UNIVERSIDAD DE ANTIOQUIA

Master of Engineering (Bioengineering)

Graduate Program Faculty of Engineering

Machine Learning model for the classification of individuals at risk of dementia type Alzheimer from multimodal databases of EEG and clinical information

By

Verónica Henao Isaza

******

Advisor:

Ph.D. John Fredy Ochoa Gomez

Universidad de Antioquia

2023

Citation	(Henao Isaza V, 2023)
Reference **APA 7 (2020)**	Henao Isaza V. (2023). Machine Learning model for the classification of individuals at risk of type dementia Alzheimer from multimodal databases of EEG and clinical information. Master's report. Universidad de Antioquia, Medellín, Colombia.

creative commons CC BY NC SA

Investigation Group Neuropsicología y Conducta (GRUNECO).
Director: Carlos Andrés Tobón Quintero

UNIVERSIDAD DE ANTIOQUIA
Vicerrectoría de Docencia

Sistema de Bibliotecas

Universidad de Antioquia - www.udea.edu.co

Chancellor: John Jairo Arboleda Céspedes

Dean/Director: Julio César Saldarriaga

Head of department: John Fredy Ochoa Gómez.

The content of this work corresponds to the right of expression of the authors and does not compromise the institutional thought of the University of Antioquia nor does it unleash its responsibility towards third parties. The authors assume responsibility for copyright and related rights.

Declaration

I hereby declare that the contents and organization of this dissertation constitute my own original work and do not compromise in any way the rights of third parties, including those relating to the security of personal data.

Verónica Henao Isaza

2023

* This dissertation is presented in partial fulfillment of the requirements for an **M.Sc. degree in Engineering** at the Graduate School of Universidad de Antioquia.

Acknowledgment

I am appreciative of my advisor for his unwavering support throughout the development of this project. His guidance and understanding during the dedicated time towards achieving my objectives of international collaboration have been invaluable. Not only he has facilitated my professional growth, but he has also played a significant role in my personal development.

I would also like to thank my colleagues and friends in the laboratory for their invaluable support at various stages of the project. Their help was instrumental in overcoming challenges and ensuring the success of this endeavor. Their collective expertise and willingness to lend a helping hand have truly made a difference in my research journey.

Abstract

Alzheimer's disease (AD) poses a significant challenge in Colombia due to the growing aging population. Detecting early signs of cognitive alterations is crucial, and electroencephalography (EEG) has emerged as a valuable tool for studying AD-related brain activity. However, challenges exist in obtaining comparable and high-quality EEG recordings. Standardized data preprocessing pipelines and harmonization efforts, such as the Brain Imaging Data Structure (BIDS) format, play a vital role in facilitating data integration and sharing. The project focused on organizing multi-site EEG data using the EEG-BIDS framework, promoting localization, accessibility, and interoperability. Open-access databases were utilized to investigate the generalizability of EEG and machine learning (ML) analysis, highlighting the need for data standardization and harmonization. A processing pipeline (Sovaharmony) with normalization and harmonization stages enabled the integration of diverse cohorts (datasets) and optimization of information extraction.

Machine learning models were employed for AD risk classification using non-invasive EEG biomarkers. Harmonization of data from multiple cohorts was crucial for increasing sample size, improving statistical power, and identifying consistent features or biomarkers across cohorts. The project aimed to develop a robust and generalizable machine learning model by harmonizing cohorts using a larger and more diverse dataset and thereby improving accuracy.

This project made significant contributions to dementia research by developing a comprehensive approach for data acquisition, processing, harmonization, and machine learning-based risk classification using EEG technology. The standardized pipelines, data harmonization, and machine learning techniques were emphasized as critical components in advancing AD research and maximizing the value of EEG data. Further research should focus on replicating the findings on larger cohorts, using techniques like the introduced in the current project, and exploring the application of machine learning models to other non-invasive biomarkers, ultimately validating the accuracy and reliability of AD classification.

Table of Contents

List of relevant publications and works.

1. **Longitudinal analysis of qEEG in subjects with autosomal dominant Alzheimer's disease due to PSEN1-E280A variant.**

   Aguillon, D., Guerrero, A., Vasquez, D., Cadavid, V., Henao, V., Suarez, X., ... & Ochoa, J. F. (2023, July). Longitudinal analysis of qEEG in subjects with autosomal dominant Alzheimer's disease due to PSEN1-E280A variant. In Alzheimer's Association International Conference. ALZ.

2. **Tackling EEG test-retest reliability with a pre-processing pipeline based on ICA and wavelet-ICA.**

   Henao Isaza V, Cadavid Castro V, Zapata Saldarriaga L, Mantilla-Ramos Y, Tobón Quintero C, Suarez Revelo J, Ochoa Gómez J. 2023 Jun; Authorea. DOI: 10.22541/au.168570191.12788016/v1

   Under review: **Biomedical Signal Processing and Control, Q1 , SJR 2022**

3. **Spectral features of resting-state EEG in Parkinson's Disease: A multicenter study using functional data analysis.**

   Jaramillo-Jiménez A, Tovar-Ríos DA, Ospina JA, et al. Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology. 2023 Apr; 151:28-40. DOI: 10.1016/j.clinph.2023.03.363. PMID: 37146531

   Clinical Neurophysiology, Q1, SJR 2022

4. **Ongoing EEG alpha rhythms reflect the abnormal wake-light sleep transitions in patients with Alzheimer’s disease mild cognitive impairment.**

Internship (from 11/07/2022 to 07/12/ 2022) at the Laboratory of " Neurosciences of Human Higher Functions" with Ph.D. Claudio Babiloni, located at the Department of Physiology and Pharmacology "Vittorio Erspamer" of the Sapienza University of Rome.

5. **Reproducible Neuronal Components found using Group Independent Component Analysis in Resting State Electroencephalographic Data**

Ochoa-Gomez, J. F., Mantilla Ramos, Y. J., Henao Isaza, V., Tobon, C. A., Lopera, F., Aguillon, D., & Suarez Revelo, J. X. (2023). Reproducible Neuronal Components found using Group Independent Component Analysis in Resting State Electroencephalographic Data. bioRxiv, 2023-11. DOI: 10.1101/2023.11.14.566952

Under review: **Brain Topography**

List of Figures

List of Tables

Abbreviation list

AD – Alzheimer disease

APP - Precursor protein

BIDS - Brain Imaging Data Structure

CHBMP - Cohort available thanks to the Project: Cuban Human Brain Mapping Project

Control - Healthy controls who voluntarily participated.

DCL and MCI - Mild Cognitive Impairment

DFT - The discrete Fourier transform

DTA - Alzheimer's disease patients that carry the PSEN1-E280A variant.

EEG – Electroencephalography

FDA - Food and Drug Administration

FFT - Fast Fourier Transform

fMRI - functional Magnetic Resonance Imaging (FMRI)

G1 - Healthy subjects that carry the PSEN1-E280A variant.

G2 - Control group of healthy subjects for G1

ICA - Independent Component Analysis

ML - Machine Learning

MMSE - Mini Mental State Examination

MoCA - Montreal Cognitive Assessment

MRI - Magnetic Resonance Imaging.

PSEN1 - Presenilin 1

PSEN1-E280A – Presenilin 1 with genetic variant E280A

PSEN2 – Presenilin 2

SRM - Cohort available thanks to the Project: Stimulus-Selective Response Modulation.

STFT - Short-Time Fourier Transform.

UdeA 1 - Cohort available thanks to the Project "Cambios en los patrones del electroencefalograma cuantitativo (reactividad alfa, theta y su índice) en reposo y tareas de memoria, en el seguimiento longitudinal de pacientes con riesgo genético para Enfermedad de Alzheimer Temprano"

UdeA 2 - Cohort available thanks to the Project “Identificación de marcadores preclínicos de la mutación E280A de la enfermedad de Alzheimer a partir de medidas de conectividad en EEG”

VFT - Verbal Fluency Test

WT - Wavelet Transform

Chapter 1

Relevance

1.1 Introduction

In Colombia, a person is considered an older adult from the age of 60 [1] and by 2021 it was estimated that there were more than 6 million older adults living there, which represents 13.3% of the population [2], with an average use of services significantly higher compared to the general population. Additionally, aging represents a relevant risk factor for the development of cognitive alterations [3]. Alzheimer's dementia (AD) is the most prevalent neurodegenerative disorder, accounting for more than 50% of all cases of dementia and affecting approximately 30% of all individuals over 85 years of age [4]. However, there is evidence that the pathophysiological processes in AD begin decades before the manifestation of clinical symptoms [5].

Electroencephalography (EEG) is one of the most important techniques for the study of brain electrical activity. It represents a non-invasive technology to study brain function and neurophysiological changes associated with AD [6]. The simplest and most common way to acquire EEG is to record spontaneous brain activity while the subject is in a resting state, with eyes open or closed, and this makes EEG recorded during rest highly reliable [7]. This project assumes that the current paucity EEG measures in biomarker studies are not due to a lack of

information about neuronal processes that can be gained from EEG, instead by the lack of access to information in a massive way [8]–[10] .

In terms of cost, if EEG-based biomarkers are identified, the financial burden of implementing widespread screening for such markers will be low compared to magnetic resonance imaging (MRI)-based screening [11]. In addition, a challenge that has hindered the large-scale application of EEG is the difficulty in obtaining EEG recordings of comparable quality between subjects [12]. The difficulty of obtaining these records is due to the EEG signal being influenced by technical factors and by features of the recorded subject. Some of the factors that have been detected are temperature and air humidity [13], factors that interact with sources of noise, such as line frequency or other sources of electromagnetic noise [14], and subject-related artifacts, typically reflecting unwanted physiological signals (such as eye movements, eye blinks, muscular noise, heart signals and sweating) [15], may differ from subject to subject and may interact in a complex manner with non-physiological artifacts. Due to the characterization of the artifacts mentioned, efforts have been focused on creating preprocessing pipelines for artifact cleaning or artifact correction [16].

In this context, machine learning (ML) models have emerged as valuable tools for leveraging the immense potential of EEG data for early AD detection [17]. These models can sift through vast amounts of data to uncover patterns and relationships that might not be immediately apparent through traditional analysis methods [18].

In the scope of this project, individuals at risk of Alzheimer's Disease (AD) pertain to those with a specific genetic variant, PSEN1-E280A [19]. The primary goal of this research is to develop an accurate and reliable machine learning (ML) model that can effectively classify individuals at risk of AD using non-invasive biomarkers extracted from multiple databases.

To achieve this objective, the project involves creating and organizing a comprehensive database by integrating diverse information sources. This process will be facilitated by employing efficient data management tools. The database will then undergo harmonization, focusing on crucial electrophysiological and clinical parameters, which will be accomplished through advanced biomedical data processing techniques. Finally, a state-of-the-art machine learning model will be designed to leverage the structured database.

This undertaking necessitates the normalization and harmonization of data. Normalization refers to the process of transforming variables to a standardized scale, facilitating data comparison and analysis [20]. Harmonization, on the other hand, involves ensuring consistent and comparable data across various sources, thereby reducing variability caused by technical and subject-related factors [21]. In the context of EEG data, normalization and harmonization ensure that data from different subjects and sources are treated uniformly, enhancing the reliability and generalizability of the ML model [22].

As part of the research methodology, classification models come into play to effectively discern patterns and associations in the EEG data that indicate potential

AD risk. By training the ML model on a diverse dataset that has been meticulously harmonized and normalized, the goal is to achieve accurate classification of individuals at risk of AD.

1.2 Problem Description

Achieving preprocessing pipelines that can be used in different databases or in longitudinal studies, preserving the processing configuration, and making the information processed (at the same time or years later) comparable would make a standard procedure based in EEG possible. In addition to the urge of standardized preprocessing, there is a need for a format to organize, harmonize, and share data.

In the recent years, EEG datasets have been made increasingly openly available, some of them can be found on repositories like GitHub [23] or OpenNeuro [24] and it has been shown that integrating EEG datasets across studies offers unique insights [25]. The Brain Imaging Data Structure (BIDS) is at this moment the principled way of data sharing that has been successfully adopted in the functional Magnetic Resonance Imaging (fMRI) data [26] and various extensions of the BIDS format (including extensions for EEG data [27]) have been proposed that not only provide a standard for the respective data modality but moreover facilitate the integration between data of different modalities (e.g. simultaneous fMRI and EEG recordings) and also neuropsychological test data [16]. To make EEG data sharing simple and intuitive, it is beneficial that preprocessing pipeline supports BIDS format as in- and output.

In the current landscape, the pursuit of biomarkers encounters distinct challenges. The process of diagnosing Alzheimer's Disease (AD) through neurological examinations and medical record reviews is time-intensive and subject to inconsistencies, necessitating skilled clinicians and protracted assessments. To surmount these challenges, the prominence of developing and utilizing biomarkers has risen, offering an objective and efficient avenue for AD diagnosis. A biomarker in the context of AD refers to a quantifiable biological trait that reflects normal or pathological brain activity [28]. As EEG signals capture functional alterations in the cerebral cortex, they hold the potential to assess neuronal degeneration linked to AD progression prior to discernible tissue loss or behavioral manifestations [5]. Incorporating machine learning models, such as analyzing complex patterns within EEG data, can enable the identification of subtle patterns associated with late-life cognitive decline, such as Mild Cognitive Impairment (MCI). These models can play a pivotal role in uncovering non-linear relationships between EEG signals and early cognitive deterioration symptoms, potentially leading to early and precise detection of conditions like MCI and AD, thus offering opportunities for more effective interventions and treatments [29].

To tackle the issue of sample size and demographic disparities, we need to explore normalization and harmonization techniques that enhance model performance. With data organized and processing streamlined, the next step is feature extraction. This sets the stage for using machine learning to train classifiers for Alzheimer's risk prediction [30].

EEG studies face data variability from various sources, hindering comparability and biomarker accuracy. Implementing effective strategies is crucial for reliable EEG biomarkers in cognitive decline research. Among these sources, the amplitude of derivatives measurements of EEG is key [22]. It's influenced by various factors and directly impacts extracted features, which fuel subsequent analysis. Standardizing amplitude across acquisition settings is vital for consistency [22]. Advanced preprocessing techniques addressing amplitude differences and ensuring accurate scaling enhance EEG biomarker validity [31]. This approach strengthens biomarker development and overall EEG study integrity.

1.3 Justification

Several limitations are mentioned in articles discussing EEG studies of early Alzheimer's disease, for example, the small sample size of the cohort [32] and mismatched demographic variables could lead to inconclusive results [33], making the generalization of the model unrealistic. Other limitations include the non-recruitment of participants with severe AD in resting-state experiments and the negative effect of a small number of electrodes on spatial resolution in source localization studies [34]. In addition, manual selection of clean EEG epochs may introduce human bias and limit reproducibility [35].

To address one of the main issues mentioned above regarding the sample size and demographic variables of the cohorts, it is necessary to delve into normalization and harmonization methods dedicated to improving the models.

Now that the data is organized and a unique processing pipeline is in place, the next step is to consider the feature extraction derived from the processing, in this way it is possible to use machine learning (ML) techniques to train a classifier to recognize the best features and improve the accuracy of the models for classifying subjects at risk of Alzheimer's disease.

1.4 Hypothesis

The resulting machine learning model is expected to produce an output with a higher accuracy than the state of the art. By incorporating more data and applying additional preprocessing steps.

Understand that harmonization is primarily about extracting information by using libraries that facilitate data processing, normalization, and enhancement while effectively managing the variables present in the records.

Figure 1 illustrates a comparison between two pipelines of the processing steps, along with the anticipated outcomes in terms of the accuracy of the classification

model. It highlights the evolution of the processing pipeline, showcasing the modifications and improvements proposed to enhance the accuracy of the model.

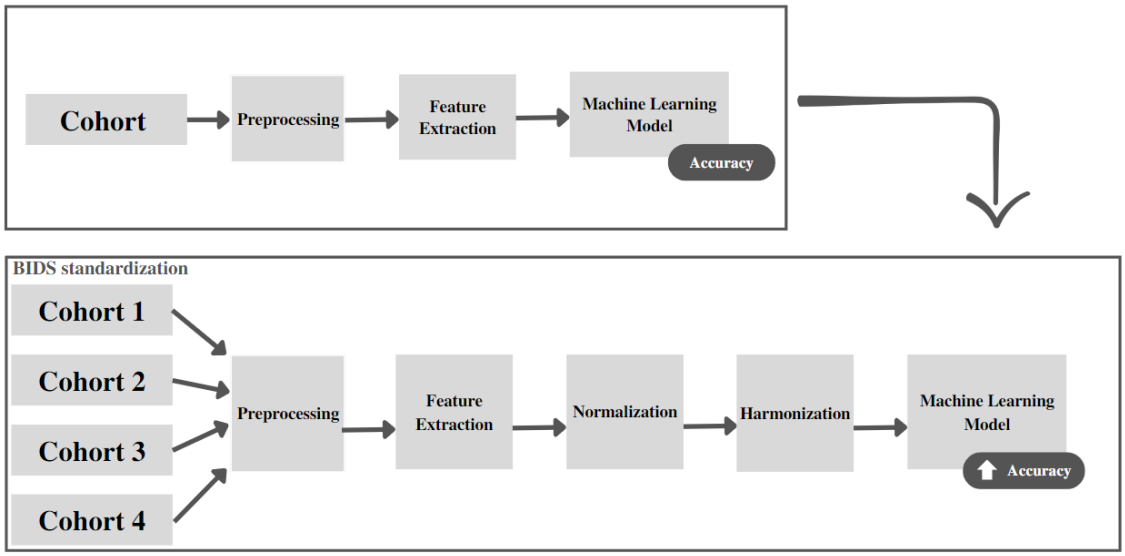


Figure 1 Proposed Hypothesis Scheme. The inclusion of normalization and harmonization steps is shown, with an arrow in the "Accuracy" text symbolizing the hypothesis that the addition of these steps will lead to improved accuracy.

Based on the hypothesis that harmonizing different electroencephalogram (EEG) databases will result in a large enough database to train a reliable machine learning (ML) model for the classification of subjects at risk of Alzheimer's Disease (AD), the following research question is generated:

“What is the effectiveness of harmonizing different electroencephalogram (EEG) databases in generating a large enough database for training a reliable machine learning (ML) model for the classification of subjects at risk of Alzheimer's Disease (AD)?”

1.5 Objectives

1.5.1 General objective

The objective of this work is to develop an accurate and reliable machine learning (ML) model that can effectively classify subjects at risk of Alzheimer's Disease (AD) using non-invasive biomarkers extracted from multiple databases.

“To build an ML model that allows classifying subjects at risk of AD using non-invasive biomarkers from multiple databases.”

1.5.2 Specific objectives

1. **To build and standardize a database with multimodal information, taking multisite databases, using tools that facilitate data storage and manipulation before and during processing.**

Chapter 2, "Database Construction and Standardization," focuses on the objective of developing a comprehensive database with multimodal information, which is critical for building an accurate and reliable machine learning (ML) model. In this chapter, we describe the methodology used to select the databases, providing a detailed description of each data source, and discuss the methods utilized to standardize and harmonize the data. Through this chapter, we demonstrate our commitment to ensuring the quality and consistency of the data used in our study, and how we can overcome the challenges posed by using multiple data sources.

2. **To harmonize the database to obtain comparable relevant electrophysiological and clinical parameters among healthy subjects using biomedical data processing techniques.**

Chapter 3, "Processing Pipeline and Harmonization" focuses on adding two additional steps to the processing pipeline previously reported in our laboratory and automating the execution of the pipeline on the comprehensive database resulting from the first objective. The primary focus of this chapter is on the critical steps of harmonization and processing that are necessary to ensure the accuracy and reliability of the data utilized in the machine learning (ML) model. Through this chapter, we aim to provide a comprehensive and transparent description of our methods, enabling replication and validation of our findings by other researchers.

3. **To design a machine learning (ML) model that, using the database built with neuropsychological and neurophysiological information, allows the classification of subjects at risk of AD.**

Chapter 4, "Machine Learning Model" represents a significant milestone in this work, as we aim to design and implement a machine learning (ML) model that can accurately classify subjects at risk of Alzheimer's Disease (AD) using the comprehensive database developed in previous chapters. In this chapter, we present the methodology and technical details of our ML model, including the feature selection process, model architecture, and evaluation metrics. We also discuss the results of our experiments and validate the effectiveness of our model in accurately

identifying subjects at risk of AD. Through this chapter, we hope to make a significant contribution to the field of AD research and aid in the development of effective early detection and intervention strategies.

1.6 Theoretical Framework

1.6.1 Electroencephalography

Electroencephalography (EEG) is a non-invasive method of measuring the electrical activity of the brain. Electrodes are placed on the scalp to record electrical activity produced by populations of brain cells called neurons. When neurons are activated, they generate time-varying electrical currents [36].

Since the first measurements by Hans Berger, we have known that the brain produces rhythmic electrophysiological activity that can be measured by EEG [37]. This has led to a large knowledge about the types of rhythmic activity that can be recorded, the circumstances under which they occur. Circumstances in which brain rhythms can occur include spontaneous activity and related or evoked events. This spontaneous activity is studied as activity at rest (resting-state).

The study of resting-state brain activity becomes particularly interesting if neural processes are view as primarily intrinsic - the weighting, gating, and subsequent integration of new and external information into the brain - as opposed to a more absolute resting state that contrasts with momentary activity driven by external demands [38]. Unless one creates a contextual setting in which 'rest' is defined [39], paradigmatic repetitive stimulation precludes rest. This suggests that the method of

choice is to analyze ongoing spontaneous brain activity rather than averaged or induced brain activity.

Recording of EEG rhythms is an experiment on brain neurophysiological mechanisms underpinning the control and maintenance of cerebral arousal [9].

In different studies the bands have been segmented a little more for different purposes in this project the segmentation used comes from in pharmaco-EEG studies [9]. These frequency bands have been defined based on factorial analysis of EEG recordings and therefore provide a very robust framework. It does not mean that other frequency ranges should not be used for specific purposes [40](Figure 2).

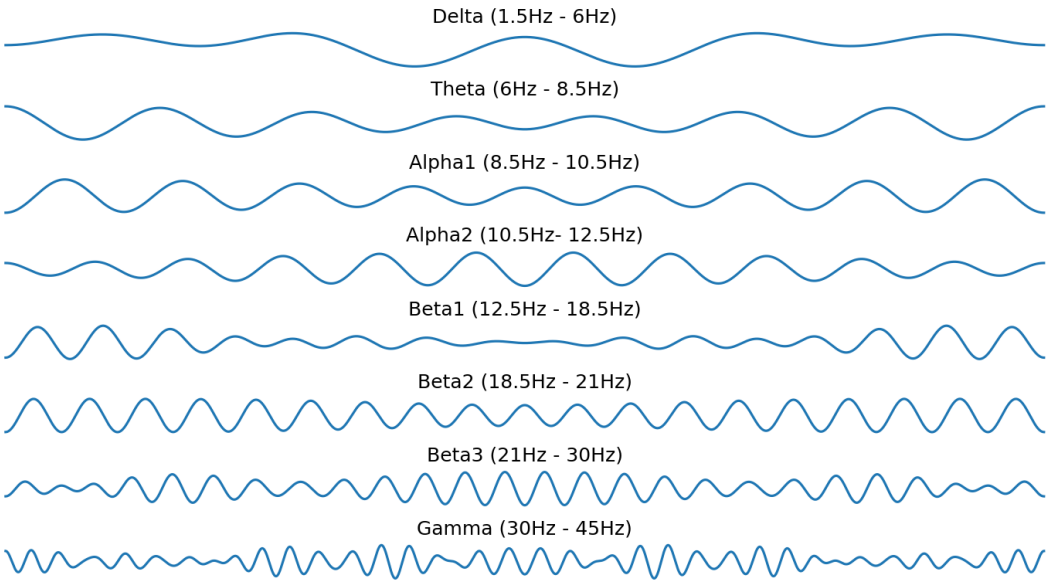


Figure 2 Frequency Ranges of EEG Waves Proposed by [40], These frequency bands have been defined on the basis of factor analysis of EEG recordings and therefore provide a very robust framework to ensure that the results of a study can be compared with other published studies and thus provide reference material useful to other scientists.

