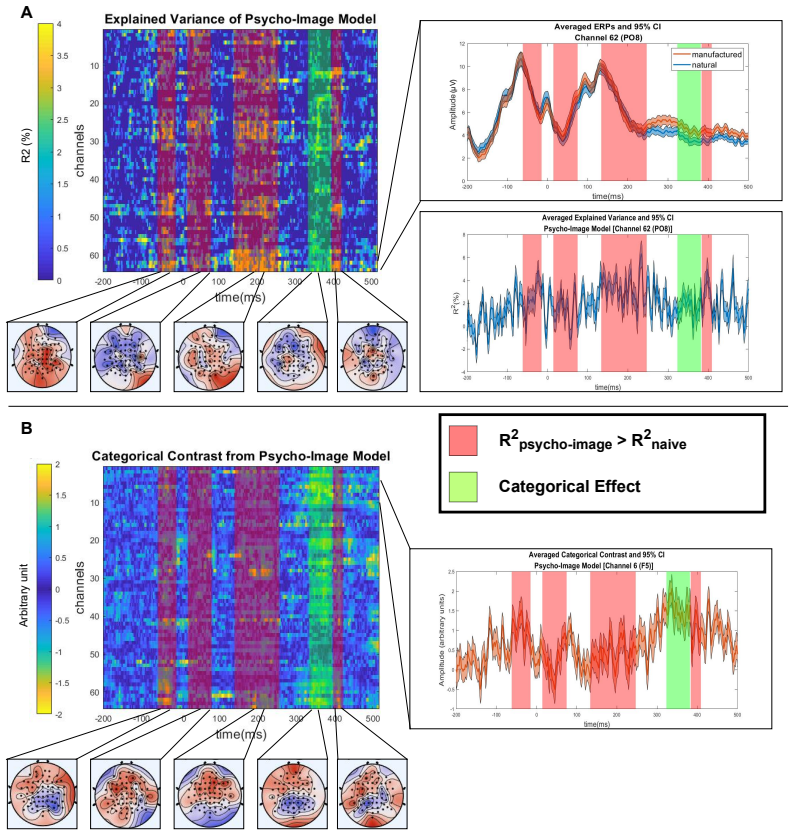



**Figure 4.10.** Statistical analysis of psycho-image model. **(A)** Trimmed mean of the explained variance ($R^2 - R^2_{naive}$) across subjects with corresponding regions of significant explained variance (red bands) and significant categorical contrast (green band). For each highlighted area, the topological view is shown (bottom). On the channel corresponding to maximum $R^2$ (PO8 electrode), the averaged ERPs of both categories (top right) and the $R^2$ timecourse (bottom right) are displayed. **(B)** Trimmed mean of the categorical contrast across subjects with significant regions highlighted and the corresponding topological views (bottom). On the channel corresponding to maximum contrast (F5 electrode), the averaged contrast parameter ($\beta_{man.} - \beta_{nat.}$) is displayed.

effect is not influenced by the chosen covariates when using the psycho-image model.
To identify spatio-temporal regions of the ERP that can be wrongly interpreted as clusters of significant categorical contrast if the covariate effects are not modeled, Figure 4.11 shows the thresholded maps of the categorical contrast obtained when using the categorical model (4.11A), the psycho-image model (4.11B) and the naive psycho-image model (4.11C). We can observe that, on top of the actual region of high contrast between the studied categories, the 3-dimensional categorical model detects two more clusters: one between 43ms and 95ms (max T value 6.37 at 67.14ms on channel F1, corrected p-value = 0.01), overlapping with the second $R^2$ cluster of the full psycho-image model, and the other one between 137ms and 163ms (max T value 5.99 at 156.98ms channel FC1, corrected p-value = 0.04), overlapping with the third $R^2$ cluster of the psycho-image model. When using the 26-dimensional naive model, similar clusters in excess appear with the first cluster ranging from 43ms to 75ms (max T value 6.01 at 51.51ms on channel FCz, corrected p-value = 0.02) and the second one from 147ms to 167ms (max T value 4.49 at 160.89ms on channel FCz, corrected p-value = 0.02), but an additional region between 446ms and 489ms (max T value 5.63 at 483.15ms on channel F3, corrected p-value = 0.01) is also considered as a cluster of significant categorical contrast. These results show that existing biases in the dataset can be badly exploited in the regression process and this biased effect becomes higher as the complexity of the model used increases, as discussed in Section 4.3.1.

The quantization of the variance that is explained by the different types of covariates was performed by analyzing the $R^2$ distribution across the spatio-temporal regions of significant contrast identified in Figure 4.11 using the psycho and image models separately (Figure 4.12). Figure 4.12A highlights the regions of significant categorical effect on top of the explained variance maps, with the displayed $R^2$ values resulting from the difference between the $R^2$ of the considered model and that of the corresponding naive model. A one sample t-test followed by the MCC run on the $R^2$ values, gives us the spatio-temporal regions of the ERP where the variance is significantly explained by the focused type of covariates. For the psycho model, the first significant cluster appears between 151ms and 177ms (max T value 5.61 at 158.9ms on channel FCz, corrected p-value = 0.01) and the second significant cluster ranges from 319ms to 490ms (max T value 9.01 at 367.9ms on channel

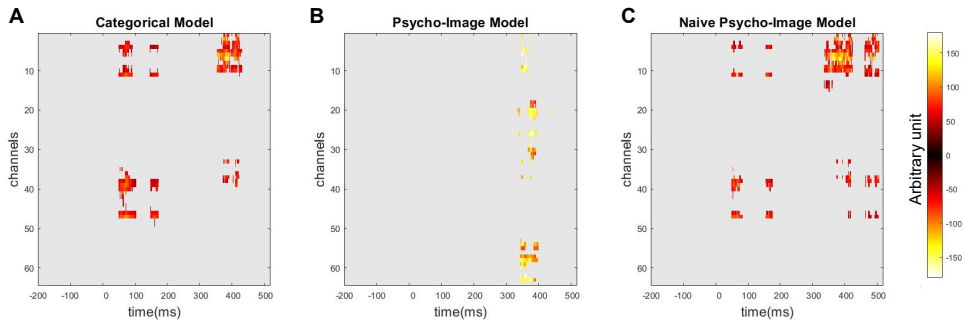



**Figure 4.11.** Thresholded maps of the categorical contrast showing spatio-temporal regions of significant categorical contrast using a one-sample t-test followed by an MCC with spatio-temporal clustering. These regions are extracted from the categorical model **(A)**, the psycho-image model **(B)** and the naive psycho-image model **(C)**.

F7, corrected p-value = 0.01). For the image model, the first significant cluster appears between -52ms and -9ms (max T value 4.71 at -22ms on channel O2, corrected p-value = 0.02), the second significant cluster ranges from 16ms to 38ms (max T value 4.47 at 18.3ms on channel O2, corrected p-value = 0.01) and the third significant cluster starts at 174ms and ends at 210ms (max T value 6.79 at 203.8ms on channel POz, corrected p-value = 0.02). Comparing the spatio-temporal regions of significant explained variance with the clusters considered of high categorical contrast by the model used allows areas where the classification can be biased by the experimental covariates to be detected. In fact, if a spatio-temporal region whose variance is mainly explained by the covariate values overlaps a cluster of high categorical contrast, an algorithm could use the covariate information to perform the classification instead of the actual categorical information.

To measure this overlapping effect, Figure 4.12B shows the distribution of the explained variance of each model as well as the confidence interval of the variance explained by the corresponding naive model within each categorical cluster. The explained variance from which the distribution is displayed is computed as the difference between the $R^2$ of the psycho-image model and the $R^2$ of the model not concerned, e.g. $R^2_{psycho} = R^2_{psycho-image} - R^2_{image}$. In this way, the part of the variance that is explained by the categorical effect is

excluded from the comparison, allowing a focus on the covariates effect only. The 95% confidence interval of the corresponding naive $R^2$ values shows the part of the variance that is explained by the increase of the model dimensionality. The categorical effect is displayed separately to provide a reference point. In the first cluster of significant categorical contrast obtained from the categorical model, the inter-quartile range of the $R^2$ values from the image model (from 15.41% to 16.65%) stands above the 95% confidence intervals of the $R^2$ values from the corresponding naive model (14.99% to 15.36%). The same behavior is observed in the second categorical region of interests (ROIs) (categorical model) for the psycho model with the inter-quartile range spreading from 15.88% to 17.24% and the confidence interval from the naive model between 15.21% and 15.58%. When focusing on the third categorical cluster (categorical model) or on the ROI extracted from the psycho-image model, none of the covariates explain a significant part of the variance as the 95% confidence intervals of the $R^2$ of both naive models are included in the inter-quartile ranges of the $R^2$ values from the covariate models.
To validate that the categorical cluster found in the late ERP response can be used to perform a reliable classification between categories, the separability between the categorical and the covariate effects should be proven. As described in Section 4.2.4, the separability can be evaluated by quantifying the part of $R^2$ that is lost due to the correlation between the categories and the covariates in the regressed signals. We computed the loss in $R^2$ using Equation 4.4 within each categorical cluster separately and by adapting the computations to each model. The results are shown in Table 4.2. As the part of lost

$R^2$ is significantly lower than the explained variance of the categorical model, the covariate effect can be considered as almost orthogonal to the categorical effect, meaning their effects can be easily separated when properly modeled. We can note that negative values in Table 4.2 are mainly due to the correlation between the psycho-linguistic variables and the image features, knowing that that the lost $R^2$ is computed as the combination of both.

The full analysis of the explained variance highlights the influence of the image features on the regressed ERP around 70ms and the influence of the psycho-linguistic variables around 150ms. The late evoked response around 350ms exhibits a high independence to the covariate effects, making it a good candidate to perform a reliable classification between living and non-living entities from the EEG trials.

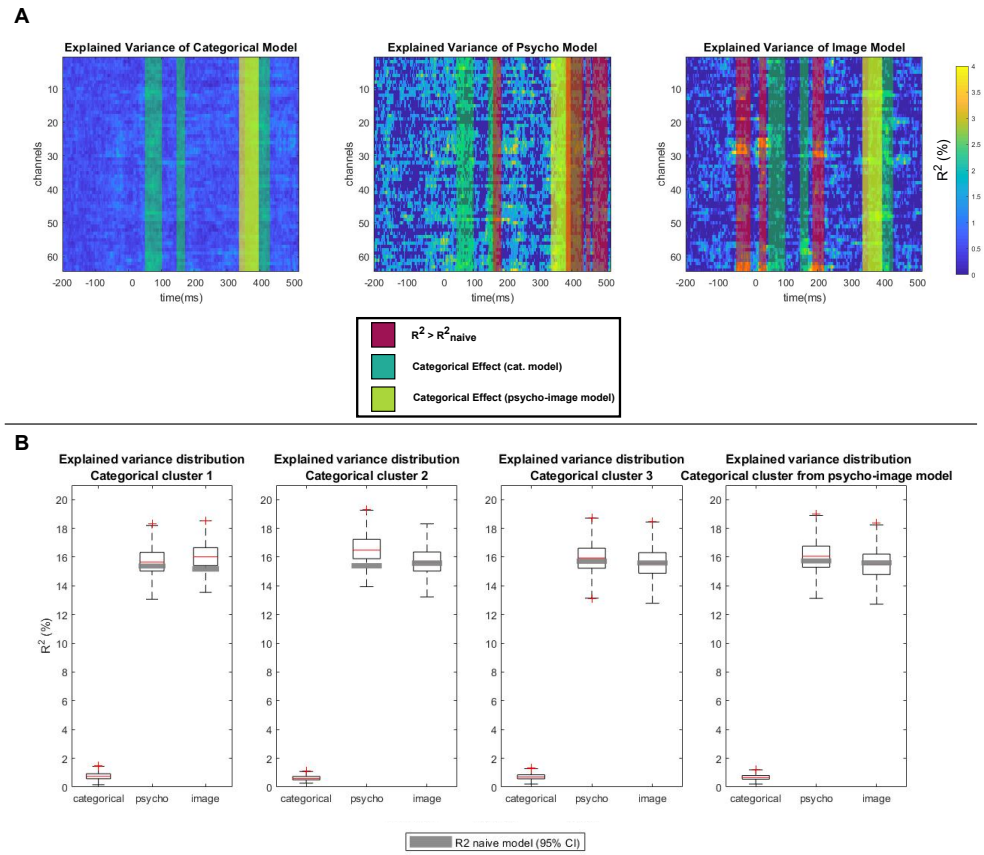



**Figure 4.12.** Explained variance quantization. **(A)** Trimmed mean of the explained variance ($R^2 - R^2_{naive}$) across subjects for categorical, psycho and image models. Regions of significant explained variance (red bands), significant categorical contrast when using categorical model (dark green bands) and significant categorical contrast when using psycho-image model (light green band) are highlighted. **(B)** Explained variance distribution for each of the identified categorical clusters. The gray zones within the box plots represent the confidence interval of the variance explained when using the corresponding naive model.

| R2 (%) (95% CI) | cluster 1 (cat.model) | cluster 2 (cat.model) | cluster 3 (cat.model) | cluster 1 (psycho-image model) |
|---|---|---|---|---|
| categorical model | 0.72 - 0.82 | 0.60 - 0.67 | 0.69 - 0.79 | 0.66 - 0.75 |
| psycho model | -0.14 - 0.01 | -0.17 - 0.08 | -0.18 - 0.04 | -0.07 - 0.02 |
| image model | -0.05 - 0.11 | -0.12 - 0.01 | -0.18 - 0.06 | -0.10 - 0.03 |

**Table 4.2.** 95% confidence intervals of the explained variance of the categorical model compared with the part of the explained variance that is lost in psycho and image models due to the correlation between the covariates and the categories.

### 4.3.1 Discussion

In this chapter, we present a solution aimed at mitigating confounding bias encountered during the experiment planning phase. Our proposed approach encompasses an end-to-end framework designed to quantify the impact of uncontrolled variables on the neural processes of interest. This framework spans from the selection of confounders to the assessment of the separability between the studied effects and spurious influences.

Our work is rooted in the fundamental concept that the influence of existing confounding variables may vary depending on the choice of the modeling process, as demonstrated by Holm *et al.* [120]. In response to this challenge, our proposed methodology, which leverages the LIMO EEG toolbox [114], aligns with the approach advocated by Pernet *et al.* [121]. They explored the confounding impact of the time delay between two presentations of the same facial image on ERP signals, aimed at distinguishing responses to famous versus unfamiliar faces. In this work, we adapted this framework to the dataset of the visual priming experiment described in Section 3.1 and demonstrated the capability to distinguish the effects in the EEG caused by stimuli from different origins (natural vs. manufactured) from the confounding effects arising from psycho-linguistic and image properties. The novelty of our approach lies primarily in quantifying the separability between categorical and confounding effects.

When the covariate information is not considered, any dataset bias, such as an imbalanced distribution of these variables between categories, can be exploited by the BCI algorithm. This leads to an increased probability of misclassification, as the algorithm relies on covariate values instead of the categorical effect itself. This effect becomes more pronounced with the complexity of the model, as observed in Figure 4.11, which compares clusters incorrectly considered as regions of high categorical contrast. In the late evoked response, the cluster of significant contrast is wider when using the naive psycho-image model (26 dimensions) than with the categorical model (3 dimensions).

Given that all the models used in this study are linear models, the biasing effect could be even more pronounced in BCI applications that employ more complex algorithms. Since achieving an equal balance of covariate values across categories while maintaining a diverse set of stimuli in the experiment is impractical, we propose a solution: modeling covariate effects through our hierarchical linear modeling approach, then evaluate the separability between categorical and covariate effects. While we established separability for visual stimulation by natural vs. manufactured items in this study, this demonstration should be repeated for other experimental conditions.

Quantifying the variance explained by each of the covariate types separately, as shown in Figure 4.12, revealed that both psycho-linguistic variables and image features influence the ERP, albeit outside of the spatio-temporal regions of significant categorical contrast between both categories. A BCI experiment aiming to classify EEG trials into evoked responses induced by natural or manufactured item visual stimulation can leverage this finding for reliable classification based on regions of dominant categorical information.

This research adheres to reproducibility standards, with code developed using the open-source FieldTrip [122] and LIMO EEG [114] toolboxes available in the following GitHub repository: `https://github.com/numediart/Covariates_Analysis.git`. As such, researchers can apply this method to analyze any task that elicits an evoked response, following the presented methodology to identify experimental features that affect the distinguishability of EEG differences induced by stimuli of distinct categories. Depending on the stimuli and studied categories, other covariates can be considered, although covariate selection should be carried out judiciously, considering the psychological effects related to the experimental task. The presented experimental design, being a semantic task, necessitated the study of psycho-linguistic features. The selected

variables align with those proposed by Alario *et al.* [116], who demonstrated their influence on ERPs related to picture naming tasks. Additionally, since the experiment involved visual stimulation via displayed pictures, we examined inner image properties. Traditional spatial and spectral features such as entropy, energy, and maximal frequency were computed and included in the analysis. Care should be taken regarding the number of selected covariates, as more variables can capture more uncontrolled effects but can also decrease the significance of the effect of increased dimensionality compared to the covariate effect itself. Conversely, fewer covariates may lead to the omission of significant covariate effects. The model dimension is limited by the number of trials available per subject, as conducting regression with more parameters than observations results in overfitting. Since the categories are exclusive, the minimum number of trials can be found by multiplying the number of covariates by the number of categories. The number of different subjects will affect the significance of the computed statistical values, as these computations occur at the second level of hierarchical modeling.

To integrate the proposed method into the design of a BCI experiment, researchers should conduct a preliminary test with an initial group of participants. This test aims to identify spatio-temporal regions of significant categorical contrast and assess separability between the covariate and categorical effects in these regions. In cases of proven separability, the BCI classifier can be trained using the identified regions of interest only. Otherwise, an additional step involving balancing the biasing covariates across categories should be performed before commencing the training procedure.

We emphasize the scope of this study. The provided method does not offer insights into the internal brain processes responsible for discriminating information from specific categories or covariates. Instead, it focuses on revealing the impact of uncontrolled variables on the ability to identify parts of an ERP that exhibit high contrast between experimental conditions.

Given the inherent complexity of interpreting EEG signals, which are linked to actual brain processes, future work will involve conducting covariate analysis on brain source activity derived from the recorded EEG signal using source reconstruction techniques. Additionally, since the proposed framework currently limits the analysis to the temporal response, further spectral analysis could unveil unknown frequency domain or non-timelocked effects.

We anticipate that this research will serve as a foundation for uncovering as many confounding biases as possible and optimizing the process of designing BCI protocols. To this end, future work will entail the creation of a shared database comprising results from various applications.

## 4.4 In Brief

Summary of Chapter 4

- This chapter introduces a comprehensive framework designed to evaluate the influence of specific uncontrolled confounding factors on the interpretation of ERP data.
- The selection of confounding variables is carefully determined through rigorous correlation analysis, with a particular focus on psycho-linguistic variables and image features.
- At the first-level analysis, linear regression of ERP data is conducted while considering the presence of diverse confounders through the utilization of design matrices.
- The second-level analysis involves statistical examinations aimed at identifying spatio-temporal regions with significant categorical differences and regions particularly sensitive to imbalanced confounders among categories.
- The separability of the confounding effects from categorical effects is assessed by the correlation between their effects.

Perspective for Chapter 4

- Extend the study to encompass source neural activity by employing source reconstruction and source connectivity techniques.
- Explore spectral components of ERP data, moving beyond temporal responses for a more comprehensive analysis.

# Chapter 5

# Data Collection Phase: ERP Preprocessing

## Contents


**5.1 Scope** . . . . . **90**
**5.2 Proposed Framework** . . . . . **91**
5.2.1 FieldTrip . . . . . 91
5.2.2 Visual Inspection . . . . . 92
5.2.3 Ocular Artifacts Reduction . . . . . 93
5.2.4 Detrending and Filtering . . . . . 94
5.2.5 Segmentation and Downsampling . . . . . 95
5.2.6 Line Noise Removal . . . . . 96
5.2.7 Muscle Artifacts Reduction . . . . . 96
5.2.8 Baseline Correction and Re-referencing . . . . . 98
**5.3 In Brief** . . . . . **101**


In the realm of biomedical signal analysis, the data collection phase plays a pivotal role in shaping the quality and reliability of research outcomes. This phase, however, is susceptible to measurement bias, which may arise from various sources, including imprecise sensors, suboptimal data acquisition procedures, or the presence of unwanted artifacts. This chapter is dedicated to describing our solution for mitigating measurement bias through a standardized preprocessing framework tailored specifically for ERP data.

Understanding the critical role of data preprocessing in EEG studies is essential to ensure the reproducibility and validity of research outcomes. Existing literature underscores the need for standardized preprocessing procedures due to the significant impact that variability in methods and parameters can have on the reliability of findings. Notable recommendations, such as those provided by C. Pernet [123], offer valuable guidance for preprocessing magnetoencephalographic and electroencephalographic (MEEG) data. Additionally, approaches like preregistration, as proposed by Paul *et al.* [124], aim to enhance transparency within preprocessing pipelines.

In our pursuit of optimal preprocessing methods, we advocate for benchmarking standardized and reproducible frameworks specifically designed for particular types of experiments. To encourage this approach within the scientific community, we have developed a dedicated preprocessing framework tailored to mitigate measurement biases in ERP data. We not only provide the open-source code but also offer the specific configuration used for this framework, customized for visual ERP experiments in which participants respond to stimuli displayed on a computer by pressing fixed buttons, as described in Section 3.1.

## 5.1 Scope

While preprocessing encompasses a broad field of research, we introduce a customized pipeline optimized for our specific experiment's challenges. Fine-tuning the proposed framework requires a detailed analysis of each algorithm's parameters, but this is beyond the scope of this study. Our framework serves as an example of the standardized preprocessing pipeline recommended for enabling transparent benchmarking across existing methods.

The primary objective is to maximize data retention, especially in the context of visual ERP experiments, which frequently entail extensive recording sessions with a limited number of participants. This limitation results in a scarcity of recorded data. In this context, the rejection of entire trials, when effective artifact cleaning is an alternative, is an impractical option. Consequently, artifact reduction assumes critical significance.

Throughout this chapter, we explore each step of the proposed preprocessing pipeline, from visual inspection to advanced techniques like muscle artifact reduction. Our overarching goal is to obtain ERP trials with minimal noise and artifacts, ready for in-depth analysis.

## 5.2 Proposed Framework

The entire preprocessing pipeline is executed using the open-source FieldTrip software [122], described in Section 5.2.1, and follows best practice recommendations set forth by the OHBM COBIDAS MEEG committee [39]. A core feature of this framework, fundamental for any preprocessing workflow, is its emphasis on reproducibility.

In addition to detailing each processing step in the following sections, we offer open-source code that enables the replication of the entire pipeline for any ERP experiment. Alongside this code, we provide a configuration file containing the key parameters utilized in the proposed process. These resources are accessible on our GitHub repository: `https://github.com/numediart/PreprocEEG.git`.

### 5.2.1 FieldTrip

The FieldTrip toolbox is a versatile and comprehensive software package widely utilized in the field of EEG signal analysis. FieldTrip has become an indispensable resource for neuroscientists, engineers, and researchers worldwide. The key features of the FieldTrip toolbox include:

- **Modularity**: FieldTrip is designed with a modular structure, allowing users to customize and combine various functions to suit their specific research needs. Its flexibility makes it adaptable to a wide range of EEG analysis tasks, from basic preprocessing to sophisticated statistical analyses.
- **Data Preprocessing**: FieldTrip provides a rich set of tools for EEG data preprocessing, including filtering, artifact removal, and epoching. These functions ensure that the data is prepared optimally for subsequent analyses.

- **Source Localization**: FieldTrip enables accurate source localization of EEG signals, offering methods such as beamforming and distributed source modeling. This capability is crucial for understanding the neural generators of recorded brain activity.
- **Statistical Analysis**: The toolbox offers a wide array of statistical methods, making it suitable for hypothesis testing, group-level comparisons, and advanced multivariate analyses. FieldTrip facilitates rigorous statistical assessment of EEG results.
- **Visualization**: FieldTrip includes powerful visualization tools for creating topographical maps, time-frequency representations, and source reconstructions. These visualizations enhance the interpretation of EEG findings and facilitate the communication of results.
- **Integration with Other Software**: FieldTrip seamlessly integrates with other popular EEG analysis tools, such as EEGLAB [125], allowing users to combine their strengths and benefit from a broader set of capabilities.
- **Community and Documentation**: FieldTrip benefits from an active user community and extensive documentation. Users can access tutorials, forums, and expert guidance to maximize their proficiency with the toolbox.
- **Open Source**: FieldTrip is an open-source project, fostering collaboration and innovation in EEG research. Its open nature encourages the sharing of code and methods across the scientific community.

In this thesis, FieldTrip served as an indispensable tool for the processing, analysis, and visualization of EEG data, contributing to the robustness and reliability of the presented results. Its flexibility and extensive capabilities make FieldTrip a valuable asset for any researcher engaged in EEG-based investigations.

### 5.2.2 Visual Inspection

The first step of our preprocessing framework involves visual inspection to remove bad channels and trials from the dataset. This step is performed using the *ft_rejectvisual* function provided by FieldTrip. Figure 5.1 showcases

examples of good and bad channels and trials that can be easily identified through this method.

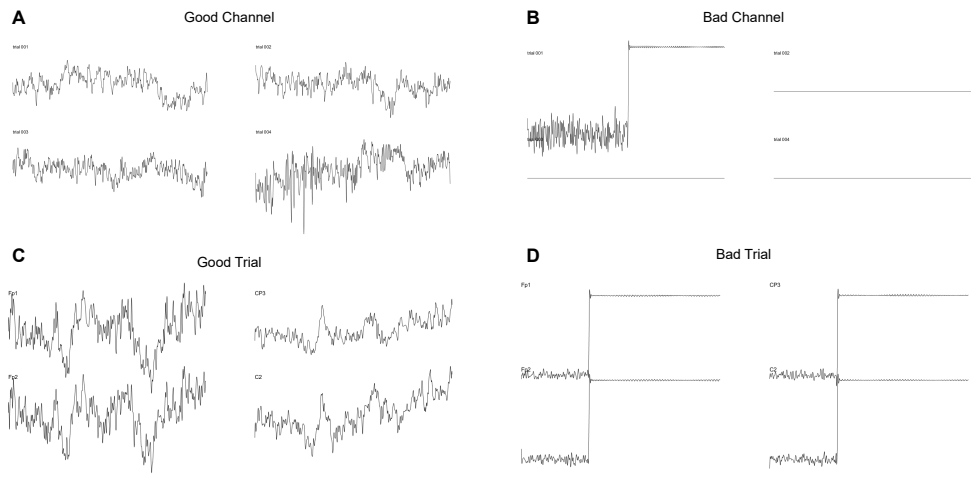



**Figure 5.1.** Visual Inspection. Examples of good (A) and bad (B) channels, as well as good (C) and bad (D) trials. Rejected samples typically exhibit flat signals with no useful information.

### 5.2.3 Ocular Artifacts Reduction

The second step involves mitigating ocular artifacts, which include blinks and eye movements, while retaining the valuable signal components. Traditionally, this task is tackled using blind source decomposition methods such as ICA [29]. However, this approach can yield variable outcomes, heavily dependent on the expertise of the experimenter in identifying artifact components post-decomposition.

To address this challenge, certain tools have been developed to automatically select components for retention or removal, such as ICLabel [126]. However, employing such algorithms may result in the loss of an opportunity to consider contextual information specific to the use-case under study. This may lead to the exclusion of components that are relevant in one context but might

be considered artifacts in another. Alternatively, it could also result in the retention of components that are contextually irrelevant.

To ensure a standardized preprocessing framework that can be consistently applied across studies, we have opted for the Multichannel Wiener Filter (MWF) technique, as proposed by Somers *et al.* [127]. Although the MWF method does require manual intervention to select a few (typically 5 to 10) segments of the signal that clearly represent ocular artifacts for algorithm initialization, this process is made simpler due to the straightforward nature of the task. Instead of sifting through numerous components to decide which ones to remove, Somers' approach involves extracting a low-rank approximation of the covariance matrix from the segments of the signal that have been annotated as artifacts. This allows for the minimization of artifact power while preserving the signal related to brain activity.

### 5.2.4 Detrending and Filtering

This stage of the pipeline involves two common steps: detrending and low-pass filtering.

Detrending aims to eliminate low-frequency trends or drifts in the EEG data. These trends may originate from various sources, including electrode drift, slow changes in electrode impedance, or physiological processes unrelated to neural activity. This process entails regressing the data with a $1^{st}$-order polynomial and subtracting it from the signal. It's important to note that this subtraction operation also serves as a demeaning operation.

Low-pass filtering is employed to remove high-frequency noise and physiological artifacts from the EEG data. In our case, the EEG signals of interest fall below 100Hz (end of the gamma band), while high-frequency noise can be filtered out to enhance the signal-to-noise ratio. We utilize a default $4^{th}$-order Butterworth filter, known for its phase response that is slightly non-linear. The filter's cutoff frequency is set at 200Hz, facilitating subsequent downsampling without encountering aliasing effects.

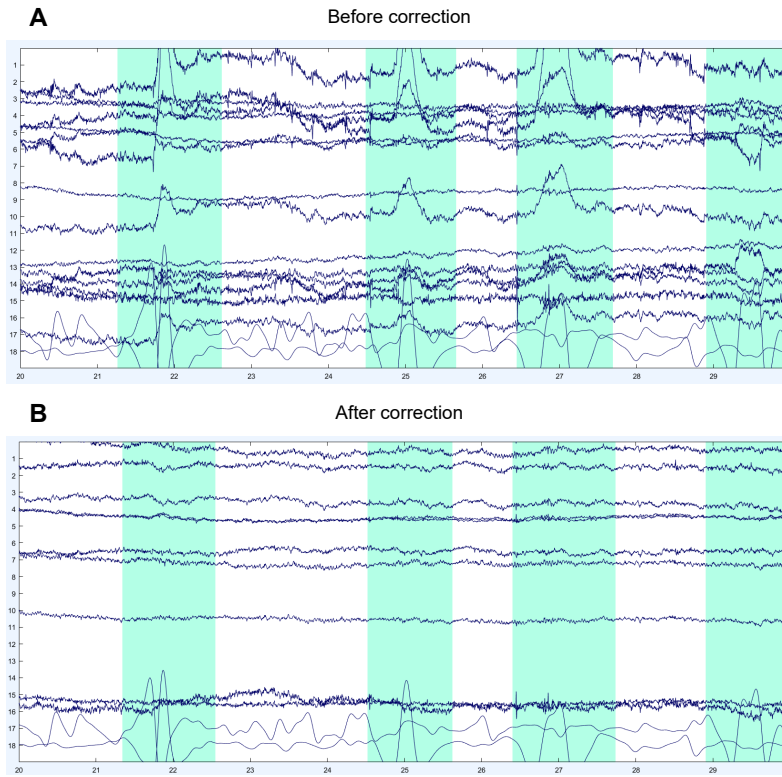



**Figure 5.2.** Ocular Artifact Reduction. (A) illustrates EEG signals affected by ocular artifacts, with the green bands indicating the segments containing artifacts. (B) presents the corrected version of the same signal after applying the Multi-channel Wiener Filter (MWF). Notably, the segments displayed here are not part of the manually annotated segments. The bottom two signals correspond to the horizontal and vertical EOGchannels, which serve as visual references for annotating the artifact segments but are not processed by the MWF. For clarity, only a subset of EEG channels is shown here.

### 5.2.5 Segmentation and Downsampling

We segment the data around stimulus onsets, ensuring that only relevant information is retained for analysis. Segmentation covers the period from

500ms before the target onset to 1000ms after it. Downsampling reduces the original 2048Hz data to 512Hz, maintaining signal quality without aliasing.

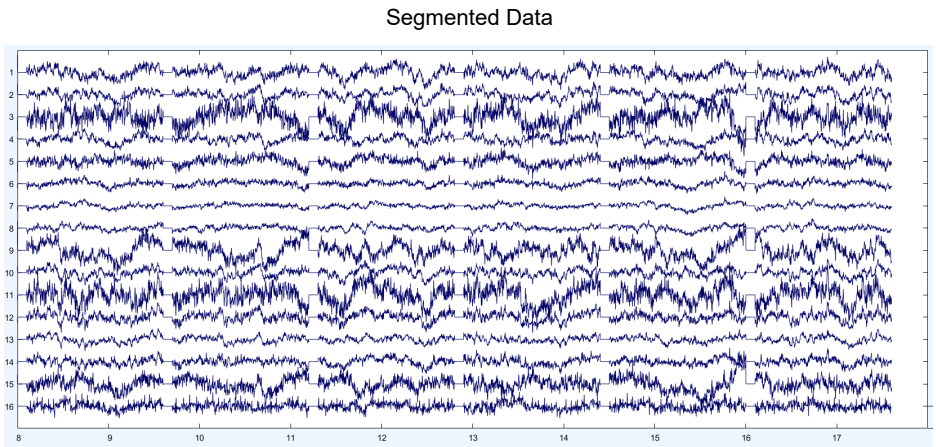



**Figure 5.3.** Data Segmentation. The figure illustrates the segmented data. For clarity, we have separated each segment with a flat zero signal, displaying only a subset of EEG channels.