Delta waves lie within the range of 1.5–6 Hz: These waves are primarily associated with deep sleep and may be present during wakefulness [41]. Theta waves are in the range of 6-8.5 Hz [42]. Alpha waves appear in the back of the head and are usually found over the occipital region of the brain. They can be detected in all parts of the posterior lobes of the brain. Alpha1 waves have a frequency of 8.5-10.5 Hz and Alpha2 waves have a frequency of 10.5-12.5 Hz and usually appear as a round or sinusoidal signal [43]. A Beta wave is the electrical activity of the brain that can be divided into beta1 waves with a frequency of 12.5-18.5 Hz, beta2 waves with a frequency of 18.5-21 Hz, and beta3 waves with a frequency of 21-30 Hz [44]. A beta wave is the normal waking rhythm of the brain associated with active thinking, active attention, focusing on the outside world or solving concrete problems and is found in normal adults [45]. The frequencies above 30 Hz (mainly up to 45 Hz) correspond to the gamma range (sometimes called the fast beta wave). Although the amplitudes of these rhythms are very small and their occurrence is rare, the detection of these rhythms can be used to confirm certain brain diseases [46].

1.6.2 Neurodegenerative diseases

Neurodegenerative diseases such as Parkinson's disease (PD) Alzheimer's disease (AD) amyotrophic lateral sclerosis (ALS) and Huntington's disease (HD) affect millions of people worldwide [47].

Dementia is a syndrome that consists of a decline in intellectual and cognitive abilities. This consequently affects normal social activities and relationships and

interaction with other people. According to the World Health Organization, AD accounts for 60-70 percent of senile dementia characterized by severe cognitive decline, and the neuronal death [48].

AD is the primary cause of dementia globally and is characterized by the abnormal accumulation of beta-amyloid (Aβ) protein and hyperphosphorylated tau protein [49]. Pathogenic genetic variants of complete penetrance in genes such as amyloid precursor protein (APP), presenilin 1 (PSEN1), and presenilin 2 (PSEN2) are responsible for 5-10% of early-onset AD cases, with pathological genetic variants in PSEN1 being the most prevalent cause of familial Alzheimer's disease [50].

AD is thought to initiate its pathological process up to two decades prior to the onset of noticeable clinical symptoms [51]. However, in the past 20 years, it has become clear that there is not always a direct relationship between the pathology of the disease and the clinical symptoms experienced by patients [52]. Instead, the pathology and clinical symptoms of Alzheimer's are better understood as separate continuums that may evolve independently but with a temporal delay [53]. As a result, AD is currently perceived as a gradual continuum rather than a series of distinct stages.

The degenerative brain disorder of AD starts with progressive memory loss, and the loss of cholinergic cells in the basal forebrain is responsible for its first stage of development [54]. The cholinergic hypothesis of AD suggests that cognitive decline in patients results from a deficiency in cholinergic neurotransmission [55].

The Neurosciences Group of Antioquia has been studying an extended family with the genetic variant PSEN1-E280A for 30 years [56]. The variant has almost 100% penetrance, with an amnestic presentation and an onset age of dementia at 49 years [57].

Synaptic dysfunction is a pathophysiological event that impacts neuronal connections at various levels, including molecular, cellular, brain networks, and cerebral cortex, among others [58].

The gold standard for AD diagnosis is the Amyloid/Tau/Neurodegeneration (ATN) framework proposed by the National Institute on Aging and the Alzheimer's Association in 2018 [59]. In the ATN framework, the biological state of AD is classified by identifying three biomarkers (i.e., amyloid, tau, and neurodegeneration) measured from cerebrospinal fluid (CSF) and positron emission tomography (PET) imaging [60]. However, this approach is typically performed by lumbar puncture or PET, which is costly, invasive, and highly dependent on clinical infrastructure, severely limiting its availability in clinical practice [61].

The term "biomarker" (an acronym for "biological marker") is used in this project to refer to a broad subcategory of medical signs, i.e., objective indications of a patient's externally observed medical status that can be accurately and reproducibly measured. Thus, a biomarker is defined as "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention" as defined in

1998 by the Biomarker Definitions Working Group of the National Institutes of Health [62].

Electrophysiological data, such as EEG, provide valuable insights into the brain's activity [63]. However, to gain a more comprehensive understanding of a patient's condition and disease progression, it is often desirable to complement this data with neuropsychological tests [64]. These tests serve as essential tools for healthcare professionals, enabling them to interpret the electrophysiological data within the broader context of cognitive functioning and cognitive decline in patients [65]. By combining electrophysiological data with neuropsychological assessments, a more nuanced and accurate assessment of the patient's condition can be achieved.

1.6.2.1 Neuropsychological tests

The European Federation of the Neurological Societies (EFNS) also developed a guideline to diagnose and monitor AD [66]. The most used test to measure cognitive ability for AD diagnosis is the Mini Mental State Examination (MMSE) [67]. The Montreal Cognitive Assessment (MoCA) [68] and Addenbrooke's Cognitive Examination revised (ACE-R) [69].

The evaluation of the mental state is crucial in assessing psychiatric patients. To supplement the standard examination, many investigators have incorporated quantitative assessment of cognitive performance, documenting the reliability and validity of various "clinical tests of the sensorium" [70]. However, the available batteries for cognitive assessment are often lengthy, which can be problematic for

elderly patients, particularly those with delirium or dementia syndromes, who may only cooperate for short periods of time. To address this issue, the "Mini-Mental State Examination" (MMSE) was created as a scored form of the cognitive mental status examination. The MMSE includes only 11 questions and requires only 5-10 minutes to administer, making it practical to use serially and routinely [71].

Another widely used test is the verbal fluency test (VFT) [72], this is a useful tool to assess frontal lobe function and semantic memory, as it measures the ability to generate examples in different categories, which depends on the integrity of the semantic network, efficient retrieval, and organization. VFT has been widely used to evaluate various psychiatric and neurological disorders [73], [74].

It should be noted that there is uncertainty in AD diagnosis when using MMSE and other neuropsychological tests [75]. While several studies measure the classification accuracy between AD and healthy controls using the results from these tests, neuropsychological tests cannot provide a 100% certain diagnosis.

In particular, the use of approaches based on resting-state EEG and neuropsychological test could be beneficial in neurology or even primary care [76]. To achieve this kind of support in Alzheimer's disease detection, it is necessary to generate a reliable processing pipeline that can provide information about the characteristics of the signals and their relationship to the disease [77].

1.6.3 Multi-site database harmonization (Cohorts)

Harmonization encompasses the development of pipelines aimed at integrating neurophysiological databases originating from diverse cohorts [78] try to solve the integration problems discussed in this section 1.6.3. Its primary objective is to optimize information extraction through the utilization of purpose-built libraries. These libraries facilitate data processing, normalization, and enhancement by effectively managing variables present within the records.

Integrating multiple cohorts poses challenges beyond technological aspects. The diversity in data content introduces additional complexities, as the same medical procedure can be described and conceptualized differently across countries, institutions, and even studies. Although there are guidelines to assist in the design of clinical studies, they often overlook the technological aspects. Consequently, the lack of harmonization in data structure and clinical concepts becomes a major obstacle to health data sharing, significantly delaying or even preventing multi-cohort analysis. Recognizing the potential impact of these studies, researchers are driven to seek more robust and reusable solutions for aggregating knowledge from distributed health datasets [78]. This motivation has led to the establishment of organizations and the development of new methodologies for exploring clinical databases.

The Cuban Human Brain Mapping Project (CHBMP) was developed through multiple stages, with its initial phase focused on establishing norms (means and

standard deviations) for narrow-band (NB) log-spectral DP based on a dataset of 211 individuals aged 5 to 97 years from a single country [79]. However, the relatively small sample size may limit the statistical power of comparing data from different countries, particularly when compared to larger-scale neuroimaging efforts such as ENIGMA [80].

The Dementia ConnEEGtome project is a crucial study that aims to improve the reliability and validity of EEG data in dementia research by harmonizing EEG connectivity measurements across multiple centers [81]. This project's focus on multicenter harmonization addresses the lack of consistency in data collection and analysis methods, which is a significant challenge in dementia research [78].

Establishing harmonization in multinational EEG standards poses a more formidable challenge compared to MRI due to the considerable variability in recording systems across different manufacturers, further compounded by the lack of standardized protocols [82]–[84]. Differences in amplifier transfer functions, electrode placement systems, and preprocessing methods give rise to concerns about the presence of EEG batch effects [85]. Such sources of variability may also emerge from different conceptual frameworks employed in quantifying EEG connectivity, including various connectivity metrics and methodological procedures [86].

Addressing these challenges is critical to achieving robust and standardized multinational EEG standards and requires the development of effective strategies for minimizing the impact of these sources of variability.

1.6.3.1 BIDS format

The Brain Imaging Data Structure (BIDS) is a standardized format for organizing both data and metadata generated by neuroimaging experiments [26] BIDS has gained popularity within the EEG community in recent years [27] as it facilitates the sharing and reuse of data. Although converting EEG data to BIDS is not technically complex, it is a time-consuming task when performed manually [87]. Existing software solutions for automated conversion of EEG to BIDS require either programming skills or extensive user input [88]. One of the activities necessary to meet the objectives of this study's project is the development of user-friendly software that automates the process of converting EEG data into BIDS, see Annex 1.

1.6.3.2 Data Normalization (Record-specific constant)

The analysis of EEG data is highly customizable, allowing research teams to adopt their own processing strategies. However, when combining samples across centers, dataset variability must be considered. Efforts have been made to enable the joint connectivity analysis of raw data from different multicentric studies, and harmonization of raw EEG data has proven to be essential in eliminating technical and methodological sources of variability that impact the interpretation of EEG meta-analysis. Based on previous studies, it is proposed that between-dataset

variability can be reduced by multiple normalizations to improve comparability across recordings.

“Data normalization is a process employed in data analysis and statistics to transform variables onto a consistent scale or comparable range. Its primary aim is to mitigate scale effects and ensure that variables exhibit a similar distribution, simplifying data comparison and analysis.”

Normalization in the context of EEG refers to the process of standardizing or scaling the amplitude values of EEG signals to a common range or scale. This ensures that EEG data from different channels and recordings are comparable and can be analyzed effectively. The goal of normalization is to eliminate the impact of differences in signal amplitudes, which can vary due to factors such as electrode placement, hardware variations, and subject-specific characteristics.

Data normalization or rescaling can be achieved through various methods of data alignment, including within-electrode and across-electrode transformations [81]. Within-electrode transformations involve normalizing the data for each electrode separately, while across-electrode transformations use linear transformations of the EEG data to reduce between-subject variability. Across-electrode weighting factors include the mean, Huber mean, and Euclidean (L2) norm [81]. The Huber mean is more robust to outliers compared to the mean and Euclidean norm, and these methods capture the central tendency of the EEG amplitude [25].

The Huber mean is an iterative technique used for robustly approximating the mean in the presence of outliers [89]. In our study, normalization factors were computed column-wise on the amplitude matrix (across channels) as shown in Figure 3 for the Huber mean. The resulting amplitude matrix was then divided column-wise by the 1xR vector of resulting recording normalization factors to produce a normalized amplitude matrix.

To characterize the comparability of channel amplitudes across a data collection, a dispersion vector was computed by taking the robust standard deviation of each row (across recordings in the collection) of an amplitude matrix. This vector was then divided by the row median (across recordings) to obtain a 58x1 channel dispersion vector representing the collection variability for each N.

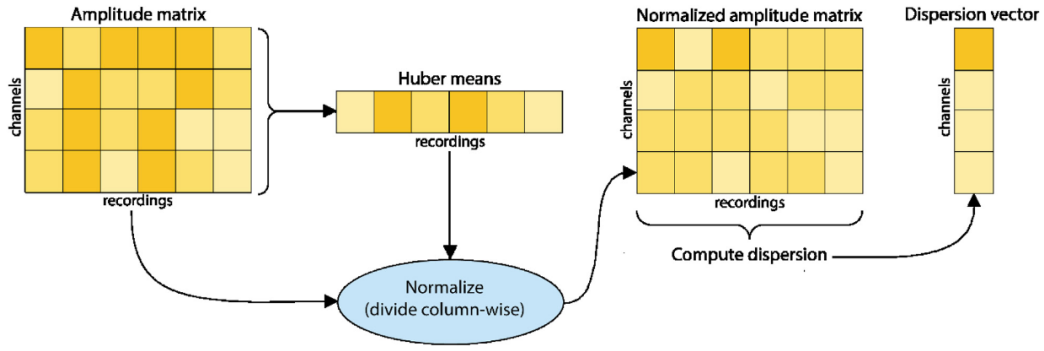


Figure 3 Steps for computing the normalized amplitude matrix and the channel dispersion vector for Huber mean normalization. Source: N. Bigdely-Shamlo et al. NeuroImage [25], showcases an illustrative instance involving a 4-channel (row) by 6-column (record) matrix, symbolizing the amplitude matrix. This project adheres closely to the methodology depicted in figure, though with an expanded scope featuring 58 channels and 457 records.

In Figure 3, we present an illustrative example featuring a matrix with 4 channels (rows) and 6 columns (records) to represent the amplitude matrix. In this project,

we diligently followed the methodology depicted in Figure 3, albeit employing 58 channels and 457 records. Within the scope of this study, we faithfully adopt the same methodology, initiating the harmonization process with Huber's normalization method. This strategic approach is aimed at attenuating inter-dataset variability by aligning EEG data across different recordings. Moreover, it serves to compensate for divergences among datasets when amalgamating samples sourced from disparate research centers.

1.6.3.3 Harmonization of extracted features

As has been discussed since the beginning of this Chapter 1, there is a growing trend of large-scale initiatives that aim to gather diverse EEG datasets for sharing and dissemination. In such studies, multiple records are crucial due to logistical challenges and the geographical differences in the subjects or cohorts being studied. However, a significant drawback of combining EEG data from multiple sites is the potential introduction of non-biological sources of variability, primarily arising from differences in EEG acquisition protocols and hardware used across different locations.

"Harmonization primarily aims to extract information by utilizing libraries that facilitate data processing, normalization, and improvement while effectively managing variables present in the records."

Harmonization, in the context of EEG analysis, goes beyond normalization. It involves aligning EEG data from different recordings or datasets to minimize

variations caused by differences in recording conditions, equipment, and other confounding factors. Harmonization techniques aim to create a unified dataset that mitigates the influence of inter-dataset variability, making it easier to analyze EEG data across different subjects, centers, or studies.

Harmonization methods for EEG signals are in their developmental stages [90]. These techniques, while crucial, are not widely adopted as standardized protocols in multicentric studies of resting-state EEG and neurodegenerative phenomena [85]. However, our current project is actively addressing these gaps by innovatively implementing ComBat as part of our research on resting-state EEG. Rooted in the framework of Generalized Additive Linear Mixed-effects Models, these strategies have the potential to effectively manage confounding factors and determinants stemming from various recording centers and headset configurations [91].

The ComBat method, originally developed for batch-effect correction in genomics research by Johnson et al.[92], has been modified to address site-related effects in multi-site DTI studies, as reported by Jean-Philippe et al.[93].

$$y_{ijv} = \alpha_v + X_{ij}^T \beta_v + \gamma_{iv} + \delta_{iv} \in_{ijv}$$

Equation 1

Where α_v is the feature for the reference site for feature v, the procedure for the estimation of the site parameters γ_{iv} and δ_{iv} uses Empirical Bayes and where β_v is the p × 1 vector of coefficients associated with X for feature v. ComBat assumes

that the residual terms $\in_{ijv}$ have mean 0. The parameters δ_{iv} describe the multiplicative site effect of the j-th site on voxel v.

This technique has demonstrated efficacy in harmonizing the data by removing unwanted variations associated with site while retaining biological associations. The success of ComBat has led to the development of other methods like neuroComBat [94] and neuroHarmonize [95] in neuroimage, which aim to further refine and improve the harmonization process.

It has previously been done harmonization of large MRI datasets for the analysis of brain imaging patterns by Raymond Pomponio [95]. That study discusses the challenges of harmonizing different MRI datasets and proposes a framework for harmonizing these datasets, which involves preprocessing, quality control, and statistical analysis. The approach can help to reduce confounding factors and improve the accuracy of results.

The proposed method by Pomponio [95] involves the harmonization of individual ROI features through a model based on the statistical harmonization technique presented by Johnson [92], which facilitates adjustments to the data for location and scale (L/S) variations. This approach involves the estimation of location (mean) and scale (variance) differences in ROI features across multiple sites (or cohorts), along with the preservation of variations due to other biologically relevant covariates present in the data. Once the estimates are obtained, the standardized ROI features

are obtained by removing the location and scale effects caused by differences between cohorts.

For cohort, i, subject j, region k, a general framework for an LS-adjustment of an ROI features, Y_{ijk}, is:

$$Y_{ijk}^{*} = \frac{(Y_{ijk} - f_k(X_{ij}) - g_{ik})}{d_{ik} + d_k(X_{ij})}$$

Equation 2

where $f_k(X_{ij})$ denotes the variation of Y captured by the biologically relevant covariates (age and sex) X, g_{ik} is the estimated location effect for cohort i and region k, and d_{ik} is the estimated scale effect for cohort i and region k. In the linear case, $f_k(X_{ij}) = a_k + X_{ij} * b_k$ and the corresponding adjustment is

$$Y_{ijk}^{*} = \frac{(Y_{ijk} - a_k - X_{ij} * b_k - g_{ik})}{d_{ik} + a_k + X_{ij} * b_k}$$

Equation 3

In the context of neuroHarmonize [95], the Generalized Additive Model (GAM) is used to replace $f_k(X_{ij})$, where the covariates age, sex, and ICV (Intercept of Covariates) are represented by X_{ij}, Z_{ij}, and W_{ij}, respectively. This approach enables the modeling of nonlinear age trends in ROI features using a basis expansion to allow for flexible nonlinearity in X_{ij}.

$$f_k(X_{ij}, Z_{ij}, W_{ij}) = a_k + f_{(X_{ij})} + b_k * Z_{ij} + c_k * W_{ij}$$

Equation 4

To integrate the non-linear GAM model with ComBat, neuroHarmonize leverages the previously proposed framework of ComBat [92] for the multivariate harmonization of multiple ROIs. The ComBat approach assumes that location and scale effects for multivariate outcomes are drawn from a common parametric prior distribution. In this study, a normal distribution is used as the prior for g_{ik}, while an inverse-gamma distribution is used for d_{ik}. Empirical Bayes framework is employed to estimate the hyperparameters of the prior distributions from the data. These hyperparameters are then used to compute the conditional posterior estimates of all location and scale effects, as detailed in [92]. ComBat adjusts an ROI feature, Y^*_{ijk}, using these conditional posterior estimates. The ComBat-GAM adjustment, together with the non-linear GAM model in neuroHarmonize, provides a robust method for ROI harmonization.

$$Y^*_{ijk} = \left(\frac{Y_{ijk} - f_k(X_{ij}, Z_{ij}, W_{ij}) - g^*_{ik}}{d_{ik} + f_k(X_{ij}, Z_{ij}, W_{ij})}\right)$$

Equation 5

where g^*_{ik} is the posterior estimate of the location effect for cohort i and region k, and d_{ik} is the conditional posterior estimate of the scale effect for site i and region k. For more details of the ComBat-GAM algorithm see [95].

1.6.4 Acquisition and Signal processing

1.6.4.1 Conventional Electrode Positioning

The 10–20 system avoids both eyeball placement and considers some constant distances by using specific anatomic landmarks from which the measurement would be made and then uses 10 or 20% of that specified distance as the electrode interval [46]. The odd electrodes are on the left and the even ones on the right. Extra electrodes are sometimes used for the measurement of as electro-oculogram (EOG), electrocardiograph (ECG), and electromyography (EMG) of the eyelid and eye surrounding muscles [46].

1.6.4.2 Preprocessing

In a broad sense, EEG signal preprocessing stands for the manipulations performed on the raw acquired data in order to prepare it for feature extraction in the next processing phases [96]. Most of these techniques are common to almost all neuroscience EEG studies, not only to AD diagnosis [97]. When an EEG signal is acquired, the data is usually not clean, so some preprocessing is required [98]. This often includes the application of filters such as a high pass filter to remove the DC components of the signals and the drifts, usually a frequency cutoff of 0.5 Hz is sufficient [99]. A low pass filter can also be applied to remove the high frequency components [100]. In EEG, we rarely look at frequencies above 70 Hz, which is the gamma range. There are many other preprocessing techniques such as EOG artifact correction that may need to be applied if the subject is recorded with their eyes open. This is because blinking and eye movements create strong electrical fields

that interfere with our EEG recordings, and the filters are designed not to change or distort the signals [100].

On the other hand, high-frequency noise is reduced by using low-pass filters with a cutoff frequency of about 50 or 60 Hz (Depending on the country) [101].

The goal of the processing techniques is to characterize the signal by a set of model parameters that best describe the signal generation system [102].

The EEG signal can be considered as the output of a nonlinear system that can be characterized deterministically and non-stationarity of the signals can be quantified by measuring some statistics of the signals at different time lags [103]. It is necessary to label the EEG signals into segments of similar characteristics that are most meaningful to the clinician and for evaluation by the neurophysiologist. Within each segment, the signals are statistically stationary, usually with similar time and frequency statistics. If the signals are statistically stationary it is straightforward to characterize them in either the time or frequency domains. The most common epoch duration is 2 s according to a systematic review by Cassani R et al. [7].

The concept of independent component analysis (ICA) lies in the fact that the signals may be decomposed into their constituent independent components [104]. In places where the combined source signals can be assumed independent from each other this concept plays a crucial role in separation and denoising the signals [105].

ICA is usually able to concentrate the artifactual information into a single component, but in most cases this component also carries non-artifactual information, so rejecting it may cause information loss [106].

Moreover, ICA performance depends on the size of the dataset (number of samples): the larger the dataset processed, the higher the probability that the effective number of sources will overcome the number of channels (overcomplete ICA) [107], because the number of channels is fixed over time, but the number of contributions from different neural sources is likely to increase with the length of the recording. In this case redundancy is not sufficient to estimate the sources and an independent component might account for more than one contribution [108], in other words, the algorithm might not be able to separate the artifactual signals from the rest.

On the contrary, the smaller the number of samples, the more difficult the estimation of the parameters and thus the performance of ICA suffers. The best choice is a tradeoff between a small dataset and a high performance [109]. To overcome this limitation, the proposed methodology includes a step prior to ICA that increases the redundancy of the dataset, thanks to wavelet decomposition, bypassing the possible problem of overcomplete ICA [110].

Another limitation is that ICA cannot take advantage of the features of the artifacts in frequency domain: artifacts have a typical frequency range, and their spectrum is overlapped to the spectrum of the EEG, thus filtering the dataset is not an optimal

solution because this would lead to a great information loss [111]. But we can make the most of this limitation in frequency domain performing ICA in the range where the artifact is concentrated [112].

Once our signals are clean, i.e., pre-processed, it is quite common to cut them into epochs of a few seconds and then extract features from each of these epochs. This allows us to have many features from a single EEG recording, which is preferable when doing statistics or applying classifiers [113].

1.6.4.3 Feature Extraction

Over the last thirty years, the dimensionality of the data involved in machine learning and data mining tasks has exploded. Data with extremely high dimensionality has posed serious challenges to existing learning methods, i.e., the curse of dimensionality [114].

EEG signals are complex, which makes it very difficult to extract information using raw data. Nowadays, thanks to computers, we can apply complex automatic processing algorithms that allow to extract "hidden" information from EEG signals [115]. There are several techniques, such as time domain features (mean, standard deviation, Entropy, etc.), frequency domain features (Fourier transform, wavelets, etc.) and finally synchrony features, which look at the relationship between 2 or more EEG signals (Coherence, correlation, mutual information, etc.), just to mention a few [116].

There are other feature extraction methods, such as EEG tomography, which allows us to compute the active regions inside the brain (using the so-called inverse

problem approach) [117]. This, in turn, usually requires many electrodes (at least 32) to achieve a decent spatial resolution [118]. Also, if possible, more advanced methods are used, such as converting the EEG recording into a graph where each node represents an electrode [119], and the connections of these nodes depend on the similarity of the EEG signals from each electrode to analyze properties using analysis techniques [7], [120].

Some features aim at measuring one major effect of AD in the EEG signal with slowing, complexity reduction, synchronization decrement, and neuromodulatory deficit and others includes data-driven features which are not necessarily driven by known biological processes [121].

Measurement of the slowing effect of AD on EEG signals typically relies on spectral features derived either from each of the EEG channels or from the average of the channels, being the most common the Power Spectral Density (PSD) and Wavelet [122]. On the other hand, the complexity of EEG signals is typically evaluated with Entropy measures [123].

The various metrics used to measure synchronization of EEG signals can be classified according to two criteria: (1) the presence or absence of directional (causal) information, and (2) whether the metric assumes a linear relationship between the analyzed signals (model-based) or no assumption of a linear relationship (model-free) [86]. This type of metric includes Coherence,

Synchronization Likelihood, Graph theory metrics, Correlation (amplitude envelopes) and Permutation disalignment index.

1.6.4.3.1 Relative Power

It is often used in EEG analysis to examine changes in the power of specific frequency bands in response to different stimuli or conditions [124]. In routine use, electrical potentials are acquired indirectly from the scalp surface and include waveform analysis of frequency, voltage, morphology, and topography, in addition, the amplitude of the EEG recorded in a particular subject depends on many factors, including neurophysiological, anatomical and physical properties of the brain and surrounding tissues (skin, bone, dura mater, and pia mater) [125], but these parameters vary from one subject to another and are basically unknown.

These variations result in large variations in the absolute EEG spectra, but to compensate for this variation, the relative EEG power is calculated so that the variability in absolute power is greater than the variability in Relative Power [126].

The mathematical formula for Relative Power is:

$$RS(f) = \frac{S(f)}{\sum S(f)}$$

Equation 6

where RS(f) is the Relative Power at frequency f, S(f) is the power at frequency f, and ΣS(f) is the total power in the EEG signal. Relative Power is typically expressed as a percentage or decibel (dB) value.

Where S is the Power spectral density, and the mathematical formula is:

$$S(f) = |X(f)|^2 = \left| \int_0^T h(t)x(t)e^{2\pi i f t} d\tau \right|^2$$

Equation 7

Where:

- $S(f)$: Represents the convolved power spectral density as a function of frequency "f". It measures how the power of a signal is distributed across different frequencies.

- $X(f)$: Is the Fourier transform of the signal "x(t)", representing the signal in the frequency domain.

- $h(t)$: Is a taper or windowing function.

T: Is the time interval over which the integration is performed.

- $x(t)$: Is the original signal in the time domain.

- $e^{2\pi i f t}$: Is a complex exponential function used to decompose the signal into its frequency components.

- f: Represents the frequency at which the power distribution is being analyzed.

Two key performance metrics of spectral estimators are bias and variance. Bias can be decomposed into two types: local and broadband. Local bias arises from the bandwidth of the main lobe of a spectral window, while broadband bias is a function of its side lobes [127].

The Multitaper Method (MTM) further reduces bias by obtaining statistically independent estimates that are effectively averaged to reduce uncertainty, like the Welch WPM. Each window of MTM is pairwise orthogonal to all other windows, providing a statistically independent set of spectral estimates that are averaged (weighted) to provide the final spectrum.

In this project, an adaptation in python of the Matlab Chronux module [128] was used to use the MTM, which is represented mathematically by the following formula:

$$S_{MT} = \frac{1}{K}\sum_{k=1}^{k}|X_k(f)|^2 = \frac{1}{K}\sum_{k=1}^{k}\left|\int_0^T u_k(t)x(t)e^{-2\pi ift}dt\right|^2$$

Equation 8

where $K = 2TW - 1$ is a taper or Slepian sequences function for duration T.

S_{MT}: Represents the spectral estimation using the Multitaper Method.

K: Denotes the number of tapers used in the method.

$|X_k(f)|^2$: Represents the squared magnitude of the Fourier transform of the data using the k-th taper at frequency 'f'.

$u_k(t)$: Represents the k-th taper function in the time domain.

$x(t)$: Denotes the original signal in the time domain.

$e^{-2\pi ift}$: Represents the complex exponential function that allows the signal to be analyzed in the frequency domain.

T: Denotes the time interval over which the integration is performed.

f: Represents the frequency being analyzed.

The physiological interpretation of Relative Power in EEG analysis is that it reflects the degree of neural activity in different frequency bands in the brain and this frequency bands have been associated with different cognitive and physiological processes.

1.6.4.3.2 Entropy

Shannon Entropy is a measure of the uncertainty or randomness in a signal, named after the mathematician Claude Shannon [129]. In the context of EEG analysis, Shannon Entropy can be used to quantify the complexity of the EEG signal, based on the distribution of amplitudes across the signal [130]. It measures the degree to which the signal deviates from a uniform distribution of amplitudes.

The mathematical formula for Shannon Entropy is:

$$H(p) = -\sum_{i=1}^{m} p_i \log(p_i)$$

Equation 9

The calculation of the Shannon Entropy value (H) is given as minus the sum of the probabilities of the event(i) multiplied by the logarithm to base two of the probabilities of the event(i) and p is the probability of each amplitude value in the signal. Shannon Entropy is measured in bits, and its value ranges from 0 (no uncertainty) to the maximum Entropy of the signal, which is determined by the number of possible amplitude values.

The physiological interpretation of Shannon Entropy in EEG analysis is not completely clear, but it is thought to reflect the complexity of neural activity in the brain [131]. A higher Entropy signal may indicate greater variability or complexity in the neural activity underlying the EEG signal, while a lower Entropy signal may indicate more uniform or simple neural activity [132].

There are several studies that have investigated the relationship between Shannon Entropy and EEG signals. For example, one study found that EEG signals from patients with Alzheimer's disease had lower Entropy than those from healthy controls, suggesting reduced complexity of neural activity in Alzheimer's disease [133]. Another study found that Shannon Entropy was positively correlated with the complexity of the cognitive task being performed by participants, suggesting that Entropy may reflect cognitive demand [134].

1.6.4.3.3 Coherence

From a structural (anatomical) point of view, the most striking feature of the brain is the abundant connectivity between neurons. From a functional point of view, this connectivity is reflected in synchronous activities within the brain: neurons in anatomically connected structures tend to fire synchronously [135].

Electrophysiological data show that this synchronicity is performed in bursts repeating at different frequencies [136]. The frequency of the synchronization seems to define the functional meaning of connectivity through consistency since it is a measure of synchronization between EEG recorded in different scalp locations and reflects a correlation between EEG powers computed for these two locations in the same frequency band [137].

For example, alpha frequencies are idling rhythms of sensory systems and synchronization at 10 Hz frequency indicates the state of the sensory system when neurons do not relay sensory information but ready to commence when a relevant stimulus will appear [138]. Oscillatory synchronization in gamma band has been proposed as a binding mechanism for combining different features of an object into a single percept [139]. Synchronization at 40 Hz frequency indicates the synchronous activation of neurons responsible for detecting different features of the same stimulus [140].

The disruption of “normal” synchronization may be a sign of neurological or psychiatric dysfunction. For example, an abnormal pattern of synchronization

between different parts of the basal ganglia seems to be responsible for tremor and dyskinesia in Parkinson's disease [141].

Mathematically, Coherence is calculated by dividing the cross-spectral density of two signals by the product of their individual power spectra [142]. The resulting value ranges from 0 (indicating no correlation) to 1 (indicating perfect correlation) and is often expressed as a percentage and is computed as follows:

$$C_{xy} = \frac{\left|P_{xy}\right|^2}{P_{xx} * P_{yy}}$$

Equation 10

Where Pxx and Pyy are power spectral density estimates of X and Y, and Pxy is the cross spectral density estimate of X and Y.

Physiologically, Coherence reflects the degree of communication between different brain regions. When two brain regions are functionally connected, their neural oscillations will be synchronized, and their Coherence value will be high. In contrast, when there is no functional connection between two brain regions, their neural oscillations will be independent, and their Coherence value will be low.

There have been numerous studies that have used Coherence analysis to investigate the functional connectivity of different brain regions and networks. For example, one study found that Coherence between the prefrontal cortex and the hippocampus was higher during working memory tasks, suggesting that these brain regions are

functionally connected during this cognitive process [143]. Another study found that Coherence between the primary motor cortex and the cerebellum was increased during motor learning, indicating enhanced functional connectivity between these regions during this process [144].

1.6.4.3.4 Cross Frequency

Cross Frequency refers to the phenomenon in which the amplitude of high-frequency oscillations is modulated by the phase of low-frequency oscillations [145]. It has been observed in a variety of brain regions and is thought to play an important role in cognitive processes such as attention, memory, and perception [146].

Mathematically, Cross Frequency is typically measured following the steps (Figure 4):

1. The full band EEG signal is broken down into sub-bands.

$$s_i(n) = s(n) * h_i(n)$$

Equation 11

Where s(n) is the full band signal and hi(n), i = 1, 2, …, k are the responses of the band-pass filters used to separate each of the sub-bands.

2. Using the Hilbert transform (H {*}) the time amplitude envelope of each of the sub-bands is calculated.

$$e_i(n) = \sqrt{s_i(n)^2 + H\{s_i(n)\}^2}$$

Equation 12

e_i: Represents the time amplitude envelope of the i-th sub-band at time instance 'n'.

$s_i(n)^2$: Denotes the instantaneous amplitude of the i-th sub-band signal at time 'n'.

3. The representation by “spectral modulation” is obtained for each sub-band by applying the Fourier transform (F {*}) with a Hamming window of 5 seconds and movement of 500ms, on the temporal envelopes.

$$\varepsilon_i(m; f) = |F\{e_i(m, n)\}|$$

Equation 13

Where m are the frames and f the modulated frequencies.

4. Bearing in mind that, with the Hilbert transform, the envelope signal can only contain frequencies (i.e., modulation frequencies) up to the maximum frequency of its source signal (Bedrosian's theorem). Therefore, for each TF result, only the modulated sub-bands with lower frequencies than the original sub-band of the time envelope are taken.

5. Finally, a modulation energy “ratio” parameter called percentage modulation energy (PME) is calculated, which has given good results in classification tasks and is given by:

$$PME_{i,j} = \frac{\bar{\varepsilon}_{i,j}}{\sum_{i=1}^{k}\sum_{j=1}^{k-1}\bar{\varepsilon}_{i,j}} x100\%$$

Equation 14

Where $\bar{\varepsilon}_{i,j}$ is the average modulation energy (in all frames) of the sub-band i grouped by the modulated sub-band j.

As shown in Figure 4, the Hilbert transform definition limits the frequency range of the extracted envelopes $e_i(n)$ to the modulation frequencies present in the original signal $s_i(n)$, as demonstrated by Bedrosian's theorem [147]–[149]. This implies that Gamma can modulate all frequency bands, whereas Delta is limited to modulating only itself.

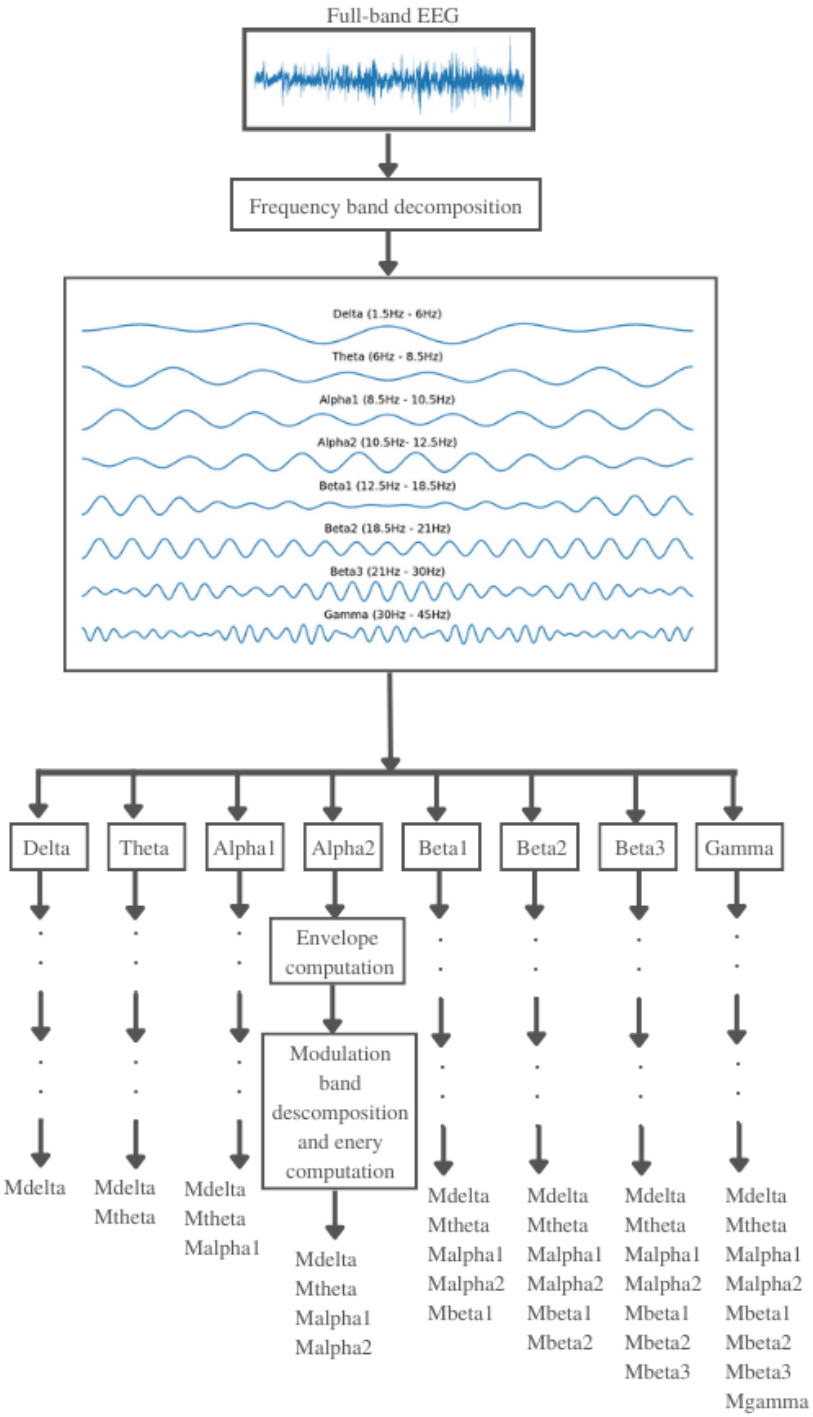


Figure 4 Hilbert transform definition limits the frequency range of the extracted envelopes $\boldsymbol{e_i(n)}$ to the modulation frequencies present in the original signal $\boldsymbol{s_i(n)}$,, as demonstrated by Bedrosian's theorem. Adapted Image Source: [147]

Physiologically, Cross Frequency reflects the functional interactions between different neural oscillations [150]. The phase of low-frequency oscillations is thought to modulate the excitability of neural populations, which in turn affects the amplitude of high-frequency oscillations. This Cross Frequency interaction is believed to be important for coordinating the activity of different brain regions and for regulating neural processing [151].

There have been numerous studies that have used Cross Frequency analysis to investigate the functional connectivity and neural processing in the brain. For example, one study found that Cross Frequency between theta and gamma oscillations in the hippocampus was increased during memory retrieval, suggesting that this coupling is important for memory processing [152]. Another study found that Cross Frequency between alpha and beta oscillations in the motor cortex was increased during motor planning, indicating that this coupling is important for motor processing [153].

Finally, another study said although both within-frequency and Cross Frequency networks can be used to predict AD with high accuracy, the bispectrum-based functional connectivity outperforms cross-spectrum suggesting an important role of Cross Frequency functional connectivity [154].

1.6.4.3.5 Synchronization Likelihood

Synchronization Likelihood (SL) is a measure of the non-linear synchronization between two signals [155]. It was developed to study the functional connectivity between different brain regions by measuring the degree of synchronization

between their respective neural oscillations. Unlike Coherence, which is a measure of linear synchronization, SL can capture non-linear synchronization, which is thought to be more common in the brain [156].

SL is calculated by computing the probability that two signals will remain within a certain phase difference range over time. The resulting value ranges from 0 (indicating no synchronization) to 1 (indicating perfect synchronization) and is often expressed as a percentage.

Mathematically, the Synchronization Likelihood $S_{k,i,j}$ for each channel k and each discrete time pair (i, j) where each of the M is time series embedded vectors $X_{k,i}$,and the number $H_{i,j}$ of channels, for which the distance of embedded vectors $X_{k,i}$ and $X_{k,j}$ is smaller than εk,i, defined as:

$$if\ \left|X_{k,i} - X_{k,j}\right| < \varepsilon_{k,i}: S_{k,i,j} = \frac{H_{i,j} - 1}{M - 1}$$

$$if\ \left|X_{k,i} - X_{k,j}\right| \geq \varepsilon_{k,i}: S_{k,i,j} = 0$$

Equation 15

by averaging over all j, we finally obtain the Synchronization Likelihood $S_{k,i}$:

$$S_{k,i} = \frac{1}{2(w_2 - w_1)} \sum_{\substack{j=1 \\ w_1<|j-i|<w_2}}^{N} S_{k,i,j}$$

Equation 16

Therefore, the Synchronization Likelihood $S_{k,i}$ is a measure that describes how strongly channel k at time i is synchronized to all the other M − 1 channels.

Physiologically, SL reflects the degree of synchronization between different brain regions [157]. When two brain regions are functionally connected, their neural oscillations will be synchronized, and their SL value will be high. In contrast, when there is no functional connection between two brain regions, their neural oscillations will be independent, and their SL value will be low [158].

There have been numerous studies that have used SL analysis to investigate the functional connectivity of different brain regions and networks. For example, one study found a significant heritability that suggests that SL can be used to examine genetic susceptibility [159]. Another study found that SL between the prefrontal cortex and the motor cortex was increased during motor imagery, indicating enhanced functional connectivity between these regions during this process [160]. Finally, another study found the relationship between functional connectivity and complexity exhibited various temporal-scale-and-regional-specific dependencies in both control participants and patients with AD and the combination of functional connectivity and complexity might reflect the complex pathological process of AD [161].

1.6.5 Machine Learning

Using machine learning (ML) techniques, we can train a classifier to recognize the best features or, from among select features, which ones belong to one class (or

condition, i.e., AD,) or to another (i.e., subject healthy). This is a very powerful technique, and it is extensively used in EEG data analysis [162]. ML has the potential to support and perhaps accelerate the neurophysiological diagnostic or monitoring pathway, but with the adoption of any new technology there will be difficulties [163].

Among classification algorithms, the Support Vector Machine (SVM) algorithm is the most widely used, where classification accuracy is widely used as a performance metric. However, in AD studies, given the differences experimental setup, EEG processing pipeline, and cross-validation paradigms, there is no way to directly compare the results [7].

On the other hand, in the application of machine learning techniques to the selection of features, some tools appear, such as The Tree-Based Pipeline Optimization Tool (TPOT) [164] or Boruta [165], these tools are based on automated machine learning (AutoML).

AutoML algorithms are not as simple as fitting a model to the data; they consider multiple machine learning algorithms (random forests, linear models, SVMs, etc.) in a pipeline with multiple preprocessing steps (missing value imputation, scaling, feature selection, etc.), the hyperparameters for all models and preprocessing steps, and multiple ways to ensemble or stack the algorithms within the pipeline.

TPOT's optimization algorithm is stochastic in nature, i.e., it uses randomness (in part) to search the possible pipeline space. If two TPOT runs recommend different

pipelines, it means that the TPOT runs didn't converge due to lack of time, or that multiple pipelines perform the same on the dataset [166].

Boruta iteratively compares the importance of features with the importance of shadow features created by shuffling the original attributes. Features that have significantly worse importance than the shadows are successively discarded. On the other hand, attributes that are significantly better than the shadows are allowed to be confirmed. Shadows are created in each iteration[167].

Therefore, to better represent the domain, many candidate features are introduced, resulting in the existence of irrelevant redundant features for the target concept. A relevant feature is neither irrelevant nor redundant to the target concept, an irrelevant feature is not directly associated with the target concept but affects the learning process, and a redundant feature does not add anything new to the target concept [168]. Reducing the number of irrelevant redundant features can drastically reduce the running time of the learning algorithms and yield a more general classifier [169].

Chapter 2

Database construction and standardization

2.1 Introduction

Currently, there are a variety of pipelines for EEG analysis, so it is common to find a processing strategy in each repository or public database [170]. Additionally, it is necessary to apply organizational standards for the security and organization of EEG data, which also protect the personal data of patients [171]. This is the case of EEG-BIDS, an extension of the brain imaging data structure for EEG [27], which addresses the organization of multimodal data following localization, accessibility, and interoperability principles. A multimodal database is a data processing tool that supports multiple data models and defines the parameters of how information is organized and accommodated in a database [172]. While the approaches presented above section 1.6.1 and 1.6.5 for EEG and ML analysis show promising performance, the validation methods used are generally limited to relatively small, controlled, and mostly local and private data sets [173]. Therefore, the question arises as to whether the detection capabilities of these algorithms generalize to larger samples, considering the different databases, and whether they could subsequently be scaled to a clinical (uncontrolled) setting.

The possibility of answering this question has become a common interest at the scientific level and is motivated by the publication of open access databases [174]. The Cuban Center for Neurosciences has shared a set of resting-state EEG normative data collected between 1988-1990 [175], as has the CHBMP repository which shares an open-label multimodal neuroimaging and cognitive dataset of 282 healthy young and middle-aged participants [79]. This data set was acquired from 2004 to 2008 as a subset of a larger stratified random sample of 2019 participants from the municipality of La Lisa in Havana, Cuba. However, these efforts have also sparked a debate about the importance of moving beyond data pooling, toward data standardization to facilitate use of aggregated data sets that also share methodology in terms of pre-processing and quality control. The use of complementary multimodal databases results in a standardized data state. EEG-IP is a platform developed to advance biomarker discovery by enhancing large-scale integration of data from multiple sites [176]. Where lossless signal processing implementation algorithms are shared on this platform [96] maximizing signal isolation and minimizing data loss. In addition, it provides a unified and standardized output data status.