### 5.2.6 Line Noise Removal

To eliminate line noise originating from the electrical grid (typically 50Hz in Europe), we employ the *Zapline* algorithm proposed by de Cheveigné [128]. This algorithm combines spectral and spatial filtering to effectively remove line noise while preserving the signal's overall morphology.

### 5.2.7 Muscle Artifacts Reduction

Muscle artifacts, often random and exhibiting low temporal auto-correlation, can occur due to jaw clenching, smiling, chewing, swallowing, or head movement. To mitigate these artifacts, we employ the Ensemble Empirical Mode Decomposition-Canonical Component Analysis (EEMD-CCA) algorithm proposed by Chen et al. [129] and included in the ReMAE toolbox [130].

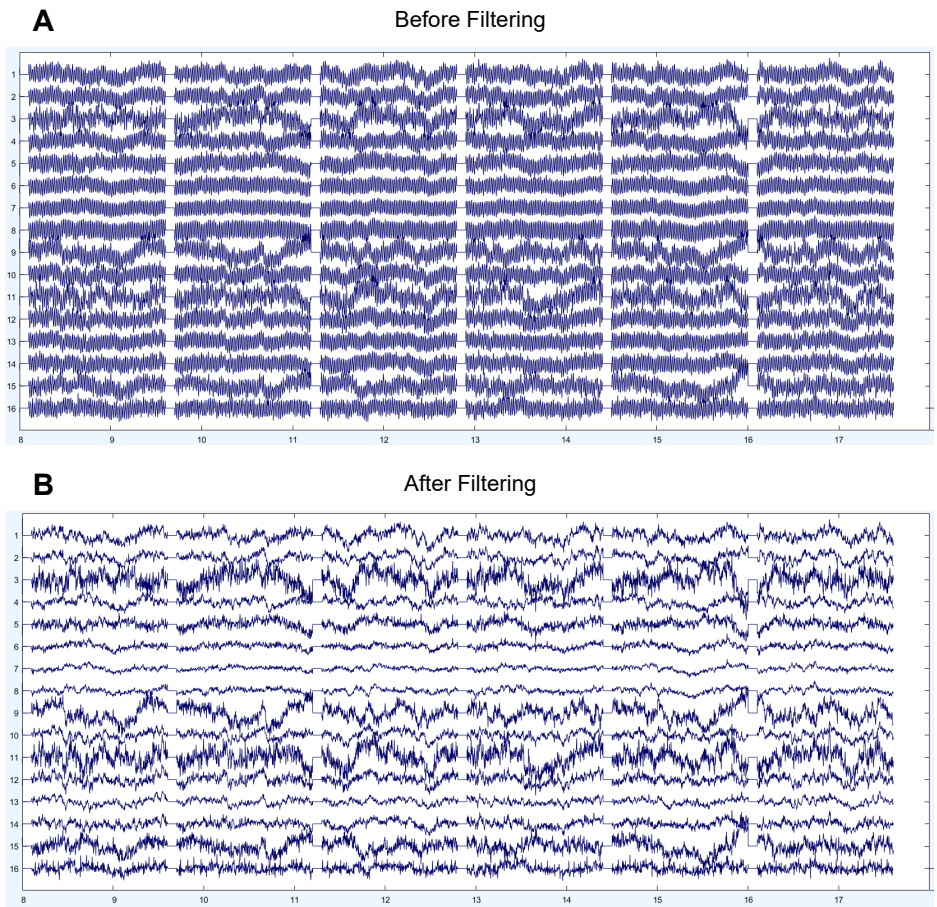



**Figure 5.4.** Line Noise Removal. (A) Segmented ERP data affected by line noise and (B) the filtered version using *Zapline*. The filtering process removes the 50Hz component while preserving the general signal morphology.

Ensemble Empirical Mode Decomposition (EEMD) decomposes the signal into intrinsic modes, allowing the distinction of different oscillatory behaviors. Autocorrelation values for each mode serve as indicators to identify potential artifactual components. Any mode with an autocorrelation below a specific threshold is considered a potential artifact (a higher threshold, such as 0.9, is recommended). Subsequently, a Canonical Component Analysis (CCA) is

applied to these potential artifacts to estimate sources that are maximally autocorrelated and mutually uncorrelated. All sources with autocorrelation values lower than a chosen threshold (in this experiment, we employ a threshold of 0.5 based on experimental constraints) are treated as artifacts and set to zero. The final step involves an inverse CCA followed by an inverse EEMD to obtain the cleaned EEG signal [129]. Figure 5.5 provides an example of the muscle artifact reduction efficiency.

### 5.2.8 Baseline Correction and Re-referencing

The final stages of our preprocessing, akin to conventional ERP pipelines, involve two steps: baseline correction and electrode re-referencing.

Baseline correction is a fundamental operation that involves subtracting the average EEG amplitude, computed for each electrode and at each time point during a baseline period across all trials, from the signal in each trial. Typically, this baseline period corresponds to the resting-state period before the stimulus onset. The primary objectives of baseline correction are twofold: firstly, to eliminate any DC offset, and secondly, to improve the signal-to-noise ratio (SNR) by amplifying the differences in EEG amplitudes relative to this baseline. This process ensures consistent signal amplitudes across trials, facilitating their meaningful comparison. In our approach, we employ a specific baseline window, ranging from 500ms before the target onset to 200ms before this onset. This window effectively captures the period preceding the primer onset when no stimulation occurs (see Section 3.1).

Re-referencing plays a pivotal role in mitigating common sources of noise in the EEG data. It establishes a standardized reference scheme that simplifies the comparison of ERP data across different subjects and enables the aggregation of data for group-level analyses. Within our framework, we have chosen to implement the average reference method. This method involves subtracting the mean signal value computed across all electrodes from each individual electrode's signal. This choice proves to be valuable within a standardized preprocessing framework as it ensures consistency across setups that may employ different reference electrodes.

It is important to note that the timing of re-referencing can influence the outcome of preprocessing. In our approach, we perform re-referencing as the

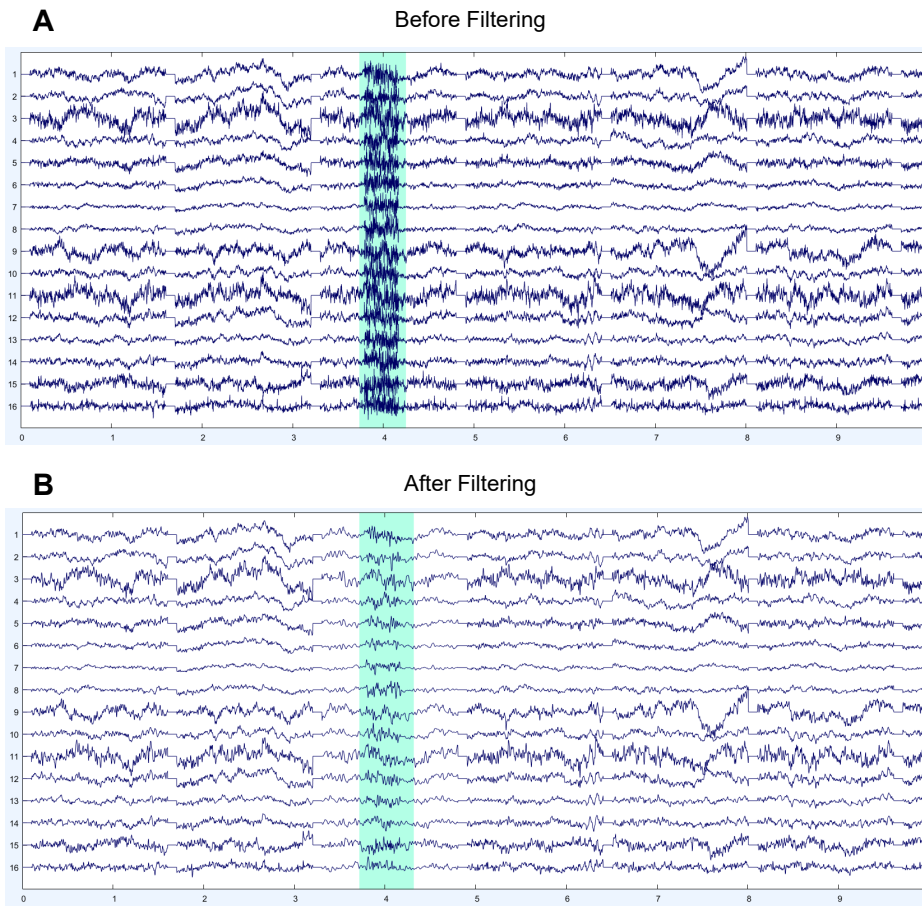



**Figure 5.5.** Muscle Artifact Reduction. (A) Segmented ERP data affected by a muscle artifact, highlighted with the green band. (B) The cleaned data after applying the EEMD-CCA algorithm, effectively reducing muscle artifacts.

final step to preserve the initial morphology of the recorded EEG signals. This is particularly relevant when EEG data are recorded with a predefined reference, such as the mastoid reference. However, in cases where a recording system like the BioSemi Active-Two system is used, no referencing is applied during recording. In such instances, it may be more suitable to perform the

re-referencing step at the beginning of the analysis to effectively remove a significant portion of common noise.

It's also worth noting that, theoretically, in cases where the primary focus of data analysis is source space inference, re-referencing may not be deemed necessary. This is because source reconstruction techniques primarily focus on the relative differences in signal amplitudes between sensors rather than their absolute values. Consequently, the choice of absolute reference often gets canceled out during the estimation process. Nevertheless, re-referencing can still be advantageous for facilitating comparisons with existing literature and promoting consistency in data processing pipelines.

## 5.3 In Brief

Summary of Chapter 5

- Establishing reliable benchmarks for preprocessing methods requires standardization and reproducibility in preprocessing pipelines.
- The proposed framework is tailored to a specific experiment type: visual ERP experiments involving participants seated in front of a computer and interacting via fixed button presses.
- The framework encompasses various stages, ranging from initial visual inspection to electrode re-referencing, with specific artifact reduction techniques and filtering methods in between.
- Every decision made in designing the framework aims to maximize its adaptability to other datasets of similar experiments. Enhancing reproducibility is facilitated through the provision of open-source code and comprehensive parameter configurations.

Perspective for Chapter 5

- Conduct a thorough fine-tuning of each algorithm's parameters to optimize the framework's performance.
- Develop a benchmarking framework to compare this pipeline with others designed for similar experiments.

# Chapter 6

# Analysis Phase: Source Localization Benchmarking

## Contents


**6.1 Source Reconstruction** . . . . . **104**
6.1.1 Forward Modeling . . . . . 104
6.1.2 Inverse Modeling . . . . . 107
**6.2 Related Work** . . . . . **109**
**6.3 Proposed Framework** . . . . . **111**
6.3.1 Parameters Selection . . . . . 112
6.3.2 Source Selection . . . . . 113
6.3.3 Pseudo-Source Signal Generation . . . . . 113
6.3.4 Pseudo-EEG Signal Generation . . . . . 117
6.3.5 Performance Evaluation . . . . . 118
6.3.6 Benchmark . . . . . 120
**6.4 Discussion** . . . . . **122**
**6.5 In Brief** . . . . . **123**


During the analysis phase of an experiment, modeling bias can arise from the misrepresentation of data when transformed into different representations. In EEG studies, these transformations include source reconstruction, connectivity analysis, spectrogram analysis, topographical representation, or even basic statistical feature derivation like mean, standard deviation, skewness, kurtosis, etc. In this chapter, we present a solution to address modeling bias in brain source activity obtained from EEG signals. The proposed framework serves as a benchmark, evaluating the accuracy of source localization methods.

In Section 6.1, we provide a detailed overview of the classical pipeline commonly employed for reconstructing source signals. Section 6.2 discusses the current state of the art in methods for validating source reconstruction. Following that, in Section 6.3, we outline the details of our benchmarking framework and its application in evaluating a classical source reconstruction approach [131]. Finally, in Section 6.4, we engage in a comprehensive discussion regarding the practicality and limitations of the proposed framework.

## 6.1 Source Reconstruction

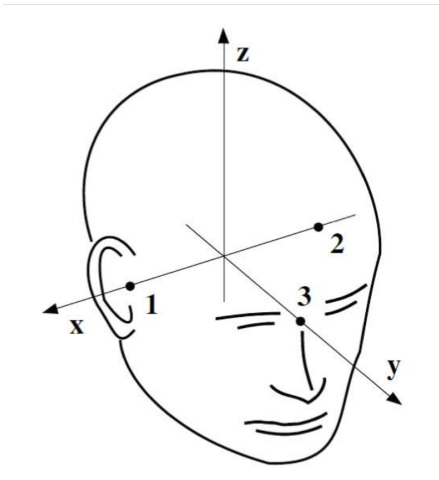
The source reconstruction consists of an estimation of brain region activations from the EEG signal. This process is composed of two main steps: *forward* and *inverse* modeling.

### 6.1.1 Forward Modeling

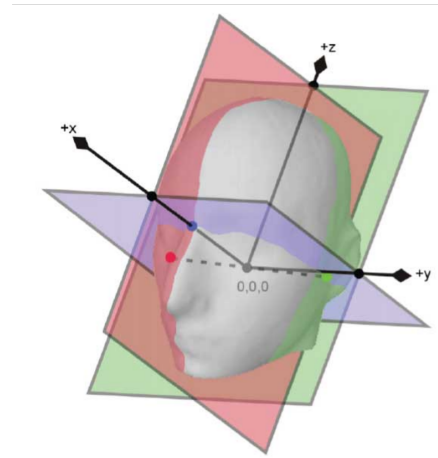
The forward modeling aims to build a leadfield representing the flow of the electrical field from each predefined brain source to the electrode positions. Different information are needed to compute the final field for each participant:

- **Electrode Positions:** the precise location of each electrode must be properly recorded relatively to the head position. First, the coordinate system has to be defined uniformly across participants. This homogeneity is reached through the use of anatomical landmarks on the outside of the head, called *fiducials*. Their position is shown on Figure 6.1. Several coordinates system are derived from those landmarks as described in `https://www.fieldtriptoolbox.org/faq/coordsys/`. In this thesis, we use the CTF system (cf. Figure 6.1b) as it is the one used by EEGLAB [125], the EEG toolbox with the wider community.

  The recording of the electrode positions can be done either by 3D scan of the participant's head with the cap or using a 3D digitizer, such as the Polhemus Fastrak [132]. Figure 6.2 illustrates both techniques. The 3D scan requires the experimenter to further localize each electrode on




**(a)** Fiducials position: 1.Right Pre-Auricular (RPA), 2.Left Pre-Auricular (LPA), 3.nasion Coordinates system = Neuromag




**(b)** Coordinates system = CTF

**Figure 6.1.** Examples of coordinates system derived from fiducials

the scanned image introducing resolution errors, while the 3D digitizer directly stores the relative electrode positions in the desired coordinate system. We therefore use the 3D digitizer in our experiment to record the electrode positions.

- **Head Model:** the geometrical and electrical/magnetic properties of the head is described in a volume conduction model, also called head model. This description is ideally derived from the anatomical MRI of the participant's head. However, when not available, the individual MRI can be approximated by a standard MRI reshaped according to the fiducials position previously measured. The latter technique is the one we used during the building of our dataset. The different tissue types of the brain are detected and linked to their theoretical conductivity value. We used a model composed of 5 different tissue types (skin, skull, white matter, gray matter and Cerebro-Spinal Fluid (CSF)).

**(a)** 3D scan using structure core device [133]

**(b)** Polhemus Fastrak 3D digitizer

**Figure 6.2.** Devices to record electrode positions

- **Source Model:** the location and orientation of the current generators, called dipoles, in the brain are defined in the source model. Two main techniques are used to obtain the source model: uniform and distributed source locations. The choice of the model depends on the method of inverse modeling used. The type of targeted activation is determinant in this choice. For oscillatory source reconstruction, beamforming [134] is the most common technique, while minimum-norm estimation (MNE) [135] is mainly applied to ERP.
  The preferred source model for beamformers is the uniform model where where dipoles are defined on a regular 3D grid, with a regular spacing between the dipole locations. Using MNE, the sources are distributed and only the strength at all possible cortical locations is to be estimated. In the latter case, sources should only be placed in regions where generators might be present which implies the dipoles to be placed at gray matter location only.
  As we study ERP, we use a distributed source model.

From the electrode positions, head model and source model, the leadfield can be computed using Finite Element Method (FEM) [136]. Figure 6.3 illustrates the forward modeling process.

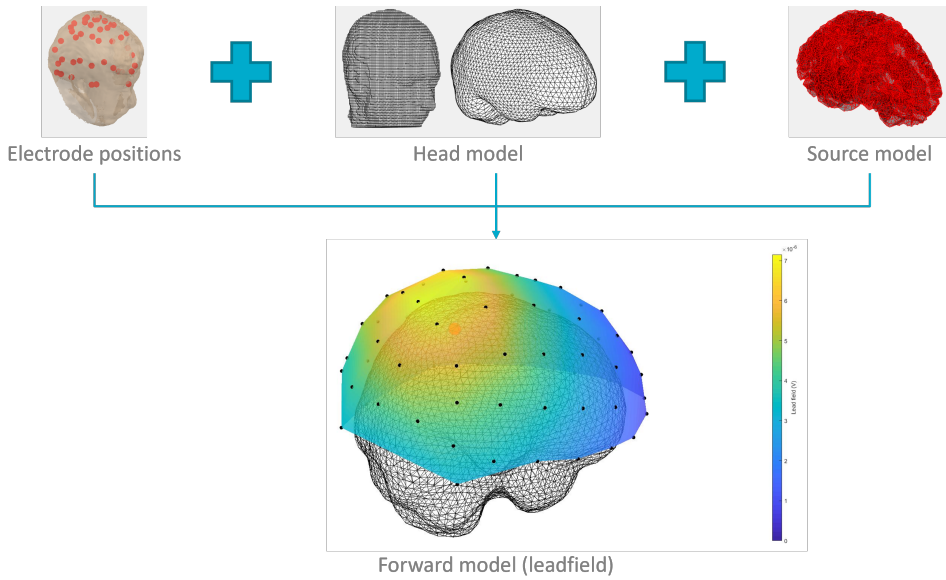



**Figure 6.3.** Forward modeling process

### 6.1.2 Inverse Modeling

As mentioned in Section 6.1.1, beamforming and MNE are the two main source reconstruction methods. Knowing the leadfield, the reconstruction is performed by computing the inverse solution of the forward model.

The beamformer technique is a spatially adaptive filtering method. It estimates the activity in specific source locations by minimizing the source power or variance at those locations. This minimization is achieved by adaptively weighting the contributions of different EEG sensors in such a way that it enhances the signal coming from the source of interest while attenuating the signals from other sources. This technique relies on the assumption that sources originating from different regions of the brain are not correlated in time [137]. In other words, it assumes that the activities in different parts of the brain are statistically independent from each other. By leveraging this assumption, beamforming allows for the localization and isolation of neural sources.

MNE is a distributed inverse solution that estimates the amplitude of all modeled source locations simultaneously and recovers a source distribution with minimum overall energy consistently to the measured EEG signal [138]. As MNE is the most suited technique for ERP signal, we applied it to our preprocessed timelocked EEG signal for each participant.

To combine source activity information by region, we employed the Singular-Value Decomposition (SVD) technique, as proposed by Rubega *et al.* [139]. The objective is to focus on the principal direction of current flow within each brain region defined by the Anatomical Automatic Labeling (AAL) atlas version 4 [140]. This technique allows us to reduce the dimensionality of the source moments, emphasizing the primary mode of activity within each region.

Mathematically, SVD decomposes a matrix into three component matrices, as described in Equation 6.1:

$$\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \tag{6.1}$$

Where:

- $\mathbf{M}$ represents the original 3-dimensional moments matrix for a given dipole over time.
- $\mathbf{U}$ is the left singular vectors matrix.
- $\mathbf{\Sigma}$ is a diagonal matrix containing the singular values, ordered in descending order.
- $\mathbf{V}^T$ is the transpose of the right singular vectors matrix.

In our application, we focused on the principal direction by considering only the first singular vector, $\mathbf{u}_1$, which corresponds to the maximum singular value $\sigma_1$. This singular vector captures the dominant mode of variability in the source activity. The moments in each region were projected onto the principal direction using the dot product, as shown in Equation 6.2:

$$\mathbf{m} = \mathbf{u}_1^T\mathbf{M} \tag{6.2}$$

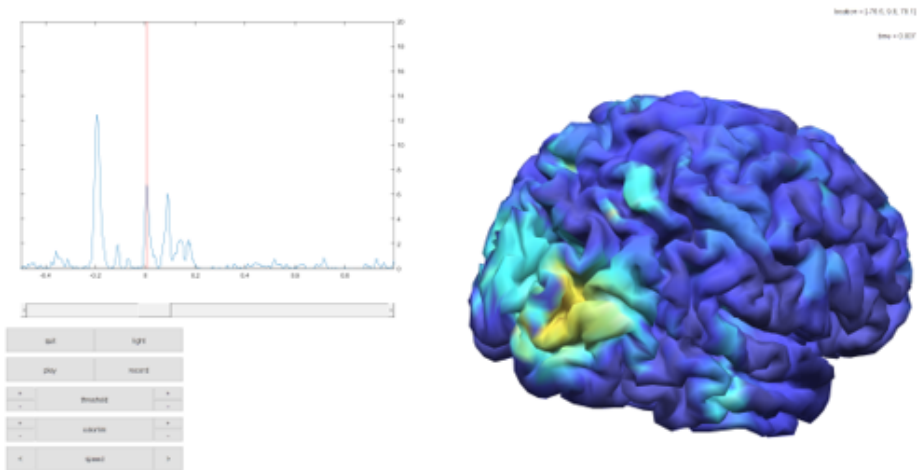


**Figure 6.4.** Result of the source reconstruction process

Where:

**m** represents the 1-dimensional moment, emphasizing the primary mode of activity within the region.

This process was performed for each dipole within each region defined by the AAL atlas, resulting in a concise representation of the neural activity for further analysis.

In summary, the SVD technique was utilized to reduce the dimensionality of source moments, focusing on the principal direction of neural activity within specific brain regions. This approach provided a more compact and informative representation of activity patterns for subsequent investigations.

Figure 6.4 shows the result of the reconstruction at the moment of appearance of the target picture for one participant.

## 6.2 Related Work

The lack of ground truth brain activity makes validating reconstruction algorithms a complex task. This issue impacts the forward problem as well, as

we have limited tools to accurately characterize head shapes and conductivity along the path from neurons to EEG sensors. Variations in head shape significantly affect the accuracy of the forward model [97]. Even if we could obtain a perfect forward model, the absence of ground truth activity forces us to rely on simulated EEG data to assess the reliability of reconstructed sources.

In the literature, benchmark frameworks are proposed for evaluating reconstruction using fixed pseudo-EEG data, while other tools enable the generation of custom data. However, to the best of our knowledge, none of the available pipelines offers a standardized approach for validating a reconstruction method using custom pseudo-data. Therefore, we have developed a versatile validation framework for EEG-based source localization [141]. This new tool is an all-in-one validation pipeline designed to assess source localization using pseudo-EEG data that closely mimics the experimental environment of the study. The framework comprises five steps, from configuration to evaluation.

The question of evaluating source reconstruction methods has been addressed since the late 90s with phantom studies that controlled inverse method accuracy using highly detailed volume conduction models [142,143]. Subsequently, EEG simulation gained importance for validating source reconstruction in the literature [144,145]. Most evaluation methods are custom-made and rely on different assumptions (linear models, spatial dependencies, etc.) or have been designed for specific cases, such as the Source Information Flow Toolbox (SIFT) [146] for connectivity and blind source separation evaluation, or simBCI for studying Brain-Computer Interface (BCI) methods [147]. Haufe and Ewald [148] proposed a more general benchmark framework for EEG-based source localization and connectivity. However, they limited their analysis to only two activated sources and eight brain regions (octants), which may not be sufficient to ensure the reliability of a specific source reconstruction pipeline. Furthermore, they did not provide users the opportunity to customize the generated pseudo-EEG signal, restricting the analysis to oscillatory signals in the alpha band.

To address the lack of custom EEG data simulation tools in the literature, Krol *et al.* [149] introduced the Simulating Event-Related EEG Activity (SEREEGA) toolbox. This toolbox's purpose is to generate custom pseudo-event-related EEG data, allowing users to generate EEG signals with known ground truth according to their own signal patterns, head models, source localizations, and event timestamps.

However, a gap still exists between signal generation and validation. Our aim is to bridge this gap by offering a comprehensive validation framework built upon SEREEGA. We have extended SEREEGA to provide additional features, creating a complete validation framework, which includes:

- Artifact generation
- Evaluation methods for source localization
- Adaptive region-based ground truth

These features empower users to evaluate their source reconstruction pipelines using realistic signals with precision tailored to their chosen brain atlas.

For the sake of standardization, we have developed this framework in Matlab, as many major toolboxes for EEG data analysis are based on this platform. To ensure accessibility to the widest possible audience, we offer easy configuration via a JSON file. Additionally, our open-source FieldTrip-based codes are available in the GitHub repository: `https://github.com/numediart/ValidEEG.git`.

## 6.3 Proposed Framework

The proposed framework is based on the following publication:

- "*A Versatile Validation Framework for ERP and Oscillatory Brain Source Localization Using FieldTrip*", In 4th International Conference on Biometric Engineering and Applications (ICBEA'21), May 25–27, 2021, Taiyuan, China. [141]

The goal of our work is to provide an easy-to-configure validation framework for brain source localization. The versatile aspect allows the users to validate their pipelines on pseudo-data closer to their own use case. The framework is divided into 5 steps:

- Parameters selection: the custom parameters are defined in a configuration (.json) file allowing the user to design the framework with specific considerations such as the number of pseudo-sources (n_dipoles) or the number of trials within one session (n_trials).

- Source selection: n_dipoles are selected from the predefined atlas so that the region corresponding to each dipole is not a neighbor of the other selected dipole's regions.
- Pseudo-source signal generation: a signal containing n_trials occurrences of the desired pattern (ERP or oscillatory) is generated for each of the selected sources.
- Pseudo-EEG data generation: the final pseudo-EEG data are first generated through FieldTrip functions and some artifacts are further added as well as noise.
- Performance evaluation: the n_dipoles non-neighboring reconstructed regions with highest power are considered as the source regions and the score for each dipole of each session is given as follow:
  - 1 if reconstructed source region = pseudo-source region.
  - 0.5 if reconstructed source region is the neighbor of a pseudosource region.
  - 0 if reconstructed source region is the second neighbor (neighbor of neighbor) of a pseudo-source region.
  - -1 otherwise.

  The final score is the mean of each individual score through all sessions.

### 6.3.1 Parameters Selection

The parameters have been chosen as a trade-off between controllability (enough parameters to fit a specific analysis) and simplicity (limited number of parameters allowing an easy handling). Those parameters can be classified in 4 types:

- **General pipeline**: definition of the number of sessions/ dipoles/trials over which we want to generate the pseudo-source signal. The session and trial lengths are also defined there as well as the number of artifacts we want to introduce in the final EEG pseudo-data. An event file can also be defined there to control the appearance time of each trial.

- **Pseudo-source definition**: selection of the source type (ERP or oscillatory) and their main features (e.g., specific peaks for ERP and frequency bands for oscillatory signals). The desired atlas is also defined there. A dipole file can be specified to control the dipole locations for each session.
- **Pseudo-EEG definition**: selection of the head model and the electrodes with respect to FieldTrip requirements.
- **Artifacts and noise**: Template artifacts and SNR, for source and EEG signals, are defined in this section.

### 6.3.2 Source Selection

Each source is defined as a 3-dimensional dipole with specific position and orientation. The dipole position is the position of a randomly chosen dipole among those of a predefined atlas. The atlas, given as a parameter, is a FieldTrip mesh structure where each dipole's region is defined. Importantly, the atlas mesh must be aligned to the head model and the electrodes. Templates of the required data are provided (template atlas is the AAL MNI atlas [140]), but custom atlas can be obtained through the function prepare_atlas. The orientation of each dipole is defined as a random unitary vector. To ensure the selected dipoles are not part of the same region or neighboring ones, we designed a neighboring matrix of the atlas regions as shown in Figure 6.5. This matrix fulfills 2 conditions:

- A region from one hemisphere cannot be a neighbor of one of the other hemisphere.
- If two regions are neighbors in one hemisphere, their corresponding regions in the other hemisphere must be neighbors too.

### 6.3.3 Pseudo-Source Signal Generation

The pseudo-source signal is generated as a series of base signal trials around a baseline (cf. Figure 6.6). This base signal is either an ERP or an oscillatory signal (OSCIL) depending on the *type* parameter. The trial samples are defined through the event parameter. We used the SEREEGA toolbox [149] to generate the base signals as follow:

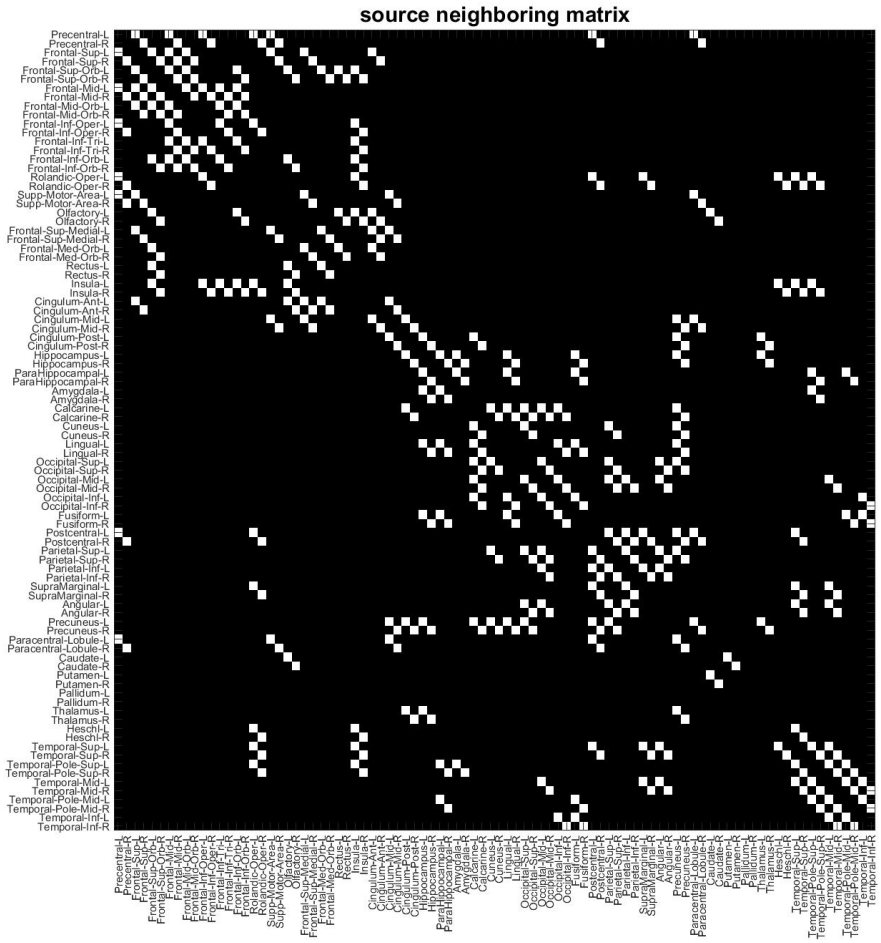



**Figure 6.5.** Source neighboring matrix. For each region in row, neighbors are defined with a white square.

- **ERP**
  An ERP trial is a series of positive and/or negative peaks defined as a normal probability density function around the specified latency (e.g., P300 is a positive peak appearing 300ms after the beginning of a trial) with the corresponding *width* ($\sigma$) covering 6 standard deviations and the maximum amplitude ($A$) being the corresponding *ampli* parameter. This can be mathematically expressed as:

  $$ERP(t) = A \cdot e^{-\frac{(t-\text{Latency})^2}{2(\sigma/6)^2}}$$

  To introduce variability between trials, we defined a latency deviation ($\Delta t$) varying between +/- 50ms, a width deviation ($\Delta\sigma$) of half the desired width, and an amplitude deviation ($\Delta A$) of a fifth of the corresponding amplitude. These deviations can be represented as:

  $$\Delta t \sim \mathcal{U}(-50, 50) \text{ ms}$$

  $$\Delta\sigma \sim \mathcal{U}(-\tfrac{\sigma}{2}, \tfrac{\sigma}{2})$$

  $$\Delta A \sim \mathcal{U}(-\tfrac{A}{5}, \tfrac{A}{5})$$

  where $\mathcal{U}$ represents a uniform distribution.