Metrics for evaluation and comparison of multiple databases include calculation of epoch rejection rate, Signal to Noise Ratio (SNR), amplitude variation in particular time windows, the susceptibility of the experimental setup to line noise, the percentage of EEG segments contaminated by artifacts, and metrics on signal stability based on autocorrelation and cross-correlation analysis [16], [25], [177].

With this approach, the effect on any independent parameter that has a potential impact on EEG database integration can be tested. Finally, the metrics that indicate the success of harmonization can be extended to the EEG source space. This is the case for the pairwise normalized difference/skewness (ND), a measure that represents the dimensionless, normalized pairwise skewness of functional connectivity matrices and graph-based derived metrics, and can be used as an indicator of variability between subjects [178]. Thus, standardizing the data offers maximum possibilities for large-scale data exploration in EEG data, substantially accelerating hypothesis testing in biomarker discovery research.

In this chapter, we describe the methodology used to select the databases, providing a detailed description of each data source, and discuss the methods utilized to standardize and harmonize the data. Through this chapter, we demonstrate our commitment to ensuring the quality and consistency of the data used in our study, and how we can overcome the challenges posed by using multiple data sources.

2.2 Methodology

In this chapter, we delineate the method employed to choose suitable databases, aligning with our first objective of data integration and standardization, a crucial step that shapes the ensuing analysis. The process unfolds in the following steps:

1. **Initial Database Identification:** The foundation of our study involves two primary databases, namely a central database and a preceding parent database, both administrated by the University of Antioquia in collaboration with the

Neuropsychology and Behavior Research Group (Gruneco). These repositories encompass diverse datasets encompassing EEG recordings during rest tasks (with eyes open and closed), demographic insights (age, gender, education), and outcomes of neuropsychological assessments (MMSE, MOCA, etc.).

2. **Exploration of External Sources:** We venture beyond our internal databases to explore external sites that provide open access to diverse databases. A meticulous evaluation of the available data is conducted, encompassing both neurophysiological (EEG) recordings and demographic/neuropsychological particulars.

3. **Database Selection Criteria:** Leveraging the comparative lens, we meticulously select databases that feature neuropsychological assessments akin to our initial databases and encompass neurophysiological data obtained during rest tasks with eyes closed.

4. **Data Acquisition and Compilation:** Selected databases are procured and systematically stored within a consistent directory on the computing system. The acquisition process adheres closely to the stipulated instructions offered by each respective website.

5. **Standardization through Sovabids Tool:** To ensure uniformity and Coherence across the collected databases, we employ the Sovabids tool. This tool adeptly transforms the databases into the BIDS format, thereby enhancing compatibility and ease of integration.

6. **Localization of Processable Signals:** Within the standardized databases, we meticulously pinpoint the precise locations of raw signals slated for processing in alignment with our research objectives.

The schematic representation of our methodology is depicted in Figure 5.

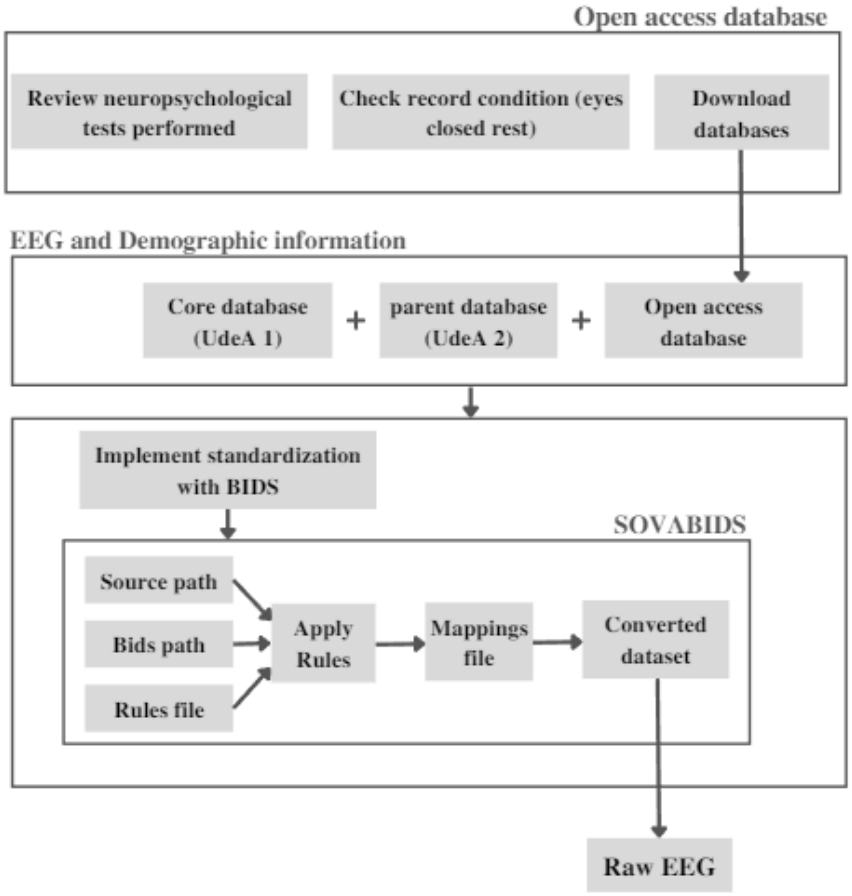


Figure 5 Illustrates the integration of cohorts from the neuropsychology and behavior group, as well as the subsequent conversion of their data to the BIDS standard.

This integration process ensures that data from different sources and formats can be harmoniously merged and organized, allowing for standardized and consistent

analysis. By adhering to the BIDS standard, the data becomes more accessible, interpretable, and easily shared among researchers and institutions.

2.3 Search criteria

The development of the search criteria was guided by the project's objectives, outlined below:

1. **Establishing a Comprehensive Database:** The primary objective was to construct and standardize a database containing multimodal information sourced from diverse sites. This involved leveraging tools that facilitate both data storage and manipulation prior to and during processing. Our approach involved an exhaustive review of prominent open-access databases commonly referenced in pertinent neuroscience literature and journals.

To devise our search criteria, we methodically explored websites frequently cited within the neuroscience domain, see Figure 6. We aspired to ensure the flexibility of our search criteria at the outset, enabling us to encompass diverse tasks, such as data from rest or eyes-closed studies, as well as data from both healthy subjects and those afflicted with Alzheimer's disease, see Figure 7.

Once we had identified databases that matched these preliminary criteria, we refined our search to target repositories featuring both EEG data and neuropsychological assessments. This intricate process led us to isolate 707 repositories housing both neurophysiological and neuropsychological records.

From this pool, we extracted 1098 EEG records, of which 1061 also boasted available neuropsychological test results.

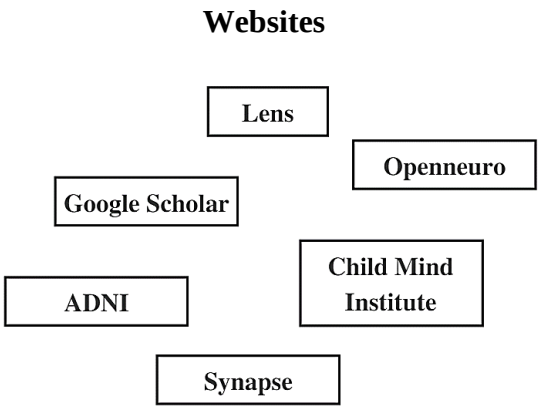


Figure 6 Query criteria for websites. Commonly referenced in pertinent neuroscience literature and journals.

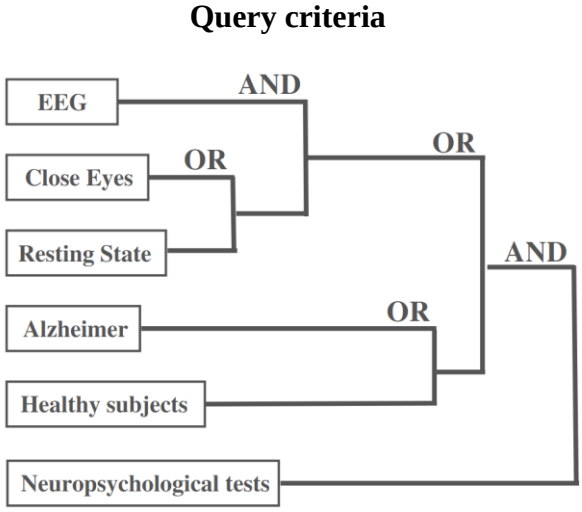


Figure 7 Query criteria for open access cohorts. Encompass diverse tasks, data from rest or eyes-closed studies, data from both healthy subjects and those afflicted with Alzheimer's disease.

2. **Achieving for Comparative Analysis:** From the 1061 records, we isolated those that contained only entries in the resting state, of which we found 617, and those that could be downloaded without any preprocessing. This pursuit culminated

in our selection of two public access databases aligning with the neuropsychological tests outlined in our initial search. Consequently, our study was fortified by a total of 457 records, a breakdown of which is depicted in Figure 8.

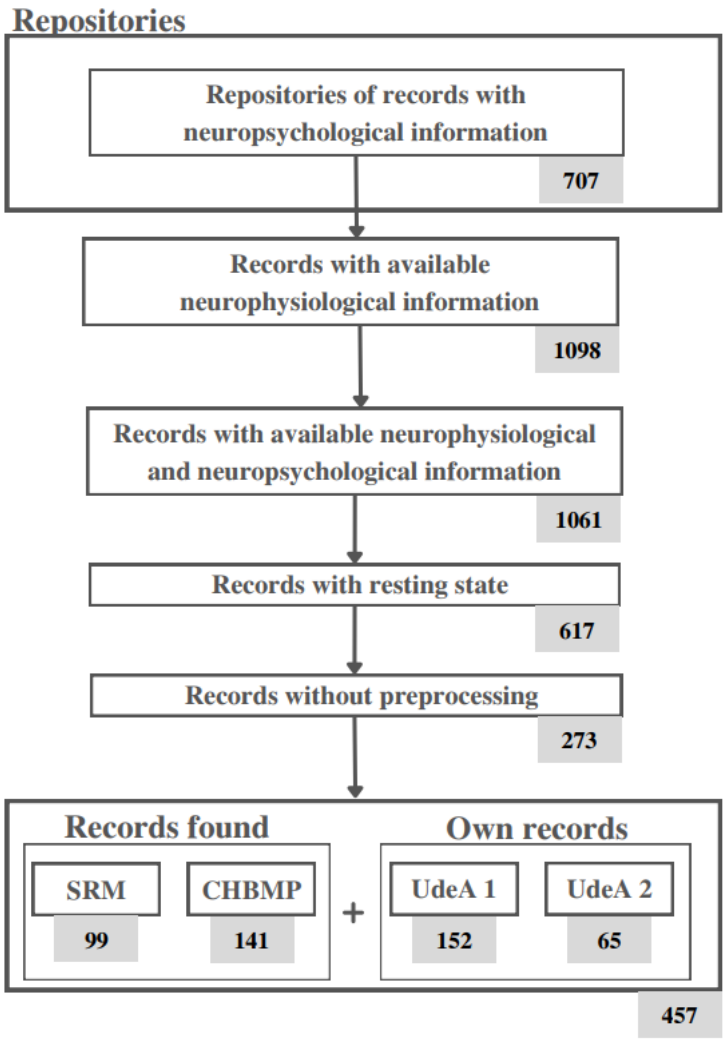


Figure 8 Illustrates the comprehensive review process conducted to gather the records utilized in this project. The transition from repositories to registries is illustrated, culminating in the selection of two open access databases and two proprietary databases.

The realization of this objective rested upon the databases provision of information on healthy subjects and their associated neuropsychological assessments, which collectively enabled the meticulous selection of records.

3. **Crafting a Machine Learning (ML) Model for Risk Assessment:** The third objective revolved around the design of an ML model capable of classifying individuals at risk of Alzheimer's disease (AD), utilizing the comprehensive database of neuropsychological and neurophysiological data.

In pursuit of this goal, records sourced from the databases outlined in objectives 1 and 2 were employed. Yet, during model training, a challenge emerged due to the presence of missing data in certain records. To address this challenge, the possibility of employing an imputation process was explored, wherein missing values could potentially be substituted with informed estimations. However, it was decided to eliminate the records that did not contain all the data.

This decision was based on, the inclusion criteria relating to neuropsychological tests presented an obstacle. Although our identified databases didn't exhibit identical sets of tests, they did share at least one common test within each repository. As a result, the integration of all four databases resulted in the elimination of this test-specific data. Despite these constraints, the objective was fulfilled within the project's defined scope.

Looking ahead, it's imperative to acknowledge the considerations that arose from these limitations for subsequent projects. Lessons gleaned underscore the

importance of early consideration to circumvent the inadvertent removal of valuable records, thereby optimizing the utility of information within the model for future endeavors.

2.4 Results

2.4.1 Database standardization

The following is a description of each of the databases that are part of the project, starting with the initial databases (Gruneco Research Group) and ending with the open access databases.

2.4.1.1 UdeA 1 Database

Subjects

The study included individuals from families with the PSEN1-E280A genetic variant, as well as healthy controls who voluntarily participated. Participants were asymptomatic individuals aged between 20 and 45 years, with 32 carriers (G1) and 37 non-carriers (G2). Additionally, 19 subjects with mild cognitive impairment (DCL) and 8 with dementia, who carried the PSEN1-E280A variant (DTA) and were over 40 years old, were included. Lastly, 30 community controls were included, who volunteered and did not have any psychiatric, neurological, or systemic disorders, history of TBI, stroke, use of anticonvulsant drugs, or abuse of psychoactive substances, which could affect EEG or cognitive test performance. Participants and evaluators were unaware of the genetic status of the participants, and the groups were matched as closely as possible for age, sex, and schooling.

Acquisition protocol

During the study, EEG signals were recorded in both resting states: with eyes closed (EC) and with eyes open (EO) for a duration of 5 minutes. The EEG data was acquired using a Neuroscan amplifier (Neuroscan Medical System, Neurosoft Inc. Sterling, VA, USA) with a 58-tin channel cap placed according to the international 10-10 system. The sampling rate of the data was set to 1000 Hz, and in-line band pass filtering (0.05 to 200 Hz) and a band reject filter (60 Hz) were applied to remove any power supply noise. A reference electrode was placed on the right earlobe, and Fz electrode was used as the ground electrode. Prior to recording, a channel impedance calibration was conducted to ensure contact impedances of EEG electrodes remained below 1 KΩ. Furthermore, to minimize any external electromagnetic interference, recordings were performed inside a Faraday cage, a soundproof and electromagnetically shielded enclosure.

Number of channels: 60 (Including EEG and others)

Channels: ['FP1', 'FPZ', 'FP2', 'AF3', 'AF4', 'F7', 'F5', 'F3', 'F1', 'FZ', 'F2', 'F4', 'F6', 'F8', 'FC5', 'FC3', 'FC1', 'FCZ', 'FC2', 'FC4', 'FC6', 'T7', 'C5', 'C3', 'C1', 'CZ', 'C2', 'C4', 'C6', 'T8', 'TP7', 'CP5', 'CP3', 'CP1', 'CPZ', 'CP2', 'CP4', 'CP6', 'TP8', 'P7', 'P5', 'P3', 'P1', 'PZ', 'P2', 'P4', 'P6', 'P8', 'PO7', 'PO5', 'PO3', 'POZ', 'PO4', 'PO6', 'PO8', 'O1', 'OZ', 'O2', 'HEO', 'VEO']

Neuropsychological tests

All participants were evaluated by medical and neuropsychological experts and underwent genotyping for the PSEN1-E280A variant and cognitive status verification according to the protocol of the Neurosciences Group of Antioquia. Informed consent was obtained from all participants, and the Ethics Committee Board of the Faculty of Medicine - University of Antioquia approved the study.

Support

This work was supported provided by the Comité para el Desarrollo de la Investigación - CODI Universidad de Antioquia, through the project "Cambios en los patrones del electroencefalograma cuantitativo (reactividad alfa, theta y su índice) en reposo y tareas de memoria, en el seguimiento longitudinal de pacientes con riesgo genético para Enfermedad de Alzheimer Temprano", identified with the code 2017-16371.

2.4.1.2 UdeA 2 Database

Subjects

The study enrolled individuals belonging to the E280A mutation Colombian kindred, which included 22 asymptomatic carriers (G1), 18 healthy non-carriers (G2), 20 symptomatic carriers (DTA), and 17 healthy non-carriers (Control). To ensure comparable gender, age, and educational level across the groups, we selected healthy non-carriers that matched with the carriers in these characteristics. Moreover, we compared each carrier group with its corresponding control group to evaluate the effect of the genetic mutation. It is worth noting that 30 of the

asymptomatic subjects had previously participated in other studies involving connectivity analysis and quantitative EEG.

The study was conducted following the ethical guidelines approved by the local institutional review boards, and informed consent was obtained from the participants or their legal representatives. The study was conducted according to a general protocol approved by the Human Subjects Committee of the Sede de Investigación Universitaria (SIU) of Universidad de Antioquia, Medellin, Colombia. The genetic status of the participants was masked by the investigators collecting the data. The exclusion criteria included severe physical illness, psychiatric or neurological disorders that may affect cognitive function, and other forms of dementia. Additionally, individuals with alcohol or drug abuse and those under regular treatment with neuroleptics or antidepressants with anticholinergic activity were excluded.

Acquisition protocol

During the study, EEG recordings were obtained using a Neuroscan amplifier (Neuroscan Medical System, Neurosoft Inc. Sterling, VA, USA). Participants were seated comfortably and instructed to rest with their eyes closed for 5 minutes during the EEG data acquisition. EEG data were collected from 64 electrodes with a bandpass filter of 0.1-200 Hz and a midline reference at a sampling rate of 1000 Hz. The electrodes were positioned according to the international 10-10 system, and an electrooculogram recording (0.1±100 Hz bandpass) was performed simultaneously. All recordings were obtained in the second semester of 2012.

Number of channels: 68 (Including EEG and others)

Channels: ['FP1', 'FPZ', 'FP2', 'AF3', 'AF4', 'F7', 'F5', 'F3', 'F1', 'FZ', 'F2', 'F4', 'F6', 'F8', 'FT7', 'FC5', 'FC3', 'FC1', 'FCZ', 'FC2', 'FC4', 'FC6', 'FT8', 'T7', 'C5', 'C3', 'C1', 'CZ', 'C2', 'C4', 'C6', 'T8', 'M1', 'TP7', 'CP5', 'CP3', 'CP1', 'CPZ', 'CP2', 'CP4', 'CP6', 'TP8', 'M2', 'P7', 'P5', 'P3', 'P1', 'PZ', 'P2', 'P4', 'P6', 'P8', 'PO7', 'PO5', 'PO3', 'POZ', 'PO4', 'PO6', 'PO8', 'CB1', 'O1', 'OZ', 'O2', 'CB2', 'HEO', 'VEO', 'EKG', 'EMG']

Neuropsychological tests

All participants underwent a thorough clinical and neuropsychological assessment, which was conducted by a neurologist or a physician specially trained in dementia evaluation. The neurological examination and clinical history review were conducted to obtain a complete medical history of the subjects. The neuropsychological protocol included the widely used Mini-Mental State Examination (MMSE) and the Consortium to Establish a Registry for Alzheimer's Disease (CERAD) battery, which was adapted to the Colombian population. In addition, a cognitive composite test with high sensitivity for tracking E280A subjects was employed to examine the correlation analysis with the neurophysiological measures. This composite test consisted of the average of scores from several subtests, including Abstract Reasoning (Ravens Progressive Matrices), Orientation (MMSE Orientation to Time), Language (CERAD Boston Naming Test), Memory (CERAD Word list Recall), and Praxis (CERAD Constructional Praxis).

Support

This work was supported by Vicerrectoría de Investigación of Universidad de Antioquia (CODI), Projects “Identificación de marcadores preclínicos de la mutación E280A de la enfermedad de Alzheimer a partir de medidas de conectividad en EEG”, code PRG14-1-02.

2.4.1.3 SRM Database

Subjects

The dataset includes raw neuropsychological assessment scores, age, sex, and resting-state EEG data from 111 healthy control participants (Control) with a mean age of 37.6 years (SD = 14.0, range = 17-71). Prior to participation, written informed consent was obtained from all participants, who had normal or corrected-to-normal vision and hearing, and reported no severe psychiatric or neurological symptoms. Basic audiometry screening was conducted on all participants. Recruitment was conducted through social media platforms, such as Facebook and Instagram, as well as local advertisements.

Acquisition protocol

During the data acquisition procedure, a resting-state EEG was obtained towards the end of the session while the participants had their eyes closed. EEG data were collected from 64 electrodes. The segment was initiated with a set of standardized written instructions displayed on a 24” LCD screen (BenQ, model ID: XL2420-B). The instructions, translated from Norwegian, asked the participants to close their eyes and remain seated with their eyes closed for about four minutes. During this

time, there were no visual or auditory stimuli presented, and the participants were not required to perform any actions.

Number of channels: 64 (Including EEG and others)

Channels: ['Fp1', 'AF7', 'AF3', 'F1', 'F3', 'F5', 'F7', 'FT7', 'FC5', 'FC3', 'FC1', 'C1', 'C3', 'C5', 'T7', 'TP7', 'CP5', 'CP3', 'CP1', 'P1', 'P3', 'P5', 'P7', 'P9', 'PO7', 'PO3', 'O1', 'Iz', 'Oz', 'POz', 'Pz', 'CPz', 'Fpz', 'Fp2', 'AF8', 'AF4', 'AFz', 'Fz', 'F2', 'F4', 'F6', 'F8', 'FT8', 'FC6', 'FC4', 'FC2', 'FCz', 'Cz', 'C2', 'C4', 'C6', 'T8', 'TP8', 'CP6', 'CP4', 'CP2', 'P2', 'P4', 'P6', 'P8', 'P10', 'PO8', 'PO4', 'O2']

Neuropsychological tests

At the initial time point, all participants underwent a comprehensive neuropsychological assessment. The assessment battery consisted of multiple tests to evaluate different cognitive domains. These included the Rey Auditory Verbal Learning Test, which measured verbal learning and memory, the Wechsler Adult Intelligence Scale-IV Digit Span, which assessed attention span and working memory, and the Delis-Kaplan Executive Function System tests. The D-KEFS tests comprised of the Trail Making Test, which measured psychomotor speed and executive functioning, the Colour-Word Interference Test, which evaluated reading speed and executive functioning, and the Verbal Fluency Test, which assessed phonemic and semantic processing abilities.

Support

The dataset from the Stimulus-Selective Response Modulation (SRM) project at the Department of Psychology, University of Oslo, Norway, has been made publicly available on the internet and can be accessed freely. The project team has divulged this dataset to facilitate wider access and promote research in this field.

2.4.1.4 CHBMP Database

Subjects

The CHBMP repository is a publicly accessible, multimodal neuroimaging and cognitive dataset that includes data from 282 healthy participants (Control) (age range 18-68 years, mean age 31.9 ± 9.3 years). This dataset was obtained between 2004 and 2008 and is a subset of a larger stratified random sample of 2,019 participants from La Lisa municipality in La Habana, Cuba. Participants with any signs of disease or brain dysfunction were excluded from the study.

Acquisition protocol

Resting-state EEG for 10 minutes was recorded using the digital electroencephalograph system MEDICID 5-with 64 and 128 electrodes with differential amplifiers and gain of 10,000. The amplifiers used three filters: 1) Low cutoff (-3dB, high-pass): first order (6 dB/octave) 2) High Cutoff (-3dB, low-pass): Butterworth, second order (12 dB/octave) and 3) Line filter with a unit frequency response. Electrodes were placed according to the 10–10 International System with a customized electrode cap. Linked earlobes were used as the EEG reference. Electrode impedances were considered acceptable if less than 5 KΩ. The bandpass

filter parameters were 0.5–50 Hz and 60 Hz notch, and a sampling period of 200 Hz.

The EEG was recorded in a temperature and noise-controlled room while the participant was sitting in a reclined chair. All individuals were asked to relax and remain at rest during the test to minimize artifacts produced by movements and to avoid excessive blinking. The participants received instructions to have enough sleep the previous night, take breakfast, and wash the hair before attending this appointment. The structure of raw EEG recording was generated in the default format of the MEDICID neurometrics system (*.plg extension), which later is converted to standard BIDS format.

Number of channels: 123 (Including EEG and others)

Channels: ['Fp1', 'Fp2', 'F3', 'F4', 'C3', 'C4 ', 'P3 ', 'P4 ', 'O1 ', 'O2 ', 'F7 ', 'F8 ', 'T7 ', 'T8 ', 'P7 ', 'P8 ', 'Fz ', 'Cz ', 'Pz ', 'F1 ', 'F2 ', '22', '23', 'P1 ', 'P2 ', 'AF3', 'AF4', '28', '29', '30', '31', '32', '33', 'FT7', 'FT8', '36', '37', 'P5 ', 'P6 ', 'FC5', 'FC6', '42', '43', 'C5 ', 'C6 ', '46', '47', '48', '49', '50', '51', 'TP7', 'TP8', 'PO5', 'PO6', '56', '57', 'AF7', 'AF8', '60', '61', 'FpZ', '63', 'FCZ', 'CPZ', 'POZ', 'OZ ', '68', '69', '70', '71', '72', '73', '74', '75', '76', '77', 'PO3', 'PO4', '80', '81', 'CP1', 'CP2', '84', '85', '86', '87', '88', '89', 'CP3', 'CP4', '92', '93', '94', '95', 'C1 ', 'C2 ', 'F5 ', 'F6 ', 'FC3', 'FC4', 'FC1', 'FC2', '104', '105', '106', '107', '108', '109', '110', '111', 'CP5', 'CP6', 'PO7', 'PO8', '116', '117', '118', '119', '120', 'EOI', 'EOD', 'ECG']

Neuropsychological tests

The psychological test results (including MMSE, Wechsler Adult Intelligence Scale - WAIS III, and computerized reaction time tests using a go no-go paradigm), as well as general information about each participant (age, gender, education, ethnicity, handedness, and weight).

The Mini-Mental State Examination MMSE is a quick and easy measure of cognitive functioning that has been widely used in clinical evaluation and research involving patients with dementia. In our study, the MMSE was employed as a screening test to exclude participants with cognitive impairment. The total score of the participants is available in the file MMSE.csv with also the individual items for 52 subjects.

Support

The dataset from the The Cuban Human Brain Mapping Project (CHBMP) project at the Cuban Ministry of Public Health (MINSAP) and coordinated by the Cuban Neuroscience Center (CNEURO), has been made publicly available on the internet and can be accessed freely. The project team has divulged this dataset to facilitate wider access and promote research in this field.

2.4.2 Sovabids tool implementation

Following meticulous database selection and a thorough grasp of their distinctive features, a seamless integration process ensued within the Sovabids tool, an accessible open-source solution [179].

The genesis of the Sovabids tool stems from the realization that typical EEG experiments produce a spectrum of data structures that are uniformly organized across participants. However, there may be subtle differences in the organization of these structures for individual participants.

The Sovabids tool masterfully navigates these complexities by embracing an approach that capitalizes on overarching commonalities among participants while maintaining a dynamic adaptability to outliers. This intricate equilibrium was meticulously achieved through the instantiation of two discrete configuration files:

1. **The Rules File**: This repository encodes the bedrock conversion principles that extend across an expansive EEG dataset. It functions as a standardized framework that illumines the trajectory of the conversion process.

2. **The Mappings File**: Evolving from the Rules File, the Mappings File assumes a personalized mantle. It houses nuanced conversion directives, meticulously tailored to the distinct traits of each participant. This bespoke approach guarantees that even participants with variances in data organization find a harmonious inclusion.

This dual-tiered configuration architecture encapsulates a deliberate fusion of harmony and flexibility, empowering the Sovabids tool to seamlessly harmonize with the intricate topography of diverse EEG datasets. (as shown in the Figure 9).

To increase the level of automation, sovabids incorporates heuristics that exploit the common file path patterns used in EEG research. This further streamlines the process of converting EEG data to BIDS.

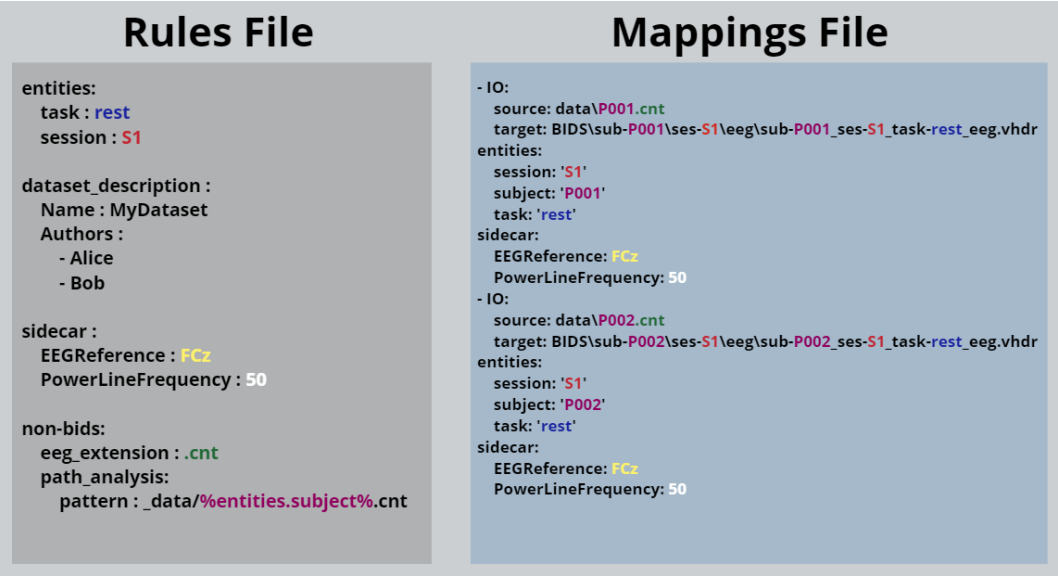


Figure 9 From a Rules File, a mapping for each file in the dataset can be generated and saved in the Mappings File. The colors illustrate how the information in both files is related. Source: sovabids.readthedocs.io

Finally, the result is observed in Figure 10, since it shows how a data set made up of different tasks, different sessions and different patients is transformed into the ordered BIDS structure.

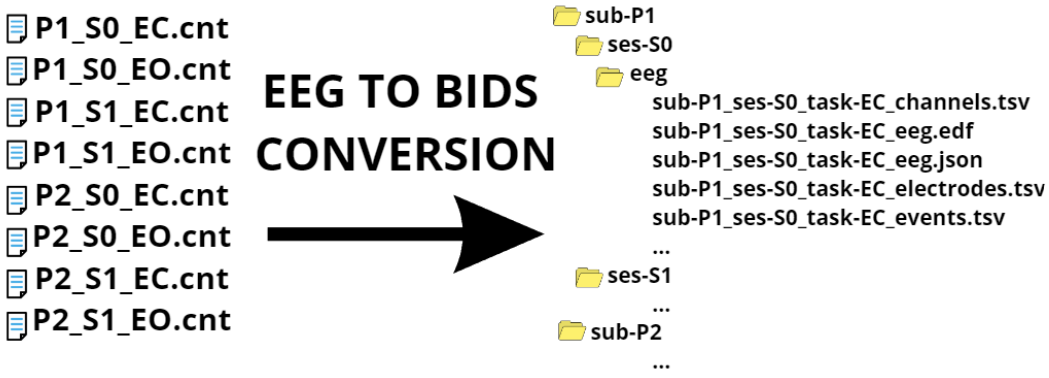


Figure 10 Graphical representation of the conversion to BIDS

2.4.3 Data Description

Ultimately, the data that has been meticulously organized in the BIDS format is consolidated and stored within a singular dataframe—a tabular data structure wherein information is systematically arranged into rows and columns. This coherent arrangement empowers us to seamlessly extract comprehensive insights from the entirety of the dataset, a representation vividly exemplified in Table 1.

Table 1 Description of total subjects in selected cohorts

	Age			Sex
	n	mean	std	F/M
Group				
Control	281	35.61	14.43	122/159
G1	49	30.18	5.50	29/20
G2	52	31.37	5.90	31/21
Total	457			

The information used consists of 457 records, divided into several groups: Healthy subjects (Controls and G2), and subjects carrying the genetic mutation PSEN1-E280A (G1). Basic demographic information is available for almost all the subjects, presented in a Table 1 that includes their respective averages and standard deviations. Table 2 provides a segregated presentation of the identical information, distinctly categorized by each of the four cohorts.

Table 2 Provides the statistical description of each selected cohort.

		Age			Sex
		count	mean	std	F/M
Database	Group				
UdeA 1	Control	28	49.36	7.35	15/13
	G1	27	30.16	5.86	15/12
	G2	34	32.09	5.82	20/14
CHBMP	Control	141	31.16	9.32	39/102
UdeA 2	Control	13	46.23	8.89	9/4
	G1	22	29.54	5.10	14/8
	G2	18	30.00	5.96	11/7
SRM	Control	99	36.66	13.92	59/40
	Total	457			

2.5 Discussion

EEG databases are an essential resource for researchers studying the brain's electrical activity. However, the lack of standardization in the format of these databases has been a significant barrier to their accessibility and usability. This is where the BIDS enters as alternative, which is a standardized format for organizing and sharing neuroimaging data.

Converting EEG databases to the BIDS format can have significant benefits for the neuroscience community. First and foremost, it allows researchers to easily access and use the data in a standardized format, enabling reproducible and transparent research [27]. Additionally, it facilitates the integration of EEG data with other neuroimaging modalities, such as MRI and fMRI, which are already commonly organized in BIDS format [174], [180].

Moreover, the availability of easy-to-use tools for converting EEG data to the BIDS format is critical. Without such tools, the conversion process can be time-consuming and error-prone, which can hinder the adoption of BIDS by the neuroscience community. Therefore, the development and dissemination of user-friendly conversion tools can significantly accelerate the adoption of BIDS and promote open and collaborative research in the field of neuroscience [179].

Overall, the use of BIDS format for EEG databases and the availability of conversion tools is essential for promoting open science and enabling transparent, reproducible, and collaborative research in the field of neuroscience.

Our methodology for database selection prominently featured neuropsychological tests, serving as a pivotal criterion for inclusion. These tests facilitated the classification of subjects across distinct groups of interest, including Controls, G1, and G2. It is noteworthy, however, that the absence of certain evidentiary elements within the selected databases curtailed the integration of some longitudinal follow-ups or data encompassing healthy controls and AD participants. Consequently, the

results stemming from these tests were not integrated into the subsequent data analysis showcased in Chapters 3 and 4.

Finally, disseminating and implementing tools like Sovabids in projects that utilize databases from multiple sites can promote their adoption, increase their visibility within the scientific community, and ultimately facilitate data exchange for maximal benefit of these databases.

2.6 Conclusions

In summary, Chapter 2 has successfully achieved its goal of creating and standardizing a comprehensive multimodal database from multiple databases. We've gone through the process of selecting databases, explaining their contents and the tools used for data standardization in the field of neuroscience.

Of particular importance is the Sovabids tool, which we've adopted for wider using by the scientific community. It's important to emphasize our overarching mission: to create a processing pipeline that unifies datasets across different cohorts and databases. Our goal is to use harmonization techniques to build a machine learning model capable of identifying Alzheimer's disease risk using noninvasive biomarkers extracted through semi-automated processing.

As we move into Chapter 3, our focus shifts to practical implementation. We'll delve into the intricacies of our generalized processing pipeline designed specifically for feature extraction. This leads to the construction of a harmonized

and augmented database that serves as the cornerstone of our machine learning model. The culmination of this process is the unveiling of the results, which illuminate the classification effectiveness of the model and the insights it provides. These insights pave the way for significant advances in scientific progress.

Chapter 3

Processing pipeline and Harmonization

3.1 Introduction

The analysis of EEG data has become more varied and flexible with the availability of different pipelines, allowing research teams to adopt their own processing strategy. However, the choice of algorithms used in different processing steps, such as artifact removal, filtering, and time-domain transformations, can have significant effects on the estimation of the power spectral density of different EEG frequency-bands, ultimately affecting scientific conclusions [46]. Researchers generally overestimate the likelihood of significant results across hypotheses, and reproducibility of results obtained using a single analysis pipeline is hard to estimate. To increase the statistical power and sensitivity of multicentric studies, it is important to have standardized data preprocessing pipelines in addition to standard collection procedures [81]. As discussed in the previous Chapter 2, organizational standards for EEG data such as BIDS can help with the security and organization of data, as well as protecting patients' personal data. Therefore, it is crucial to plan and carefully report the selection of tools, the sequence of processing steps, and the analysis parameters.

As we delve further into the concept of harmonization in this Chapter 3, it is essential to note that the development of a pipeline for harmonizing EEG-related multi-feature in neurodegenerative research presents substantial challenges. Nevertheless, there is promising progress in the field, with several tools available for harmonizing preprocessing steps that need to be comparable within basic common processing pipelines [81], [181]. These advancements have the potential to revolutionize current EEG approaches in neurodegenerative research, leading to a new generation of objective, computer-based tools for the diagnosis, characterization, and treatment of neurodegenerative diseases and other disorders.

Aligned with this trend, a processing pipeline has been developed, demonstrating favorable outcomes in single-site databases [182], [183]. Nevertheless, the pipeline has been enhanced with normalization and harmonization stages to facilitate its applicability across multiple databases sourced from diverse websites. Chapter 3 showcases the processing pipeline implemented and elucidates the undertaken measures to harmonize it for seamless integration with various cohorts.

3.2 Methodology

Prior to delving into this Chapter 3, it is essential to remember the definition of harmonization presented in the theoretical framework to establish the scope of its relevance to this project: **"Harmonization primarily aims to extract information by utilizing libraries that facilitate data processing, normalization, and improvement while effectively managing variables present in the records."**

Figure 11 depicts a step-by-step processing diagram, which includes four essential steps. The first step is the pre-processing pipeline, followed by the normalization of the data using the "Record-specific constant." Next, spectral analysis is conducted, and finally, data harmonization is performed. In the subsequent sections 3.3.3, will discuss the execution of each step-in detail, this preprocessing pipeline is referred to as "**Sovaharmony**".

Sovaharmony is a proprietary package developed within the scope of this project, encompassing the `harmonize` function designed to process EEG data within a BIDS-format dataset. The function sequentially processes EEG files, executing preliminary stages such as artifact detection, signal filtering, and scaling. Processed data, along with its pertinent details, is stored in both derived files and JSON formats. The function adeptly manages event-related operations, enforces exclusion criteria, and scales data. Notably, it also efficiently handles and documents errors arising from files facing difficulties during processing.

The creation of this processing pipeline stands as an outcome of the project; however, it falls outside the purview of the project's objectives. As a result, an illustration of the pipeline is provided and expounded upon in an annex section, specifically labeled as Annex 2. In this annex, one can access the requisite tools to comprehensively grasp the workflow and thereby facilitate its replication by various interested groups and research teams engaged in the field.

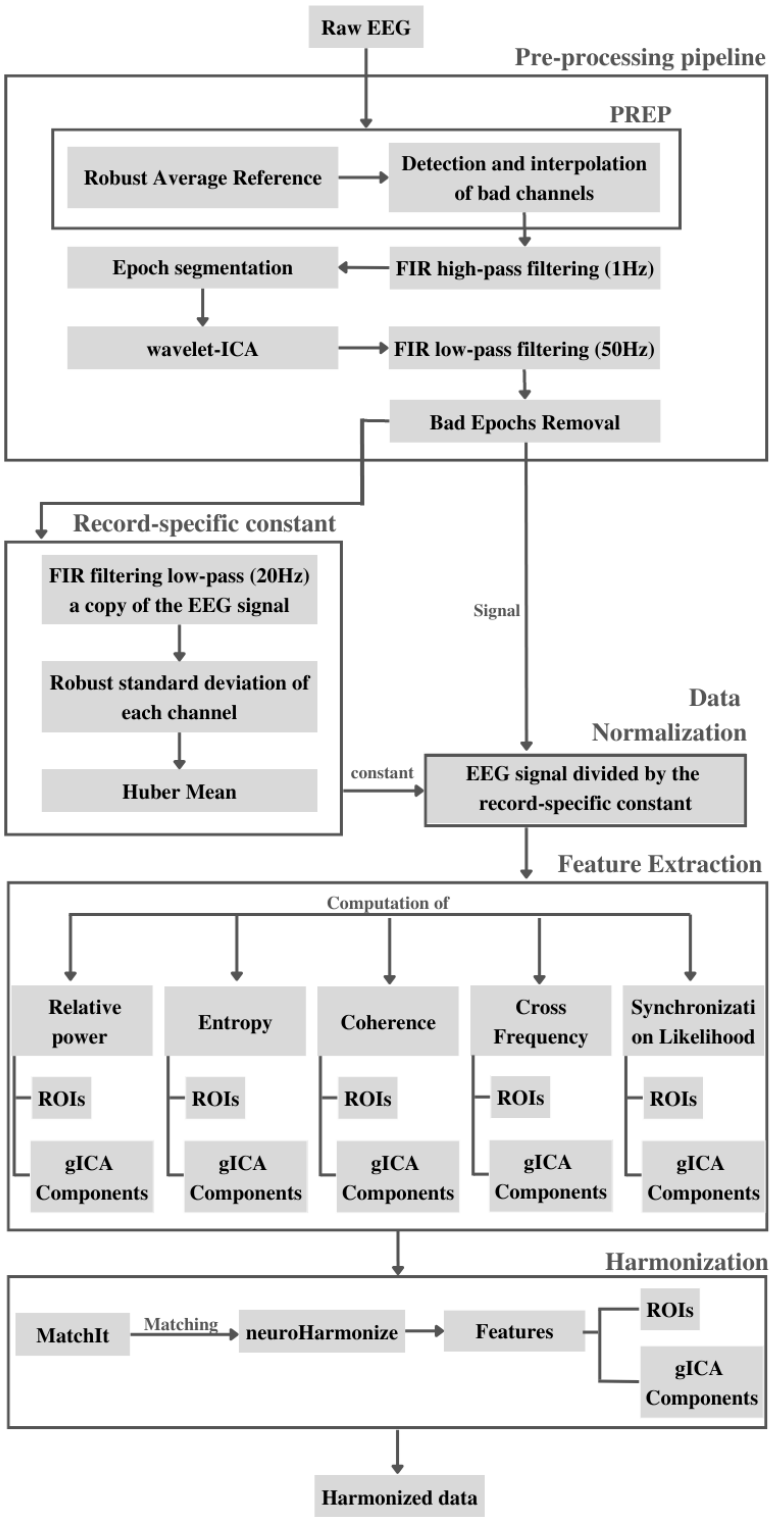


Figure 11 Pre-processing Pipeline Sovaharmony

The methodology comprised several essential steps aimed at achieving specific objectives. The second objective was to harmonize the database. This harmonization process was executed to ensure that the features extracted from different cohorts of healthy subjects were comparable and aligned. Various biomedical data processing techniques, such as spatial representation, channel interception, and preprocessing, were employed. The aim was to establish a standardized foundation for data analysis and interpretation, facilitating meaningful comparisons across the healthy cohorts (Controls).

As the data from healthy individuals were compared, the need for harmonization procedures among other groups became evident. These procedures facilitate the accomplishment of Objective 3, which revolves around the creation of a machine learning (ML) model for classifying subjects at risk of developing Alzheimer's disease.

Consequently, the results presented in this chapter exclusively pertain to the outcomes of the control subjects, thereby ensuring the fulfillment of Objective 2 (Section 3.3.5 Feature Extraction). Additionally, the results encompass the group carrying the Alzheimer's PSEN1-E280A gene mutation, completing the Matching and NeuroHarmonize phases (Sections 3.3.6 Matching and 3.3.7 NeuroHarmonize), setting the stage for the introduction of Objective 3 in the subsequent chapter.

3.3 Results

3.3.1 Spatial representations

Relative Power spectral density was calculated following EEG frequency bands: delta (1.5-6 Hz), theta (6-8.5 Hz), alpha 1 (8.5-10.5 Hz), alpha 2 (10.5-12.5 Hz), beta 1 (12.5-18.5 Hz), beta 2 (18.5-21 Hz), beta 3 (21 -30 Hz) and gamma (30-45 Hz) [40].

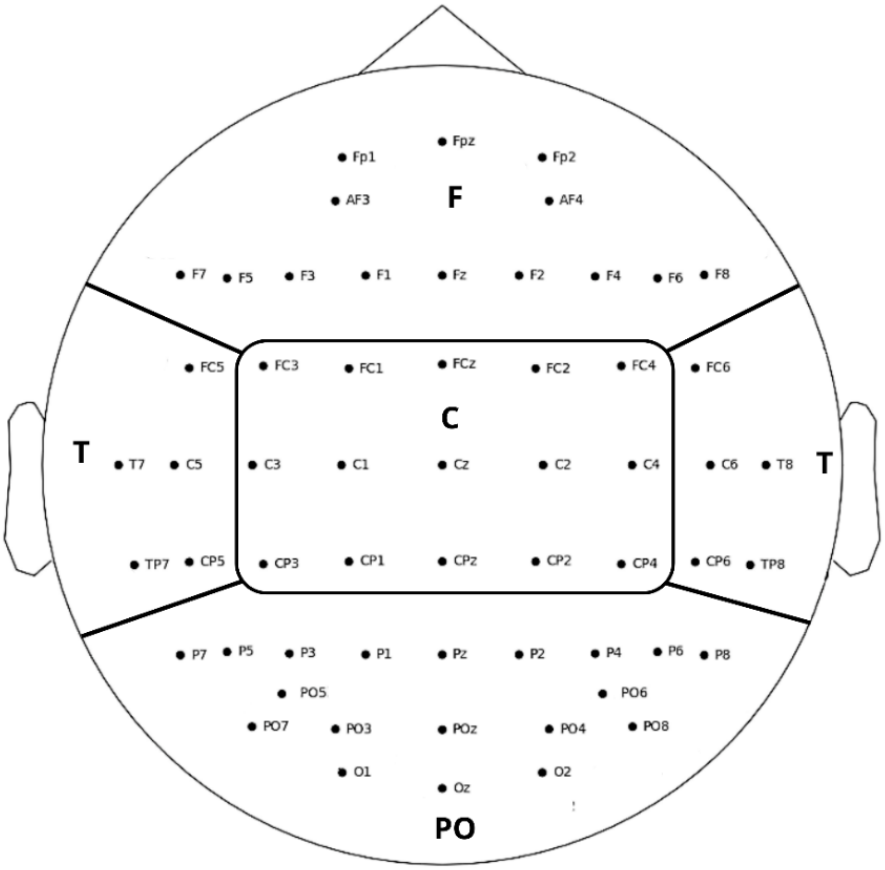


Figure 12 Schematic picture of the 58-electrodes system 10-10 and the ROIs generated. F: frontal; T: temporal; C: central; PO: parieto-occipital.