  An additional parameter introduces habituation to the stimulus along the session through a decaying slope in amplitude. This slope leads the last trial amplitude to be a fourth of the initial amplitude. The amplitude decay can be modeled as:

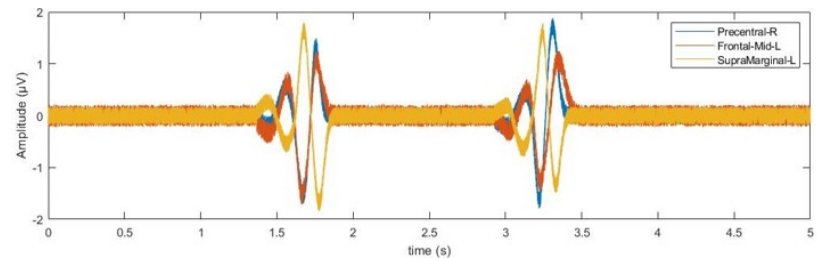
  $$A(t) = A_0 \left(1 - \tfrac{t}{T}\right)$$

  where $A(t)$ is the amplitude at time $t$, $A_0$ is the initial amplitude, and $T$ is the session duration.

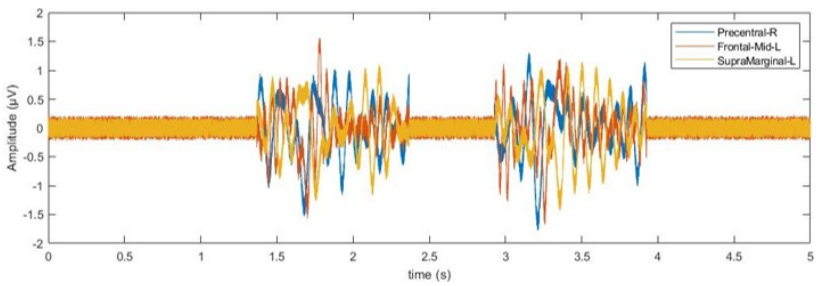
  To consider the polarity inversion between anterior and posterior brain regions [150], signals from anterior sources are reverted.

  The last step is the addition of pink noise. This colored noise, inversely proportional to the frequency, is generated following Zhivomirov method [151] with respect to the *snr_source* parameter. An example of a 3-dipoles pseudo-ERP source signal is given in Figure 6.6a.

- **OSCIL**
  An oscillatory trial is defined as an event-related spectral perturbation




**(a)**




**(b)**

**Figure 6.6.** Example of 3-dipoles pseudo-source signals with the 2 first dipoles being on anterior regions and the 3rd once on the posterior region. (a) pseudo-ERP defined as a series of P100, N200, P300 and N400. (b) pseudo-oscillatory signal with frequency band of each dipole defined as: 8-12Hz (blue), 16-24Hz (orange), 9-13Hz (yellow).

(ERSP) [152]. This signal is obtained by band-pass filtering a uniform white noise in a predefined frequency band (*freq* parameter) using a Kaiser window-based finite impulse response filter [153] with a specific amplitude (*ampli* parameter) and a random phase. Finally, pink noise is added to the signal with respect to the *snr_source* parameter. An example of a 3-dipoles pseudo-oscillatory signal is shown in Figure 6.6b.

### 6.3.4 Pseudo-EEG Signal Generation

The pseudo-EEG signal generation consists of 2 steps: the first one creates the EEG signal on each channel as a FieldTrip *raw* structure, the second step introduces artifacts within the data:

1. **From source to EEG**: FieldTrip offers the opportunity to simulate channel-level time-series data from one or multiple dipole signals considering a specific volume conduction model, that geometrically defines the head model and carry information about the different tissues through which the electrical signal will spread (i.e., white/grey matter matters, cerebro-spinal fluid, skull, and scalp), and a particular electrode montage. White noise with a relative level to data signal corresponding to *snr_eeg* parameter is also added to the generated pseudo-signal. The resulting EEG data are then normalized. An example of FieldTrip-generated EEG data is shown in Figure 6.7.A. To provide a clearer representation of the signal characteristics, we conducted a timelock analysis on the FieldTrip-generated signals. This analysis involved averaging across all trials while ensuring alignment with the stimulus onset, starting at 0 seconds for the simulated data (cf. Figure 6.7.B).

2. **Artifact generation**: On top of the FieldTrip-generated EEG signal, we introduce artifactual signals. Those artifacts have been chosen within the annotated corpus of Hamid et al. [154] . This dataset consists of 310 EEG recordings in which every artifact has been annotated as one of the five following types: electrode, eye movement, muscle, chewing or shiver artifacts. We decided to only use the first three types in our pseudo-data as chewing and shiver artifacts are so rare that their effect on the result of timelock-based source reconstruction algorithms is negligible. We have extracted all the artifactual segments from recordings of the patient number 100 to represent our template artifacts. The artifact trials, being recorded with a sampling rate of 256 Hz on a 19-channel set-up, are first linearly interpolated to the pseudo- EEG sampling rate, at 2048 Hz. Then, a second interpolation is conducted to generate artifact signals on channels of the pseudo-EEG set-up that are not present in the template artifact dataset. This latter interpolation is based on a neighboring matrix computed from the pseudo-EEG channel positions similarly to Figure 1. Every missing channel signal is computed as the

mean of the neighbor's signals. Finally, the artifact trials are normalized in order to keep comparable scales between pseudo-EEG and artifact signals. Following the chosen configuration, *n_artifacts* are randomly selected from the adapted trials and added to the pseudo-EEG signals at random timestamps. An example of the final pseudo-EEG data is shown in Figure 6.7C. It is noteworthy that some studies propose the automatic generation of realistic EEG signals using generative models. For instance, Macke et al. employed diffusion probabilistic models for this purpose [155]. Their approach involves a noising-denoising process, where they successively introduce noise to their initial EEG recordings and then employ a deep neural network to reverse this operation by denoising the signal. By incrementally increasing the level of noise, they develop a model capable of generating an EEG signal from a noise-only input. However, such techniques do not afford control over the specific types of artifacts introduced into the data, which is a desirable feature in our particular case.

### 6.3.5 Performance Evaluation

We propose a qualitative evaluation of the performance of a timelock-based source reconstruction algorithm applied to the pseudo-EEG signals of each session. This evaluation is based on the closeness of reconstructed regions with highest root mean square values to the ground truth regions (i.e., regions to which the pseudo-source belongs). To avoid a dipole to be considered twice, we ensure the selected regions not to be neighbors using the previously computed source neighboring matrix. For each session, once the *n_dipoles* highest power regions selected, we compare each ground truth region with the selected regions and a qualitative score is assigned as follow: if one of the reconstructed region is the same as the ground truth, the score is 1; if one region is a neighbor of the ground truth, the score is 0.5; the score is 0 if one of the selected regions is a neighbor of the ground truth's neighbors; otherwise, the score is -1. This simple qualitative assignment gives the opportunity to be less penalized if the ground truth dipole is at the limit of several regions while greatly penalizing totally wrong reconstruction. The score is therefore represented by an *n_session*n_dipoles* matrix with the global mean being the final score.

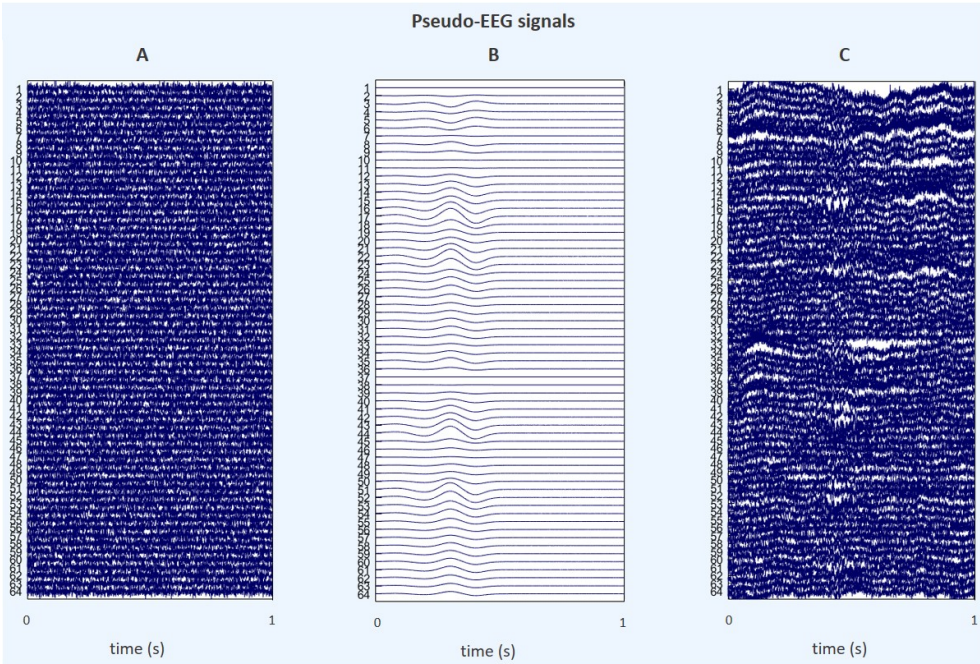



**Figure 6.7.** Example of a 3-dipoles pseudo-EEG signal. (a) 64-channels EEG generated from a 3-dipoles pseudo-ERP signal using forward modeling. (b) timelock analysis of signals in (a). (c) final pseudo-EEG signals after having added muscle artifact to the signal in (a).

We performed the reconstruction, considering the benchmarking configuration described in Section 6.3.6, using the MNE method proposed by Hansen *et al.* [131]. Prior to reconstruction, we preprocessed the pseudo-EEG data following the framework described in Chapter 5. Additionally, we utilized the template head and source models provided by FieldTrip, along with their recommended default conductivity values, to obtain the forward solution. Figure 6.8a and 6.8b shows how the evaluation is represented in our framework through this example. As the ground truth dipole positions and regions are automatically saved during generation, the users can visualize the reconstructed vs. ground truth sources on their atlas source model as shown on Figure 6.8c.

### 6.3.6 Benchmark

As we provide template data and configuration, the proposed validation framework can be used as a benchmark generator. The chosen template configuration is a 10-sessions 3-dipoles pseudo-source signals carrying, within each session of 15 minutes, 200 one-second trials and 400 artifacts, lasting less than 10 seconds, with a sampling rate of 2048 Hz. The artifact trials were randomly chosen among a set of 184 artifact segments composed of 13 electrode artifacts, 54 eye movement artifacts and 117 muscle artifacts. The template electrode is the 10/20 standard set-up from FieldTrip template over which we only kept 64 electrodes with respect to the 10/20 standard. The template volume conduction model (i.e., head model) were built following FieldTrip pipeline from their standard MRI template head model that we have segmented using five tissue types (gray matter, white matter, cerebro-spinal fluid, skull, and scalp) using the SimBio finite element method to build the forward model. The template atlas is the AAL MNI atlas provided by FieldTrip from which we kept only the 90 first regions as the cerebellum and the vermis are not part of our head model. We then realigned the electrodes, head model and atlas together with respect to the CTF coordinate system. The pseudo-ERP template is a 4 peaks ERP composed of P100, N200, P300 and N400 with corresponding amplitudes of 0.2, 0.4, 1 and 0.8 microvolt and widths of 300, 300, 200 and 200 milliseconds, respectively. The pseudo-oscillatory template is defined by a specific frequency band for each dipole: 8-12 Hz (dipole 1), 16-24 Hz (dipole 2) and 9-13 Hz (dipole 3) with a maximum amplitude of 1 microvolt for all of them. The source SNR is set to 1, while the EEG SNR is set to 2.

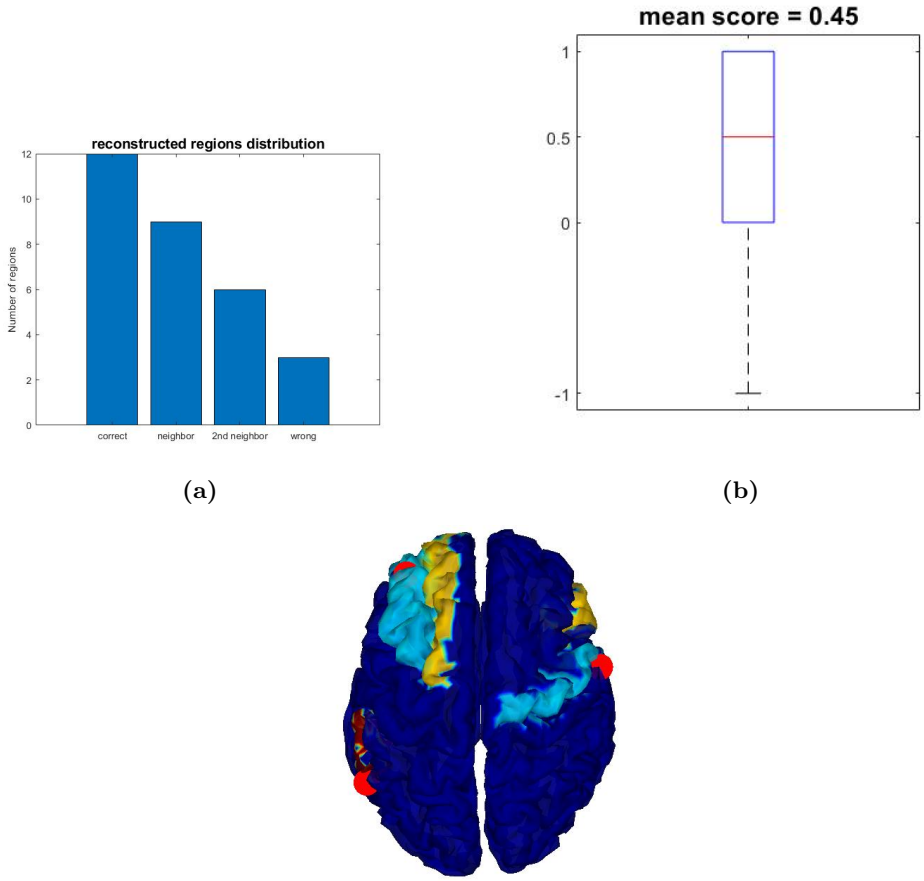



**Figure 6.8.** Example of the evaluation of a 10-sessions 3-dipoles source reconstruction of pseudo-EEG signals. A: distribution of the correctness of reconstructed regions through their relative position to the initial pseudo-regions (i.e., correct, neighbor, second neighbor or wrong position) (left) and the corresponding score statistics (right). B: mean accuracy score computed following the neighbor-based evaluation rules. C: top view of reconstructed regions (yellow) from one session in comparison with the ground truth pseudo-dipoles (red spheres) and their corresponding region (light blue). The region in red is a properly reconstructed source.

## 6.4 Discussion

From EEG simulation to localization evaluation, the proposed framework offers the possibility to customize multiple features so that the users can validate their method in a very specific way to fit their experimental data. The provided configuration file gives an easy way to modulate the framework while the open-source FieldTrip-based code allows more sophisticated analyses. The addition of artifacts makes the generated signals closer to real EEG data. We therefore offer a large set of artifactual segments composed of electrode, eye movement and muscle artifacts, but expert Matlab users will easily be able to select specific artifacts from the dataset provided by Hamid *et al.* [154]. Future work may involve validating the realism of synthetic signals. One approach would be to train a Generative Adversarial Network (GAN) to distinguish between real and synthetic EEG signals. The GAN's performance can serve as a reference for assessing the realism of the synthetic signal.

The evaluation process is based on neighboring matrices computed on source and sensor domains. Those matrices can be adapted to the desired accuracy through the provided template functions. The chosen atlas also influences the way the validation is performed. Our template atlas is composed of 90 regions, but some users may want to work with different brain regions. For this purpose, it is possible to either combine and/or remove regions from an existing atlas or to transform an atlas in NIfTI format to the required FieldTrip-like source model atlas while realigning it to the selected volume conduction model and the EEG-cap to CTF coordinates. Importantly, the final reconstructed source activity must be given as one signal per region with regions order corresponding to the chosen atlas.

We emphasize the scope of the proposed validation framework, which focuses on evaluating the accuracy of a reconstruction algorithm in performing the inverse solution from predefined pseudo-EEG data, given a known forward model. It's important to note that this study does not aim to validate the proper representation of a subject's head or the corresponding conductivity values. Instead, it concentrates solely on the preprocessing procedure and the reconstruction method. The primary objective of this study is to demonstrate the utility of the proposed framework for conducting a fair comparison between different source reconstruction methods, specifically in terms of their accuracy in localizing brain source regions. It's important to note that this

study does not encompass the validation of existing methods. This is because the outcomes of each method can be significantly influenced by the chosen parameters, requiring a detailed investigation of each method beforehand to ensure a truly equitable comparison.

The framework's versatility and adaptability make it suitable for assessing the efficiency of a specific method in various situations. The provided template configuration serves as a benchmark for a typical scenario and can be expanded to accommodate specific needs within the research community.

## 6.5 In Brief

Summary of Chapter 6

- Source reconstruction can be influenced by various factors, including inaccuracies in head shape representation, conductivity patterns, and signal processing procedures. This chapter focuses primarily on addressing issues related to signal processing.
- The proposed validation framework relies on the use of synthetically pseudo-EEG signals, which can be easily customized via a configuration file.
- Performance evaluation is conducted with reference to a specific target brain atlas, utilizing a defined neighboring matrix.
- A template configuration is provided to facilitate the benchmarking of source reconstruction algorithms.

Perspective for Chapter 6

- Extend the validation framework to encompass connectivity analysis, broadening the scope of source localization assessment.
- Enhance the framework by allowing users to customize artifact shapes using recorded signals, thereby increasing its flexibility and applicability.

# Chapter 7

# Human-Centered Explainable AI

## Contents


7.1 Genesis . . . . . 126
7.2 Model Architecture . . . . . 132
7.3 Use Case: Cat/Dog Classification . . . . . 136
7.4 Discussion . . . . . 139
7.5 In Brief . . . . . 142


In the realm of machine learning, biases often affect models in ways that are imperceptible to human experimenters. This can create the illusion that the model performs a given task as proficiently as a human expert. Understanding these biases is of utmost importance.

This chapter introduces our novel method for addressing the *confounder exploitation bias*. This bias emerges from the tendency of machine learning models to leverage confounding factors that exhibit heterogeneity across categories. To tackle this issue, we propose a strategy that makes these biasing factors more apparent to experimenters. Our approach falls within the domain of xAI and enhances the transparency of deep learning models. It does so by adopting a human-centered perspective that involves comparing the analyzed input data among themselves.

We begin by delineating the genesis and development of the proposed *human-centered xAI* approach in Section 7.1. Section 7.2 provides the detailed model architecture, while Section 7.3 illustrates its benefits with a cat-and-dog image classification case, considering bias from image rotation. Section 7.4 discusses the results and their implications.

## 7.1 Genesis

In order to address scenarios where current inherent and post-hoc approaches encounter difficulties, as discussed in Section 1.3, we propose a novel *human-centered xAI* approach. This principle is rooted in the way humans commonly make decisions through comparisons with familiar situations or items. For instance, when encountering a small dog on the street, we identify it as a small dog because we have seen many dogs and can judge that this one is smaller in comparison.

This approach, which emphasizes item comparison to generate explanations, stands in contrast to existing post-hoc explanations that focus on highlighting specific aspects of an item as most relevant for the model's decision. It bridges the interpretability gap present in heatmap-based explainability methods. These methods often leave it to humans to determine what was essential for the model's decision within a region highlighted by the explanation algorithm. This includes aspects like the shape, edges, pixel values, or texture.

This comparison-based principle was previously leveraged by Chen *et al.* in their work, where they proposed comparing the input of interest with prototypical examples representing important components of each studied category [83]. When the closest prototype is identified, any component of the input can be compared with the prototype to ascertain which components were influential for the classification. Nevertheless, this method still faces challenges regarding how prototypes are constructed and the model's handling of human-imperceptible features.

To tackle this issue, Wang *et al.* proposed an approach that eliminates the model's focus on inhuman features by removing them from the training dataset [84]. However, this method necessitates the identification of all imperceptible features that could affect the training process in advance, which is task-specific and relies on the experimenter's expertise.

Therefore, we aim to introduce an approach that can be universally applied to any deep learning model and allows for a posteriori comparisons based on the specific task at hand.

A promising avenue for examining how the model represents input data is delving into the latent representation within the neural network. The latent

space has been extensively explored in tasks involving image generation. It enables the transition from one class to another or the addition/removal of specific features in an image by selecting samples from particular locations where the model internally represents the desired class or feature.

This technique is especially effective when dealing with Variational Auto-Encoders (VAEs), as these models learn specific distributions that represent classes and features [156–158]. Navigating through a latent distribution corresponds to transitioning from samples highly correlated with a specific class or feature to another class or feature with a smooth transition in between, as illustrated in Figure 7.1 in the context of facial transitions.


**Figure 7.1.** Example of Latent Space Exploration in VAEs on Smooth Transition Between Faces. By manipulating continuous variables in the latent space, we can transform the attributes of the same face. This enables us to age a young person gracefully or rejuvenate an elderly individual, enhance masculine or feminine features, all while retaining the original essence. Left images represent the starting point of these attribute transitions. Reproduced from [159].

Drawing inspiration from the field of image generation, we can explore the latent space obtained after the training of a model for a specific task. This latent space allows us to visualize the variations in input data as we transition from one class's location in the latent space to another. The changing components along this trajectory correspond to crucial features influencing the model's decision-making process. The explanation task, in this context, involves identifying these essential components.

Similar to image generation tasks, this exploration necessitates a non-sparse distribution in the latent space to achieve a smooth transition. In other words, a non-sparse latent space means that there are no empty or unoccupied regions within it. Every part of the latent space contributes meaningfully to the understanding of data. A non-sparse latent space is akin to a well-connected and informative map that allows for smooth transitions between different data variations, ensuring that the resulting insights are coherent and gradual throughout the transition. In contrast, sparse spaces have gaps or areas where there is little to no information, leading to abrupt changes during transitions. However, classical deep learning classifiers do not inherently enforce non-sparse distribution. To address this challenge, specific architectural designs, such as VAEs, are introduced to impose the required distribution.

One challenge with VAEs is the introduction of a random variable $\epsilon$ via the reparameterization trick [156], which makes the latent space stochastic, disrupting the deterministic relationship between input images and output classification, as shown in Equation 7.1.

$$\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \cdot \boldsymbol{\epsilon} \tag{7.1}$$

Where:

$\mathbf{z}$ is the resulting latent variable

$\boldsymbol{\mu}$ is the mean vector of the latent variable distribution

$\boldsymbol{\sigma}$ is the standard deviation vector of the latent variable distribution

$\boldsymbol{\epsilon}$ is a random sample from a standard normal distribution

Without a deterministic relationship between the input and the latent space, for a specific sample, the corresponding latent representation is not exclusive, which makes it inappropriate for classification. To overcome this, Zhang *et al.* proposed the introduction of another latent vector before applying the reparameterization trick, resulting in a deterministic latent space [160]. This deterministic latent space is further encouraged to replicate the distribution of the variational latent space through an adversarial training phase. A discriminator module is trained to distinguish between samples originating from the deterministic latent space and those derived from the variational one, while the encoder is trained to deceive the discriminator by generating a determin-

istic latent space with a distribution closely resembling that of the variational latent space.

The deterministic nature of the added latent vector makes it well-suited for classification tasks. Figure 7.2 and Algorithm 1 illustrate how this deterministic latent vector, denoted as $z_I$, is incorporated into the standard VAE architecture.

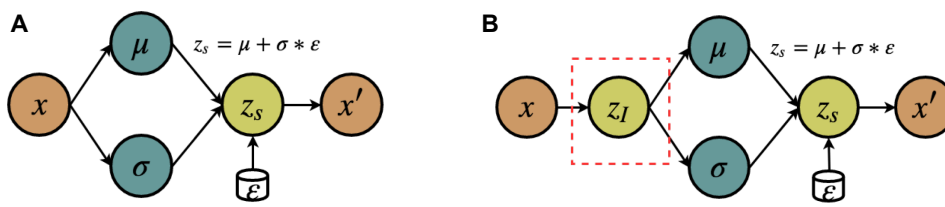



**Figure 7.2.** Introduction of a Deterministic Latent Vector $z_I$ into Standard VAE. In this comparison between the standard VAE (A) and VAE++ (B), proposed by Zhang *et al.*, x and x' represent the input and reconstructed data, while $\mu$ and $\sigma$ signify the learned expectation and standard deviation. In the standard VAE, $z_S$ is considered the learned representation, composed of $\mu$, $\sigma$, and $\epsilon$, where $\epsilon$ is randomly sampled from $\mathcal{N}(0, 1)$. VAE++ introduces a deterministic latent vector $z_I$, highlighted here, which can be employed for classification purposes. Reproduced from [160].

**Algorithm 1:** Adversarial Variational Embedding Proposed by Zhang *et al.* [160]

```
Input: labelled observations (X^L, Y^L)
Output: Deterministic Latent Representation z_I
for x in {X^L} do
    z_I ← x
    μ, σ ← z_I
    Sampling ϵ from N(0, I)
    z_s = μ(z_I) + σ(z_I) * ϵ
    x' ← z_s
    L_VAE ← x, x', p(z_s|x)
    for z_I, z_s, y in Y^L do
        y_GAN ← z_I, z_s
        L_GAN ← y_GAN
    end
    Minimize L_VAE and L_GAN
end
return z_I
```

Where:

$z_s$ is the latent representation randomly sampled from a Gaussian distribution

$z_I$ is the latent representation directly learned from the input data

$\mathcal{L}_{VAE}$ is the common loss used to train variational auto-encoders (cf. Equation 8.1)

$\mathcal{L}_{GAN}$ is the common loss used to train adversarial networks (cf. Equations 7.3, 7.4)

$y_{GAN}$ is the output of the discriminator

To enhance the performance of distribution-constrained models and expand their applicability beyond Gaussian distributions, Makhzani *et al.* introduced the concept of Adversarial Auto-Encoders (AAEs) [161]. The key idea is to disentangle the distribution learning process from the reconstruction process.

This is achieved by utilizing a standard Auto-Encoder for reconstruction and incorporating a discrimination step that distinguishes the latent vector from a vector directly sampled from a desired distribution, which operates independently of the rest of the neural network. The architecture of the AAE they proposed is visually depicted in Figure 7.3. It allows for the addition of a classification module, which takes the latent vector $z$ as input to perform the desired task.

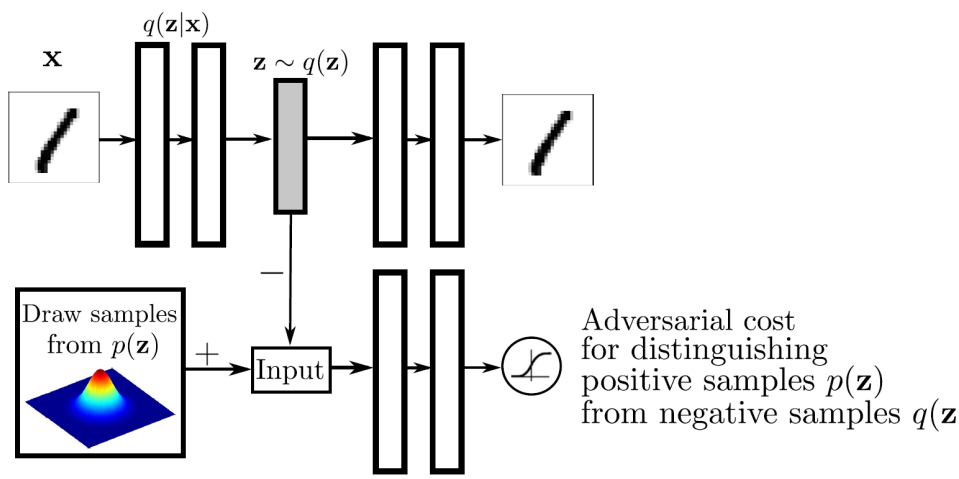



**Figure 7.3.** Adversarial Autoencoder Design. The upper row depicts a standard autoencoder's reconstruction of an image from a latent code, while the lower row illustrates a second network trained to discern between hidden code and a user-specified distribution sample. Reproduced from [161].

The essence of our proposed *human-centered xAI* approach involves training an AAE to execute the designated task. Subsequently, we navigate through the latent space $z$ to identify the features that have been most instrumental in facilitating accurate model classification. To complete the process, we scrutinize the origins of these features to assess the extent to which the *confounder exploitation bias* has influenced the classification.

## 7.2 Model Architecture

In standard deep learning classifiers, the architecture primarily consists of two components: an encoder and a classifier. The encoder, built with hidden layers, is responsible for *feature extraction* from the input preprocessed data. It outputs a latent vector $z$ that is optimized to be highly discriminative for the target categories. The classifier, usually a fully connected layer, then processes this latent vector $Z$ to yield a probability score for each of the categories under study. This typical architecture is depicted in Figure 7.4A and is applicable not just to classification tasks but also to regression.

In our approach, we seek to modify the distribution of the latent space. To achieve this, we incorporate two crucial modules to the standard deep learning classifier: a decoder and a discriminator. The decoder endows the model with properties characteristic of an AE. This ensures that the latent vector encapsulates all the requisite independent information, enabling the original input data to be reconstructed from it. The discriminator, on the other hand, enforces the preferred distribution on the latent space. It achieves this through adversarial training, wherein the encoder assumes the role of a GAN generator. The discriminator's task is to distinguish between the latent samples generated by the encoder and random samples drawn from the desired distribution. This enhanced architecture is illustrated in Figure 7.4B. A notable feature of the latent space produced by this model is its non-sparsity. This characteristic facilitates the exploration of transitions between classes by navigating along the most discriminative directions, a concept further elaborated in Section 7.3. We have named this new model *xAAEnet* for eXplainable Adversarial Auto-Encoder network.

The proposed model follows a curriculum learning approach for training [162]. The training process is divided into three main phases: Autoencoder (AE), Generative Adversarial Network (GAN), and Classifier. The loss functions used for each training step consider the previous ones to build upon the previously acquired properties.

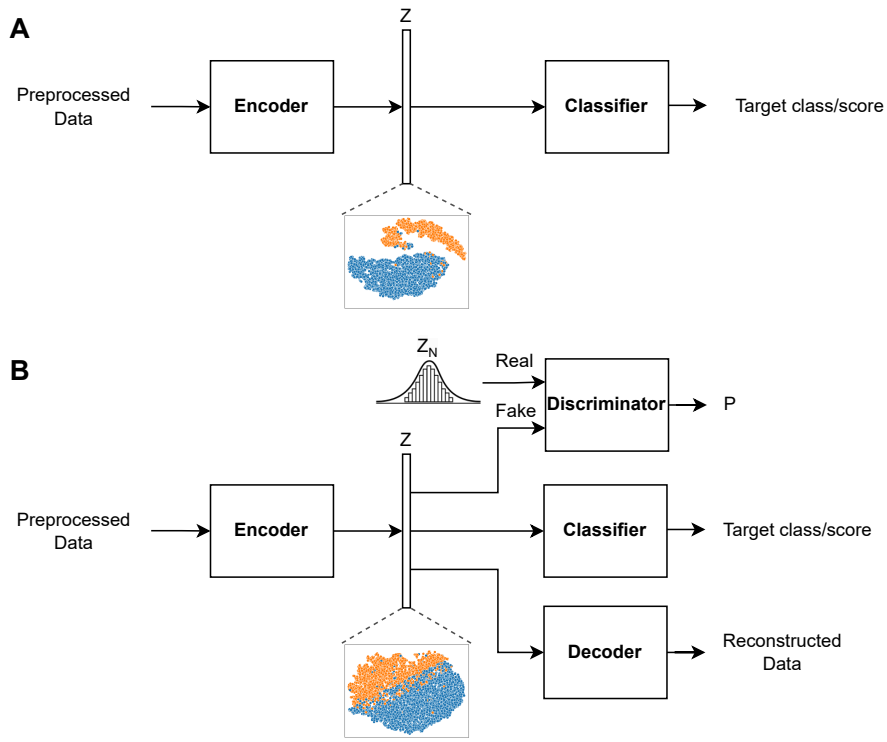



**Figure 7.4.** Transformation from Standard Architecture to Explainable Adversarial Auto-Encoder Network (xAAEnet). (A) Depicts the traditional structure consisting of an encoder for *feature extraction*, leading to a sparse latent space configuration maximizing the discriminability between classes, and a classifier for category prediction. (B) Illustrates the architecture required for our human-centered xAI approach, integrating a decoder for data reconstruction and a discriminator to regulate and modify the latent space distribution. This controlled latent space enables more interpretable transitions between classes.

## Autoencoder (AE)

The AE aims to learn a compact representation of the input data through encoding and decoding. The AE produces a low-dimensional, discriminative representation that is useful for clustering and classification [163]. The latent representation $Z$ is generated by the encoder, while the decoder, which is a

mirrored version of the encoder, provides a reconstructed version of the input data. To train the AE, we use the Huber loss, as described in Equation 7.2:

$$\mathcal{L}_{AE} = \mathcal{L}_{Huber}(y, \hat{y}) = \begin{cases} \frac{1}{2}(y - \hat{y})^2 & \text{if } |y - \hat{y}| \leq \delta \\ \delta \cdot |y - \hat{y}| - \frac{1}{2}\delta^2 & \text{otherwise} \end{cases} \tag{7.2}$$

Where $y$ is the true input value, $\hat{y}$ is the predicted value, and $\delta$ is a positive scalar that adjusts the sensitivity to outliers. The Huber loss balances the benefits of both Mean Squared Error (MSE) and Mean Absolute Error (MAE) [164]. It has the robustness of MAE for large errors while maintaining the mathematical properties of MSE for small errors.

The integration of a decoder module into the model enables the preservation of a maximum amount of independent information from the input data within the latent representation. This preservation of information empowers researchers to subsequently explore specific features of interest. The primary objective of this exploration is to determine whether the presence or alteration of the identified feature influences the decision-making process of the model.

### Generative Adversarial Network (GAN)

The goal of the GAN training phase is to generate a latent vector $Z$ that conforms to a specified target distribution. This is achieved by training the GAN to produce a latent vector $Z$ that closely resembles another vector, $Z_N$, which is sampled from the desired target distribution. Within the GAN framework, $Z_N$ is treated as the "real" latent representation while $Z$ is considered as the "false" representation. This adversarial network consists of a generator (encoder) shared with the AE and a discriminator (MLP). The generator's loss function is a weighted sum of the AE loss and the mean of correct predictions by the discriminator within a batch, as described in Equation 7.3:

$$\begin{aligned} \mathcal{L}_{adversarial} &= \frac{1}{bs}\sum_{i=1}^{bs}(1 - fake_i) \\ \mathcal{L}_{generator} &= (1-\alpha)\cdot\mathcal{L}_{AE} + \alpha\cdot\mathcal{L}_{adversarial} \end{aligned} \tag{7.3}$$

Where $bs$ is the batch size, $\alpha$ is the weight of the adversarial loss, and *fake* is the output of the discriminator when the "fake" latent representation ($Z$) is given as input.

The discriminator's loss function is the difference between the mean fake predictions and the mean real predictions, as shown in Equation 7.4:

$$\mathcal{L}_{discriminator} = \frac{1}{bs}\sum_{i=1}^{bs} fake_i - \frac{1}{bs}\sum_{i=1}^{bs} real_i \tag{7.4}$$

Where *real* is the output of the discriminator when the "real" latent representation ($Z_N$) is given as input.

The GAN training phase, in addition to ensuring non-sparsity in the latent space, equips it with generative capabilities. This property is utilized in image generation to produce samples that belong to the same category as those in the training dataset but are distinct from them. This is achieved by selecting latent vectors from the obtained distribution. In our scenario, we confer generative ability to the latent space by enforcing that the input data adheres to the same specific distribution. This ensures that future unseen inputs of the same category as the training dataset will also align with the latent space distribution established during training. Consequently, they can be subjected to analysis in a consistent manner.

### Classifier

The classifier module encourages the encoder to output a latent vector that maximizes the discriminability between classes. It consists of a single-layer perceptron that performs a linear combination of the latent vector values,

outputting a probability for the sample to belong to each class via a softmax activation function. The loss function used for this step combines the cross-entropy loss (CE) with the previous loss functions in the curriculum learning philosophy, as shown in Equation 7.5:

$$\mathcal{L}_{classifier} = \alpha \cdot \mathcal{L}_{AE} + \beta \cdot \mathcal{L}_{adversarial} + \gamma \cdot CE(predicted, target) \tag{7.5}$$

Where *target* is the ground truth class, *predicted* is the output of the classifier module, and $\alpha, \beta, \gamma$ are the weights allocated to each module.

## 7.3 Use Case: Cat/Dog Classification

This section demonstrates the application of the proposed *human-centered xAI* approach to identify *confounder exploitation biases* influencing decisions made by deep learning models. To accomplish this, we have selected a straightforward standard task: classifying images of cats and dogs. We intentionally introduced bias into the dataset by applying different rotation angles to cat and dog images. The objective of this study is to assess the extent to which the rotation confounder is utilized by the neural network.

We utilized the *Oxford-IIIT Pet Dataset* [165], which comprises images of cats and dogs from various species, each with different scales, poses, and lighting conditions. Our bias manipulation involved applying a random rotation angle between 0 and 90° to dog images and a random angle between 90° and 180° to cat images. This rotation transformation serves as the confounder under investigation.

Given the nature of this image processing task, we employed a CNN for classification, specifically the Resnet34 network [166]. Initially, we trained the Resnet model for classification without modifications. Upon reaching convergence, we extracted the 128-dimensional latent vector produced by the penultimate layer for the entire dataset, representing the latent space. The 2D version of this latent space, obtained through t-distributed stochastic neighbor embedding (t-SNE) transformation [167], is illustrated in Figure 7.5A.

Subsequently, we incorporated the decoder and discriminator modules, as detailed in Section 7.2, to create the eXplainable Adversarial Auto-Encoder network (xAAEnet) architecture necessary for our xAI approach. The decoder, symmetric to Resnet34 encoder, functions to accurately reconstruct the input image. Meanwhile, the discriminator consists of a 3-layer MLP designed to encourage the latent space to adhere to a Gaussian distribution $\mathcal{N}(0, 1)$, selected as the reference distribution for $Z_N$. After training xAAEnet following the curriculum learning process outlined in Section 7.2, we extracted the 2D t-SNE representation of the resulting latent space, as shown in Figure 7.5B.

Comparing the latent space of the Resnet model with that of xAAEnet in Figure 7.5, a noticeable difference in sparsity is evident. The sparse space obtained through training the Resnet network clearly fails to allow a smooth transition between cat and dog samples, as they are entirely separated. In contrast, xAAEnet enables us to make use of the most discriminative direction, defined as the axis of highest explained variance when performing a Linear Discriminant Analysis (LDA) [168], represented by an arrow in Figure 7.5B. Along this direction, we can analyze how the model transitions between images classified as dogs and images of cats. Along this direction, we can conduct sensitivity analyses to scrutinize the behavior of any desired feature during the transition, thereby evaluating its impact on the classification process, as shown in Figure 7.6.

For instance, as shown in Figure 7.5, to assess the influence of image rotation, we initially sampled images positioned at specific locations within each latent space, providing a qualitative illustration of how the images rotate along the direction of interest. The smooth transition observed in xAAEnet latent space provides clear insights into the influence of rotation angles, while the sparsity of Resnet's latent space hinders any meaningful conclusions.

To quantitatively assess the impact of rotation angles, we reorganized the samples along the most discriminative direction to characterize the transition from samples most considered as cats to those most considered as dogs. Next, we calculated the mean difference in angles between samples at the same distance within this sorted sequence, as illustrate in Figure 7.6C. By progressively increasing this distance, we obtained insights into the behavior of the rotation angle along the most discriminative direction. If the rotation angle consistently varies as we increase the distance between samples, it indicates that

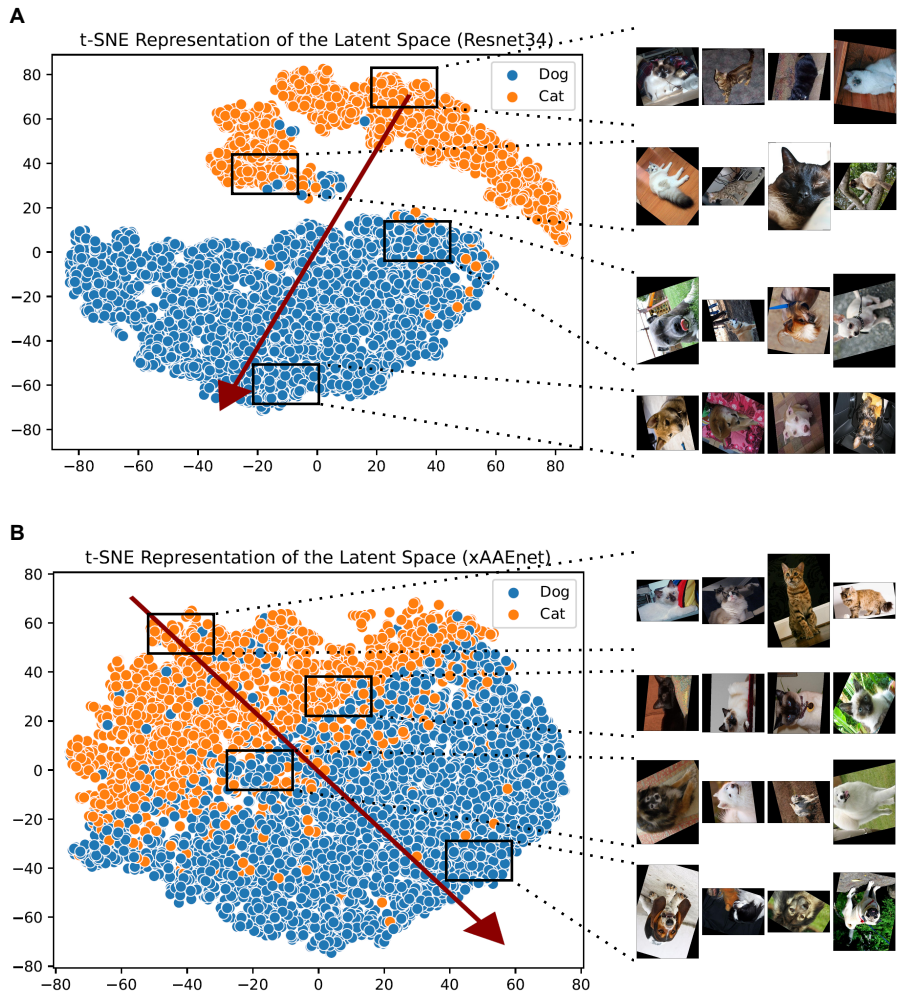



**Figure 7.5.** t-SNE visualizations of the latent spaces generated by two architectures: (A) Resnet34 and (B) xAAEnet. Each dot represents an image of either a cat (in orange) or a dog (in blue). The sparsity and separation between classes in the Resnet34 latent space contrast with the smoother transitions observed in the xAAEnet space along the most discriminant direction (red arrow). Specific image samples are provided for highlighted regions in both visualizations, indicating the influence of rotation angles on classification.

this feature significantly influences the classification process performed by the deep learning model.

When applying this technique to the Resnet latent space, the results in Figure 7.6A reveal that the angle difference remains non-significantly different from 0° until the distance becomes so large that it predominantly involves comparisons between dog and cat samples. Beyond this point, the angle differences approach 90°, which corresponds to the typical difference angle between a cat and a dog image, as imposed by the biasing effect. This lack of a discernible trend highlights the impracticality of comparing samples along the most discriminant direction in the latent space of conventional classifiers. In Figure 7.6B, we observe that the difference in rotation angles gradually increases with the distance between two samples. This behavior indicates that the model encodes the rotation angle as a discriminative feature. The insights gained from this sensitivity analysis lead us to the conclusion that the model does not effectively fulfill the intended classification task, as it leverages a confounding factor, namely the rotation angle in this case.

## 7.4 Discussion

The proposed approach of *human-centered xAI* involves conducting sensitivity analyses to assess how specific features behave when transitioning between different classes within the latent space. Our sensitivity analysis relies on key properties of the latent space, including its reconstruction ability acquired during the training phase of the AE, as well as its non-sparsity and generative capability obtained during the training phase of the GAN.

It's important to note that the introduction of these constraints into the latent space may reduce the model's classification performance. However, our proposed technique doesn't aim to achieve the highest classification accuracy. Instead, its primary goal is to gain insights into how confounding factors influence the encoding of input data during classification.

For clarity, the figures presented in Section 7.3 are displayed in 2D, but it's important to note that the sorting of input items is based on the entire latent space, which contains 128 dimensions in the provided use-case. As highlighted by Wattenberg and Viegas [169], a t-SNE plot, computed with a perplexity

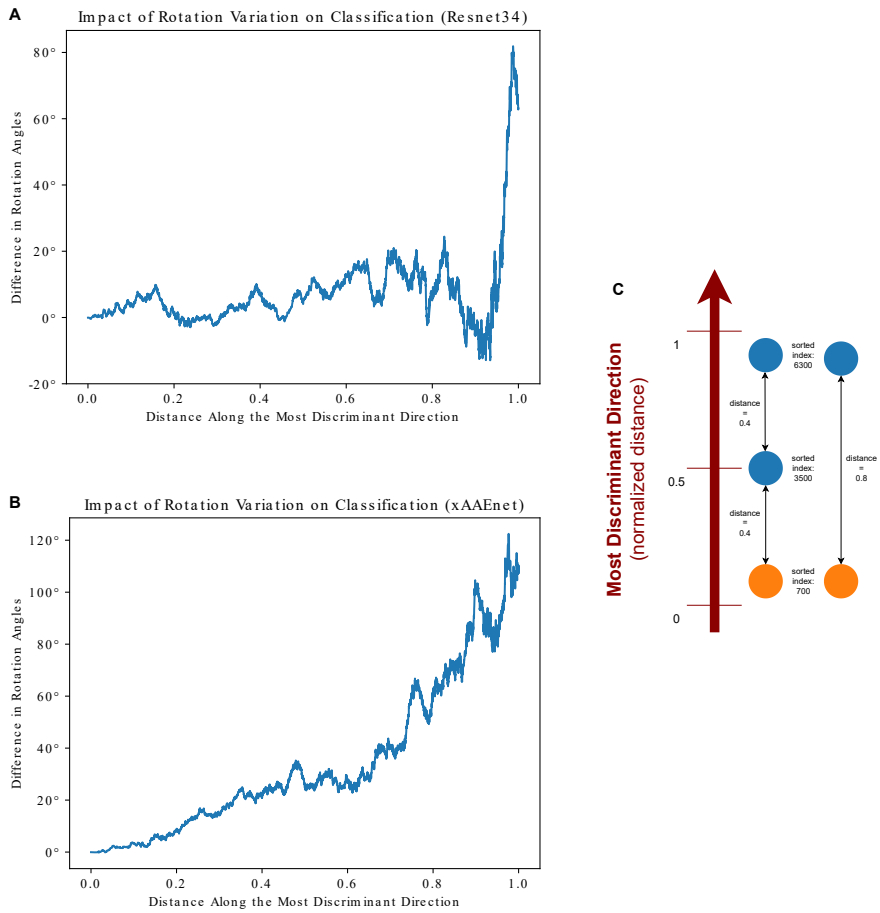



**Figure 7.6.** Sensitivity Analysis of Rotation Angles in Classification Models. (A) & (B) Display the difference in rotation angles as a function of the distance along the most discriminant direction of the Resnet34 (A) and xAAEnet (B) latent spaces. In (A), the Resnet latent space reveals non-significant variations in angles, underscoring the challenge of making meaningful comparisons in this latent space. In contrast, (B) demonstrates that within the xAAEnet model's latent space, a consistent rise in rotation angle differences indicates the angle's role as a discriminative feature. (C) Provides a visual depiction of samples sorted along the most discriminative axis, elucidating the method for determining normalized distances.

parameter falling within the range of 30-50, can provide valuable insights into the topography of high-dimensional data distribution. This aids in visualizing complex data structures effectively.

To gain a comprehensive understanding of the essential features that govern the model's decision-making process, the selection of these features becomes a pivotal step in obtaining a meaningful explanation of the model's decisions. In fact, the purpose of the proposed method is not to extract which features among all the possible ones influence the classification process, but rather to enable anyone interested in the study to directly assess the impact of an unexplored feature. Notably, unlike the approach proposed by Wang *et al.* [84], this feature selection can occur retrospectively. This flexibility enhances the accessibility and applicability of the method.

## 7.5 In Brief

Summary of Chapter 7

- This chapter introduces a pioneering approach to explainable AI that relies on a human-centered perspective. It involves comparing the input data while transitioning between different categories.
- Our approach hinges on specific properties within the latent space of deep learning models, including reconstruction ability, non-sparsity, and generative capacity. These properties are provided through a sequential training process that involves multiple neural networks sharing their encoding components: auto-encoder, generative adversarial network, and classifier. The resulting model is called eXplainable Adversarial Auto-Encoder network (xAAEnet).
- To exemplify our approach, we apply it to a straightforward use case involving cat-dog classification. This classification task is deliberately biased by the rotation of the images. Remarkably, xAAEnet autonomously identifies and rectifies this biasing confounder.

Perspective for Chapter 7

- Enhance the accuracy of the classification task per se.
- Delve into the intrinsic features of images from the original dataset, devoid of rotation-induced bias. This investigation seeks to uncover critical features for cat-dog classification, providing a more comprehensive understanding of the AI system's decision-making process.

# Chapter 8

# xAI for Obstructive Sleep Apnea Assessment

## Contents


**8.1 Obstructive Sleep Apnea Assessment . . . . . 144**
**8.2 Model Architecture . . . . . 145**
8.2.1 xVAEnet . . . . . 146
8.2.2 xAAEnet . . . . . 151
**8.3 Biomarkers Identification . . . . . 156**
**8.4 Obstructive Sleep Apnea Severity Scoring . . . . 163**
**8.5 In Brief . . . . . 175**


This chapter leverages the *human-centered xAI* approach, described in Chapter 7, within the context of Obstructive Sleep Apneas (OSAs) to enhance the assessment of their severity. The key contributions associated with this research are outlined in the following publications:

- "*Explainable AI for EEG Biomarkers Identification in Obstructive Sleep Apnea Severity Scoring Task*", In 11th International IEEE EMBS Conference on Neural Engineering (NER 2023), April 25–27, 2023, Baltimore, MD, USA. [170]
- "*Enhancing OSA Assessment with Explainable AI*", In 45th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2023), July 24-28, 2023, Sydney, Australia. [171]

In Section 8.1, we provide an overview of the sleep apnea assessment problem, a discussion of the specific use case for the proposed framework, and insights into addressing related issues. Section 8.2 delves into the architecture of the adapted model, which includes its initial version (xVAEnet) and its improved iteration (xAAEnet). Section 8.3 provides an explanation of the methodology used to identify EEG biomarkers associated with severe sleep apnea events. Lastly, Section 8.4 explores how the proposed approach contributes to enhancing the severity scoring process.

## 8.1 Obstructive Sleep Apnea Assessment

OSA is a common sleep disorder associated with multiple medical conditions, from excessive daytime sleepiness to cognitive or cardiovascular disorders [172]. The assessment of how OSAs affect patients' health, i.e., its severity, is currently based on the number of apnea and hypopnea events occurring overnight, measured by the AHI [173]. However, the use of AHI has faced significant criticism over the last decade, as it fails to estimate the impact of OSAs on related medical conditions [174]. This issue has prompted extensive research efforts to find better metrics for characterizing OSAs severity. These metrics include hypoxic burden, arousal intensity, duration of apneic events, odds ratio product, heart rate variability, and cardiopulmonary coupling [175–182]. Despite numerous studies aimed at discovering the most efficient metric for OSAs severity, no consensus has emerged within the sleep research community [183].

In response to this lack of consensus, we propose a novel approach that leverages xAI to pave the way for a common severity metric. Our aim is to address the subjective biases associated with the metrics proposed by clinicians. We demonstrate the relevance of using explainable DL models to identify the essential features for assessing OSAs severity. This demonstration involves identifying EEG biomarkers that our DL model considers highly important for a severity classification task. Importantly, this task is defined using PSG-derived features not directly related to EEG signals. Previous research has shown that OSA events trigger specific EEG power variations that differ between patients with severe OSA syndrome and those with a moderate form [184]. For example, Sforza *et al.* identified an increased $\theta/\alpha$ ratio in patients with excessive

daytime sleepiness [185], and Dingli *et al.* demonstrated the relationship between non-REM arousals and EEG power [186].

To the best of our knowledge, the principles of xAI have never been applied to assess OSAs severity. Most studies involving DL algorithms on sleep data focus on tasks such as automatically detecting sleep stages or apnea-hypopnea events from PSG signals [67, 187], distinguishing OSAs from other sleep conditions like insomnia [188], estimating specific underlying symptoms like excessive daytime sleepiness [189], or estimating AHI from arbitrarily chosen signals, such as the oxygen saturation signal [190]. In contrast, our work aims to reduce the subjectivity of diagnosis by examining how a DL algorithm makes its decisions.

## 8.2 Model Architecture

The application of the *human-centered xAI* approach to assess the severity of OSA involves training a classification model using various metrics associated with OSA severity. These metrics serve as proxies for the ground-truth severity score, which remains unavailable in OSA research. Achieving convergence across all these metrics using the same model establishes a robust latent representation. In this context, instead of focusing on bias detection, we employ the explainability technique to unveil the biomarkers that underlie the severity of OSAs. This process involves conducting a sensitivity analysis, similar to the method described in Section 7.3, within the derived latent space.

In this work, the latent space must still possess specific properties: 1) *Reconstruction Ability* ensuring a direct relationship between the feature space and all independent input channels, allowing clinicians to evaluate the relevance of the proposed classification; 2) *Non-Sparsity* ensuring a fair comparison between trials when conducting directional studies within the feature space; 3) *Generative Ability* ensuring that the model remains usable for new patients by avoiding characterization gaps in any relevant part of the latent space.

Throughout this thesis, we have explored two versions of the model:

1. *xVAEnet*: This model is based on the VAE++ architecture proposed by Zhang *et al.* [160]. xVAEnet aims to showcase the potential of xAI in identifying biomarkers.
2. *xAAEnet*: This model is built on the AAE architecture proposed by Makhzani *et al.* [161]. xAAEnet seeks to improve the accuracy of predicted scores while maximizing the interpretability of the decision-making process.

### 8.2.1 xVAEnet

As depicted in Figure 7.2, Zhang *et al.* proposed working with two latent spaces, $Z_I$ and $Z_S$. In the adapted model utilized in this work, we refer to these latent spaces as $Z_e$, which corresponds to the encoder latent space, and $Z_d$, which corresponds to the decoder latent space. Figure 8.1 illustrates the final architecture, named *eXplainable Variational Auto-Encoder network (xVAEnet)*, which comprises three sub-networks trained sequentially: a VAE, a GAN, and classifiers. The detailed architecture is provided in Table.

The initial VAE training phase imparts the reconstruction ability to the encoder latent space ($Z_e$) and instills non-sparsity and generative properties in the decoder latent space ($Z_d$). Subsequently, during the GAN training phase, these properties are transferred from $Z_d$ to $Z_e$. Finally, the classifier enforces the separation of samples from different conditions within the latent space of interest ($Z_e$).

For accessing the open-source code, please refer to our GitHub repository available at the following link: `https://github.com/numediart/xVAEnet.git`.

#### VAE

As described by Kingma and Welling [156], a VAE model is designed to learn a latent representation of the input data with a desired probability distribution and generative ability. In this paper, the VAE model used is inspired by the *Stagernet* model proposed by Banville *et al.* to analyze long EEG sequences of

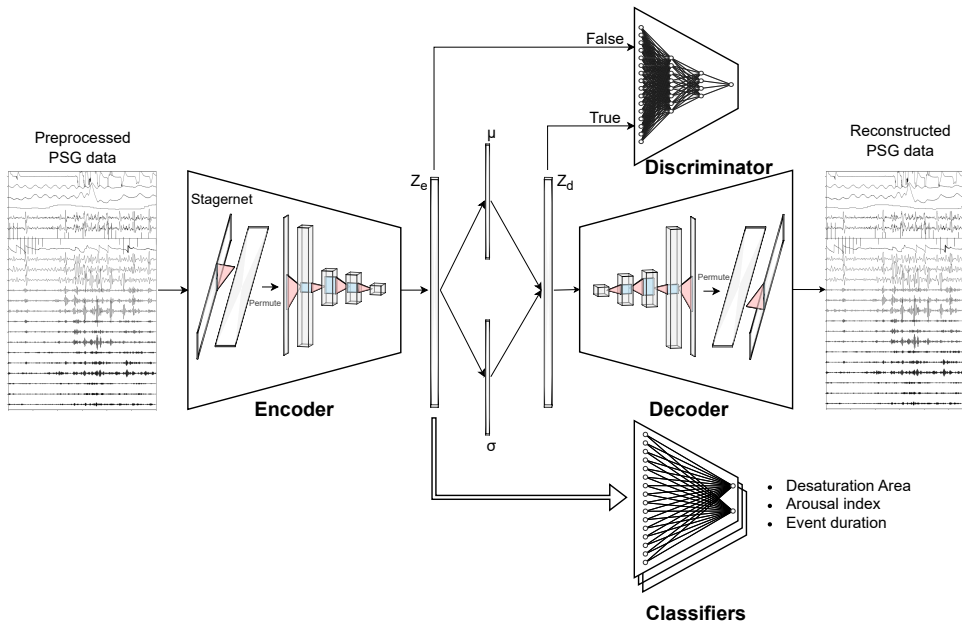



**Figure 8.1.** xVAEnet Architecture. The model is composed of 3 parts: a VAE, a GAN and a classifier, all of them making use of the convolutional encoder that encodes the input data into an embedding, the encoder latent space ($Z_e$). The **VAE** (center) first estimates the mean ($\mu$) and the standard deviation ($\sigma$) of the dataset distribution from $Z_e$ using dense layers to obtain the decoder latent space ($Z_d$), then $Z_d$ is decoded to derive a reconstructed version of the input data using a deconvolutional decoder. The **GAN** (top) exploits the encoder as a generator and discriminates $Z_d$ (real distribution) from $Z_e$ (fake distribution) using an MLP discriminator. The **Classifier** (bottom) uses the features extracted by the encoder in $Z_e$ to classify the desaturation area, the arousal events and the respiratory event duration with a unique single-layer perceptron.

| Block | Layer | # filters | kernel size | # params | Output | Activation | Options |
|---|---|---|---|---|---|---|---|
| Encoder | Input | | | | (23,3001) | | |
| | Reshape | | | | (1,23,3001) | | |
| | Conv2D | 23 | (23,1) | 23*23 | (23,1,3001) | | pad(0,0),stride(1,1) |
| | Permute | | | | (1,3001,23) | | |
| | Conv2D | 16 | (50,1) | 16*50 | (16,2952,23) | | pad(0,0),stride(1,1) |
| | MaxPool2D | | (13,1) | | (16,227,23) | | return_indices(in1) |
| | BatchNorm | | | 2*16 | (16,227,23) | | |
| | Activation | | | | (16,227,23) | ReLU | |
| | DepthwiseConv2D | 16 | (50,1) | 16*50 | (16,178,23) | | pad(0,0),stride(1,1) |
| | MaxPool2D | | (13,1) | | (16,13,23) | | return_indices(in2) |
| | BatchNorm | | | 2*16 | (16,13,23) | | |
| | Activation | | | | (16,13,23) | ReLU | |
| | Flatten | | | | (4784) | | |
| | Dropout | | | | (4784) | | p=0.5 |
| Latent | Dense | | | 128*4784 | (128) | ReLU | |
| | BatchNorm | | | 2*128 | (128) | | |
| | Dense (mu) | | | 128*128 | (128) | ReLU | |
| | BatchNorm | | | 2*128 | (128) | | |
| | Dense (sigma) | | | 128*128 | (128) | ReLU | |
| | BatchNorm | | | 2*128 | (128) | | |
| | Reparameterize | | | | (128) | | |
| Decoder | Dense | | | 128*4784 | (4784) | | |
| | Unflatten | | | | (16,13,23) | | |
| | MaxUnpool2D | | (13,1) | 2*16 | (16,178,23) | | indices=in2 |
| | BatchNorm | | | | (16,178,23) | | |
| | Activation | | | | (16,178,23) | ReLU | |
| | DepthConvTrans2D | 16 | (50,1) | 16*50 | (16,227,23) | | pad(0,0),stride(1,1) |
| | MaxUnpool2D | | (13,1) | 2*16 | (16,2952,23) | | indices=in1 |
| | BatchNorm | | | | (16,2952,23) | | |
| | Activation | | | | (16,2952,23) | ReLU | |
| | ConvTranspose2D | 16 | (50,1) | 16*50 | (1,3001,23) | | pad(0,0),stride(1,1) |
| | Permute | | | | (23,1,3001) | | |
| | ConvTranspose2D | 23 | (23,1) | 23*23 | (1,23,3001) | | pad(0,0),stride(1,1) |
| | Reshape | | | | (23,3001) | | |
| | Output | | | | (23,3001) | | |
| Discrim. | Dense | | | 128*32 | (32) | LeakyReLU | negSlope=0.2 |
| | BatchNorm | | | 2*128 | (32) | | |
| | Dense | | | 32*8 | (8) | LeakyReLU | negSlope=0.2 |
| | BatchNorm | | | 2*32 | (8) | | |
| | Dense | | | 8*1 | (1) | LeakyReLU | negSlope=0.2 |
| | BatchNorm | | | 2*1 | (1) | | |
| | Activation | | | | (1) | Sigmoid | |
| Classif. | Dense | | | 128*2 | (2) | | |
| | Activation | | | | (2) | Softmax | |

**Table 8.1.** xVAEnet architecture details

sleep recordings [191]. The encoder part of our VAE is therefore a replica of the Stagernet CNN adapted to the 23x3001 format of our input data. The decoder is the mirrored version of the encoder where the convolutions are replaced by transposed convolutions and the max pooling layers by max unpooling ones. Contrarily to classical VAE models, the mean ($\mu$) and standard deviation ($\sigma$) are not directly derived from the last convolutional layer of the encoder, but an intermediary latent vector is added to the encoder side ($Z_e$). In fact, as stated by Zhang *et al.*, the latent representations in VAEs are stochastically sampled from the prior distribution instead of being directly rendered from the input data [160]. This property compromises the further classification of the input data from their latent representation. The purpose of adding $Z_e$ is therefore to make available a deterministic latent representation of the input data for the classification task. The decoder latent space ($Z_d$) is obtained by the reparameterization trick classically used in VAEs, i.e. $Z_d = \mu + \sigma\epsilon$ where $\epsilon$ is a random variable with Gaussian distribution. The loss function used to train the VAE part of our model is a combination of reconstruction loss, using Huber loss, and Kullback-Leibler divergence, as described in Equation 8.1:

$$\mathcal{L}_{VAE} = 0.5 \cdot \mathcal{L}_{Huber}(output, input) + \\ 0.5 \cdot \frac{1}{bs} \sum_{i=1}^{bs} -0.5 \cdot \sum_{latent\_dim} (1 + log(\sigma) - \mu^2 - \sigma) \tag{8.1}$$

with $bs$ the batch size. As only $Z_d$ acquires the generative ability and follows a non-sparse Gaussian distribution during the VAE training phase, another process should transfer these properties to $Z_e$ that is the purpose of the GAN module.