The analysis was approached in two spatial representations. The first one focuses on four regions of interest (ROIs) presented in Figure 12: frontal, temporal, central and parietal-occipital, where each region is defined by a set of electrodes.

The second one consisted of a spatial filter used previously in the work of García-Pretelt et al.[184] through the temporally concatenated group-ICA (gICA) methodology. The methodology employed by García-Pretelt et al. [184]with a single database has consistently shown good results.

In relevant preceding studies [184], Ochoa J et al. [185] analyzed differences in the frequency domain between a group of asymptomatic carriers of the PSEN-1 E280A mutation of familial Alzheimer's (ACr) and a group of symptomatic carriers of the same mutation, using data from the UdeA 2 database of this project. García-Pretelt et al. [184], on the other hand, utilized data from the UdeA 1 database to obtain independent components (spatial filters).

Figure 13 shows an adaptation of the methodology used by Garcia to select the neural gICA Components.

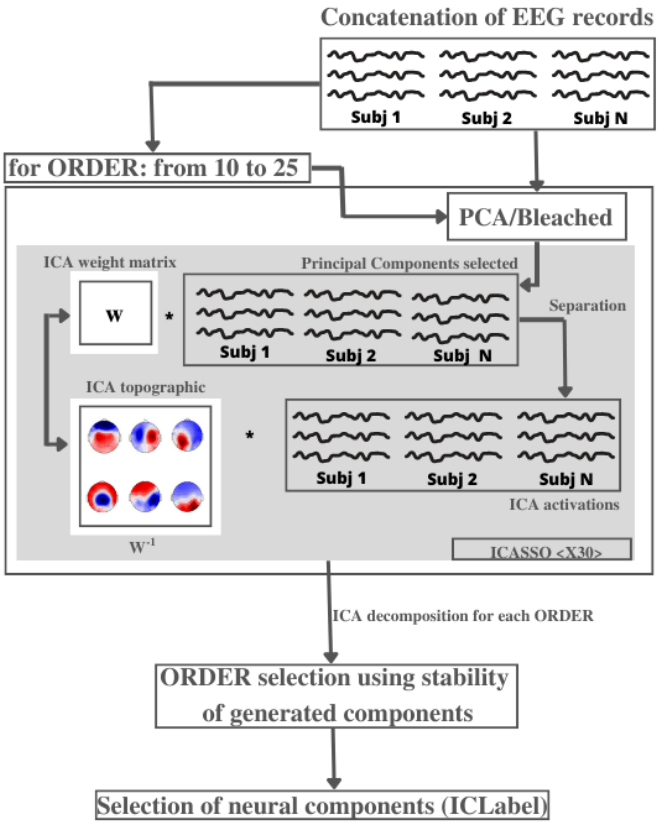


Figure 13 The calculation of the gICA components was performed with MATLAB (V.2017 a) using the FourerICA algorithm. Pipeline applied for the calculation of gICA components. Taken and modified from [184], [186].

The gICA components were calculated using MATLAB and the FourerICA algorithm. The procedure included concatenation of EEG records, optimal order testing, data laundering, and calculation of gICA components using ICASSO x30 [184]. The resulting components were evaluated for stability, and the weight matrix was applied to both groups.

Here, it is important to clarify that the spatial filter used was the one directly derived by García-Pretelt et al. [184], that is, the recordings used for the gICA procedure correspond to the subjects of that study, where the weight matrices were taken and

multiplied by the clean data to extract the activations. The gICA components used were the eight labeled as neural. Figure 14 shows the scalp-map for each component of the spatial filter.

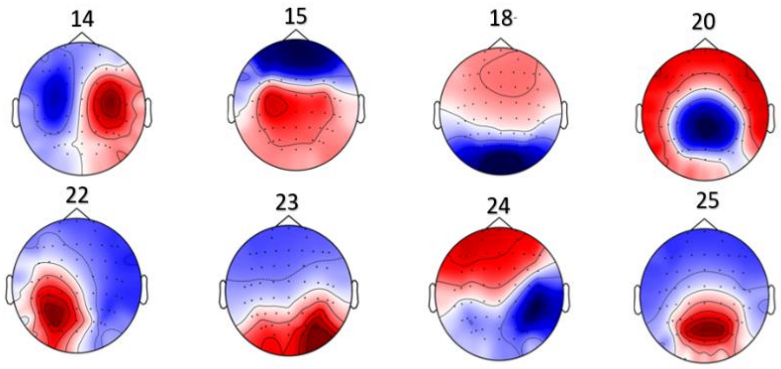


Figure 14 Scalp maps of the group-ICA components used.

3.3.2 Interception of EEG montages

The results presented in Chapter 2 provide a comprehensive breakdown of the registered channels for each cohort. The first stage of our pre-processing involves intercepting the channels of each cohort in relation to those used to construct the gICA matrix. It's worth noting that all cohorts consistently included the full set of 58 channels in the matrix, which simplifies the interception process. This strategic measure ensures consistency across all cohorts, and consequently the channels used throughout the project correspond to those shown in Figure 12, which illustrates the Regions of Interest (ROIs).

3.3.3 Pre-processing pipeline

The raw data underwent pre-processing using the pipeline proposed by Suarez et al. [183], which was entirely implemented in Python. The standardized early-stage

EEG (PREP) processing pipeline was first applied, which involved signal detrending and robust referencing to the average [184]. Bad channels were excluded and interpolated after referencing. The Fast ICA algorithm from Scikit-learn library was then used to obtain both artifactual and neural gICA Components after applying a high-pass filter at 1 Hz (FIR filters zero phase sinc, with Hamming window, order=3300, transition bandwidth= 1Hz) [187].

The records were subsequently segmented into 5-second epochs and subjected to wavelet-ICA for smoothing any remaining eye blink artifacts [188]. A 50 Hz low-pass filter (FIR filters zero phase sinc, with Hamming window, order=264, transition bandwidth=12.5 Hz) was then applied[184], and any remaining noisy epochs were detected and removed based on the following criteria: abnormal linear trends, statistically atypical activity, extreme kurtosis values, abnormal power spectra, and extreme signal amplitudes.

Moreover, to account for the variability introduced by the subject's hair, scalp, and skull, the normalization stage proposed by N. Bigdely-Shamlo et al.[25] was also included. This stage involved dividing the signal by a specific constant obtained from each EEG record that represents the overall channel amplitude. The constant was calculated by applying a 20 Hz low-pass filter, then calculating the channel-wise robust standard deviation and finally aggregating the values into a single constant using Huber's mean. While Relative Powers are unaffected by this

division, it is included as a standard stage since the pipeline is generalized for the calculation of other metrics.

The preprocessed and normalized data is utilized to extract spectral, connectivity, and amplitude modulation features (Relative Power, Entropy, Coherence, Cross Frequency, and Synchronization Likelihood) and make matches based on the age and sex of the subjects in the database using the MatchIt tool. This tool implements the suggestions of Ho D et al. [189] to improve parametric statistical models for estimating treatment effects in observational studies [190]. It achieves this by reducing model dependence through preprocessing data with semi-parametric and non-parametric matching methods.

Next, data harmonization methods are employed to eliminate unwanted variability arising from site or vendor differences while retaining the genuine biological variability within the measures. ComBat is an empirical Bayesian method for data harmonization initially developed for harmonizing gene expression data [92]. ComBat is applied directly to the extracted features from the signals without the need to retrieve the signals. It estimates and corrects site effects directly from the available signal feature values measured at different sites, which theoretically improves the alignment of the mean and standard deviation of the distributions based on the method's optimized criterion [191].

To execute the preprocessing, the sova packages, which are hosted on GitHub, are installed (For installation see Annex 2).

If the objective is to harmonize databases coming from different repositories, acquired by different devices, with a different sampling frequency and different channels, the processing must be done from the Sovaharmony package, otherwise the processing can be done only by taking the sovapipeline package. However, it is recommended to use Sovaharmony which has built-in both pre-processing and post-processing routines.

3.3.4 Quality control

Data quality control is a crucial step in workflow development, providing validity and oversight for executed processes. Researchers define monitoring protocols tailored to their study and processing. For EEG data, three main stages[192] are suggested: annotating movements or incidents during recordings, visually inspecting data for repetitions, excluding non-neuronal segments, and selecting EEG segments for analysis. However, these reviews often lack automation.

To address this, tools like 'sovaviolin' have been introduced to facilitate quality control Zapata [193] integrated it into the 'sova' packages, streamlining quality assessment. Nevertheless, methodologies primarily rely on visual inspection, equipment-specific protocols, and workflow validations. For post-processing data quality evaluation, metrics in Table 3 were adopted, quantitatively assessing key processing stages like PREP, Wavelet-ICA, and epoch rejection. Project aimed to compare data quality between cohorts before additional harmonization. Violin and

box-and-whisker plots illustrated metrics from Table 3, with a focus on cohort comparisons to assess variations tied to acquisition differences.

Different analysis approaches visualized data transition post-PREP, revealing patterns of decreased signal mean during interpolation. This stage identifies potentially damaged channels, favoring zero length in evaluated metrics. Afterward, a high-pass filter removed low-frequency trends, followed by wICA for artifact filtering. The percentage of filtered components highlighted processing quality, expected to be low due to previous signal enhancements.

Finally, a low-pass filter attenuated frequency ranges, and noisy epochs were rejected post-filtering. Segmented with a five-second window, epochs' size considered signal stationarity. This preprocessing pipeline automated metric generation, aiding in data quality assessment.

Table 3 Metrics used for the quantitative evaluation of each processing stage.

Early-Stage Data Processing pipeline (PREP)	
Metrics	**Justification**
Bad by NAN	Detection of channels that contain NAN type data.
Bad by flat	Identifying Channels with Flat Signals in Comparison to Others on a Scroll Graph
Bad by deviation	The estimation of bad channels based on deviation primarily focuses on detecting amplitude differences between data

	sets. However, it does not effectively identify channels that are contaminated by blinks and muscle activity, which can introduce noise into the data.
Bad by hf noise	Those channels that have high frequency noise are detected.
Bad by correlation	Bad Channel Detection based on Maximum Correlation Thresholding.
Bad by SNR	Identification of Defective Channels based on Low Recording Signal-to-Noise Ratio (SNR)
Bad by dropout	Default Identification of Defective Channels
Bad by ransac	Defective Channel Detection using RANSAC (Random Sample Consensus) Iterative Method
Bad all	Detection of Channels with General Faults

Wavelet Cleaning and Independent Component Analysis (ICA) Technique in Combination	
Metrics	**Justification**
Ratio of Filtered Components to Total Components	Detection and Estimation of Filtered Component Percentage in Relation to Total Components

Noisy Time Rejection

Metrics	Justification
Kurtosis	Detection and Estimation of Filtered Component Percentage in Relation to Total Components
Amplitude	Maximum Variation of Displacement in Physical Measurements
Linear Trends in Data Analysis	Utilizing Straight Lines for Linear Data Sets and Identifying Linear Trends
Spectral power	The measurement you are referring to is known as a Power Spectrum, which reports the distribution of spectral power across the frequency rhythms of a signal.

Figure 15 corresponds to the metrics of Early-Stage Data Processing pipeline (PREP) described in Table 3.

The graph represents the original signal, which refers to the raw data mentioned in previous Chapter 2. illustrates that none of the cohorts had channels with NaN values. However, the CHBMP cohort exhibited some outliers with flat channels in the original signal. Regarding channels with shunts, most cohorts displayed variability in the original signal, except for UdeA 2, which only had outliers. After interpolation, the variability decreased, but outliers still persisted across all cohorts.

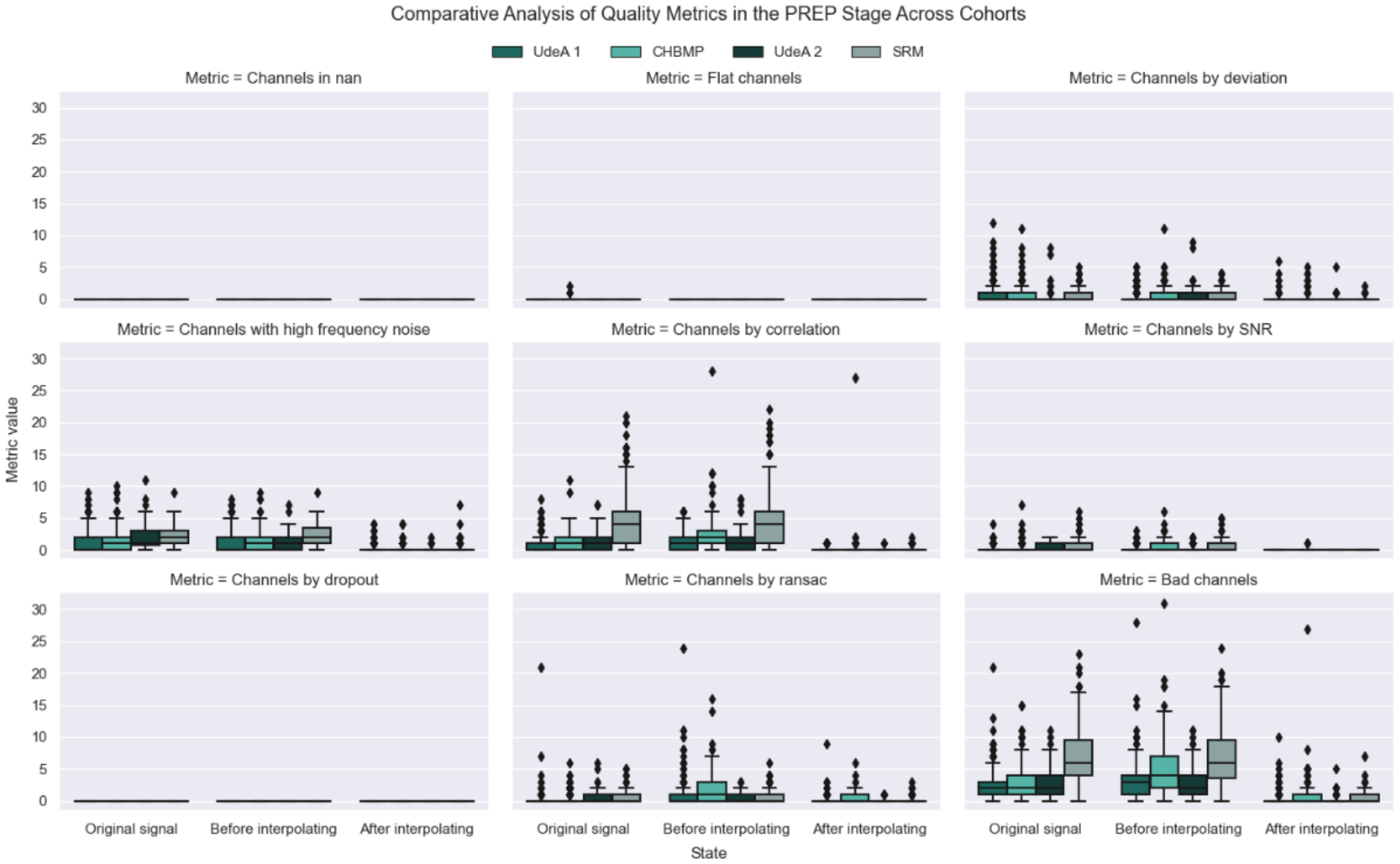


Figure 15 Comparative analysis of quality metrics in the PREP

The behavior of channels affected by high-frequency noise was similar to the previous metrics, where only outliers were observed after interpolation. Analyzing channels based on correlation, it was noted that the SRM cohort exhibited greater variation and a higher number of outliers in the original signal. Before interpolation, the CHBMP cohort displayed notably distant outliers that persisted even after interpolation. Other cohorts showed significant improvement, although a few outliers remained.

Considering the signal-to-noise ratio, most cohorts demonstrated an almost perfect behavior with metric values close to zero after interpolation. None of the cohorts experienced dropouts or carcasses in the original signal. Regarding carcasses identified using the RANSAC algorithm, UdeA2 and SRM cohorts exhibited a reduction in variation, while the variation in the CHBMP cohort increased.

Lastly, in terms of bad channels, all cohorts showed variation in the original signal, with the SRM cohort displaying the most significant variability. After interpolation, some outliers persisted, and variation was still observed in the CHBMP and SRM cohorts.

Figure 16 corresponds to the metrics of Wavelet Cleaning and Independent Component Analysis (ICA) Technique in Combination described in Table 3.

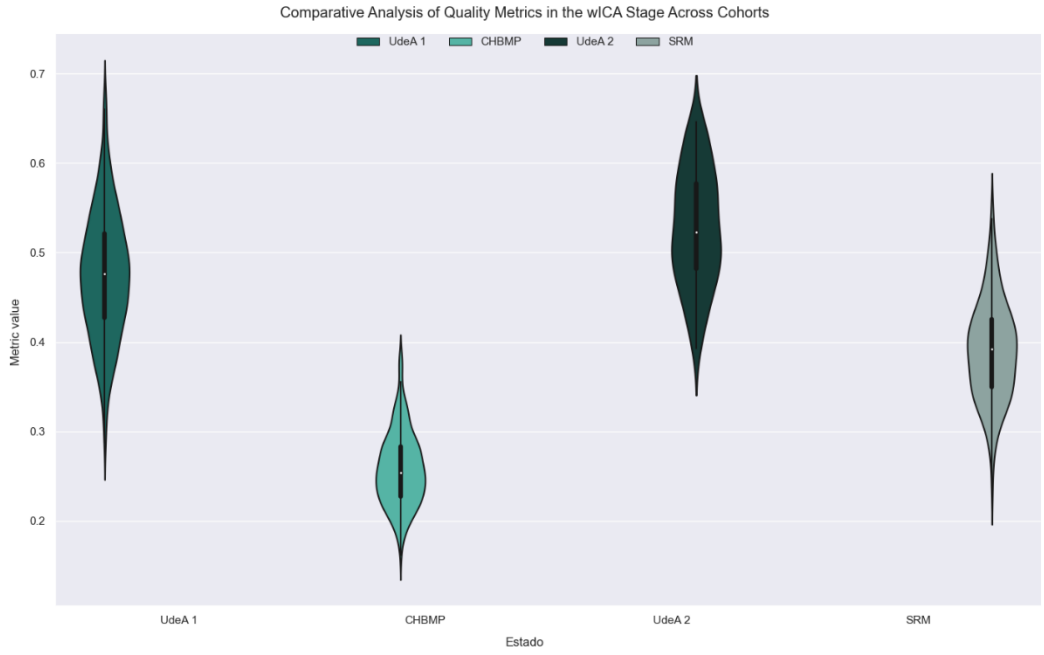


Figure 16 Wavelet Cleaning and Independent Component Analysis (ICA) Technique in Combination

The graph consists of violin diagrams representing each of the project cohorts, namely UdeA 1, UdeA 2, SRM, and CHBMP. The violin diagrams depict the values obtained from the metric derived from the combination of wavelet cleaning techniques and independent component analysis (ICA).

Analyzing the UdeA 1 violin, it exhibits a greater dispersion of values ranging from 0.25 to 0.7. The median, indicated by the point inside the violin, appears to be around 0.45. The width of the violin is similar to that of UdeA 2, but UdeA 1 shows a higher concentration of data closer to the mean value. This suggests that UdeA 1 has a wider range of values but a higher density around its median.

In contrast, the CHBMP violin demonstrates the smallest dispersion, spanning from 0.15 to 0.4. The median value is approximately 0.25, and the width of the violin is observed to be wider at the center compared to the other cohorts. This implies that CHBMP has a more concentrated distribution of values, with less variability overall.

The UdeA 2 violin differs from the other cohorts as its ends are not as sharp, indicating a relatively consistent width from end to end. The metric values for UdeA 2 range from 0.33 to 0.7, and the median value is observed to be around 0.51. This suggests that UdeA 2 has a moderate range of values with a relatively even distribution across the metric scale.

Lastly, the SRM violin exhibits metric values between 0.2 and 0.59. The majority of the values cluster around the median value of 0.4. This indicates that SRM has a relatively narrow range of values, with a significant concentration close to the median.

Overall, the violin diagrams provide insights into the distribution and variation of metric values among the different cohorts. The differences in dispersion, concentration, and range of values observed in the violin plots contribute to understanding the distinct features and patterns exhibited by each cohort in relation to the combined wavelet cleaning and ICA metric.

Figure 17 corresponds to the metrics of Noisy Time Rejection described in Table 3

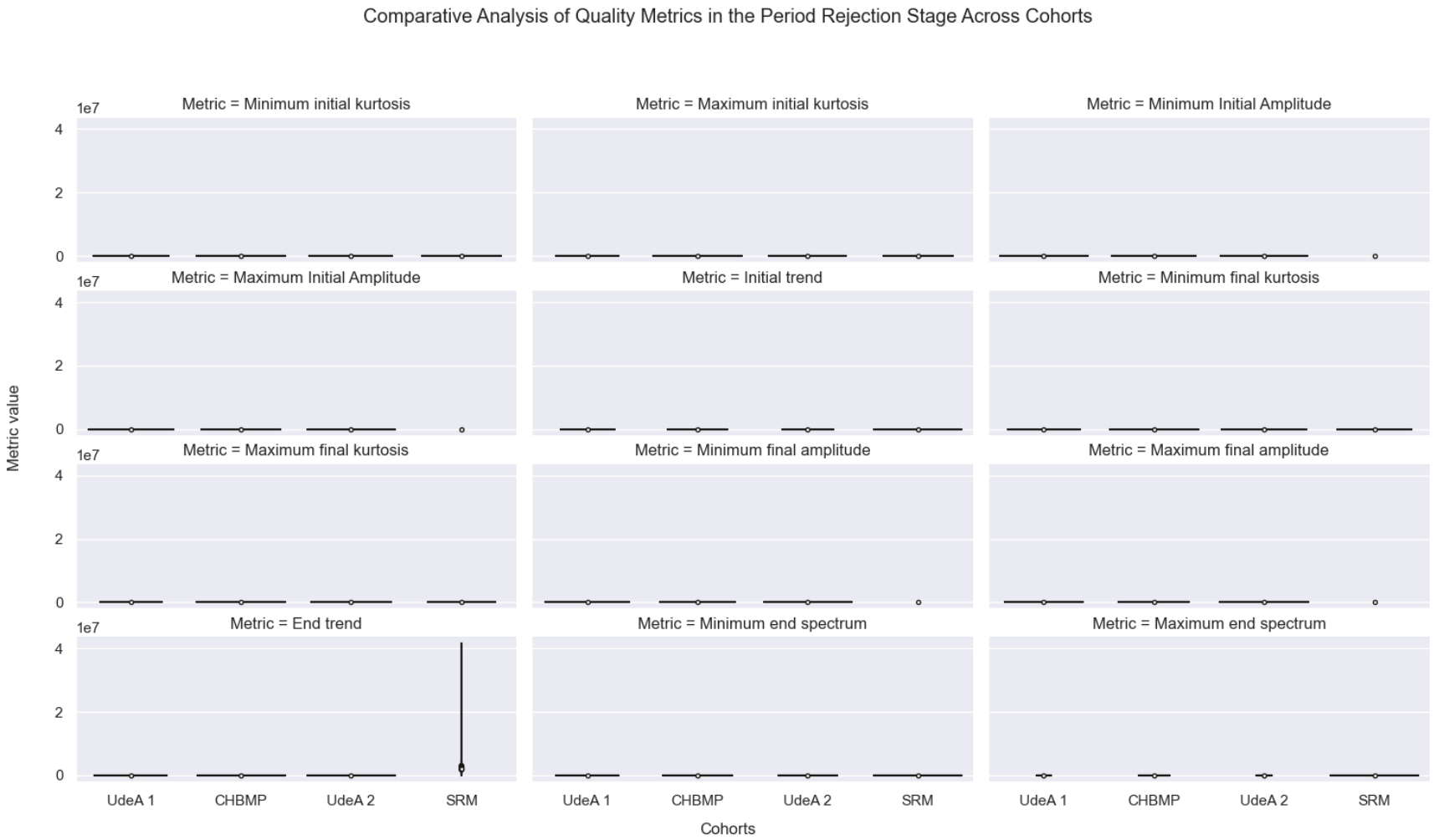


Figure 17 Noisy Time Rejection

Finally, the rejection graph illustrates that the majority of the metrics for all cohorts remain at zero, indicating that most epochs are not rejected based on the applied criteria. However, there is a notable exception in the "End Trend" metric, where the SRM cohort exhibits a significant peak reaching a metric value of $4x10^7$. This indicates that the rejection pattern or direction of epochs considered "bad" in the EEG data analysis has experienced a substantial increase in the SRM cohort.