### GAN

By considering $Z_d$ as the "real" latent representation and $Z_e$ as the "fake" one, the training of the GAN module forces the encoder to generate a latent vector $Z_e$ that mimics the properties of $Z_d$, as inspired by Zhang *et al.* [160]. Adversarial networks requires a generator that generates fake data as close as possible to real data and a discriminator that differentiates between real and fake data. In the proposed xVAEnet architecture, the generator is the encoder shared with the VAE part and the discriminator consists of a 3-layer

MLP each of them using leaky ReLU activation function with a negative slope of 0.2 and batch normalization. The output activation function is a sigmoid function. The loss function of the generator is a weighted sum of the VAE loss and the mean of correct predictions by the discriminator within a batch, as described in Equation 8.2.

$$\mathcal{L}_{gen} = (1-\alpha) \cdot \mathcal{L}_{VAE} + \alpha \cdot \frac{1}{bs} \sum_{i=1}^{bs} (1 - fake_i) \tag{8.2}$$

with $\alpha$ the weight of the adversarial loss, and *fake* the output of the discriminator when the "fake" latent representation ($Z_e$) is given as input, which should be equal to 1 for an ideal generator. The loss function of the discriminator is the difference between the mean fake predictions and the mean real predictions, as shown in Equation 8.3.

$$\mathcal{L}_{discrim} = \frac{1}{bs} \sum_{i=1}^{bs} fake_i - \frac{1}{bs} \sum_{i=1}^{bs} real_i \tag{8.3}$$

with *real* the output of the discriminator when the "real" latent representation ($Z_d$) is given as input. For an ideal discriminator, $real = 1$ and $fake = 0$. The combination of both loss functions is described in Section 8.3.

### Classifiers

The classifier modules compel the encoder to produce a latent vector where trials with a low severity level are maximally distant from trials with a high severity level. In this work, severity is characterized by a combination of metrics, including the desaturation area (i.e., the area under the curve of the $SAO_2$ signal [177]), arousal events (i.e., the presence or absence of an arousal occurring just after the OSA), and the duration of the respiratory event. For clarity and simplicity in this proof-of-concept research, we defined two severity levels: low and high. We determine these levels of severity by comparison to the median values across trials, with the exception of arousal events, which are binary by definition. This approach helps to maintain a balanced representation of both high and low values, ensuring that neither dominates the dataset.

This categorization results in four severity levels:

1. *Very Low*: low severity level on all 3 metrics.
2. *Low*: high level on 1 metric.
3. *High*: high level on 2 metrics.
4. *Very High*: high level on all 3 metrics.

A classifier is assigned to each metric, with each classifier block consisting of a single-layer perceptron that computes a linear combination of the latent vector values (with 128 dimensions). This process generates a probability score for each sample, indicating whether it belongs to the high or low severity category, using a softmax activation function. The simplicity of the classifier modules promotes the encoder's primary role in performing the classification task, with the classifiers themselves playing a secondary role.

During training, we follow a curriculum learning process by sequentially performing classification on each severity metric using a binary cross-entropy loss function: $\mathcal{L}_{classif} = BCE(predicted, target)$, as described in Section 8.3.

### 8.2.2 xAAEnet

The architecture of xAAEnet, as described in Section 7.2, has been adapted to consider PSG signals as input and to perform regressions instead of the classification performed in the initial model. These regressions aim to determine a severity score for each PSG trial.

The encoder and decoder modules are similar to the xVAEnet, as they draw inspiration from the Stagernet model proposed by Banville *et al.* [191]. The discriminator remains the same MLP as described for xVAEnet, with the difference that the "real" samples considered here are directly sampled from an independent Gaussian distribution instead of being part of another latent space. These modules allow the latent space to maintain the necessary properties, including reconstruction ability, non-sparsity, and generative ability.

Regarding the regressor modules, we continue to focus on the severity metrics used to train xVAEnet, as described in Section 8.2.1. Additionally, we have added another regression module aimed at predicting the value of a hand-

made severity score $S_h$ described in Section Regressors. The adapted xAAEnet architecture is illustrated in Figure 8.2 and detailed in Table 8.2.

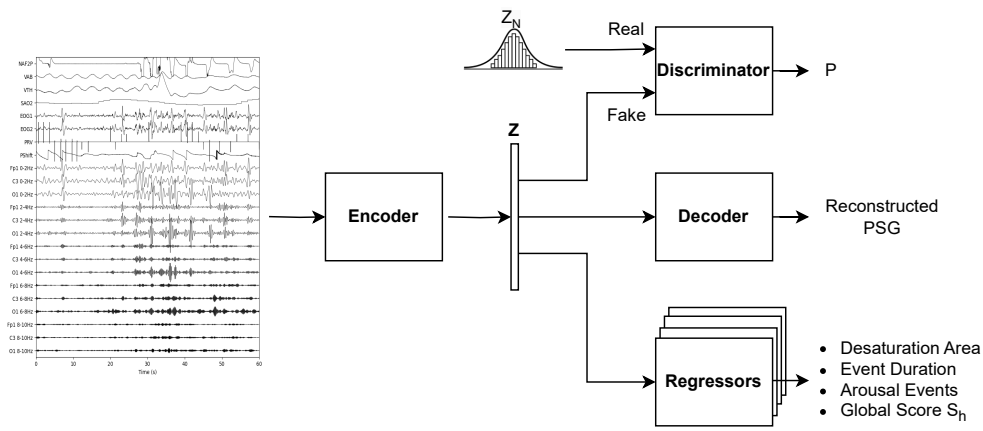



**Figure 8.2.** xAAEnet Architecture adapted to OSA severity scoring. The model is composed of 3 parts: an AE, a GAN and regressors. A shared convolutional encoder encodes input data into a latent representation $Z$. The **AE** (encoder + decoder) decodes $Z$ to derive a reconstructed version of the input data using a deconvolutional decoder. The **GAN** (encoder + discriminator) exploits the encoder as a generator and discriminates $Z$ (fake distribution) from a random Gaussian sampled batch $Z_{\mathcal{N}}$ (real distribution) using an MLP discriminator. The **regressors** (encoder + Single-Layer Perceptron (SLP)) use the features extracted by the encoder in $Z$ to predict the value of the hand-made score $S_h$, the desaturation area, the respiratory event duration, or to detect the presence of arousal events, with a SLP.

| Block | Layer | # filters | kernel size | # params | Output | Activation | Options |
|---|---|---|---|---|---|---|---|
| Encoder | Input | | | | (23,3001) | | |
| | Reshape | | | | (1,23,3001) | | |
| | Conv2D | 23 | (23,1) | 23*23 | (23,1,3001) | | pad(0,0),stride(1,1) |
| | Permute | | | | (1,3001,23) | | |
| | Conv2D | 16 | (50,1) | 16*50 | (16,2952,23) | | pad(0,0),stride(1,1) |
| | MaxPool2D | | (13,1) | | (16,227,23) | | return_indices(in1) |
| | BatchNorm | | | 2*16 | (16,227,23) | | |
| | Activation | | | | (16,227,23) | ReLU | |
| | DepthwiseConv2D | 16 | (50,1) | 16*50 | (16,178,23) | | pad(0,0),stride(1,1) |
| | MaxPool2D | | (13,1) | | (16,13,23) | | return_indices(in2) |
| | BatchNorm | | | 2*16 | (16,13,23) | | |
| | Activation | | | | (16,13,23) | ReLU | |
| | Flatten | | | | (4784) | | |
| | Dropout | | | | (4784) | | p=0.5 |
| Latent | Dense | | | 128*4784 | (128) | ReLU | |
| | BatchNorm | | | 2*128 | (128) | | |
| Decoder | Dense | | | 128*4784 | (4784) | | |
| | Unflatten | | | | (16,13,23) | | |
| | MaxUnpool2D | | (13,1) | 2*16 | (16,178,23) | | indices=in2 |
| | BatchNorm | | | | (16,178,23) | | |
| | Activation | | | | (16,178,23) | ReLU | |
| | DepthConvTrans2D | 16 | (50,1) | 16*50 | (16,227,23) | | pad(0,0),stride(1,1) |
| | MaxUnpool2D | | (13,1) | 2*16 | (16,2952,23) | | indices=in1 |
| | BatchNorm | | | | (16,2952,23) | | |
| | Activation | | | | (16,2952,23) | ReLU | |
| | ConvTranspose2D | 16 | (50,1) | 16*50 | (1,3001,23) | | pad(0,0),stride(1,1) |
| | Permute | | | | (23,1,3001) | | |
| | ConvTranspose2D | 23 | (23,1) | 23*23 | (1,23,3001) | | pad(0,0),stride(1,1) |
| | Reshape | | | | (23,3001) | | |
| | Output | | | | (23,3001) | | |
| Discrim. | Dense | | | 128*32 | (32) | LeakyReLU | negSlope=0.2 |
| | BatchNorm | | | 2*128 | (32) | | |
| | Dense | | | 32*8 | (8) | LeakyReLU | negSlope=0.2 |
| | BatchNorm | | | 2*32 | (8) | | |
| | Dense | | | 8*1 | (1) | LeakyReLU | negSlope=0.2 |
| | BatchNorm | | | 2*1 | (1) | | |
| | Activation | | | | (1) | Sigmoid | |
| Regress. | Dense | | | 128 | (1) | | |
| | Activation | | | | (1) | | |

**Table 8.2.** xAAEnet architecture details

As explained in Section 7.2, we follow a curriculum learning approach to train each module sequentially. However, since the regression task is more complex than the classification one, we have changed the order of the training phases by advancing the regression training phase to the second position:

1. AE training phase: $\mathcal{L}_{AE} = \mathcal{L}_{Huber}(target, pred)$
2. Regression training phase: $\mathcal{L}_{regression} = \alpha \cdot \mathcal{L}_S + (1 - \alpha) \cdot \mathcal{L}_{AE}$
3. GAN training phase: $\mathcal{L}_{gen} = \alpha \cdot \mathcal{L}_{adversarial} + \beta \cdot \mathcal{L}_{regression} + \gamma \cdot \mathcal{L}_{AE}$

Here, only $\mathcal{L}_{gen}$ considers the other loss functions during the GAN training phase, as the discriminator works independently to the rest of the network, $\mathcal{L}_{Huber}$ and $\mathcal{L}_{adversarial}$ are described in Section 7.2, $\mathcal{L}_S$ is described in Section Regressors, and $\alpha, \beta, \gamma$ are empirical weights that determine the contribution of each loss function.

While the general training process has already been described in Section 7.2, we now focus on the part of the model for which the training process has been modified: the regression phase. Comprehensive architecture details and the open-source code of xAAEnet can also be found in our GitHub repository: `https://github.com/numediart/xVAEnet.git`.

### Regressors

The regression modules are responsible for predicting the value of the severity metrics for each trial as well as a hand-made severity score $S_h$, which is computed as the mean of the severity metrics, as shown in Equation 8.4:

$$S_h = \frac{1}{3}(DA_{\text{norm}} + Duration_{\text{norm}} + Arousal_{\text{norm}}) \tag{8.4}$$

where *norm* stands for normalized values. To perform this normalization, the target values for Desaturation Area (DA) and duration are mapped to the range [-1,1]. The arousal event is binary by definition, but it is mapped to keep a similar distribution of scores as if we only consider the mean between the DA and the duration to avoid biasing $S_h$. This mapping is represented by Equation 8.5.

$$\text{Arousal\_norm} = \begin{cases} +\frac{std_{DA}+std_{Duration}}{2} & \text{if arousal detected} \\ -\frac{std_{DA}+std_{Duration}}{2} & \text{else} \end{cases} \tag{8.5}$$

The hand-made severity score $S_h$ was created to address the absence of a ground truth score. This $S_h$ is used to guide the encoder to arrange the latent space so that OSA trials of obviously high severity (high value for each attribute) are maximally distant from trials of obviously low severity. To achieve this, we have defined a custom loss function called ordinal loss ($\mathcal{L}_{ord}$) which specifically takes into account the ordinal relationships between the predicted scores of a batch of samples, rather than just their absolute values. This function compares the model's predictions to the target values using pairwise subtraction and sign comparison. The target differences serve as weights, and the mean of the weighted sign comparison represents the final loss value, as described in Equation 8.6:

$$\mathcal{L}_{ord}(target, pred) = \frac{1}{n}\sum_{i,j}[w_i \cdot (sign(pred_i - pred_j) \neq sign(target_i - target_j))] \tag{8.6}$$

where $n$ is the number of elements in the comparison matrix, $w_i$ is the weight of the i-th element, calculated as the absolute value of the target difference, $pred_i$ and $pred_j$ are elements from the predictions vector, and $target_i$ and $target_j$ are elements from the target values vector. Note that this equation is calculated only for the lower triangle of the comparison matrix, excluding the diagonal.

In this study, we first employed a curriculum learning approach to make independent predictions on the three severity attributes iteratively. To train on the DA and duration targets, we combined the Huber loss and ordinal loss, and for arousal detection training, we use binary cross-entropy (BCE) loss, as described in Equations 8.7, 8.8, and 8.9.

$$\mathcal{L}_{DA} = \alpha \cdot \mathcal{L}_{Huber}(\text{target}_{DA}, \text{pred}_{DA}) + \beta \cdot \mathcal{L}_{ord}(\text{target}_{DA}, \text{pred}_{DA}) \quad (8.7)$$

$$\mathcal{L}_{duration} = \alpha \cdot \mathcal{L}_{Huber}(\text{target}_{dur}, \text{pred}_{dur}) + \beta \cdot \mathcal{L}_{ord}(\text{target}_{dur}, \text{pred}_{dur}) \quad (8.8)$$

$$\mathcal{L}_{arousal} = BCE(\text{target}_{arousal}, \text{pred}_{arousal}) \quad (8.9)$$

Where $\alpha$ and $\beta$ are empirical weights weighted for each loss component.

Then, a global scoring loss ($\mathcal{L}_S$) is used to predict the final score as presented in Equation 8.10.

$$\begin{aligned} \mathcal{L}_S = & \; w_{ord} \cdot \mathcal{L}_{ord}(\text{target}, \text{pred}) \\ & + w_{Hub} \cdot \mathcal{L}_{Huber}(\text{target}, \text{pred}) \\ & + w_{DA} \cdot \mathcal{L}_{DA} + w_{dur} \cdot \mathcal{L}_{duration} + w_{ar} \cdot \mathcal{L}_{arousal} \end{aligned} \quad (8.10)$$

where *target* is the ground truth $S_h$, *pred* is the predicted $S_h$, and $w_{xx}$ are empirical weights.

## 8.3 Biomarkers Identification

In this section, we leverage the *human-centeredxAI* approach to identify EEG biomarkers associated with severe OSA events. The process of identifying these biomarkers involves several components: the training of the xVAEnet model, the explainability methodology, the results acquired, and a subsequent discussion.

### Training

The training process of xVAEnet consists in a semi-supervised curriculum learning framework. In fact, every block of the architecture described in Section 8.2.1 is trained separately, and the initialisation of the following block's training process is done using the updated weights obtained at the end of the previous stage. The VAE and the GAN blocks are trained with non-supervised learning, while the classifier is trained in a supervised manner, making the

whole model training semi-supervised. The training parameters are detailed in the provided GitHub repository.

The VAE module has been trained using a random initialization until convergence. Then, the GAN module has been trained by initializing the generator with the best weights of the encoder obtained during the VAE training phase and the discriminator has been randomly initialized. At each batch, the discriminator is first trained by freezing the generator and using the loss function of the discriminator described in Section 8.2.1, then the generator is trained by freezing the discriminator and using the corresponding loss function. Every 15 epochs, the updated network is used in inference to compute a new $Z_d$ vector given as real input for the 15 following epochs in order to avoid the deterioration of the "real" space to be responsible for the increase of the GAN performance.

Finally, for the classification phase, the encoder is initialized with the weights of the best generator previously obtained and the single-layer perceptrons are randomly initialized. In the philosophy of curriculum learning, the classifier is trained on each severity metric sequentially using a BCE loss function, starting with the low vs. high severity classification on the DA, then on the arousal events and finally on the event duration. The learning rates were $10^{-3}$, $5 \cdot 10^{-4}$, and $2 \cdot 10^{-4}$ for the first, second, and third stages, respectively. For each classification stage, a global loss, calculated every 5 epochs, combines the VAE, GAN, and classifier losses, as described in Equation 8.11.

$$\mathcal{L}_{global} = \frac{1}{3}\mathcal{L}_{VAE} + \frac{1}{3}\frac{1}{bs}\sum_{i=1}^{bs}(1 - fake_i) + \frac{1}{3}\mathcal{L}_{classif} \tag{8.11}$$

On the first stage, DA was classified without considering the other severity metrics. On the second stage, the classification has been performed on both the DA and the arousal events using Equation 8.12.

$$\mathcal{L}_{\text{classif}_2} = \begin{cases} \mathcal{L}_{\text{arousal}}, & \text{if epoch\_number } \%2 = 0 \\ \frac{1}{2} \cdot \mathcal{L}_{DA} + \frac{1}{2} \cdot \mathcal{L}_{arousal}, & \text{otherwise} \end{cases} \tag{8.12}$$

where the Modulo operation "%" allows the loss function to change at each epoch.

On the third stage, the classification has been performed on all the severity metrics:

$$\mathcal{L}_{\text{classif}_3} = \begin{cases} \mathcal{L}_{\text{duration}}, & \text{if epoch\_number } \%3 = 0 \\ \frac{1}{3} \cdot \mathcal{L}_{DA} + \frac{1}{3} \cdot \mathcal{L}_{arousal} + \frac{1}{3} \cdot \mathcal{L}_{duration}, & \text{otherwise} \end{cases} \tag{8.13}$$

## Explainability

Using the *human-centered xAI* described in Chapter 7, we can navigate through the latent space $Z_e$ to characterize the evolution of each signal when transitioning from one level of severity to another. The most discriminant direction for the severity levels is found by performing a LDA on $Z_e$ that maximizes the discrimination between the 4 classes of severity. The result of this process is a vector giving the direction of the severity encoding, namely the severity direction. By comparing input samples along the severity direction, we can highlight the channels and the time windows that are the most affected by the OSAs severity. By analyzing non-EEG channels, we can validate that the model actually looks at the important features for severity scoring. By analyzing EEG channels, we can identify the best biomarkers of OSA in the EEG signals.

## Results

The results of this research, essentially qualitative, can be divided in two parts: 1) the severity scoring efficiency and 2) the EEG biomarkers identification. Some secondary quantitative results are detailed in the provided GitHub repository.

In Figure 8.3, we can observe the evolution of the latent space distribution across the different training phases, allowing the qualitative evaluation of how $Z_e$ acquires the required properties. For illustration purpose, the 128-

dimensional latent space has been projected to a 2D space using the t-SNE transform.

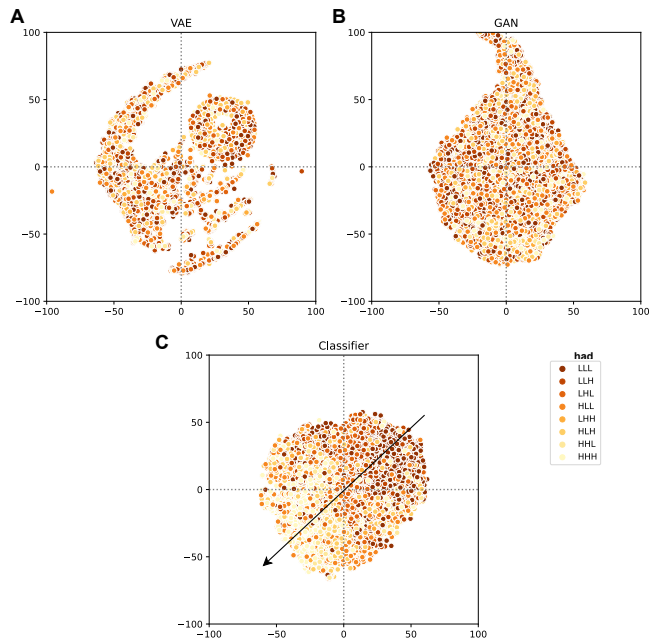



**Figure 8.3.** 2D representations of the encoder latent space ($Z_e$) using t-SNE. Each sample represents one of the 6992 OSA trials. (**A**) $Z_e$ with the training phase of the VAE module completed. (**B**) $Z_e$ with the training phase of the GAN module completed. (**C**) $Z_e$ with the training phase of the classifier module completed on every severity metrics. The arrow represents the severity direction obtained using LDA. In the legend, each letter of "**had**" represents a severity feature: "**h**ypoxic burden", "**a**rousal event", and "**d**uration of the respiratory event". The "L" means "Low-level severity", the "H" means "High-level severity"

As shown in Figure 8.3A, the training process of the VAE module leads to a sparse encoder latent space ($Z_e$). The training process of the GAN module leads to a non-sparse $Z_e$ getting closer to a Gaussian distribution, but with the samples of different severity scores randomly distributed (Figure 8.3B). The training process of the classifier module leads to a non-sparse, quasi-Gaussian $Z_e$ where the samples of the same severity levels tend to be gathered together

and separated from samples of different severity levels, as illustrated in Figure 8.3C.

From this well-designed latent space, we have performed an LDA aiming at classifying the 4 severity levels. With the classifier module trained, this LDA reaches a mean accuracy of 54.0% (trainset) and 48.8% (testset), and a mean F1-score of 56.5% (trainset) and 48.4% (testset). The direction of highest severity score variance, namely the *severity direction*, is represented with an arrow in Figure 8.3C and is responsible for 78.14% of the explained variance. This ability to estimate the severity score from a trial representation in $Z_e$ is the first proof of the relevance of the proposed framework in severity scoring task.

By navigating along the severity direction, we can sort the OSA trials by severity score to generate a severity scale and compare the trials depending on their position on this scale. Figure 8.4A provides a summary of the influence of the severity score on each PSG channels based on their power signal. This comparison is performed by computing the mean power difference of each channel separately as described in Equation 8.17:

$$Pdiff_{c[i]\ dist[j]} = \frac{1}{N-j}\sum_{k=0}^{N-j} P_{c[i]\ t[k+j]} - P_{c[i]\ t[k]}$$
$$Pdiff_{c[i]} = \frac{1}{N}\sum_{j=0}^{N} Pdiff_{c[i]\ dist[j]} \tag{8.14}$$

with trials being sorted based on their severity score, $N$ the number of trials, $t$ the trial number, $c$ the channel and *dist* being the distance on the severity scale.

The second operation allowing the evaluation of the severity scoring efficiency consists in identifying PSG channels and time windows that vary the most with the severity score. The non-EEG PSG are used to evaluate the consistency between clinical studies and the proposed framework, while the EEG channels allow the biomarkers discovery.

In Figures 8.4A and B, the high positive power difference on the $SAO_2$ signal suggests deeper and/or longer desaturations of severe OSA trials, as stated by Kulkas *et al.* [177]. The high negative power difference on the EOG signal

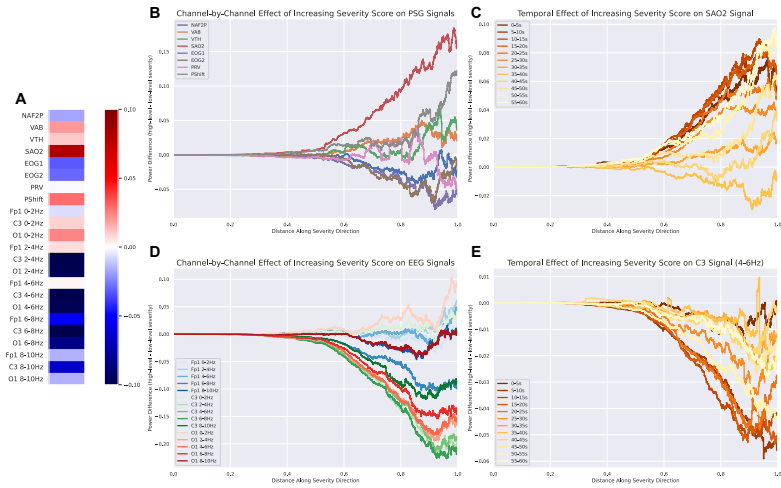



**Figure 8.4.** Biomarkers identification performed by comparing the power signal, by channel, of the OSA trials sorted by severity score ($\in$ [0,1]) along the severity direction obtained using LDA. (**A**) Mean power difference across OSA trials obtained by subtracting, for each channel separately, the power signal of each trial from the power signal of trials of higher severity scores. (**B**) Channel-by-channel mean power difference of PSG channels excluding EEG channels. The x axis represents the distance, along the severity direction, between the trials being compared. A distance of 0 means a trial is compared to itself, a distance of 1 means the comparison between the trial of lowest severity score and the trial of highest severity score. (**C**) Time window-by-time window mean power difference of the SAO2 channel (channel of highest absolute mean power difference). (**D**) Channel-by-channel mean power difference of EEG channels. (**E**) Time window-by-time window mean power difference of the C3 channel on the 4-6Hz frequency band (channel of highest absolute mean power difference).

is consistent with the works of Eiseman *et al.* who showed the dependence of apnea severity on REM vs. non-REM sleep stage (that highly affects the eye movement) [192]. Furthermore, Figure 8.4C shows that the SAO2 effect mainly appears during the respiratory events (beginning of the trial) with a spurious peak effect around 50 seconds after the start of the event (note that OSA event starts after 4s as described in Section 3.2). The aforementioned results provide the desired second proof that the proposed framework actually extracts severity information.

The EEG biomarkers identification task is based on the information provided by Figures 8.4D and E where we can observe that the central electrode (C3) is the most affected by the severity of the respiratory event in the 2-8Hz frequency range, this effect being maximal in the 5-25s trial time window (corresponding to the mean respiratory event duration). The occipital electrode (O1) also varies with the severity score in the 2-8Hz frequency range, but the frontal one (FP1) does not seem to be influenced by the OSA severity. The findings indicate a decrease in EEG power in parieto-occipital regions as the severity score increases. Further investigation, utilizing high-density EEG studies, may support the interpretation of this decrease as a reduction in brain activity during severe OSA events.

## Discussion

This research serves as a proof-of-concept, demonstrating the potential of the *human-centered xAI* approach in identifying EEG biomarkers associated with specific tasks. Specifically, our study focuses on the task of severity scoring for OSAs using PSG signals. Our proposed framework adopts a human-centered explainability approach, leveraging sample comparisons to enhance result interpretability.

The xVAEnet model comprises three modules (VAE, GAN, and classifier) trained sequentially via a semi-supervised curriculum learning process. The objective is to encode input data into a latent feature space that maximizes the discriminability between samples across varying OSA severity levels. This framework achieves several key outcomes: 1) It retains the majority of information from the input signals; 2) It is readily applicable to new patients; 3) It

facilitates equitable comparisons across OSA trials through a directional study approach.

Within this encoded space, the "severity direction" accounts for 78% of the variance in severity scores, affirming the model's ability to construct an appropriate distribution for the given task. Notably, the EEG features identified as OSA severity biomarkers pertain to central and occipital electrodes in the 2-8Hz frequency range. This finding aligns with previous research by Jones *et al.*, who demonstrated decreased EEG power in the parietal region, particularly in slow-wave activity (1-4.5Hz) and the $\theta$ band (4.5-8Hz) [193].

In this proof-of-concept study, our emphasis lies on qualitative rather than quantitative results. The primary objective is to showcase the relevance of our proposed xAI approach for biomarker discovery, rather than striving for the optimal model performance within this specific task.

Future endeavors may delve into more comprehensive EEG biomarker investigations. This could involve: 1) Incorporating time-frequency EEG representations, such as Fourier and Wavelet transforms; 2) Expanding the analysis to include all available electrodes (see Section 3.2); 3) Introducing additional severity metrics, such as the Apnea-Hypopnea Index (AHI). Furthermore, enhancing scoring efficiency could be explored through the examination of alternative encoder architectures and the fine-tuning of hyperparameters like latent space dimension, batch size, dropout rates, and weight decay, among others.

## 8.4 Obstructive Sleep Apnea Severity Scoring

In this section, we aim to enhance the performance of our proposed *human-centered xAI* approach for estimating the severity score of OSA events. We achieve this by utilizing the second version of the model we have developed, referred to as xAAEnet, and adopting a score regression approach instead of classification. A key feature of this version, compared to xVAEnet, is the disentangling of the reconstruction process from the adversarial one. We will begin by detailing the training process, followed by a description on the use of the explainability approach. Subsequently, we will present the results. Extending our analysis to a multimodal procedure, we will incorporate patient

information into the pipeline. Finally, we will engage in a comprehensive discussion.

### Training

The training process of xAAEnet still adheres to a semi-supervised curriculum learning framework. As described in Section 8.2.2, each block undergoes individual training, and the weights acquired at the end of each stage serve as the initialization for the subsequent block's training.

The model's training process combines elements of supervised and unsupervised learning. The AE and GAN components undergo unsupervised learning, while the regression module undergoes supervised learning, resulting in a semi-supervised training approach for the entire model. You can find more details on the training parameters in the accompanying GitHub repository.

The AE module was initially trained with random initialization until convergence. Subsequently, the weights from the best-performing AE model were used to initialize the encoder for the regression phase, whereas the regressors were initialized randomly. The curriculum learning approach was employed for training the regression module, beginning with the prediction of DA values, followed by event duration and, finally, arousal events. The learning rates for the first, second, and third stages were set at $10^{-3}$, $5{\cdot}10^{-4}$, and $2{\cdot}10^{-4}$, respectively. The final training phase employed the global regression loss $\mathcal{L}_{regression}$, as explained in Section 8.2.2, to optimize the regression for all severity metrics while maintaining the model's reconstruction capability.

The GAN module was trained by initializing the generator with the best encoder weights obtained during the regression unit's training phase, while the discriminator was initialized randomly. During each epoch iteration, the discriminator was updated first, keeping the generator's weights fixed and utilizing the $\mathcal{L}_{discrim}$ loss function defined in Section 8.2.2. Subsequently, the generator was updated with the discriminator's weights held constant, using the corresponding $\mathcal{L}_{gen}$ loss function.

For the sake of comparison, we designed several simpler models based on components of xAAEnet and trained them using a similar process. These models include: 1) *xVAEnet*, the initial version of the model; 2) *xAEnet*,

which omits the GAN component from xAAEnet; and 3) *xClassifnet*, which employs only the encoder component of xAAEnet for direct scoring.

## Explainability

In order to comprehend the decision-making process of our xAAEnet model, we adopt our human-centered approach within the latent space, consisting of latent vectors represented as $Z$. These vectors encode the input data into a 128-dimensional space. The non-sparse and generative characteristics of this space allow for navigation, providing insights into the factors influencing the severity score.

To identify the primary encoding direction for severity, we perform linear regression on $Z$, minimizing the error with the manually crafted severity score $S_h$. This process yields a vector representing the direction of severity encoding, denoted as the "severity direction." Using this direction, we derive a latent severity scale, ranging from 0 (indicating the least severe trial) to 1 (indicating the most severe one). This scale enables us to compute an enhanced severity score, denoted as $S_E$.

By comparing input samples along the severity direction, we can pinpoint the channels and time windows most influenced by OSA severity, as newly defined. Furthermore, we assess the capacity of xAAEnet to provide an objective OSA severity score by comparing channels that exhibit continuous power increases or decreases along the severity direction with commonly accepted OSA severity biomarkers.

## Results

The results of this study can be categorized into two main aspects: 1) model performance comparison, and 2) evaluation of the effectiveness of the proposed method in providing an objective OSA severity scoring.

Figure 8.5A illustrates the comparison of the distributions of latent representations generated by each model. We achieve this through 2D versions of the original 128-dimensional latent vectors obtained using the t-SNE transformation.

Figure 8.5B shifts our focus to the ordering capabilities of each model. It quantifies this by computing the mean $S_h$ score of neighboring trials on the latent severity scale. You can see this scale represented by an arrow in Figure 8.5A.

It becomes evident that xAAEnet produces a distribution that is closest to a Gaussian distribution and is the least sparse among the models. It also demonstrates good capabilities in ordering the latent OSA trials based on the manually-assigned severity score $S_h$. On the other hand, both xVAEnet and xClassifnet result in sparse distributions, with xVAEnet having a slight advantage in terms of trial ordering. Finally, xAEnet generates a distribution that is closer to a Gaussian, although it still contains some spurious samples, and performs adequately in ordering the trials.

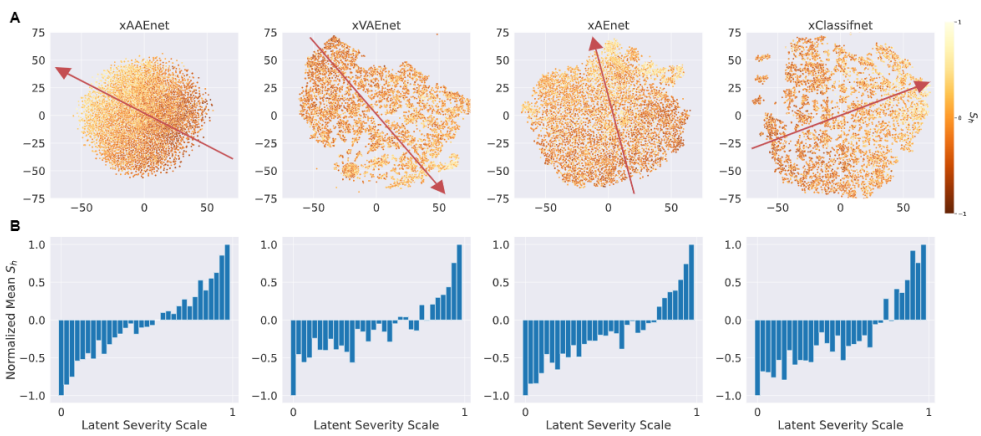



**Figure 8.5.** Comparison of the models' performance. **(A)** 2D representation, using t-SNE, of the latent space $Z$ obtained with the different networks. Each point represents an OSA trial and the color indicates the corresponding $S_h$ score. The arrow gives the severity direction. **(B)** Hand-made score distribution along the latent severity scale. Each bar represents the mean $S_h$ of the OSA trials in a specific latent severity scale range. The mean $S_h$ values have been normalized in [-1,1] for comparison purpose.

As we aim to quantify the model's ability to sort trials by severity, we employ Kendall's rank correlation coefficient to compare the $S_E$ scores (determined by each trial's position along the severity axis) with the $S_h$ scores. Kendall's coefficient assesses the similarity in the ordering of data points between two variables. The Kendall Tau distance is computed as a proportion of pairs

of data points in the good order (using concordant and discordant pairs), as described in Equation 8.15. A perfect ordering would lead to $\tau = 1$ and a random one would lead to $\tau = 0$.

$$\tau = \frac{\text{number of concordant pairs} - \text{number of discordant pairs}}{\text{total number of pairs}} \tag{8.15}$$

To focus on wrongly sorted trials with an emphasis on those significantly distant from the desired position, we have developed a custom metric, *ordinal error*, defined in Equation 8.16.

$$Ord_{Err} = \frac{1}{N-1} \sum_{i=1}^{N-1} \left(-\min\left(0, \text{mean}(S_{h,i+1:N}) - S_{h,i}\right)\right) \tag{8.16}$$

In this equation, $S_{h,i}$ represents the hand-made severity score for the $i$-th trial, and $N$ is the total number of trials. This metric evaluates the model's ability to order trials in the latent space according to the hand-made severity score. It calculates the negative difference between the mean of the hand-made scores $S_h$ of the remaining trials and the score of the current trial, with a minimum of 0 to ensure that only negative differences, indicative of an incorrect order, contribute to the metric. Finally, this error is divided by the ordinal error computed on randomly sorted trials (averaged over 30 repetitions) to obtain a final error ranging between 0 (perfect order) and 1 (random order).

Table 8.3 presents quantitative results for different models using the Leave One Subject Out (LOSO) method. This specific cross-validation method minimizes dependency between the training and test sets by preventing bias due to the intra-subject correlation that exists among trials from the same subject. It evaluates the model's ability to generalize its learning to previously unseen subjects, as discussed by Varoquaux *et al.* [194]. However, applying this method to all subjects can become computationally expensive. Therefore, we have chosen to apply it to specific subjects only.

We conducted three training sessions, each involving the transfer of three patients from the validation set to a test set not used during the training

process. This combination of multiple subjects is essential because a single patient does not experience a sufficient number of OSA events to provide statistical significance for this analysis. The selected test patients for each session were as follows: 1) patients with the lowest mean $S_h$ score across trials, 2) patients with the highest mean $S_h$ score, and 3) patients with mean $S_h$ scores closest to the global median.

| **Model** | **xAAEnet** | **xVAEnet** | **xAEnet** | **xClassifnet** |
|---|---|---|---|---|
| **Kendall Tau** | 0.193 ± 2.8 e-2 | 0.148 ± 3.0 e-2 | 0.246 ± 2.4 e-2 | 0.234 ± 1.7 e-2 |
| **Ordinal Error** | 1.5 e-3 ± 3 e-4 | 2.2 e-3 ± 2 e-4 | 2.1 e-3 ± 3 e-4 | 2.9 e-3 ± 4 e-4 |

**Table 8.3.** Comparison of Kendall Tau Distance and Ordinal Error for Different Models Using the LOSO Method. Three training sessions were conducted, each with specific test patients: 1) patients with the lowest mean $S_h$ score across trials, 2) patients with the highest mean $S_h$ score, 3) patients with mean $S_h$ scores closest to the global median. Kendall Tau distances and ordinal error values are reported with standard deviations.

The Kendall Tau distances are significantly greater than zero, and the ordinal error values are markedly lower than 1 for all the models. This indicates a substantial deviation from random ranking. When examining Kendall's coefficient, xAEnet and xClassifnet tend to exhibit slightly better performance. Conversely, if we consider our custom ordinal error, xAAEnet and xAEnet seem to perform better. However, drawing specific conclusions can be challenging, as we rely on the hand-made score $S_h$ as our ground truth. A further validation campaign should be led in collaboration with sleep clinics to evaluate the performance in sorting trials based on their severity.

We also aim to demonstrate the superior sensitivity of the $S_E$ scores obtained with xAAEnet over a simpler approach that directly sorts OSA trials based on their $S_h$ value and assigns scores based on their relative positions. Figure 8.6 illustrates the sensitivity of the proposed method by comparing the impact of increasing severity scores, whether defined as $S_h$ or $S_E$, on the PSG signals.

This comparison is conducted on the power signals and is carried out in two ways: by channel and by time window. The objective is to identify the PSG signals most affected by severe OSA events and verify whether the signal changes are indeed related to respiratory events. The mean power difference of each channel is calculated separately, as described by Equation 8.17.

$$Pdiff_{c[i]\ dist[j]} = \frac{1}{N-j}\sum_{k=0}^{N-j} P_{c[i]\ t[k+j]} - P_{c[i]\ t[k]}$$
$$Pdiff_{c[i]} = \frac{1}{N}\sum_{j=0}^{N} Pdiff_{c[i]\ dist[j]} \tag{8.17}$$

In this equation, trials are sorted based on their severity score, where $N$ represents the number of trials, $t$ represents the trial number, $c$ represents the channel, and *dist* represents the distance on the severity scale.

Figures 8.6A and B illustrate that the $S_h$-related method exhibits lower sensitivity compared to the proposed $S_E$ method. The $S_E$ method demonstrates greater sensitivity to channels affected by severe OSAs events, such as the NAF2P, the PRV, and the $SAO_2$ signals.

The $SAO_2$ signal exhibits the most significant difference between the two methods. The $S_E$ method takes into account deeper and/or longer desaturations, which are often associated with severe OSAs events, as documented by Kulkas *et al.* [177]. This can be observed in Figures 8.6C and D, where $S_E$ consistently detects this effect throughout the trial, while $S_h$ nearly misses it. It's important to note that the OSA event actually begins 4 seconds after the trial's start.

A cursory examination of Figures 8.6E and F may suggest that the $S_h$ method exhibits greater sensitivity on EEG channels. However, as reported by Jones *et al.*, the most significant impact of severe OSA events on EEG power occurs in the parietal region, particularly in the $\theta$ band (4.5-8Hz) [193]. By focusing on these regions of interest in the spatio-frequency domain, we once again observe that the $S_E$ method demonstrates higher sensitivity to severity-related features than the $S_h$ method, as shown in Figures 8.6G and H.

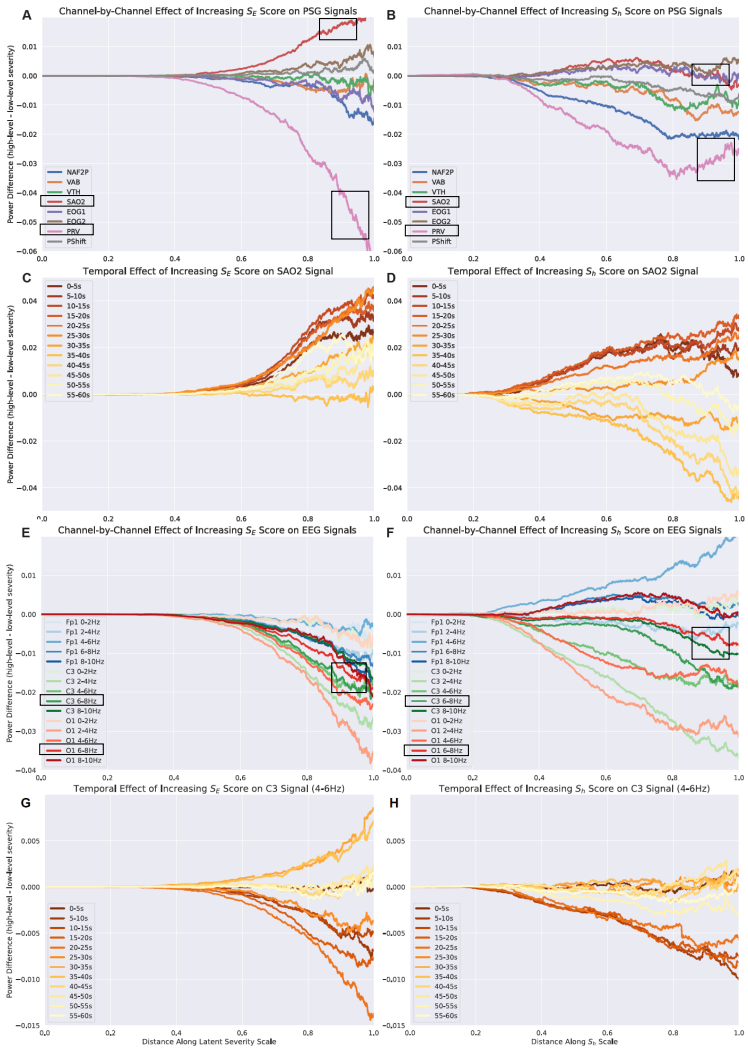



**Figure 8.6.** Severity score sensitivity comparison. The effect of an increasing severity score on input signals power is compared for severity defined as $S_E$ (**left**) vs. $S_h$ (**right**). For readability, this comparison is done separately for non-EEG PSG channels (A to D) and EEG channels (E to H). The x axis represents the distance, along the corresponding severity scale, between the trials being compared. A distance of 0 means a trial is compared to itself, a distance of 1 means the comparison between the trial of lowest severity score and the trial of highest severity score. The y axis represents the mean power difference of the PSG signals across OSA trials obtained by subtracting, for each channel/time-window separately, the power signal of each trial from the power signal of trials of higher severity scores. The figure includes the channel-by-channel mean power difference of (**A,B**) non-EEG PSG channels and (**E,F**) filtered EEG channels, as well as the time window-by-time window mean power difference of (**C,D**) the SAO2 channel and (**G,H**) the C3 channel on the 4-6Hz frequency band. The black boxes highlight the relevant signals for the comparison between $S_h$ and $S_E$.

We can also explore the direction perpendicular to the most discriminant one. This approach helps us identify features crucial for distinguishing different types of PSG trials, yet unrelated to severity scoring. These features hold significant value for signal reconstruction through the decoder. Figure 8.7 illustrates this analysis.

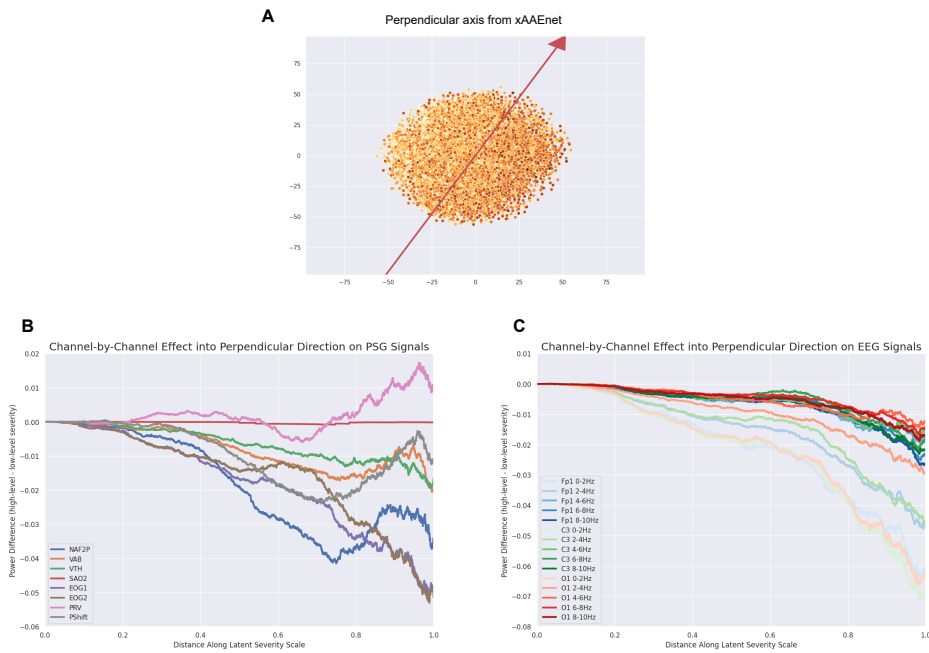



**Figure 8.7.** Sensitivity Analysis along the Perpendicular Direction. (A) Depicts a 2D representation of xAAEnet latent space using the t-SNE transform, with the perpendicular direction marked by a red arrow. (B) Shows the impact of increasing distance along this direction on PSG signal power. (C) Demonstrates the effect on EEG signal power with increasing distance along this direction.

In Figure 8.7B, we can observe a consistent decrease in power amplitudes for EOG signals. This suggests that eye movement is not a relevant feature for xAAEnet in performing the scoring task, although it plays a crucial role in discriminating between different types of PSG trials. This observation likely stems from the ability of eye movements to differentiate between sleep stages, particularly REM and NREM stages. In contrast, the curves related to $SAO_2$ and PRV remain relatively flat, indicating their high relevance to the scoring task.

Figure 8.7B emphasizes the importance of the frequency of EEG signals in the scoring task, as sub-4Hz activity appears to be irrelevant. However, further analysis could delve into these frequencies, as they seem pertinent to the overall discrimination between PSG trials.

## Patient Information

This section explains how to transform xAAEnet into a multimodal architecture in order to include patient information into the analysis pipeline. This information includes demographic data and medical conditions.

The modification is performed within the encoder by adding an embedding layer that encodes the patient information into a embedded latent vector $Z_{embed}$. Then the PSG latent representation, now denoted $Z_{PSG}$, is concatenated with $Z_{embed}$ and processed through a residual fully-connected layer, producing the final shared latent representation $Z$. This latent vector effectively incorporates both patient information and processed PSG data, with a higher emphasis placed on the temporal information derived from the PSG signals, thanks to the residual operation. The architecture of the multimodal xAAEnet is illustrated in Figure 8.8.

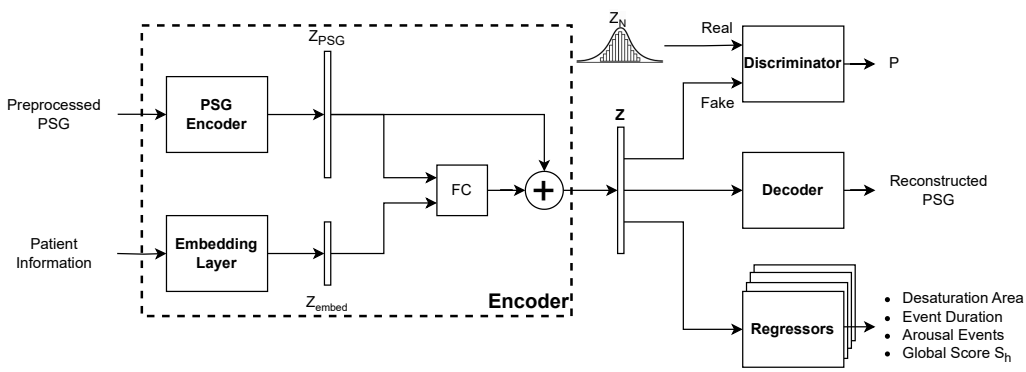



**Figure 8.8.** Multimodal xAAEnet Architecture encompassing patient information. The modified block is the encoder, which now consists of a PSG encoder and embedding layer that output independent latent vectors ($Z_{PSG}$ and $Z_{embed}$). These vectors are then concatenated and process through a residual fully-connected layer, leading to the final $Z$ latent representation that is processed the same way as the unimodal version of the model, described in Section 8.2.2.

Using the multimodal architecture, another analysis can be performed to identify patient information that may be associated with more severe respiratory

events. To achieve this, the severity scale was divided into four quartiles, and the values of the selected information (age, Body Mass Index (BMI), gender, presence of diabetes, presence of hypertension) were compared across the quartiles. The results presented in Table 8.4 indicate that there is no significant difference in these features across the severity scale. However, the limited number of patients in this study restricts the interpretation of these numbers. Despite this, the promising results obtained from the PSG data motivate us to extend our method to a larger patient population to uncover the relationship between demographic data, medical conditions and the severity of OSA events. This could enhance the early detection of patients at risk and improve patient outcomes.

| **Quartile** | **Age** | **BMI** | **Gender** (% of females) | **Diabetes** (% of presence) | **Hypertension** (% of presence) |
|---|---|---|---|---|---|
| **Q1** | 56.3 ± 15.6 | 31.5 ± 7.5 | 24.7 | 12.4 | 42.3 |
| **Q2** | 55.4 ± 14.8 | 31.7 ± 7.5 | 28.7 | 17.2 | 48.6 |
| **Q3** | 55.7 ± 14.6 | 31.8 ± 7.7 | 31.8 | 16.6 | 44.5 |
| **Q4** | 56.4 ± 14.4 | 32.2 ± 6.9 | 33.9 | 17.5 | 46.1 |

**Table 8.4.** Patient Information Comparison Across Severity Quartiles.

## Discussion

The primary objective of this study was to introduce a novel method for objectively scoring the severity of Obstructive Sleep Apnea-Hypopnea (OSA) events based on polysomnographic (PSG) signals. Our approach builds upon the human-centered explainability strategy developed in this thesis, which allows us to derive a comprehensive severity scale. This scale is represented by the most discriminant direction within the latent space, encompassing all the severity metrics investigated in this research, including desaturation area, apnea event duration, and arousal events.

Utilizing the xAAEnet architecture, we ensure that the latent space possesses essential properties for deploying our *human-centered xAI* approach effectively: 1) It retains a significant portion of the input signal information; 2) It can be

readily applied to new patients, ensuring generalizability; 3) It facilitates a fair comparison between different OSA trials through directional analysis.

In this section, we demonstrate the improved performance achieved by the second version of our xAI model, xAAEnet, in scoring the severity of OSA trials, comparing it to the initial version, xVAEnet. We assess this improvement both qualitatively, by examining the method's sensitivity to changes in severity concerning physiological variations, and quantitatively, by evaluating its ability to rank OSA trials based on their severity scores. The outcomes of this study suggest that our proposed method holds promise in addressing the key challenges associated with OSA assessment. It offers a robust, objective, and interpretable severity scoring mechanism. The multimodal version of xAAEnet also provides the opportunity to incorporate demographic information and medical conditions of the patients.

Quantifying performance necessitated the use of an approximate hand-made score since, in the current landscape of OSA severity research, no ground-truth reference is available. Consequently, for a better evaluation of the performance, we intend to design a clinical study that allows clinicians to assess the sorting performed by xAAEnet. Additionally, this clinical validation campaign will allow us to compare the efficiency of our scoring method with the AHI currently utilized in clinical practice. This campaign will also contribute to expanding our dataset and incorporating demographic data and coexisting medical conditions. This expansion aims to enhance early detection capabilities for patients at risk of OSA-related complications.

It is essential to note that the utilization of xAAEnet may result in a decrease in pure regression performance. However, since our strategy aims to move beyond simplistic hand-made scores derived from severity metrics, achieving optimal regression performance is not the primary objective. Ideally, the encoder should produce a latent representation wherein all features related to the severity of OSA events vary linearly along one specific direction - the severity scale. Consequently, representing all the severity metrics along one axis may not be optimal for pure regression tasks, but it serves as the most effective way to derive a comprehensive severity score. Future endeavors will involve further comparative studies using post-hoc explainability and the refinement of the xAAEnet model through benchmarking against state-of-the-art models.

## 8.5 In Brief

Summary of Chapter 8

- We have applied our *human-centered xAI* approach to study obstructive sleep apnea (OSA). This innovative approach leverages the inherent properties of the latent space to generate an objective severity score for each apnea event.
- Our methodology encompasses two distinct variants of explainable deep learning models. One is founded on variational auto-encoders (xVAEnet), while the other relies on adversarial auto-encoders (xAAEnet).
- xVAEnet was meticulously trained for a classification task, focusing on discrete severity levels. This endeavor aimed to identify essential EEG biomarkers linked to severe sleep apnea events.
- xAAEnet underwent rigorous training in a regression task, guided by three distinct severity metrics: desaturation area, apnea duration, and the presence of arousal events. The outcome of this process is an objective severity score, derived from the ordering of input OSA trials.

Perspective for Chapter 8

- Initiate a clinical user-study to evaluate the method's capability to categorize OSA trials based on their risk of complications for patients.
- Conduct a comparative analysis between the derived score and the Apnea-Hypopnea Index (AHI), the prevailing metric in clinical settings.

# Conclusion

This thesis represents a comprehensive exploration of the intricate domain of biomedical signal analysis, where the intersections of neuroscience, artificial intelligence, and clinical assessment have given rise to innovative solutions aimed at mitigating potential biases that could undermine the validity and reliability of biomedical research.

The infusion of machine learning into biomedical signal analysis has undoubtedly expanded the horizons of our capabilities. However, it has also brought forth the challenge of dealing with the inherent opacity of deep learning algorithms. In response to this challenge, our focus has been squarely on the field of explainable AI, where our goal is to reach a balance between accuracy and interpretability.

Throughout our journey, we have consistently underscored the omnipresence of biases that can manifest in every phase of biomedical research, from inception to dissemination.

The bedrock of our solutions lies in two meticulously constructed datasets. The priming dataset has played a pivotal role in establishing our standardized frameworks, which are tailored to address the "white-box" biases. These frameworks encompass the evaluation of confounding factors, ERP preprocessing, and the validation of brain source reconstruction. In parallel, the Obstructive Sleep Apnea (OSA) dataset, based on clinical PSG recordings, has served as an invaluable resource, offering profound insights into the variability of apnea severity among patients. It forms the cornerstone of our explainable AI approach, crafted to counter "black-box" biases.

We have unveiled a framework designed to assess confounding bias effect in the interpretation of ERP data. Employing a two-level hierarchical general linear modeling approach, we have sought to discern the separability between categorical and confounding effects, ultimately challenging the conventional

wisdom regarding the need to balance confounding factors across categories in the design of ERP experiments.

We have also built a standardized framework for ERP preprocessing. This framework, conceived with an eye toward fostering reproducibility in processing pipelines, paves the way for the creation of reliable benchmarks, thus helping to mitigate measurement bias and ensuring cleaner data for the broader biomedical community.

Further initiatives have been detailed in the realm of source localization benchmarking, which address modeling bias. To this end, we have deployed a versatile validation framework that harnesses synthetically generated pseudo-EEG signals. This framework offers customization possibilities, opening doors to its adaptation to diverse EEG datasets. Our future endeavors in this domain aim to encompass connectivity analysis within the framework while enabling automatic customization of artifact shapes.

At the heart of this thesis resides the development of a human-centered approach to explainable AI. By endowing the latent space of deep learning models with specific properties, we have ventured into this space to scrutinize the behavior of particular features of interest during transitions between conditions, guided by comparisons of input data.

Applied to the assessment of obstructive sleep apneas (OSAs), our innovative methodology has not only shed light on critical EEG biomarkers but has also yielded an objective severity score for OSA trials. The future trajectory of this research points towards a comprehensive clinical user study, a pivotal step for validating the efficiency of our method in scoring the severity of apnea events. The ultimate goal is to integrate this method into clinical setups, thus advancing patient care.

In conclusion, we hope that our contributions will pave the way towards more reproducible and reliable biomedical signal analyses, fostering transparency and excellence in the pursuit of scientific understanding.

[]

# Bibliography


[1] D. Ribatti, "William Harvey and the discovery of the circulation of the blood", *Journal of Angiogenesis Research*, vol. 1, p. 3, Sep. 2009. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2776239/

[2] F. Bigotti, "The Weight of the Air: Santorio's Thermometers and the Early History of Medical Quantification Reconsidered", *Journal of early modern studies*, vol. 7, no. 1, pp. 73–103, 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6407691/

[3] S. S. Barold, "Willem Einthoven and the birth of clinical electrocardiography a hundred years ago", *Cardiac Electrophysiology Review*, vol. 7, no. 1, pp. 99–104, Jan. 2003.

[4] M. Tubiana, "[Wilhelm Conrad Röntgen and the discovery of X-rays]", *Bulletin De l'Academie Nationale De Medecine*, vol. 180, no. 1, pp. 97–108, Jan. 1996.

[5] S. Campbell, "A Short History of Sonography in Obstetrics and Gynaecology", *Facts, Views & Vision in ObGyn*, vol. 5, no. 3, pp. 213–229, 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3987368/

[6] A. Gramfort, "Mapping, timing and tracking cortical activations with MEG and EEG: Methods and application to human vision", Ph.D. dissertation, Oct. 2009.

[7] N. Lewin, E. Aksay, and C. E. Clancy, "Computational Modeling Reveals Dendritic Origins of GABAA-Mediated Excitation in CA1 Pyramidal Neurons", *PLOS ONE*, vol. 7, no. 10, p. e47250, Oct. 2012, publisher: Public Library of Science. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0047250

[8] M. Tudor, L. Tudor, and K. I. Tudor, "[Hans Berger (1873-1941)–the history of electroencephalography]", *Acta Medica Croatica: Casopis Hravatske Akademije Medicinskih Znanosti*, vol. 59, no. 4, pp. 307–313, 2005.

[9] A. Kales and A. Rechtschaffen, *A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects: Allan Rechtschaffen and Anthony Kales, Editors.* U. S. National Institute of Neurological Diseases and Blindness, Neurological Information Network, 1968, google-Books-ID: Z41IvQEACAAJ.

[10] F. A. GIBBS, W. G. LENNOX, and E. L. GIBBS, "THE ELECTROENCEPHALOGRAM IN DIAGNOSIS AND IN LOCALIZATION OF EPILEPTIC SEIZURES", *Archives of Neurology & Psychiatry*, vol. 36, no. 6, pp. 1225–1235, Dec. 1936. [Online]. Available: https://doi.org/10.1001/archneurpsyc.1936.02260120072005

[11] M. Hamalainen and J. Sarvas, "Realistic conductivity geometry model of the human head for interpretation of neuromagnetic data", *IEEE Transactions on Biomedical Engineering*, vol. 36, no. 2, pp. 165–171, Feb. 1989, conference Name: IEEE Transactions on Biomedical Engineering.

[12] K. Glomb, J. Cabral, A. Cattani, A. Mazzoni, A. Raj, and B. Franceschiello, "Computational Models in Electroencephalography", *Brain Topography*, vol. 35, no. 1, pp. 142–161, Jan. 2022. [Online]. Available: https://doi.org/10.1007/s10548-021-00828-2

[13] M.-P. Hosseini, A. Hosseini, and K. Ahi, "A Review on Machine Learn-

ing for EEG Signal Processing in Bioengineering", *IEEE Reviews in Biomedical Engineering*, vol. 14, pp. 204–218, 2021, conference Name: IEEE Reviews in Biomedical Engineering.

[14] Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T. H. Falk, and J. Faubert, "Deep learning-based electroencephalography analysis: a systematic review", *Journal of Neural Engineering*, vol. 16, no. 5, p. 051001, Aug. 2019, publisher: IOP Publishing. [Online]. Available: https://dx.doi.org/10.1088/1741-2552/ab260c

[15] H. Marzbani, H. R. Marateb, and M. Mansourian, "Neurofeedback: A Comprehensive Review on System Design, Methodology and Clinical Applications", *Basic and Clinical Neuroscience*, vol. 7, no. 2, pp. 143–158, Apr. 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4892319/

[16] J. J. Shih, D. J. Krusienski, and J. R. Wolpaw, "Brain-Computer Interfaces in Medicine", *Mayo Clinic Proceedings*, vol. 87, no. 3, pp. 268–279, Mar. 2012. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3497935/

[17] R. Maskeliunas, R. Damaševičius, I. Martišius, and M. Vasiljevas, "Consumer-grade EEG devices: Are they usable for control tasks?" *PeerJ*, vol. 4, Apr. 2016.

[18] G. Li and W.-Y. Chung, "Electroencephalogram-Based Approaches for Driver Drowsiness Detection and Management: A Review", *Sensors*, vol. 22, pp. 1–26, Jan. 2022.

[19] G. Buzsaki, *Rhythms of the Brain*. Oxford University Press, Aug. 2006, google-Books-ID: ldz58irprjYC.

[20] E. Aserinsky and N. Kleitman, "Regularly occurring periods of eye motility, and concomitant phenomena, during sleep", *Science (New York, N.Y.)*, vol. 118, no. 3062, pp. 273–274, Sep. 1953.

[21] A. Lewicke, E. Sazonov, and S. Schuckers, "Sleep-wake identification in infants: heart rate variability compared to actigraphy", in *The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 1, Sep. 2004, pp. 442–445.

[22] O. Faust, H. Razaghi, R. Barika, E. J. Ciaccio, and U. R. Acharya, "A review of automated sleep stage scoring based on physiological signals for the new millennia", *Computer Methods and Programs in Biomedicine*, vol. 176, pp. 81–91, Jul. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169260718313865

[23] A. Crivello, P. Barsocchi, M. Girolami, and F. Palumbo, "The Meaning of Sleep Quality: A Survey of Available Technologies", *IEEE Access*, vol. PP, Nov. 2019.

[24] Y. Tsividis and J. O. Voorman, *Integrated Continuous-time Filters: Principles, Design, and Applications.* IEEE Press, 1993, google-Books-ID: QAFrQgAACAAJ.

[25] B. Hammack, S. Kranz, and B. Carpenter, *Albert Michelson's Harmonic Analyzer: A Visual Tour of a Nineteenth Century Machine That Performs Fourier Analysis.* Articulate Noise Books, Oct. 2014, google-Books-ID: 9cR3BQAAQBAJ.

[26] B. Widrow and S. D. Stearns, *Adaptive Signal Processing.* Prentice-Hall, 1985, google-Books-ID: X74QAQAAMAAJ.

[27] F. Jelinek, *Statistical Methods for Speech Recognition.* MIT Press, Jan. 1998, google-Books-ID: 1C9dzcJTWowC.

[28] I. Daubechies, *Ten Lectures on Wavelets.* Society for Industrial and Applied Mathematics, Jun. 1992, google-Books-ID: 9t5SG06AiT0C.

[29] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis.* John Wiley & Sons, Apr. 2004, google-Books-ID: 96D0ypDwAkkC.

[30] N. E. Huang, *Hilbert-Huang Transform and Its Applications*. World Scientific, 2014, google-Books-ID: aJ26CgAAQBAJ.

[31] J. B. J. Fourier, *The Analytical Theory of Heat*. Cambridge University Press, Jul. 2009, google-Books-ID: RXsLQAAACAAJ.

[32] D. Gabor, "Theory of communication", *Journal of the Institution of Electrical Engineers - Part I: General*, vol. 94, no. 73, pp. 58–58, Jan. 1947. [Online]. Available: https://digital-library.theiet.org/content/journals/10.1049/ji-1.1947.0015

[33] C. S. Herrmann, S. Rach, J. Vosskuhl, and D. Strüber, "Time–Frequency Analysis of Event-Related Potentials: A Brief Tutorial", *Brain Topography*, vol. 27, no. 4, pp. 438–450, Jul. 2014. [Online]. Available: https://doi.org/10.1007/s10548-013-0327-5

[34] T. Fernando, H. Gammulle, S. Denman, S. Sridharan, and C. Fookes, "Deep Learning for Medical Anomaly Detection – A Survey", *ACM Computing Surveys*, vol. 54, no. 7, pp. 141:1–141:37, Jul. 2021. [Online]. Available: https://doi.org/10.1145/3464423

[35] M. M. Ahsan, S. A. Luna, and Z. Siddique, "Machine-Learning-Based Disease Diagnosis: A Comprehensive Review", *Healthcare*, vol. 10, no. 3, p. 541, Mar. 2022, number: 3 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2227-9032/10/3/541

[36] H. O. Alanazi, A. H. Abdullah, and K. N. Qureshi, "A Critical Review for Developing Accurate and Dynamic Predictive Models Using Machine Learning Methods in Medicine and Health Care", *Journal of Medical Systems*, vol. 41, no. 4, p. 69, Mar. 2017. [Online]. Available: https://doi.org/10.1007/s10916-017-0715-6

[37] K. Pearson, "LIII. On lines and planes of closest fit to systems of points in space", *The London, Edinburgh, and Dublin Philosophical Magazine*

*and Journal of Science*, vol. 2, no. 11, pp. 559–572, Nov. 1901, publisher: Taylor & Francis _eprint: https://doi.org/10.1080/14786440109462720. [Online]. Available: https://doi.org/10.1080/14786440109462720

[38] P. L. Nunez and R. Srinivasan, *Electric Fields of the Brain: The Neurophysics of EEG*. Oxford University Press, 2006, google-Books-ID: fUv54as56_8C.