The rejection graph provides insights into the temporal dynamics of epoch rejection and its relationship with relevant variables. It demonstrates how the number or percentage of rejected epochs fluctuates over time or under specific conditions. In this case, the "End Trend" metric highlights a distinct pattern for the SRM cohort, suggesting a pronounced shift in the rejection of "bad" epochs compared to the other cohorts and that it is consistent with Figure 15 where SRM presented greater dispersion in the PREP in the bad channels.

The observed peak in the SRM cohort's rejection trend signifies an intensified identification and exclusion of problematic epochs during the later stages of the recording session. This may imply the presence of specific artifacts or irregularities that were more prominent in the SRM cohort's data, prompting stricter rejection criteria.

Interpreting quality control graphs is essential for gaining a comprehensive understanding of data quality, evaluating the effectiveness of rejection criteria, and

identifying potential factors that influence the rejection of epochs and removal of channels.

The graphs obtained after the completion of these processing stages will serve as a basis for evaluating the effectiveness of the applied methodologies. By comparing these updated graphs to the initial ones, we can observe any noticeable changes in the dispersion and patterns of the data.

This analysis will provide insights into the impact of the processing stages on the quality of the data, as well as the overall effectiveness of the chosen methods in mitigating noise and artifacts. Furthermore, it will allow for a better understanding of the extent to which the subsequent stages have improved the interpretability and reliability of the data.

3.3.5 Feature Extraction

3.3.5.1 Relative Power

To promptly evaluate the EEG signals for pertinent physiological insights, the power values of all individuals in good health are graphed via the power spectrum, focusing on the posterior occipital region.

This visual representation (Figure 18) encompasses channels situated in the Parieto-Occipital area (PO ROI). The resultant signal demonstrates the anticipated physiological behavior, revealing a noticeable reduction of artifacts in the processed signal when contrasted with the original signal. Additionally, the characteristic alpha peak near 10 Hz is distinctly discernible.

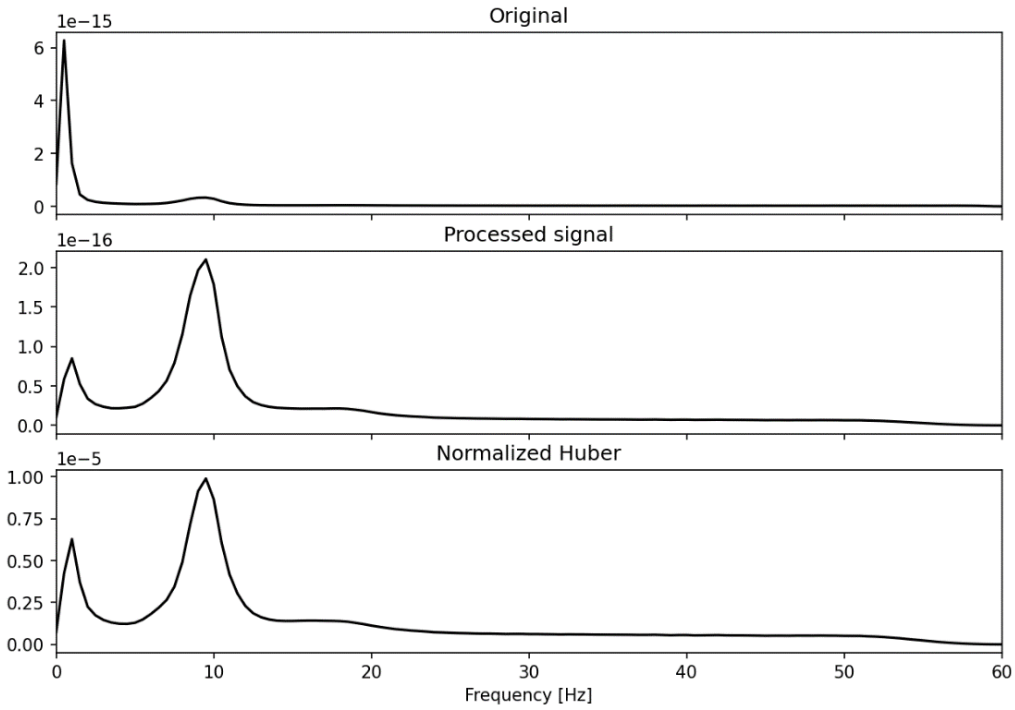


Figure 18 Power Spectrum Analysis of Subjects in the Posterior Occipital (PO) Region

Similarly, in Figure 19, the neural gICA Component (25), which is located in the posterior area, is visualized. It shows the characteristic alpha peak close to 10 Hz compared to the original signal albeit with a lower prominence than the one found for the PO ROI.

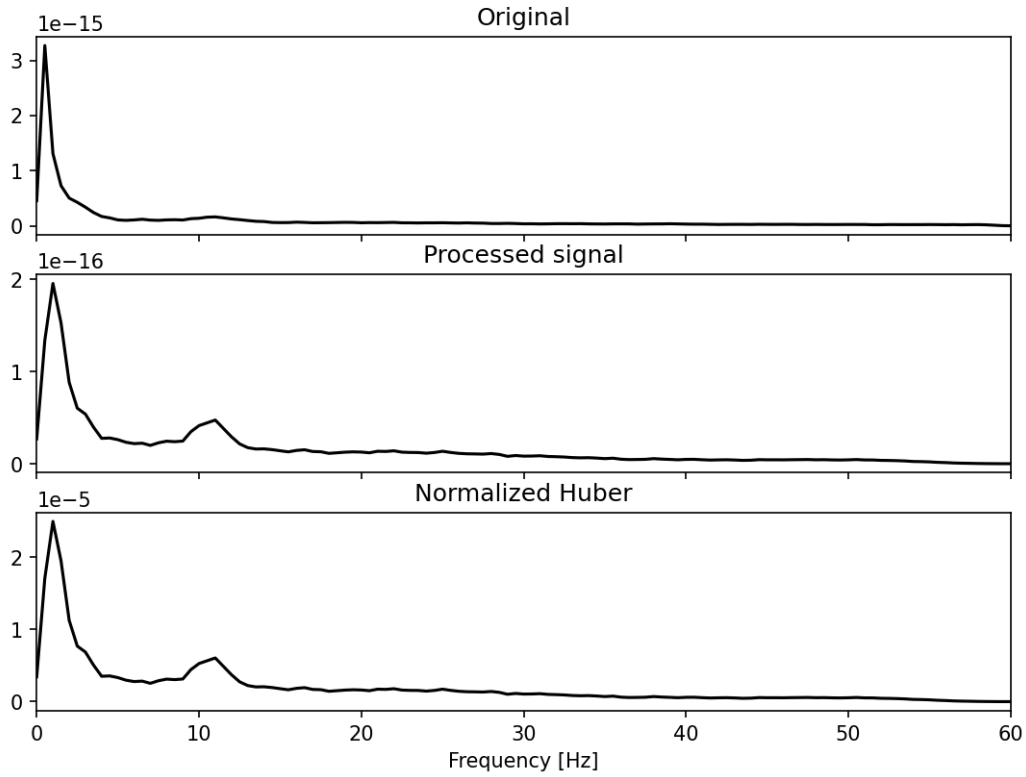


Figure 19 Power Spectrum Analysis of Subjects in the neural gICA Component 25.

Figure 20 depicts the Relative Power of the gamma brain wave across four different cohorts. The y-axis represents the Relative Power feature, while the x-axis corresponds to the different groups. Figure 20 is segmented into eight boxes, each corresponding to one of the eight gICA neural components used in the feature extraction process.

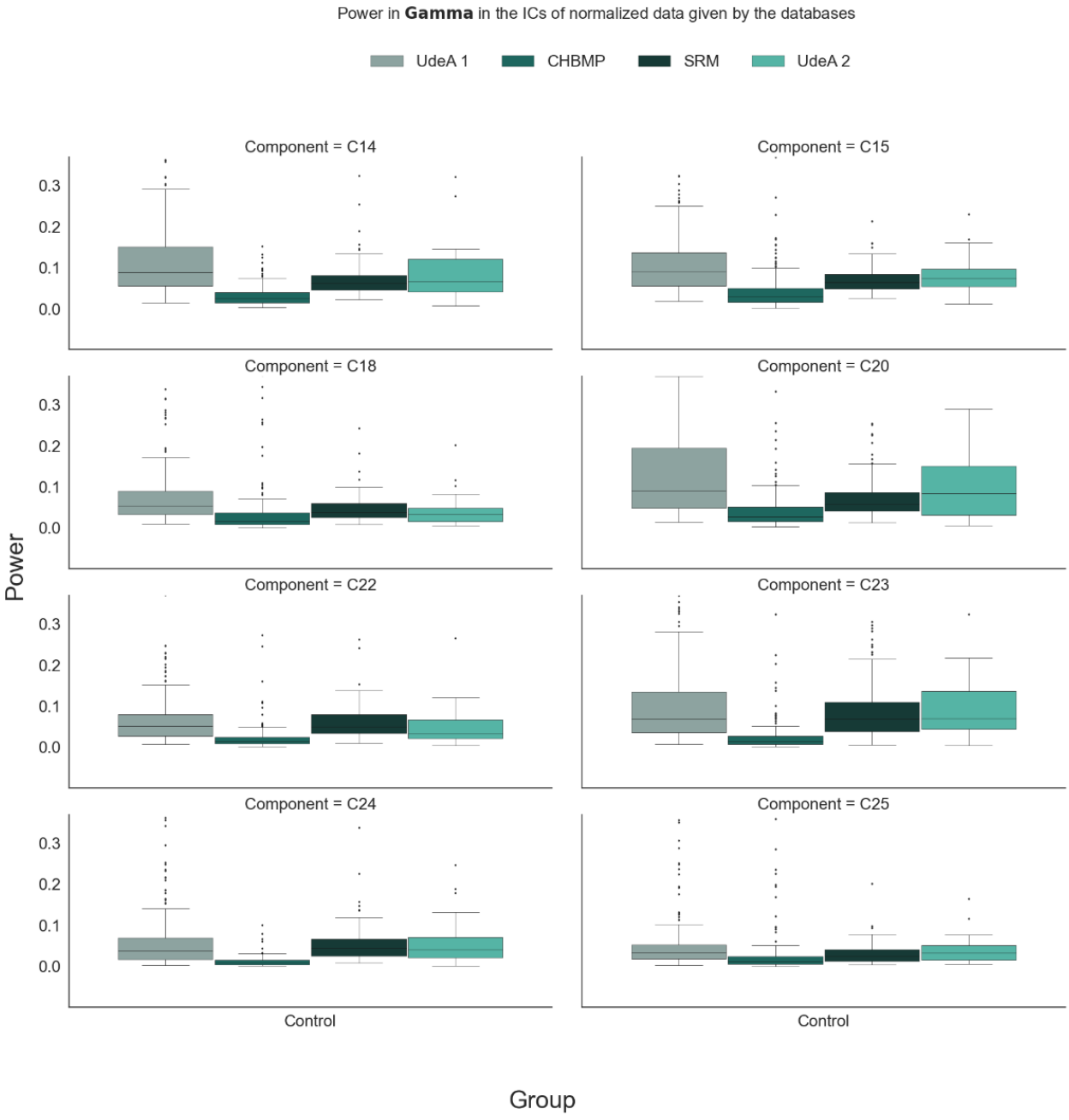


Figure 20 Relative Power of the Gamma brain wave across four cohorts.

When examining Figure 20 several important observations become apparent. First, it is evident that the controls in UdeA 1 cohort have a broader data distribution compared to most of the other cohorts. Conversely, the UdeA 2 and SRM cohorts

exhibit narrower variability than UdeA 1, with both medians closely aligned. In contrast, the CHBMP cohort shows the lowest variability of all cohorts.

Furthermore, it is noteworthy that neural gICA component 25 shows the lowest variability across all cohorts, indicating a relatively consistent pattern. Conversely, neural gICA component 20 has the highest variance between groups, indicating greater variability.

3.3.5.2 Entropy

At this stage, the data was explored by visualizing the distribution of Entropy across cohorts using a Box plot. A representative sample of these plots for neural gICA Components is displayed in Figure 21 below. For a comprehensive collection of these plots, please refer to Annex 3, located at the end of this document.

Figure 21 represents the Entropy of the Delta brain wave across four cohorts for controls. The Y-axis represents the feature of Entropy, while the X-axis represents the groups. Figure 21 is divided into eight boxes, each representing one of the eight neural gICA Components mentioned in the feature extraction.

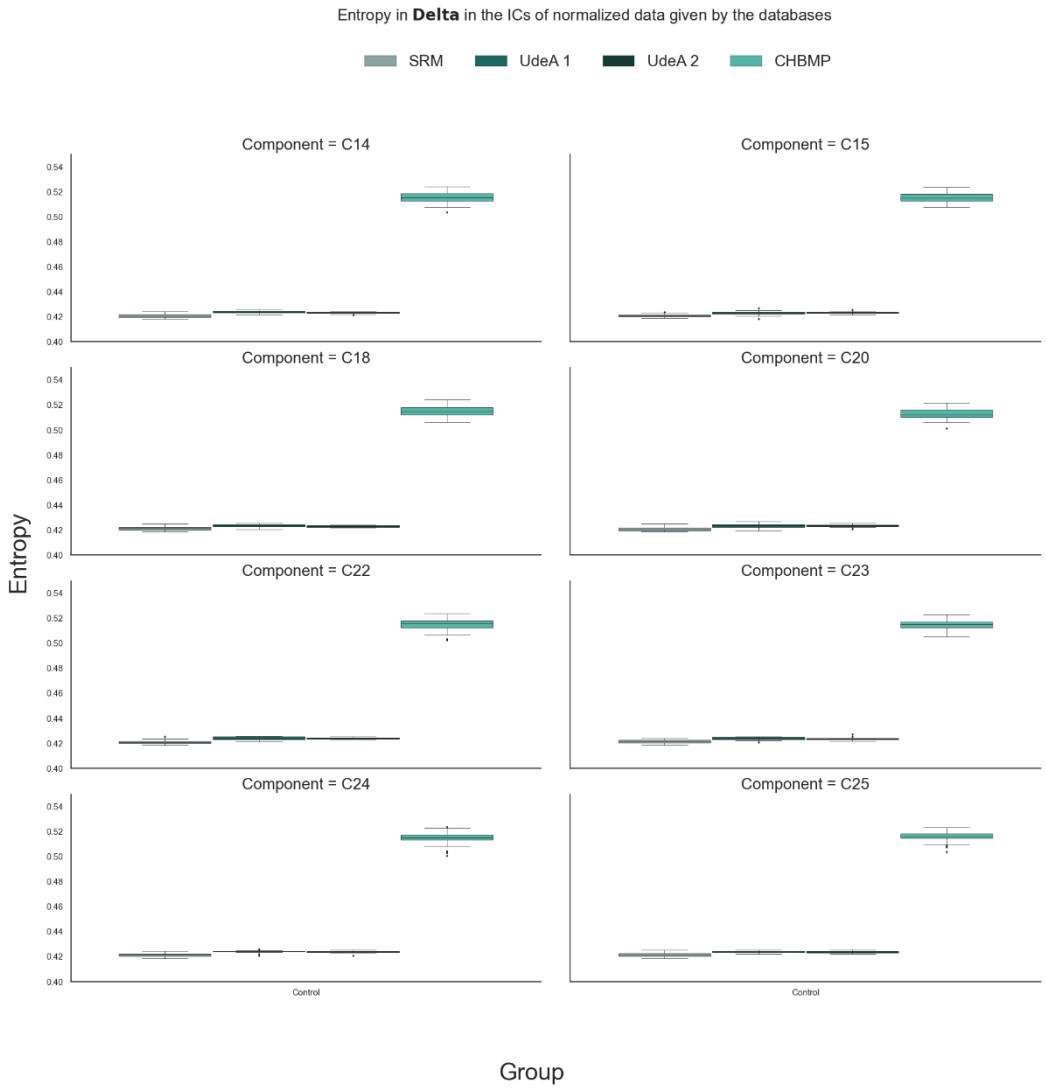


Figure 21 Entropy of the Delta brain wave across four cohorts.

The Delta band was chosen as a representative sample graph due to challenges in adequately visualizing the Gamma band, primarily for the CHBMP cohort. However, the distribution features are not clearly discernible, making it challenging to assess the dimensions of the distributions. This difficulty arises from the significant difference exhibited by the controls in CHBMP cohort concerning the Entropy result. In general, there are few notable differences within the components.

3.3.5.3 Coherence

At this stage, the data was explored by visualizing the distribution of Coherence across cohorts using a Box plot. A representative sample of these plots for neural gICA Components is displayed in Figure 22 below. For a comprehensive collection of these plots, please refer to Annex 3, located at the end of this document.

Figure 22 shows the gamma brain wave Coherence over four cohorts for controls. The y-axis represents the Coherence function, while the x-axis represents the different groups. The graph is divided into eight boxes, each symbolizing one of the eight gICA neural components used in the feature extraction process.

Looking more closely at the evaluation of Coherence within the gamma band, a distinct pattern emerges. In particular, the controls in CHBMP cohort shows a higher degree of variability compared to the other cohorts. There are also several outliers, particularly within neural gICA components 18, 22, and 23. While the medians of the boxes are generally close together, this is particularly evident in neural gICA component 24.

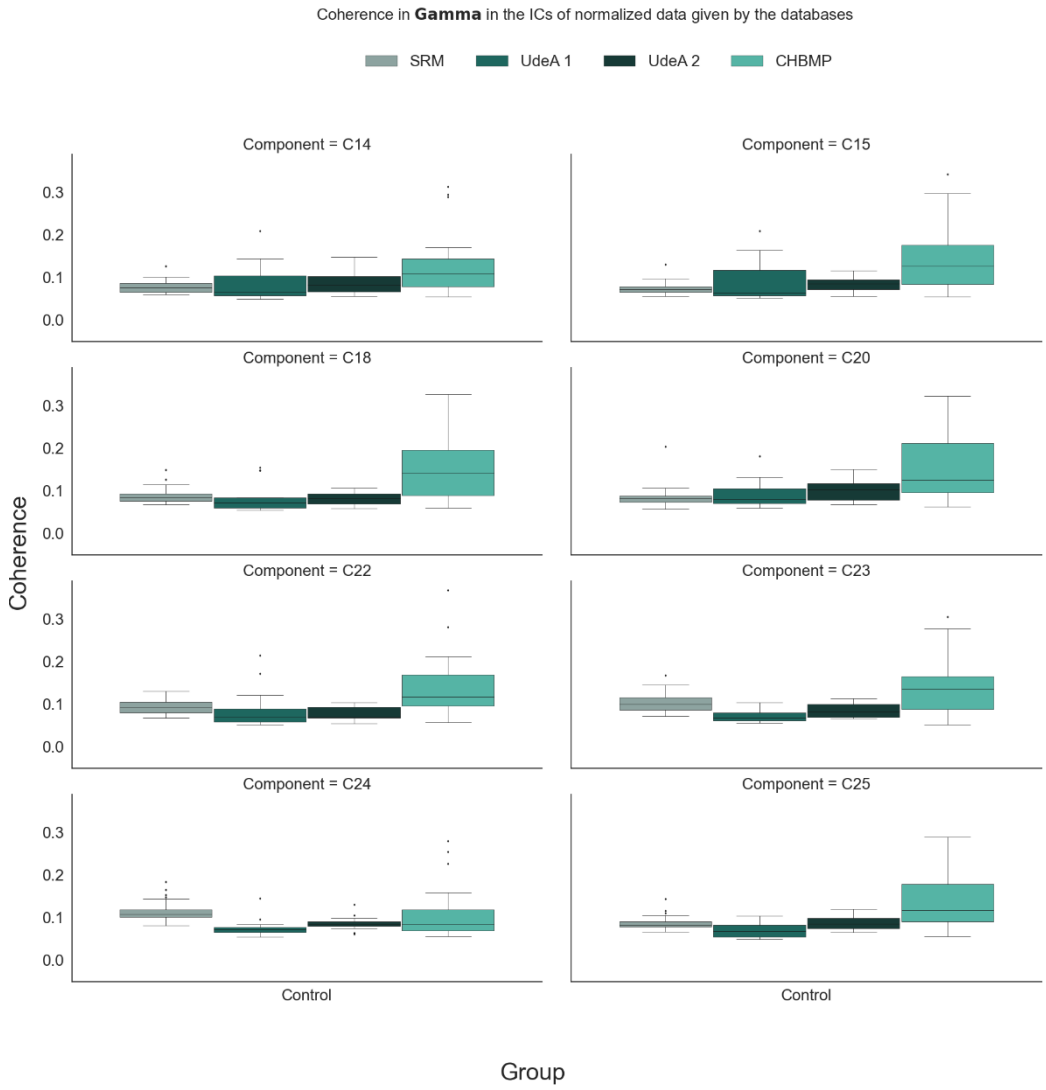


Figure 22 Coherence of the Gamma brain wave across four cohorts.

3.3.5.4 Cross Frequency

Figure 23 depicts the Cross Frequency of the gamma brain wave across four different cohorts in controls. The y-axis represents the Relative Power feature, while the x-axis corresponds to the different groups. Figure 23 is segmented into eight boxes, each corresponding to one of the eight gICA neural components used in the feature extraction process.