[39] C. Pernet, M. I. Garrido, A. Gramfort, N. Maurits, C. M. Michel, E. Pang, R. Salmelin, J. M. Schoffelen, P. A. Valdes-Sosa, and A. Puce, "Issues and recommendations from the OHBM COBIDAS MEEG committee for reproducible EEG and MEG research", *Nature Neuroscience*, vol. 23, no. 12, pp. 1473–1483, Dec. 2020, number: 12 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41593-020-00709-0

[40] R. Srinivasan, W. R. Winter, J. Ding, and P. L. Nunez, "EEG and MEG coherence: Measures of functional connectivity at distinct spatial scales of neocortical dynamics", *Journal of Neuroscience Methods*, vol. 166, no. 1, pp. 41–52, Oct. 2007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016502700700307X

[41] W. J. Freeman, "Origin, structure, and role of background EEG activity. Part 1. Analytic amplitude", *Clinical Neurophysiology*, vol. 115, no. 9, pp. 2077–2088, Sep. 2004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1388245704001695

[42] D. Abásolo, R. Hornero, P. Espino, D. Álvarez, and J. Poza, "Entropy analysis of the EEG background activity in Alzheimer's disease patients", *Physiological Measurement*, vol. 27, no. 3, p. 241, Jan. 2006. [Online]. Available: https://dx.doi.org/10.1088/0967-3334/27/3/003

[43] B. G. Luisa, *Wavelets: A Tutorial in Theory and Applications*. Academic Press, Dec. 2012, google-Books-ID: ELaaRvmgKAQC.

[44] L. Minati, G. Varotto, L. D'Incerti, F. Panzica, and D. Chan, "From brain topography to brain topology: relevance of graph theory to functional neuroscience", *NeuroReport*, vol. 24, no. 10, p. 536, Jul. 2013. [Online]. Available: https://journals.lww.com/neuroreport/abstract/2013/07100/from_brain_topography_to_brain_topology__relevance.7.aspx

[45] E. Maris and R. Oostenveld, "Nonparametric statistical testing of EEG- and MEG-data", *Journal of Neuroscience Methods*, vol. 164, no. 1, pp. 177–190, Aug. 2007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165027007001707

[46] R. D. Pascual-Marqui, C. M. Michel, and D. Lehmann, "Low resolution electromagnetic tomography: a new method for localizing electrical activity in the brain", *International Journal of Psychophysiology*, vol. 18, no. 1, pp. 49–65, Oct. 1994. [Online]. Available: https://www.sciencedirect.com/science/article/pii/016787608490014X

[47] J. Mosher and R. Leahy, "Source localization using recursively applied and projected (RAP) MUSIC", *IEEE Transactions on Signal Processing*, vol. 47, no. 2, pp. 332–340, Feb. 1999, conference Name: IEEE Transactions on Signal Processing.

[48] A. Hillebrand and G. R. Barnes, "Beamformer Analysis of MEG Data", in *International Review of Neurobiology*, ser. Magnetoencephalography. Academic Press, Jan. 2005, vol. 68, pp. 149–171. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0074774205680063

[49] F. Shaffer and J. P. Ginsberg, "An Overview of Heart Rate Variability Metrics and Norms", *Frontiers in Public Health*, vol. 5, p. 258, Sep. 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5624990/

[50] B. A. Staats, H. W. Bonekat, C. D. Harris, and K. P. Offord, "Chest wall motion in sleep apnea", *The American Review of Respiratory Disease*,

vol. 130, no. 1, pp. 59–63, Jul. 1984.

[51] M. H. Kryger, T. Roth, and W. C. Dement, *Principles and Practice of Sleep Medicine - E-Book: Expert Consult - Online and Print*. Elsevier Health Sciences, Nov. 2010, google-Books-ID: 3B52V4PnrVkC.

[52] D. Temirbekov, S. Gunes, Z. M. Yazici, and I. Sayin, "The Ignored Parameter in the Diagnosis of Obstructive Sleep Apnea Syndrome: The Oxygen Desaturation Index", *Turkish Archives of Otorhinolaryngology*, vol. 56, no. 1, pp. 1–6, Mar. 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6017211/

[53] R. Natarajan, "Review of periodic limb movement and restless leg syndrome", *Journal of Postgraduate Medicine*, vol. 56, no. 2, p. 157, Apr. 2010, company: Medknow Publications and Media Pvt. Ltd. Distributor: Medknow Publications and Media Pvt. Ltd. Institution: Medknow Publications and Media Pvt. Ltd. Label: Medknow Publications and Media Pvt. Ltd. Publisher: Medknow Publications. [Online]. Available: https://www.jpgmonline.com/article.asp?issn=0022-3859;year=2010;volume=56;issue=2;spage=157;epage=162;aulast=Nataraj;type=0

[54] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines", in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[55] S. Dreyfus, "The numerical solution of variational problems", *Journal of Mathematical Analysis and Applications*, vol. 5, no. 1, pp. 30–45, Aug. 1962. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0022247X62900045

[56] K. Hornik, "Approximation capabilities of multilayer feedforward networks", *Neural Networks*, vol. 4, no. 2, pp. 251–257, Jan. 1991. [Online]. Available: https://www.sciencedirect.com/science/article/pii/089360809190009T

[57] A. Graves, S. Fernandez, and J. Schmidhuber, "Multi-Dimensional Recurrent Neural Networks", May 2007, arXiv:0705.2011 [cs]. [Online]. Available: http://arxiv.org/abs/0705.2011

[58] Y. LeCun, K. Kavukcuoglu, and C. Farabet, "Convolutional networks and applications in vision", in *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, May 2010, pp. 253–256, iSSN: 2158-1525.

[59] G. E. Hinton and R. R. Salakhutdinov, "Reducing the Dimensionality of Data with Neural Networks", *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006, publisher: American Association for the Advancement of Science. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.1127647

[60] "Autoencoders Explained". [Online]. Available: https://www.mathworks.com/discovery/autoencoder.html

[61] N. Mashhadi, A. Z. Khuzani, M. Heidari, and D. Khaledyan, "Deep learning denoising for EOG artifacts removal from EEG signals", in *2020 IEEE Global Humanitarian Technology Conference (GHTC)*, Oct. 2020, pp. 1–6, iSSN: 2377-6919.

[62] H. Zhang, M. Zhao, C. Wei, D. Mantini, Z. Li, and Q. Liu, "EEGdenoiseNet: a benchmark dataset for deep learning solutions of EEG denoising", *Journal of Neural Engineering*, vol. 18, no. 5, p. 056057, Oct. 2021, publisher: IOP Publishing. [Online]. Available: https://dx.doi.org/10.1088/1741-2552/ac2bf8

[63] P. Kaushik, H. Yang, P. P. Roy, and M. van Vugt, "Comparing resting state and task-based EEG using machine learning to predict vulnerability to depression in a non-clinical population", *Scientific Reports*, vol. 13, no. 1, p. 7467, May 2023, number: 1 Publisher: Nature Publishing Group. [Online]. Available: https:

//www.nature.com/articles/s41598-023-34298-2

[64] X. Zheng, Z. Cao, and Q. Bai, "An Evoked Potential-Guided Deep Learning Brain Representation for Visual Classification", in *Neural Information Processing*, ser. Communications in Computer and Information Science, H. Yang, K. Pasupa, A. C.-S. Leung, J. T. Kwok, J. H. Chan, and I. King, Eds. Cham: Springer International Publishing, 2020, pp. 54–61.

[65] L. Hecker, R. Rupprecht, L. Tebartz Van Elst, and J. Kornmeier, "ConvDip: A Convolutional Neural Network for Better EEG Source Imaging", *Frontiers in Neuroscience*, vol. 15, 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnins.2021.569918

[66] S. Chambon, M. N. Galtier, P. J. Arnal, G. Wainrib, and A. Gramfort, "A Deep Learning Architecture for Temporal Sleep Stage Classification Using Multivariate and Multimodal Time Series", *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 4, pp. 758–769, Apr. 2018, conference Name: IEEE Transactions on Neural Systems and Rehabilitation Engineering.

[67] S. S. Mostafa, F. Mendonça, A. G. Ravelo-García, and F. Morgado-Dias, "A Systematic Review of Detecting Sleep Apnea Using Deep Learning", *Sensors*, vol. 19, no. 22, p. 4934, Jan. 2019, number: 22 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/1424-8220/19/22/4934

[68] F. Grün, C. Rupprecht, N. Navab, and F. Tombari, "A Taxonomy and Library for Visualizing Learned Features in Convolutional Neural Networks", Jun. 2016, arXiv:1606.07757 [cs]. [Online]. Available: http://arxiv.org/abs/1606.07757

[69] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning Deep Features for Discriminative Localization", Dec. 2015,

arXiv:1512.04150 [cs]. [Online]. Available: http://arxiv.org/abs/1512.04150

[70] "4.1. Partial Dependence and Individual Conditional Expectation plots". [Online]. Available: https://scikit-learn/stable/modules/partial_dependence.html

[71] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene Selection for Cancer Classification using Support Vector Machines", *Machine Learning*, vol. 46, no. 1, pp. 389–422, Jan. 2002. [Online]. Available: https://doi.org/10.1023/A:1012487302797

[72] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions", in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[73] V. Petsiuk, A. Das, and K. Saenko, "RISE: Randomized Input Sampling for Explanation of Black-box Models", *ArXiv*, Jun. 2018. [Online]. Available: https://www.semanticscholar.org/paper/RISE%3A-Randomized-Input-Sampling-for-Explanation-of-Petsiuk-Das/d00c7fc5201405d5411b5ad3da93c5575ce8f10e

[74] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier", Aug. 2016, arXiv:1602.04938 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1602.04938

[75] "LIME: Local Interpretable Model-Agnostic Explanations". [Online]. Available: https://c3.ai/glossary/data-science/lime-local-interpretable-model-agnostic-explanations/

[76] A. Dhurandhar, P.-Y. Chen, R. Luss, C.-C. Tu, P. Ting, K. Shanmugam, and P. Das, "Explanations based on the Missing: Towards Contrastive

Explanations with Pertinent Negatives", Feb. 2018.

[77] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale", Jun. 2021, arXiv:2010.11929 [cs]. [Online]. Available: http://arxiv.org/abs/2010.11929

[78] J. Maurício, I. Domingues, and J. Bernardino, "Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review", *Applied Sciences*, vol. 13, no. 9, p. 5521, Jan. 2023, number: 9 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2076-3417/13/9/5521

[79] J. Bastings and K. Filippova, "The elephant in the interpretability room: Why use attention as explanation when we have saliency methods?" in *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. Online: Association for Computational Linguistics, Nov. 2020, pp. 149–155. [Online]. Available: https://aclanthology.org/2020.blackboxnlp-1.14

[80] P.-J. Kindermans, S. Hooker, J. Adebayo, M. Alber, K. T. Schütt, S. Dähne, D. Erhan, and B. Kim, "The (Un)reliability of Saliency Methods", in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, ser. Lecture Notes in Computer Science, W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller, Eds. Cham: Springer International Publishing, 2019, pp. 267–280. [Online]. Available: https://doi.org/10.1007/978-3-030-28954-6_14

[81] M. Böhle, M. Fritz, and B. Schiele, "Holistically Explainable Vision Transformers", Jan. 2023, arXiv:2301.08669 [cs, stat]. [Online]. Available: http://arxiv.org/abs/2301.08669

[82] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, "The false hope of current approaches to explainable artificial intelligence in health care", *The Lancet Digital Health*, vol. 3, no. 11, pp. e745–e750, Nov. 2021.

[83] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. K. Su, "This Looks Like That: Deep Learning for Interpretable Image Recognition", in *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://papers.nips.cc/paper_files/paper/2019/hash/adf7ee2dcf142b0e11888e72b43fcb75-Abstract.html

[84] K. Wang, J. Oramas, and T. Tuytelaars, "Towards Human-Understandable Visual Explanations:Imperceptible High-frequency Cues Can Better Be Removed", Apr. 2021, arXiv:2104.07954 [cs]. [Online]. Available: http://arxiv.org/abs/2104.07954

[85] J. Oramas, K. Wang, and T. Tuytelaars, "Visual Explanation by Interpretation: Improving Visual Feedback Capabilities of Deep Neural Networks", Mar. 2019, arXiv:1712.06302 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1712.06302

[86] "The Act", Feb. 2021. [Online]. Available: https://artificialintelligenceact.eu/the-act/

[87] D. Chavalarias and J. P. A. Ioannidis, "Science mapping analysis characterizes 235 biases in biomedical research", *Journal of Clinical Epidemiology*, vol. 63, no. 11, pp. 1205–1215, Nov. 2010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0895435610000223

[88] C. J. Pannucci and E. G. Wilkins, "Identifying and Avoiding Bias in Research", *Plastic and reconstructive surgery*, vol. 126, no. 2, pp. 619–625, Aug. 2010. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2917255/

[89] T. Gerhard, "Bias: Considerations for research practice", *American*

*Journal of Health-System Pharmacy*, vol. 65, no. 22, pp. 2159–2168, Nov. 2008. [Online]. Available: https://doi.org/10.2146/ajhp070369

[90] J. Lambert, "Statistics in Brief: How to Assess Bias in Clinical Studies?" *Clinical Orthopaedics and Related Research®*, vol. 469, no. 6, pp. 1794–1796, Jun. 2011. [Online]. Available: https://doi.org/10.1007/s11999-010-1538-7

[91] D. Nunan, C. Heneghan, and E. A. Spencer, "Catalogue of bias: allocation bias", *BMJ evidence-based medicine*, vol. 23, no. 1, pp. 20–21, Feb. 2018.

[92] I. Simoes Loureiro and L. Lefebvre, "Retrogenesis of semantic knowledge: Comparative approach of acquisition and deterioration of concepts in semantic memory", *Neuropsychology*, vol. 30, no. 7, pp. 853–859, 2016, place: US Publisher: American Psychological Association.

[93] M. A. Pourhoseingholi, A. R. Baghestani, and M. Vahedi, "How to control confounding effects by statistical analysis", *Gastroenterology and Hepatology from Bed to Bench*, vol. 5, no. 2, pp. 79–83, 2012.

[94] J. Morgenstern, "Bias in medical research", Jul. 2018. [Online]. Available: https://first10em.com/bias/

[95] D. L. Sackett, "BIAS IN ANALYTIC RESEARCH", in *The Case-Control Study Consensus and Controversy*, M. A. Ibrahim, Ed. Pergamon, Jan. 1979, pp. 51–63. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780080249070500134

[96] K. Mahtani, E. A. Spencer, J. Brassey, and C. Heneghan, "Catalogue of bias: observer bias", *BMJ Evidence-Based Medicine*, vol. 23, no. 1, pp. 23–24, Feb. 2018, publisher: Royal Society of Medicine Section: EBM Learning. [Online]. Available: https://ebm.bmj.com/content/23/1/23

[97] N. von Ellenrieder, C. H. Muravchik, M. Wagner, and A. Nehorai, "Effect of head shape variations among individuals on the EEG/MEG for-

ward and inverse problems", *IEEE transactions on bio-medical engineering*, vol. 56, no. 3, pp. 587–597, Mar. 2009.

[98] A. Keil, E. M. Bernat, M. X. Cohen, M. Ding, M. Fabiani, G. Gratton, E. S. Kappenman, E. Maris, K. E. Mathewson, R. T. Ward, and N. Weisz, "Recommendations and publication guidelines for studies using frequency domain and time-frequency domain analyses of neural time series", *Psychophysiology*, vol. 59, no. 5, p. e14052, 2022, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/psyp.14052. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/psyp.14052

[99] Q. Zhao, E. Adeli, and K. M. Pohl, "Training confounder-free deep learning models for medical applications", *Nature Communications*, vol. 11, no. 1, p. 6010, Nov. 2020, number: 1 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41467-020-19784-9

[100] M. Jeng, "A selected history of expectation bias in physics", *American Journal of Physics*, vol. 74, no. 7, pp. 578–583, Jul. 2006. [Online]. Available: https://doi.org/10.1119/1.2186333

[101] N. J. DeVito and B. Goldacre, "Catalogue of bias: publication bias", *BMJ Evidence-Based Medicine*, vol. 24, no. 2, pp. 53–54, Apr. 2019, publisher: Royal Society of Medicine Section: EBM Learning. [Online]. Available: https://ebm.bmj.com/content/24/2/53

[102] M. L. Head, L. Holman, R. Lanfear, A. T. Kahn, and M. D. Jennions, "The extent and consequences of p-hacking in science", *PLoS biology*, vol. 13, no. 3, p. e1002106, Mar. 2015.

[103] E. T. Thomas and C. Heneghan, "Catalogue of bias: selective outcome reporting bias", *BMJ Evidence-Based Medicine*, vol. 27, no. 6, pp. 370–372, Dec. 2022, publisher: Royal Society of Medicine Section: EBM learning. [Online]. Available: https://ebm.bmj.com/content/27/6/370

[104] J. J. Kirkham, K. M. Dwan, D. G. Altman, C. Gamble, S. Dodd, R. Smyth, and P. R. Williamson, "The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews", *BMJ*, vol. 340, p. c365, Feb. 2010, publisher: British Medical Journal Publishing Group Section: Research Methods & Reporting. [Online]. Available: https://www.bmj.com/content/340/bmj.c365

[105] M. J. E. Urlings, B. Duyx, G. M. H. Swaen, L. M. Bouter, and M. P. Zeegers, "Citation bias and other determinants of citation in biomedical research: findings from six citation networks", *Journal of Clinical Epidemiology*, vol. 132, pp. 71–78, Apr. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0895435620311975

[106] S. Krishna Mohan, M. Mrsc, and F. Assistant, "Research Bias: A Review For Medical Students", *Journal of Clinical and Diagnostic Research*, vol. 4, pp. 2320–23 242 320, May 2010.

[107] M. Laisney, B. Giffard, S. Belliard, V. de la Sayette, B. Desgranges, and F. Eustache, "When the zebra loses its stripes: Semantic priming in early Alzheimer's disease and semantic dementia", *Cortex*, vol. 47, no. 1, pp. 35–46, Jan. 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010945209002858

[108] R. Oldfield, "The assessment and analysis of handedness: The Edinburgh inventory - ScienceDirect", 1971. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0028393271900674

[109] A. Gramfort, M. Luessi, E. Larson, D. Engemann, D. Strohmeier, C. Brodbeck, R. Goj, M. Jas, T. Brooks, L. Parkkonen, and M. Hämäläinen, "MEG and EEG data analysis with MNE-Python", *Frontiers in Neuroscience*, vol. 7, 2013.

[110] P. Varady, S. Bongar, and Z. Benyo, "Detection of airway obstructions and sleep apnea by analyzing the phase relation of respiration

movement signals", *IEEE Transactions on Instrumentation and Measurement*, vol. 52, no. 1, pp. 2–6, Feb. 2003, conference Name: IEEE Transactions on Instrumentation and Measurement.

[111] C. Shahnaz, A. T. Minhaz, and S. T. Ahamed, "Sub-frame based apnea detection exploiting delta band power ratio extracted from EEG signals", in *2016 IEEE Region 10 Conference (TENCON)*, Nov. 2016, pp. 190–193, iSSN: 2159-3450.

[112] S. Devuyst, "Analyse Automatique de Tracés Polysomnographiques d'Adultes", Ph.D. dissertation, Nov. 2011, publisher: Faculté Polytechnique de Mons. [Online]. Available: https://orbi.umons.ac.be/handle/20.500.12907/35315

[113] S. Devuyst, T. Dutoit, P. Stenuit, M. Kerkhofs, and E. Stanus, "Removal of ECG artifacts from EEG using a modified independent component analysis approach", in *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Aug. 2008, pp. 5204–5207, iSSN: 1558-4615. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/4650387?casa_token=-odOxsI7gt0AAAAA:VJ-rEPUqHK-Howbx1Zuv5zTLaJtACHjOUR7vQS5TDn5hGqbGcH-Btxcba9dGvY

[114] C. R. Pernet, N. Chauveau, C. Gaspar, and G. A. Rousselet, "LIMO EEG: A Toolbox for Hierarchical LInear MOdeling of ElectroEncephaloGraphic Data", *Computational Intelligence and Neuroscience*, vol. 2011, p. e831409, Feb. 2011, publisher: Hindawi. [Online]. Available: https://www.hindawi.com/journals/cin/2011/831409/

[115] L. La Fisca, V. Vandenbulcke, E. Wauthia, A. Miceli, I. Simoes Loureiro, L. Ris, L. Lefebvre, B. Gosselin, and C. R. Pernet, "Biases in BCI experiments: Do we really need to balance stimulus properties across categories?" *Frontiers in Computational Neuroscience*, vol. 16, 2022.

[116] F. X. Alario, L. Ferrand, M. Laganaro, B. New, U. H. Frauenfelder,

and J. Segui, "Predictors of picture naming speed", *Behavior Research Methods, Instruments, & Computers*, vol. 36, no. 1, pp. 140–155, Feb. 2004. [Online]. Available: https://doi.org/10.3758/BF03195559

[117] R. Likert, "A technique for the measurement of attitudes", *Archives of Psychology*, vol. 22 140, pp. 55–55, 1932.

[118] L. Hogonot-Diener, "Guide pratique de la consultation en gériatrie", 2007. [Online]. Available: https://www.unitheque.com/guide-pratique-consultation-geriatrie/mediguides/elsevier-masson/Livre/70494

[119] C. R. Pernet, M. Latinus, T. E. Nichols, and G. A. Rousselet, "Cluster-based computational methods for mass univariate analyses of event-related brain potentials/fields: A simulation study", *Journal of Neuroscience Methods*, vol. 250, pp. 85–93, Jul. 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165027014002878

[120] E. L. Holm, D. F. Slezak, and E. Tagliazucchi, "Contribution of image statistics and semantics in local vs. distributed EEG decoding of rapid serial visual presentation", Sep. 2023, pages: 2023.09.26.559617 Section: New Results. [Online]. Available: https://www.biorxiv.org/content/10.1101/2023.09.26.559617v1

[121] C. R. Pernet, R. Martinez-Cancino, D. Truong, S. Makeig, and A. Delorme, "From BIDS-Formatted EEG Data to Sensor-Space Group Results: A Fully Reproducible Workflow With EEGLAB and LIMO EEG", *Frontiers in Neuroscience*, vol. 14, 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnins.2020.610388

[122] R. Oostenveld, P. Fries, E. Maris, and J.-M. Schoffelen, "FieldTrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data", *Computational Intelligence and Neuroscience*, vol. 2011, p. e156869, Dec. 2010, publisher:

Hindawi. [Online]. Available: https://www.hindawi.com/journals/cin/2011/156869/

[123] C. Pernet, "4. Data Preprocessing", Aug. 2018. [Online]. Available: https://cobidasmeeg.wordpress.com/2018/08/07/preprocessing-and-processing-reporting/

[124] M. Paul, G. H. Govaart, and A. Schettino, "Making ERP research more transparent: Guidelines for preregistration", *International Journal of Psychophysiology*, vol. 164, pp. 52–63, Jun. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016787602100074X

[125] 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, Mar. 2004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165027003003479

[126] L. Pion-Tonachini, S. Makeig, and K. Kreutz-Delgado, "Crowd labeling latent Dirichlet allocation", *Knowledge and Information Systems*, vol. 53, no. 3, 2017. [Online]. Available: https://link.springer.com/epdf/10.1007/s10115-017-1053-1

[127] B. Somers, T. Francart, and A. Bertrand, "A generic EEG artifact removal algorithm based on the multi-channel Wiener filter", *Journal of Neural Engineering*, vol. 15, no. 3, p. 036007, Feb. 2018, publisher: IOP Publishing. [Online]. Available: https://doi.org/10.1088/1741-2552/aaac92

[128] A. de Cheveigné, "ZapLine: A simple and effective method to remove power line artifacts", *NeuroImage*, vol. 207, p. 116356, Feb. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053811919309474

[129] X. Chen, Q. Chen, Y. Zhang, and Z. J. Wang, "A Novel EEMD-CCA

Approach to Removing Muscle Artifacts for Pervasive EEG", *IEEE Sensors Journal*, vol. 19, no. 19, pp. 8420–8431, Oct. 2019, conference Name: IEEE Sensors Journal.

[130] X. Chen, Q. Liu, W. Tao, L. Li, S. Lee, A. Liu, Q. Chen, J. Cheng, M. J. McKeown, and Z. J. Wang, "ReMAE: User-Friendly Toolbox for Removing Muscle Artifacts From EEG", *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 5, pp. 2105–2119, May 2020, conference Name: IEEE Transactions on Instrumentation and Measurement.

[131] P. Hansen, M. Kringelbach, and R. Salmelin, Eds., *MEG: An Introduction to Methods*. Oxford University Press, Jun. 2010. [Online]. Available: https://academic.oup.com/book/9980

[132] "Polhemus Fastrak". [Online]. Available: https://polhemus.com/motion-tracking/all-trackers/fastrak

[133] "Structure Core - Depth refined". [Online]. Available: https://structure.io/structure-core

[134] J. Chen, K. Yao, and R. Hudson, "Source localization and beamforming", *IEEE Signal Processing Magazine*, vol. 19, no. 2, pp. 30–39, Mar. 2002, conference Name: IEEE Signal Processing Magazine.

[135] R. Grave de Peralta Menendez and S. G. Andino, "Distributed Source Models: Standard Solutions and New Developments", in *Analysis of Neurophysiological Brain Functioning*, ser. Springer Series in Synergetics, C. Uhl, Ed. Berlin, Heidelberg: Springer, 1999, pp. 176–201. [Online]. Available: https://doi.org/10.1007/978-3-642-60007-4_10

[136] C. H. Wolters, A. Anwander, G. Berti, and U. Hartmann, "Geometry-Adapted Hexahedral Meshes Improve Accuracy of Finite-Element-Method-Based EEG Source Analysis", *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 8, pp. 1446–1453, Aug. 2007, conference

Name: IEEE Transactions on Biomedical Engineering.

[137] B. Van Veen, W. Van Drongelen, M. Yuchtman, and A. Suzuki, "Localization of brain electrical activity via linearly constrained minimum variance spatial filtering", *IEEE Transactions on Biomedical Engineering*, vol. 44, no. 9, pp. 867–880, Sep. 1997, conference Name: IEEE Transactions on Biomedical Engineering.

[138] W. Ou, M. S. Hämäläinen, and P. Golland, "A distributed spatio-temporal EEG/MEG inverse solver", *NeuroImage*, vol. 44, no. 3, pp. 932–946, Feb. 2009. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1053811908007155

[139] M. Rubega, M. Carboni, M. Seeber, D. Pascucci, S. Tourbier, G. Toscano, P. Van Mierlo, P. Hagmann, G. Plomp, S. Vulliemoz, and C. M. Michel, "Estimating EEG Source Dipole Orientation Based on Singular-value Decomposition for Connectivity Analysis", *Brain Topography*, vol. 32, no. 4, pp. 704–719, Jul. 2019. [Online]. Available: https://doi.org/10.1007/s10548-018-0691-2

[140] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot, "Automated Anatomical Labeling of Activations in SPM Using a Macroscopic Anatomical Parcellation of the MNI MRI Single-Subject Brain", *NeuroImage*, vol. 15, no. 1, pp. 273–289, Jan. 2002. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1053811901909784

[141] L. La Fisca and B. Gosselin, "A Versatile Validation Framework for ERP and Oscillatory Brain Source Localization Using FieldTrip", in *4th International Conference on Biometric Engineering and Applications*, ser. ICBEA '21. New York, NY, USA: Association for Computing Machinery, Oct. 2021, pp. 7–12.

[142] S. Baillet, J. J. Riera, G. Marin, J. F. Mangin, J. Aubert,

and L. Garnero, "Evaluation of inverse methods and head models for EEG source localization using a human skull phantom", *Physics in Medicine and Biology*, vol. 46, no. 1, pp. 77–96, Nov. 2000, publisher: IOP Publishing. [Online]. Available: https://doi.org/10.1088/0031-9155/46/1/306

[143] R. M. Leahy, J. C. Mosher, M. E. Spencer, M. X. Huang, and J. D. Lewine, "A study of dipole localization accuracy for MEG and EEG using a human skull phantom", *Electroencephalography and Clinical Neurophysiology*, vol. 107, no. 2, pp. 159–173, Apr. 1998. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0013469498000571

[144] F. Darvas, D. Pantazis, E. Kucukaltun-Yildirim, and R. M. Leahy, "Mapping human brain function with MEG and EEG: methods and validation", *NeuroImage*, vol. 23, pp. S289–S299, Jan. 2004. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1053811904003799

[145] A. Gramfort, D. Strohmeier, J. Haueisen, M. S. Hämäläinen, and M. Kowalski, "Time-frequency mixed-norm estimates: Sparse M/EEG imaging with non-stationary source activations", *NeuroImage*, vol. 70, pp. 410–422, Apr. 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1053811912012372

[146] A. Delorme, T. Mullen, C. Kothe, Z. Akalin Acar, N. Bigdely-Shamlo, A. Vankov, and S. Makeig, "EEGLAB, SIFT, NFT, BCILAB, and ERICA: new tools for advanced EEG processing", *Computational Intelligence and Neuroscience*, vol. 2011, p. 130714, 2011.

[147] J. T. Lindgren, A. Merlini, A. Lécuyer, and F. P. Andriulli, "simBCI—A Framework for Studying BCI Methods by Simulated EEG", *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 11, pp. 2096–2105, Nov. 2018, conference Name: IEEE Trans-

actions on Neural Systems and Rehabilitation Engineering.

[148] S. Haufe and A. Ewald, "A Simulation Framework for Benchmarking EEG-Based Brain Connectivity Estimation Methodologies", *Brain Topography*, vol. 32, no. 4, pp. 625–642, Jul. 2019. [Online]. Available: https://doi.org/10.1007/s10548-016-0498-y

[149] L. R. Krol, J. Pawlitzki, F. Lotte, K. Gramann, and T. O. Zander, "SEREEGA: Simulating event-related EEG activity", *Journal of Neuroscience Methods*, vol. 309, pp. 13–24, Nov. 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0165027018302395

[150] D. Purves, K. S. LaBar, M. L. Platt, M. Woldorff, R. Cabeza, and S. A. Huettel, *Principles of Cognitive Neuroscience*, second edition ed. Oxford, New York: Oxford University Press, Nov. 2012.

[151] H. Zhivomirov, "A Method for Colored Noise Generation", *Romanian Journal of Acoustics and Vibration*, vol. 15, no. 1, pp. 14–19, Aug. 2018, number: 1. [Online]. Available: http://rjav.sra.ro/index.php/rjav/article/view/40

[152] S. Makeig, S. Debener, J. Onton, and A. Delorme, "Mining event-related brain dynamics", *Trends in Cognitive Sciences*, vol. 8, no. 5, pp. 204–210, May 2004. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1364661304000816

[153] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing, 3rd Edition*, 3rd ed., 2010. [Online]. Available: /content/one-dot-com/one-dot-com/us/en/higher-education/program.html

[154] A. Hamid, K. Gagliano, S. Rahman, N. Tulin, V. Tchiong, I. Obeid, and J. Picone, "The Temple University Artifact Corpus: An Annotated Corpus of EEG Artifacts", *IEEE Signal Processing in Medicine and Biology Symposium SPMB*, vol. 1, no. 1, Dec. 2020. [Online]. Available: https://par.nsf.gov/biblio/

10199675-temple-university-artifact-corpus-annotated-corpus-eeg-artifacts

[155] J. H. Macke, J. Vetter, and R. Gao, "Generating realistic neurophysiological time series with denoising diffusion probabilistic models", Oct. 2023, accepted: 2023-11-22T14:59:27Z Publisher: bio Rxiv. [Online]. Available: https://publikationen.uni-tuebingen.de/xmlui/handle/10900/147989

[156] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes", May 2014, arXiv:1312.6114.

[157] C. Doersch, "Tutorial on Variational Autoencoders", Jan. 2021, arXiv:1606.05908 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1606.05908

[158] "Variational autoencoders." Mar. 2018. [Online]. Available: https://www.jeremyjordan.me/variational-autoencoders/

[159] G. Lample, N. Zeghidour, N. Usunier, A. Bordes, L. DENOYER, and M. A. Ranzato, "Fader Networks:Manipulating Images by Sliding Attributes", in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/hash/3fd60983292458bf7dee75f12d5e9e05-Abstract.html

[160] X. Zhang, L. Yao, and F. Yuan, "Adversarial Variational Embedding for Robust Semi-supervised Learning", in *Proceedings of the 25th ACM International Conference on Knowledge Discovery & Data Mining*, New York, NY, USA, Jul. 2019, pp. 139–147.

[161] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey, "Adversarial Autoencoders", May 2016.

[162] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning", in *Proceedings of the 26th Annual International Conference on Machine Learning*, ser. ICML '09. New York, NY, USA: Association

for Computing Machinery, Jun. 2009, pp. 41–48. [Online]. Available: https://doi.org/10.1145/1553374.1553380

[163] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised Deep Embedding for Clustering Analysis", in *Proceedings of The 33rd International Conference on Machine Learning*. PMLR, Jun. 2016, pp. 478–487, iSSN: 1938-7228.

[164] P. J. Huber, "Robust Estimation of a Location Parameter", in *Breakthroughs in Statistics: Methodology and Distribution*, ser. Springer Series in Statistics, S. Kotz and N. L. Johnson, Eds. New York, NY: Springer, 1992, pp. 492–518. [Online]. Available: https://doi.org/10.1007/978-1-4612-4380-9_35

[165] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar, "Cats and dogs", *2012 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2012, publisher: IEEE. [Online]. Available: https://cir.nii.ac.jp/crid/1362262946028344576

[166] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition", Dec. 2015, arXiv:1512.03385 [cs]. [Online]. Available: http://arxiv.org/abs/1512.03385

[167] L. van der Maaten and G. Hinton, "Visualizing High-Dimensional Data Using t-SNE", *Journal of Machine Learning Research*, vol. 9, no. nov, pp. 2579–2605, 2008.

[168] R. A. Fisher, "The Use of Multiple Measurements in Taxonomic Problems", *Annals of Eugenics*, vol. 7, no. 2, pp. 179–188, 1936, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-1809.1936.tb02137.x. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-1809.1936.tb02137.x

[169] M. Wattenberg, F. Viégas, and I. Johnson, "How to Use t-SNE Effectively", *Distill*, vol. 1, no. 10, p. e2, Oct. 2016. [Online]. Available:

http://distill.pub/2016/misread-tsne

[170] L. La Fisca, C. Jennebauffe, M. Bruyneel, L. Ris, L. Lefebvre, X. Siebert, and B. Gosselin, *Explainable AI for EEG Biomarkers Identification in Obstructive Sleep Apnea Severity Scoring Task*, Apr. 2023, pages: 6.

[171] L. La fisca, C. Jennebauffe, M. Bruyneel, L. Ris, L. Lefebvre, X. Siebert, and B. Gosselin, "Enhancing OSA Assessment with Explainable AI", in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. Institute of Electrical and Electronics Engineers, Piscataway, United States - New Jersey, Jul. 2023, iSSN: 2375-7477. [Online]. Available: https://orbi.umons.ac.be/handle/20.500.12907/46450

[172] A. S. Jordan, D. G. McSharry, and A. Malhotra, "Adult obstructive sleep apnoea", *The Lancet*, vol. 383, no. 9918, pp. 736–747, Feb. 2014.

[173] R. B. Berry, R. Budhiraja, D. J. Gottlieb, D. Gozal, C. Iber, V. K. Kapur, C. L. Marcus, R. Mehra, S. Parthasarathy, S. F. Quan, S. Redline, K. P. Strohl, S. L. D. Ward, and M. M. Tangredi, "Rules for Scoring Respiratory Events in Sleep: Update of the 2007 AASM Manual for the Scoring of Sleep and Associated Events", *Journal of Clinical Sleep Medicine*, vol. 08, no. 05, pp. 597–619, 2012.

[174] D. A. Pevernagie, B. Gnidovec-Strazisar, L. Grote, R. Heinzer, W. T. McNicholas, T. Penzel, W. Randerath, S. Schiza, J. Verbraecken, and E. S. Arnardottir, "On the rise and fall of the apneahypopnea index: A historical review and critical appraisal", *Journal of Sleep Research*, vol. 29, no. 4, 2020.

[175] D.-H. Park, C.-J. Shin, S.-C. Hong, J. Yu, S.-H. Ryu, E.-J. Kim, H.-B. Shin, and B.-H. Shin, "Correlation between the Severity of Obstructive Sleep Apnea and Heart Rate Variability Indices", *Journal of Korean Medical Science*, vol. 23, no. 2, p. 226, 2008.

[176] G. Bachar, B. Nageris, R. Feinmesser, T. Hadar, E. Yaniv, T. Shpitzer, and L. Eidelman, "Novel grading system for quantifying upper-airway obstruction on sleep endoscopy", *Lung*, vol. 190, no. 3, pp. 313–318, Jun. 2012.

[177] A. Kulkas, P. Tiihonen, P. Julkunen, E. Mervaala, and J. Töyräs, "Novel parameters indicate significant differences in severity of obstructive sleep apnea with patients having similar apnea–hypopnea index", *Medical & Biological Engineering & Computing*, vol. 51, no. 6, pp. 697–708, 2013.

[178] A. Muraja-Murro, A. Kulkas, M. Hiltunen, S. Kupari, T. Hukkanen, P. Tiihonen, E. Mervaala, and J. Töyräs, "Adjustment of apnea-hypopnea index with severity of obstruction events enhances detection of sleep apnea patients with the highest risk of severe health consequences", *Sleep and Breathing*, vol. 18, no. 3, pp. 641–647, Sep. 2014.

[179] H. Korkalainen, J. Töyräs, S. Nikkonen, and T. Leppänen, "Mortality-risk-based apnea–hypopnea index thresholds for diagnostics of obstructive sleep apnea", *Journal of Sleep Research*, vol. 28, no. 6, p. e12855, 2019, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jsr.12855.

[180] W. Cao, J. Luo, and Y. Xiao, "A Review of Current Tools Used for Evaluating the Severity of Obstructive Sleep Apnea", *Nature and Science of Sleep*, vol. 12, pp. 1023–1031, Nov. 2020, publisher: Dove Press.

[181] A. Zinchuk and H. K. Yaggi, "Phenotypic Subtypes of OSA: A Challenge and Opportunity for Precision Medicine", *CHEST*, vol. 157, no. 2, pp. 403–420, Feb. 2020, publisher: Elsevier.

[182] G. Labarca, J. Gower, L. Lamperti, J. Dreyse, and J. Jorquera, "Chronic intermittent hypoxia in obstructive sleep apnea: a narrative review from pathophysiological pathways to a precision clinical approach", *Sleep and Breathing*, vol. 24, no. 2, pp. 751–760, Jun. 2020.

[183] A. Malhotra, I. Ayappa, N. Ayas, N. Collop, D. Kirsch, N. Mcardle, R. Mehra, A. I. Pack, N. Punjabi, D. P. White, and D. J. Gottlieb, "Metrics of sleep apnea severity: beyond the apnea-hypopnea index", *Sleep*, vol. 44, no. 7, Jul. 2021.

[184] S. Puskás, N. Kozák, D. Sulina, L. Csiba, and M. T. Magyar, "Quantitative EEG in obstructive sleep apnea syndrome: a review of the literature", *Reviews in the Neurosciences*, vol. 28, no. 3, pp. 265–270, Apr. 2017.

[185] E. Sforza, S. Grandin, C. Jouny, T. Rochat, and V. Ibanez, "Is waking electroencephalographic activity a predictor of daytime sleepiness in sleep-related breathing disorders?" *The European Respiratory Journal*, vol. 19, no. 4, pp. 645–652, Apr. 2002.

[186] K. Dingli, T. Assimakopoulos, I. Fietze, C. Witt, P. K. Wraith, and N. J. Douglas, "Electroencephalographic spectral analysis: detection of cortical activity changes in sleep apnoea patients", *The European Respiratory Journal*, vol. 20, no. 5, pp. 1246–1253, Nov. 2002.

[187] R. N. Sekkal, F. Bereksi-Reguig, D. Ruiz-Fernandez, N. Dib, and S. Sekkal, "Automatic sleep stage classification: From classical machine learning methods to deep learning", *Biomedical Signal Processing and Control*, vol. 77, p. 103751, Aug. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809422002737

[188] M. Younes, A. Azarbarzin, M. Reid, D. R. Mazzotti, and S. Redline, "Characteristics and reproducibility of novel sleep EEG biomarkers and their variation with sleep apnea and insomnia in a large community-based cohort", *Sleep*, vol. 44, no. 10, Oct. 2021.

[189] S. Nikkonen, H. Korkalainen, S. Kainulainen, S. Myllymaa, A. Leino, L. Kalevo, A. Oksenberg, T. Leppänen, and J. Töyräs, "Estimating daytime sleepiness with previous night electroencephalography, electroocu-

lography, and electromyography spectrograms in patients with suspected sleep apnea using a convolutional neural network", *Sleep*, vol. 43, no. 12, Dec. 2020.

[190] G. C. Gutiérrez-Tobal, D. Álvarez, A. Crespo, F. del Campo, and R. Hornero, "Evaluation of Machine-Learning Approaches to Estimate Sleep Apnea Severity From At-Home Oximetry Recordings", *IEEE Journal of Biomedical and Health Informatics*, Mar. 2019.

[191] H. Banville, O. Chehab, A. Hyvärinen, D.-A. Engemann, and A. Gramfort, "Uncovering the structure of clinical EEG signals with self-supervised learning", *Journal of Neural Engineering*, Mar. 2021.

[192] N. A. Eiseman, M. B. Westover, J. M. Ellenbogen, and M. T. Bianchi, "The Impact of Body Posture and Sleep Stages on Sleep Apnea Severity in Adults", *Journal of Clinical Sleep Medicine*, 2012.

[193] S. G. Jones, B. A. Riedner, R. F. Smith, F. Ferrarelli, G. Tononi, R. J. Davidson, and R. M. Benca, "Regional Reductions in Sleep Electroencephalography Power in Obstructive Sleep Apnea: A High-Density EEG Study", *Sleep*, vol. 37, no. 2, pp. 399–407, Feb. 2014.

[194] G. Varoquaux, P. R. Raamana, D. A. Engemann, A. Hoyos-Idrobo, Y. Schwartz, and B. Thirion, "Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines", *NeuroImage*, vol. 145, pp. 166–179, Jan. 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S105381191630595X

# List of Figures

1.1 Action potential (AP) and post-synaptic potential (PSP) in neuron. Action potentials traverse chemical synapses, reaching the neuron's dendrites. These interactions result in the emergence of post-synaptic potentials, whose cumulative effect gives rise to subsequent action potentials, capable of propagating along the neuron's axon. Adapted from [6]. . . . . 11

1.2 Temporal Comparison: Action Potential (AP) vs. Post-Synaptic Potential (PSP). The action potential exhibits a biphasic waveform with an initial positive peak (when excitatory), lasting approximately 1 ms. In contrast, the post-synaptic potential features a monophasic waveform (positive when excitatory), extending for about 10 ms. The PSP emerges approximately 1 ms after the peak of the action potential. . . . . 12

1.3 Pyramidal Neurons and Dipole Generation. (A) Illustration of various cortical layers with their associated brain regions. (B) Emphasis on the unique geometry of a $5^{\text{th}}$ level pyramidal neuron, highlighting the generation of a dipole due to the distinct orientation of apical dendrites relative to the soma. Adapted from [7]. . . . . 13

1.4 Electrode Placement Systems. (**A**) In the 10-20 system, electrodes are positioned based on anatomical landmarks using a grid pattern. The electrodes are placed at specific percentages (10 % and 20%) of distances between key landmarks on the scalp, providing consistent and repeatable electrode positions. (**A**) The 10-10 system further refines electrode placement by adding additional positions, allowing for more precise spatial coverage. Reproduced from [17]. . . . . 15

1.5 Comparison of EEG Cap Montages. The figure illustrates two commonly used electrode placement configurations in EEG recordings: Bipolar (A) and Unipolar (B). The bipolar montage (A) involves pairing adjacent electrodes to measure the potential difference between them, facilitating the detection of local electrical activity and providing insights into the scalp voltage gradient. In contrast, the unipolar montage (B) pairs each electrode with a common reference electrode, capturing the individual electrical activity at each electrode site and enabling a comprehensive understanding of neural dynamics across the scalp. Reproduced from [18]. . . . . 15

1.6 PSG Sensors Representation. The figure illustrates the different sensors used in PSG recordings to monitor various physiological signals during sleep: EEG electrodes provide insights into brain activity, ECG electrodes capture heart activity, EOG sensors detect eye movements, oronasal airflow sensor monitors the passage of air through the nasal passages, EMG electrodes monitor muscle tone changes, position sensors track body posture, pulse oxymetry sensor measures blood oxygen saturation levels, thoracic and abdominal belts record chest and abdominal wall movements. Adapted from [23]. . . . . 20

1.7 Time-Frequency Representation of an EEG Signal. **(A)** shows a segment of an EEG signal captured from a single channel. **(B)** illustrates the frequency domain representation of the signal obtained through the Fourier Transform, highlighting the dominant frequency components. **(C)** displays the time-frequency representation of the EEG signal, revealing how its frequency content changes over time. Reproduced from [33]. . . . . 22

1.8 EEG Preprocessing Steps. The figure illustrates the standard preprocessing workflow for EEG data, as recommended by COBIDAS. Each step impacts the data in the time (blue boxes), space (red boxes), and/or frequency (green boxes) domains. While variations in the order of these steps are permissible based on experimental considerations or specific EEG features under investigation, any deviations should be well-justified. Reproduced from [39]. . . . . 24

1.9 Three-layer Multilayer Perceptron (MLP) architecture. This MLP comprises an input layer with four input neurons, one hidden layer, and an output layer with two outputs. Each connection between neurons represents a weighted connection, and each neuron incorporates an activation function. . . . . 28

1.10 Convolutional Neural Network (CNN) Architecture for EEG Data Analysis. This figure illustrates the CNN structure for processing EEG data. The CNN comprises three main components: the input image, where raw EEG data is provided as input; the feature extractor, responsible for automatically identifying relevant patterns within the EEG data; and the classifier, which categorizes the EEG signals into distinct classes or states. This end-to-end approach eliminates the need for manual feature extraction and enables comprehensive analysis of EEG signals for various applications in biomedical signal processing. . . . . 30

1.11 Convolutional Autoencoder for EEG Data. This figure illustrates a CAE architecture designed for EEG data processing. The CAE consists of an encoder, responsible for capturing salient features from preprocessed EEG input, a decoder for reconstructing EEG data from the learned features, and a latent representation, where the compressed information about the EEG signals is stored. Adapted from [60]. . . . . 31

1.12 Class Activation Mapping (CAM) Example. CAM applied to Australian Terrier detection displays the filters from the penultimate layer and the resultant weighted sum of their activations to identify class-specific regions. This technique leverages gradients to identify important regions within images. Reproduced from [69]. . . . . 35

1.13 Example of Partial Dependence Plot (PDP). This figure illustrates the relationship between Temperature (x-axis) and Humidity (y-axis). The PDP visualizes how changes in Temperature and Humidity affect the predicted outcome, with color intensity representing the predicted values (here, the number of bike rentals). Adapted from [70]. . . . . 35

1.14 Intuition for Local Interpretable Model-Agnostic Explanations (LIME). The left side of the figure represents a complex, black-box decision function f (shown in blue/pink) that is unknown to LIME. The bold red cross indicates the instance being explained. LIME constructs a subsample of instances by making slight alterations to the target instance, obtains predictions from f, and weighs them based on their proximity to the instance being explained (indicated by size). On the right side, a linear explanation (represented by the dashed line) is learned to approximate the complex model locally, providing a more interpretable understanding of how the model behaves in the vicinity of the explained instance. Reproduced from [75]. . . . . 36

1.15 Vision Transformer (ViT) Architecture. The ViT architecture, depicted on the left, begins by partitioning the input image into consistent patches, which are subsequently linearly transformed. Position embeddings are introduced to these embeddings, creating a sequential representation of vectors. This sequence is then processed through a conventional Transformer encoder, elaborated in detail on the right. To facilitate classification, an extra learnable classification token is incorporated within the sequence. Reproduced from [77]. . . . . 38

2.1 Visualization of potential biases in medical research. This figure classifies the potential biases into four main categories retracing each step of a medical research from the study planning to the publication. . . . . 42

3.1 examples of stimuli for each answer to the task . . . . . 54

3.2 examples of pairs with the primer on the left and the target on the right . . . . . 55

3.3 Example of a 60-second OSA trial showcasing various PSG channels including respiratory (VAB, VTH), oxygen saturation (SAO2), eye movements (EOG1, EOG2), heart rate variability (PRV), phase difference between respiratory signals (Pshift), and filtered EEG signals from three electrodes (FP1, C3, O1) across five 2Hz narrow frequency bands. . . . . 59

4.1 Framework Overview. The figure illustrates the EEG recordings in blue, the design of the model to include all the needed trials information in a standardized way in green, and the statistical analysis of the regressed ERP to identify covariates influence in red. . . . . 62

4.2 Hierarchical Analysis Procedure of the LIMO EEG toolbox. At the 1$^{st}$ level (top), individual subject data, comprising all trials, undergo analysis to compute estimated beta parameters. These beta parameters capture the effects of various experimental conditions as specified within the design matrix. At the 2$^{nd}$ level of analysis (bottom), the obtained beta parameters are scrutinized concerning the experimental conditions outlined in the 1$^{st}$ level. This involves testing for statistical significance across all subjects. Reproduced from [114]. . . . . 65

4.3 Correlation analysis of covariates. **(A)** correlation between psycho-linguistic variables, **(B)** correlation between image properties, **(C)** correlation between selected covariates. . . . . 68

4.4 First-level Analysis. The figure illustrates the three-step process encompassing Data Preparation, Variable Selection, and Linear Modeling, as adapted from the LIMO EEG toolbox to our framework. . . . . 69

4.5 Design matrices. **(A)** control model (only categories and error terms, 3 dimensions), **(B)** psycho model (13 dimensions), **(C)** image model (16 dimensions), **(D)** psycho-image model (26 dimensions). The two first columns representing the categories are coded as binary values (-1 or 1), while columns corresponding to covariates have continuous values representing the z-score computed thorough all the trials. . . . . 71

4.6 Trimmed mean (20% of trimming) of beta estimates across subjects on F6 electrode using "psycho" model. The two bold lines represent the categorical variables (manufactured and natural items) while the black dashed line belongs to the constant term. All other signals are related to the covariates (cf. legend). The arrows on the x axis show the appearance of the primer and the target images. . . . . 72

4.7 Analytical framework for identifying regions of interest supporting reliable classification and detecting covariate bias using the $2^{nd}$ level analysis of LIMO EEG. . . . . 73

4.8 Geometrical representation of the combination of the different models. Left part relates to the ideal situation where the categorical effect, the psycho covariates effect and the image covariates effect are orthogonal to each other, while right part represents the real-life case of non-orthogonality. **(A)** and **(B)**: Vectorial representations of the categorical, psycho, image and psycho-image models and of the different effects resulting from their combination, with a focus on the psycho covariates effect. **(C)** and **(D)**: The corresponding projections on the $\pi$ planes where the $R^2$ values are computed for a given data set and represented as segments in their corresponding directions. The correlation effect causing the loss in explained variance is represented in red in **D**. . . . . 76

4.9 $R^2$ combination for statistical inference. Example of the study of the effect of psycho-linguistic variables on the explained variance. . . . . 77

4.10 Statistical analysis of psycho-image model. **(A)** Trimmed mean of the explained variance ($R^2 - R^2_{naive}$) across subjects with corresponding regions of significant explained variance (red bands) and significant categorical contrast (green band). For each highlighted area, the topological view is shown (bottom). On the channel corresponding to maximum $R^2$ (PO8 electrode), the averaged ERPs of both categories (top right) and the $R^2$ time-course (bottom right) are displayed. **(B)** Trimmed mean of the categorical contrast across subjects with significant regions highlighted and the corresponding topological views (bottom). On the channel corresponding to maximum contrast (F5 electrode), the averaged contrast parameter ($\beta_{man.} - \beta_{nat.}$) is displayed. . . . . 79

4.11 Thresholded maps of the categorical contrast showing spatio-temporal regions of significant categorical contrast using a one-sample t-test followed by an MCC with spatio-temporal clustering. These regions are extracted from the categorical model **(A)**, the psycho-image model **(B)** and the naive psycho-image model **(C)**. . . . . 81

4.12 Explained variance quantization. **(A)** Trimmed mean of the explained variance ($R^2 - R^2_{naive}$) across subjects for categorical, psycho and image models. Regions of significant explained variance (red bands), significant categorical contrast when using categorical model (dark green bands) and significant categorical contrast when using psycho-image model (light green band) are highlighted. **(B)** Explained variance distribution for each of the identified categorical clusters. The gray zones within the box plots represent the confidence interval of the variance explained when using the corresponding naive model. . . . . 83

5.1 Visual Inspection. Examples of good (A) and bad (B) channels, as well as good (C) and bad (D) trials. Rejected samples typically exhibit flat signals with no useful information. . . . . 93

5.2 Ocular Artifact Reduction. (A) illustrates EEG signals affected by ocular artifacts, with the green bands indicating the segments containing artifacts. (B) presents the corrected version of the same signal after applying the MWF. Notably, the segments displayed here are not part of the manually annotated segments. The bottom two signals correspond to the horizontal and vertical EOGchannels, which serve as visual references for annotating the artifact segments but are not processed by the MWF. For clarity, only a subset of EEG channels is shown here. 95

5.3 Data Segmentation. The figure illustrates the segmented data. For clarity, we have separated each segment with a flat zero signal, displaying only a subset of EEG channels. . . . . 96

5.4 Line Noise Removal. (A) Segmented ERP data affected by line noise and (B) the filtered version using *Zapline*. The filtering process removes the 50Hz component while preserving the general signal morphology. . . . . 97

5.5 Muscle Artifact Reduction. (A) Segmented ERP data affected by a muscle artifact, highlighted with the green band. (B) The cleaned data after applying the EEMD-CCA algorithm, effectively reducing muscle artifacts. . . . . 99

6.1 Examples of coordinates system derived from fiducials . . . . . 105

6.2 Devices to record electrode positions . . . . . 106

6.3 Forward modeling process . . . . . 107

6.4 Result of the source reconstruction process . . . . . 109

6.5 Source neighboring matrix. For each region in row, neighbors are defined with a white square. . . . . 114

6.6 Example of 3-dipoles pseudo-source signals with the 2 first dipoles being on anterior regions and the 3rd once on the posterior region. (a) pseudo-ERP defined as a series of P100, N200, P300 and N400. (b) pseudo-oscillatory signal with frequency band of each dipole defined as: 8-12Hz (blue), 16-24Hz (orange), 9-13Hz (yellow). . . . . 116

6.7 Example of a 3-dipoles pseudo-EEG signal. (a) 64-channels EEG generated from a 3-dipoles pseudo-ERP signal using forward modeling. (b) timelock analysis of signals in (a). (c) final pseudo-EEG signals after having added muscle artifact to the signal in (a). . . . . 119

6.8 Example of the evaluation of a 10-sessions 3-dipoles source reconstruction of pseudo-EEG signals. A: distribution of the correctness of reconstructed regions through their relative position to the initial pseudo-regions (i.e., correct, neighbor, second neighbor or wrong position) (left) and the corresponding score statistics (right). B: mean accuracy score computed following the neighbor-based evaluation rules. C: top view of reconstructed regions (yellow) from one session in comparison with the ground truth pseudo-dipoles (red spheres) and their corresponding region (light blue). The region in red is a properly reconstructed source. . . . . 121

7.1 Example of Latent Space Exploration in VAEs on Smooth Transition Between Faces. By manipulating continuous variables in the latent space, we can transform the attributes of the same face. This enables us to age a young person gracefully or rejuvenate an elderly individual, enhance masculine or feminine features, all while retaining the original essence. Left images represent the starting point of these attribute transitions. Reproduced from [159]. . . . . 127

7.2 Introduction of a Deterministic Latent Vector $z_I$ into Standard VAE. In this comparison between the standard VAE (A) and VAE++ (B), proposed by Zhang *et al.*, x and x' represent the input and reconstructed data, while $\mu$ and $\sigma$ signify the learned expectation and standard deviation. In the standard VAE, $z_S$ is considered the learned representation, composed of $\mu$, $\sigma$, and $\epsilon$, where $\epsilon$ is randomly sampled from $\mathcal{N}(0,1)$. VAE++ introduces a deterministic latent vector $z_I$, highlighted here, which can be employed for classification purposes. Reproduced from [160]. . 129

7.3 Adversarial Autoencoder Design. The upper row depicts a standard autoencoder's reconstruction of an image from a latent code, while the lower row illustrates a second network trained to discern between hidden code and a user-specified distribution sample. Reproduced from [161]. . . . . 131

7.4 Transformation from Standard Architecture to Explainable Adversarial Auto-Encoder Network (xAAEnet). (A) Depicts the traditional structure consisting of an encoder for *feature extraction*, leading to a sparse latent space configuration maximizing the discriminability between classes, and a classifier for category prediction. (B) Illustrates the architecture required for our human-centered xAI approach, integrating a decoder for data reconstruction and a discriminator to regulate and modify the latent space distribution. This controlled latent space enables more interpretable transitions between classes. . . . . 133

7.5 t-SNE visualizations of the latent spaces generated by two architectures: (A) Resnet34 and (B) xAAEnet. Each dot represents an image of either a cat (in orange) or a dog (in blue). The sparsity and separation between classes in the Resnet34 latent space contrast with the smoother transitions observed in the xAAEnet space along the most discriminant direction (red arrow). Specific image samples are provided for highlighted regions in both visualizations, indicating the influence of rotation angles on classification. . . . . 138

7.6 Sensitivity Analysis of Rotation Angles in Classification Models. (A) & (B) Display the difference in rotation angles as a function of the distance along the most discriminant direction of the Resnet34 (A) and xAAEnet (B) latent spaces. In (A), the Resnet latent space reveals non-significant variations in angles, underscoring the challenge of making meaningful comparisons in this latent space. In contrast, (B) demonstrates that within the xAAEnet model's latent space, a consistent rise in rotation angle differences indicates the angle's role as a discriminative feature. (C) Provides a visual depiction of samples sorted along the most discriminative axis, elucidating the method for determining normalized distances. . . . . 140

8.1 xVAEnet Architecture. The model is composed of 3 parts: a VAE, a GAN and a classifier, all of them making use of the convolutional encoder that encodes the input data into an embedding, the encoder latent space ($Z_e$). The **VAE** (center) first estimates the mean ($\mu$) and the standard deviation ($\sigma$) of the dataset distribution from $Z_e$ using dense layers to obtain the decoder latent space ($Z_d$), then $Z_d$ is decoded to derive a reconstructed version of the input data using a deconvolutional decoder. The **GAN** (top) exploits the encoder as a generator and discriminates $Z_d$ (real distribution) from $Z_e$ (fake distribution) using an MLP discriminator. The **Classifier** (bottom) uses the features extracted by the encoder in $Z_e$ to classify the desaturation area, the arousal events and the respiratory event duration with a unique single-layer perceptron. . . . . 147

8.2 xAAEnet Architecture adapted to OSA severity scoring. The model is composed of 3 parts: an AE, a GAN and regressors. A shared convolutional encoder encodes input data into a latent representation $Z$. The **AE** (encoder + decoder) decodes $Z$ to derive a reconstructed version of the input data using a deconvolutional decoder. The **GAN** (encoder + discriminator) exploits the encoder as a generator and discriminates $Z$ (fake distribution) from a random Gaussian sampled batch $Z_{\mathcal{N}}$ (real distribution) using an MLP discriminator. The **regressors** (encoder + SLP) use the features extracted by the encoder in $Z$ to predict the value of the hand-made score $S_h$, the desaturation area, the respiratory event duration, or to detect the presence of arousal events, with a SLP. . . . . 152

8.3 2D representations of the encoder latent space ($Z_e$) using t-SNE. Each sample represents one of the 6992 OSA trials. (**A**) $Z_e$ with the training phase of the VAE module completed. (**B**) $Z_e$ with the training phase of the GAN module completed. (**C**) $Z_e$ with the training phase of the classifier module completed on every severity metrics. The arrow represents the severity direction obtained using LDA. In the legend, each letter of "**had**" represents a severity feature: "**h**ypoxic burden", "**a**rousal event", and "**d**uration of the respiratory event". The "L" means "Low-level severity", the "H" means "High-level severity" . . . . . 159

8.4 Biomarkers identification performed by comparing the power signal, by channel, of the OSA trials sorted by severity score ($\in$ [0,1]) along the severity direction obtained using LDA. (**A**) Mean power difference across OSA trials obtained by subtracting, for each channel separately, the power signal of each trial from the power signal of trials of higher severity scores. (**B**) Channel-by-channel mean power difference of PSG channels excluding EEG channels. The x axis represents the distance, along the severity direction, between the trials being compared. A distance of 0 means a trial is compared to itself, a distance of 1 means the comparison between the trial of lowest severity score and the trial of highest severity score. (**C**) Time window-by-time window mean power difference of the SAO2 channel (channel of highest absolute mean power difference). (**D**) Channel-by-channel mean power difference of EEG channels. (**E**) Time window-by-time window mean power difference of the C3 channel on the 4-6Hz frequency band (channel of highest absolute mean power difference). . . . . 161

8.5 Comparison of the models' performance. **(A)** 2D representation, using t-SNE, of the latent space $Z$ obtained with the different networks. Each point represents an OSA trial and the color indicates the corresponding $S_h$ score. The arrow gives the severity direction. **(B)** Hand-made score distribution along the latent severity scale. Each bar represents the mean $S_h$ of the OSA trials in a specific latent severity scale range. The mean $S_h$ values have been normalized in [-1,1] for comparison purpose. 166

8.6 Severity score sensitivity comparison. The effect of an increasing severity score on input signals power is compared for severity defined as $S_E$ (**left**) vs. $S_h$ (**right**). For readability, this comparison is done separately for non-EEG PSG channels (A to D) and EEG channels (E to H). The x axis represents the distance, along the corresponding severity scale, between the trials being compared. A distance of 0 means a trial is compared to itself, a distance of 1 means the comparison between the trial of lowest severity score and the trial of highest severity score. The y axis represents the mean power difference of the PSG signals across OSA trials obtained by subtracting, for each channel/time-window separately, the power signal of each trial from the power signal of trials of higher severity scores. The figure includes the channel-by-channel mean power difference of (**A,B**) non-EEG PSG channels and (**E,F**) filtered EEG channels, as well as the time window-by-time window mean power difference of (**C,D**) the SAO2 channel and (**G,H**) the C3 channel on the 4-6Hz frequency band. The black boxes highlight the relevant signals for the comparison between $S_h$ and $S_E$. . . . . 170

8.7 Sensitivity Analysis along the Perpendicular Direction. (A) Depicts a 2D representation of xAAEnet latent space using the t-SNE transform, with the perpendicular direction marked by a red arrow. (B) Shows the impact of increasing the distance along this direction on PSG signal power. (C) Demonstrates the effect on EEG signal power with increasing distance along this direction. . . . . 171

8.8 Multimodal xAAEnet Architecture encompassing patient information. The modified block is the encoder, which now consists of a PSG encoder and embedding layer that output independent latent vectors ($Z_{PSG}$ and $Z_{embed}$). These vectors are then concatenated and process through a residual fully-connected layer, leading to the final $Z$ latent representation that is processed the same way as the unimodal version of the model, described in Section 8.2.2. . . . . 172

# List of Tables

3.1 Summary of Trial Categories . . . . . 53

4.1 Caption for LOF . . . . . 67

4.2 95% confidence intervals of the explained variance of the categorical model compared with the part of the explained variance that is lost in psycho and image models due to the correlation between the covariates and the categories. . . . . 84

8.1 xVAEnet architecture details . . . . . 148

8.2 xAAEnet architecture details . . . . . 153

8.3 Comparison of Kendall Tau Distance and Ordinal Error for Different Models Using the LOSO Method. Three training sessions were conducted, each with specific test patients: 1) patients with the lowest mean $S_h$ score across trials, 2) patients with the highest mean $S_h$ score, 3) patients with mean $S_h$ scores closest to the global median. Kendall Tau distances and ordinal error values are reported with standard deviations. . . . . 168

8.4 Patient Information Comparison Across Severity Quartiles. . . 173