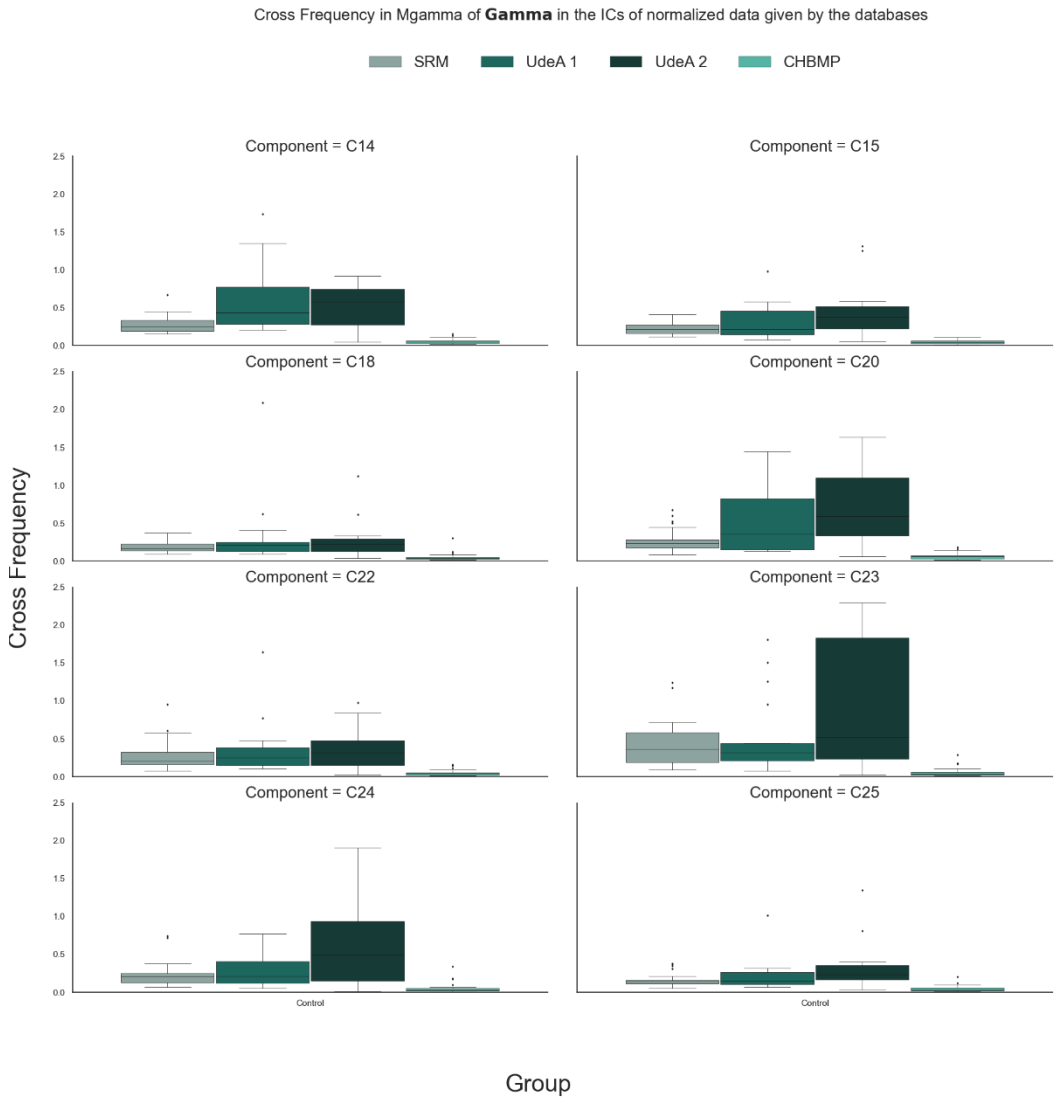


Figure 23 Cross Frequency of the Gamma brain wave in the gamma modulated band across four cohorts.

Figure 23, it becomes clear that the middle values (medians) for the different components are quite similar. However, the controls in CHBMP group stands out as having the most distinct differences from the others, which is consistent with what we saw in the earlier data features.

What's particularly interesting is what we observe in the UdeA 1 group, specifically within neural gICA Component 20. This suggests that this group shows a wider range of differences in comparison to the other groups, especially in this specific aspect. Similarly, the UdeA 2 group also shows more differences in Component 22. In contrast, both Component 18 and Component 25 show less variation across all groups, but at the same time, they have quite a few data points that are different from the norm.

3.3.5.5 Synchronization Likelihood

Figure 24 depicts the Synchronization Likelihood of the gamma brain wave across four different cohorts. The y-axis represents the Relative Power feature, while the x-axis corresponds to the different groups. Figure 24 is segmented into eight boxes, each corresponding to one of the eight gICA neural components used in the feature extraction process.

In Figure 24, it is evident that the controls in CHBMP cohort continues to exhibit the highest variation among the cohorts. This consistency in behavior across all neural gICA Components is noteworthy. Additionally, the number of outliers remains relatively constant across all components, indicating a consistent presence of extreme values in the data.

There are fewer outliers in this feature. However, the medians of the different cohorts don't show any similarity either. While these medians are close to each other, none are at the same level as the others. It's also worth noting that for most

of the components, the means follow an ascending order: SRM, UdeA1, UdeA2 and CHBMP.

In all the components, the CHBMP cohort showed greater variation.

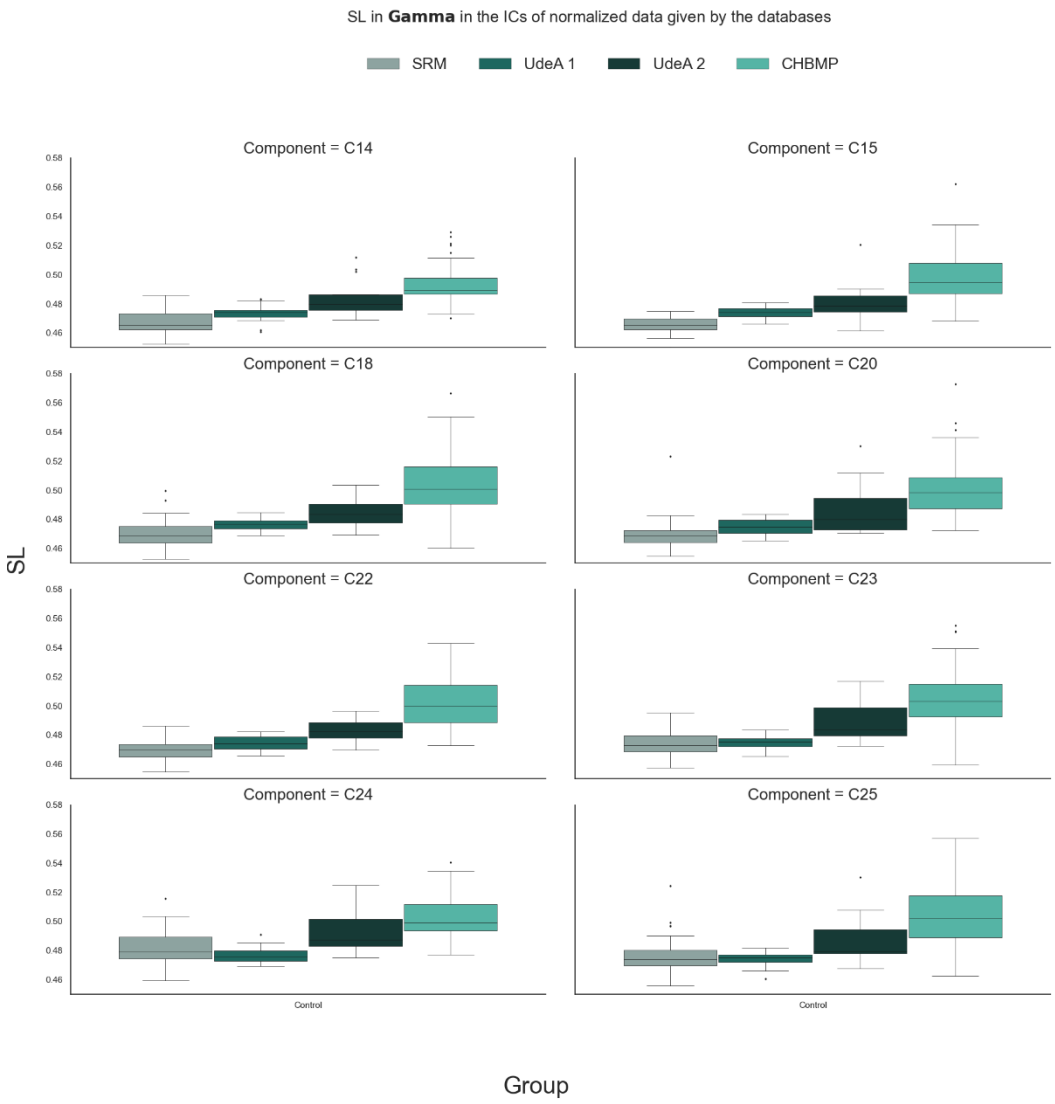


Figure 24 Synchronization Likelihood of the Gamma brain wave band across four cohorts.

3.3.6 Matching between subjects

As we have been mentioning in Chapter 1, EEG data has shown promise as a potential biomarker for Alzheimer's risk. However, analyzing such data is often complicated by differences in data collection protocols, instruments, and cohorts. Harmonizing different EEG data cohorts is essential to improve the efficiency and accuracy of machine learning models in predicting Alzheimer's risk.

Observational studies that collect EEG data are subject to confounding due to non-random treatment assignment. **To address this, MatchIt is a powerful tool for causal inference that can be implemented in R. It creates matched pairs of individuals from two groups (Alzheimer gene carriers (G1) and Control plus G2 subjects) based on their similarity in pretreatment covariates, such as sex and age. The resulting matched pairs have similar distributions of confounding variables, allowing for a more accurate estimation of the causal effect of the treatment variable (in this case, PSEN1-E280A gene carrier status) on the outcome of interest.**

Using MatchIt can improve the accuracy and efficiency of machine learning models in predicting Alzheimer's risk by controlling for confounding variables. This ensures that any differences in EEG data between the two groups are due to the treatment variable (i.e., PSEN1-E280A gene carrier status) rather than other confounding factors. This can lead to more accurate predictions of Alzheimer's risk,

which is essential for developing effective treatments and interventions for this debilitating disease.

Figure 25 displays the data entered for groups G1 (group1) and Controls plus G2 (group2) of all joined cohorts (UdeA 1, UdeA 2, SRM, CHBMP) and the application of the MatchIt algorithm in R. The resulting matched dataset contains twice as many Controls as carriers of the PSEN1-E280A gene.

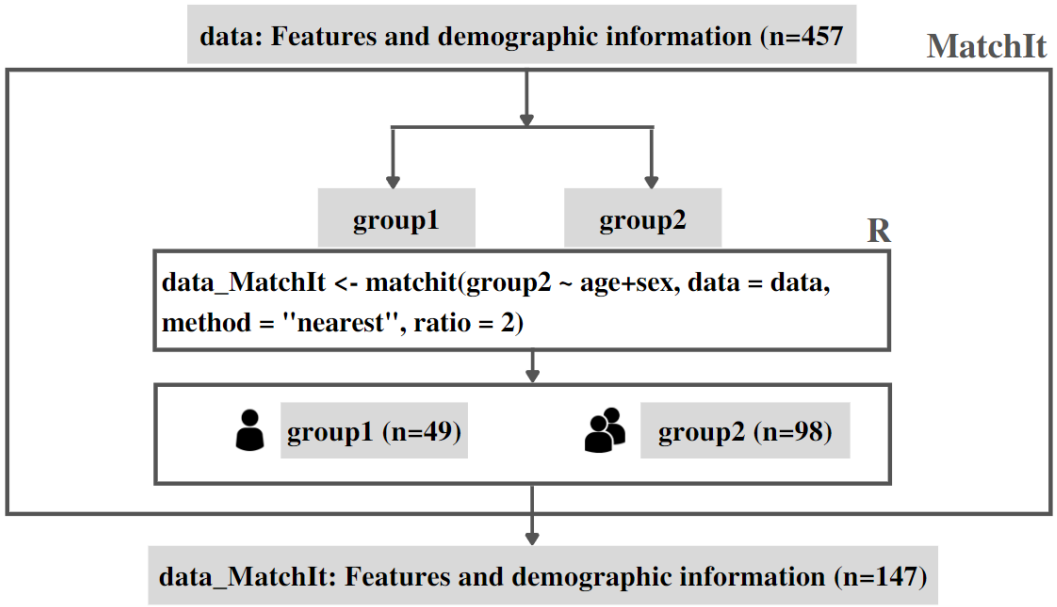


Figure 25 Application of the MatchIt algorithm in R. In the R algorithm or rpy2 in python, you can use MatchIt to include 457 age- and sex-matched records for two groups, carriers (G1) and controls plus G2, by applying the 'matchit' function at a 2:1 ratio. This process results in a G1 group of 49 subjects and a control group of 98 subjects, for a total of 147 subjects.

It's worth noting that the reduction from 457 to 147 subjects may seem extreme but it is due to the longitudinal nature of the UdeA 1 and SRM cohorts. Each subject within the studies has multiple registrations, as they attended different registration

sessions over a period of 6 months to 2 years. To prevent subject duplication, a filter was applied during MatchIt to include only one registration per subject.

Table 4 Provides the statistical description of each selected cohort after Matching.
Healthy: Control Group + G2 Group

		Age			Sex
		count	mean	std	F/M
Database	Group				
UdeA 1	Healthy	17	30.12	5.41	10/7
	G1	27	30.16	5.86	15/12
CHBMP	Healthy	38	27.63	6.67	13/25
UdeA 2	Healthy	12	31.42	7.15	10/2
	G1	22	29.54	5.10	14/8
SRM	Healthy	31	30.77	5.21	19/12
	Total	147			

Table 4 provides a statistical overview of the data used in the machine learning model. This data was matched using the MatchIt algorithm, taking into account age and gender, as shown in Figure 25.

3.3.7 neuroHarmonize Implementation

The implemented algorithm for group harmonization of neuroHarmonize between individuals with and without the PSEN1-E280A variant includes only factors that contribute to the change in acquisition. It is assumed that these factors should be similar between the two groups. Examples of such factors are the acquisition team, the city, the type of cap, the type of reference, etc.

For this project, the covariates considered in the algorithm were Cohort, Sex, and Age. These covariates were chosen based on their potential influence on the acquisition and the need to account for their effects during the harmonization process with neuroHarmonize. By including these covariates, the algorithm aims to adjust for any differences associated with cohort, sex, and age, ensuring that the harmonization process is more accurate and effective.

It is crucial to emphasize that the nature of Relative Power poses challenges for the harmonization process. Therefore, to evaluate the effectiveness of the neuroHarmonize methodology, a necessary step was taken extracting the component specific to one of the bands, in this case, the Gamma band. This extraction allowed for the harmonization process to be conducted in the other bands, after which the Gamma band was reintroduced proportionally using a relationship Equation 17.

$$\gamma_h = 1 - \sum(\delta + \theta + \alpha + \beta)$$

Equation 17

Where $\gamma_h = harmonized\ gamma, \delta = Delta,\ \theta = Theta,\ \alpha = Alpha,\ \beta = Beta$. This equation (Equation 17) allows for the calculation of the Relative Power specifically for the Gamma band, ensuring its inclusion in the analysis alongside the other frequency bands, see Figure 26.

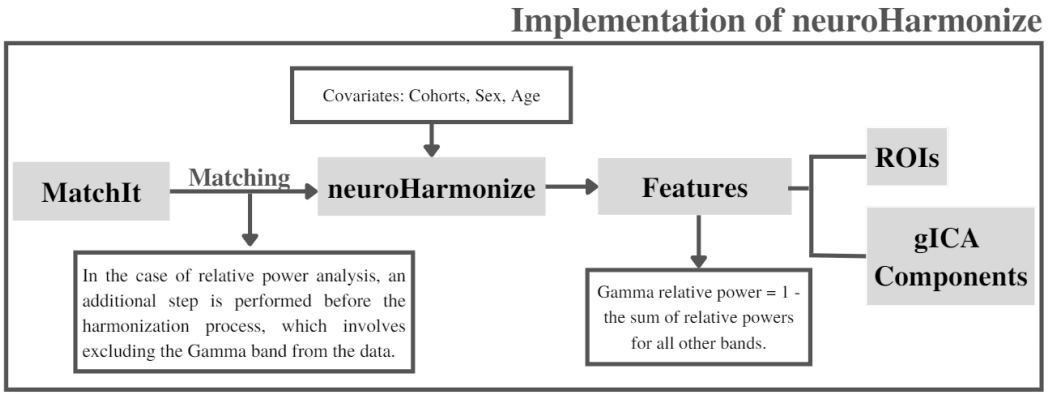


Figure 26 Incorporation and Exclusion of Gamma in the neuroHarmonize Process.

An essential step in the harmonization process with neuroHarmonize involved applying a heuristic transformation to prevent negative values from arising after harmonization. This transformation was necessary due to the presence of very small or close to zero values in the data.

To achieve this transformation, a constant value of 0.001 was added to each data point before initiating the harmonization procedure. The addition of this small constant ensured that all values remained positive, see Figure 27.

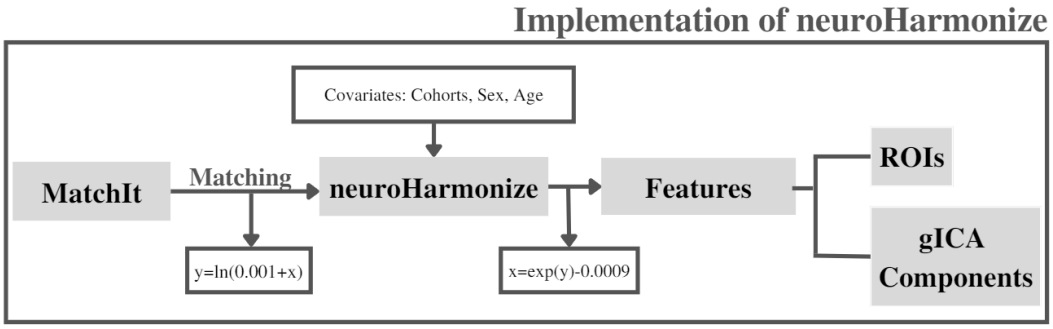
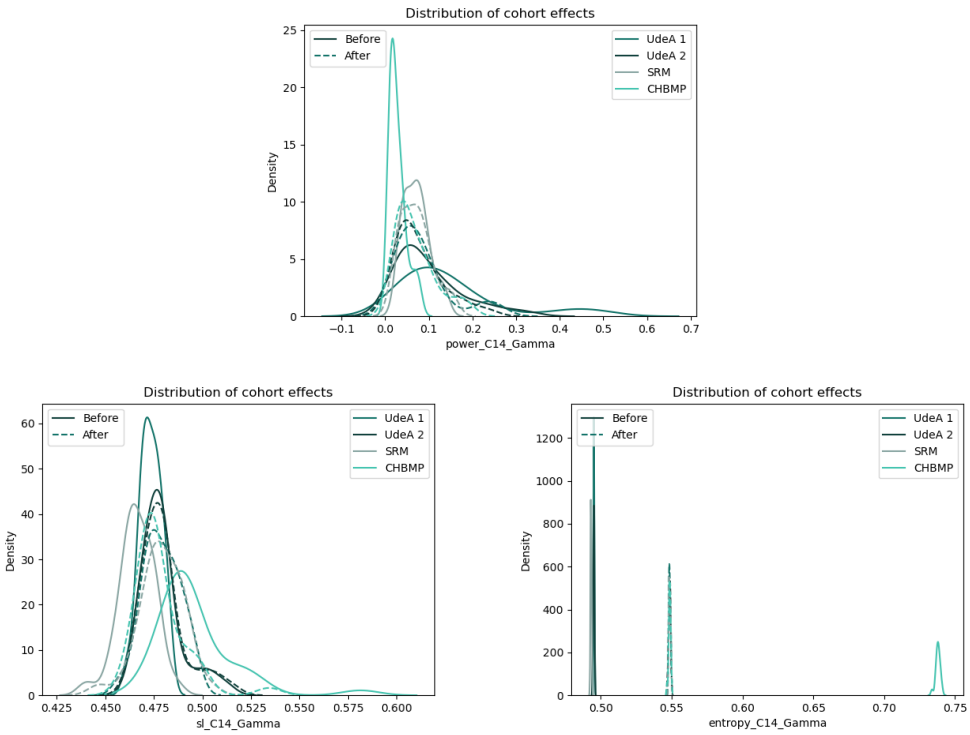


Figure 27 Heuristic transformation to prevent negative values from arising after neuroHarmonize.

The reverse transformation was then applied to restore the data to its original scale and facilitate interpretation within the context of the initial values. By utilizing the exponential function, followed by the subtraction of 0.001, the adjustment made during the initial transformation was undone, resulting in the recovery of the original values.

The same method was applied to the feature extraction process for each ROI and each neural gICA Components, resulting in a visualization of the distribution of cohort effects. A representative sample of these graphs for one neural gICA Component is presented below see Figure 28. For a complete set of graphs, please refer to Annex 4, which is located at the end of this document.



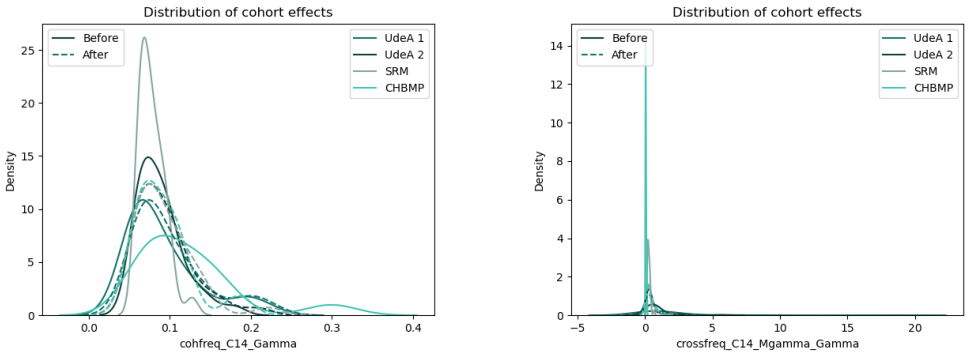
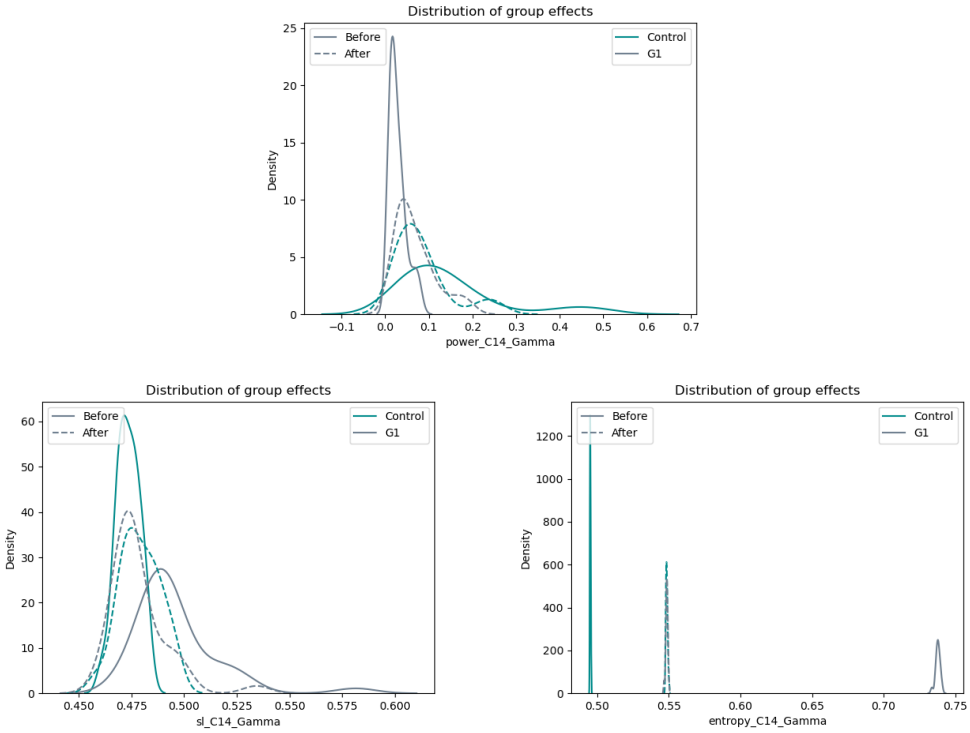


Figure 28 Comparing Pre- and Post-Cohort Effects: Analyzing Distribution Patterns



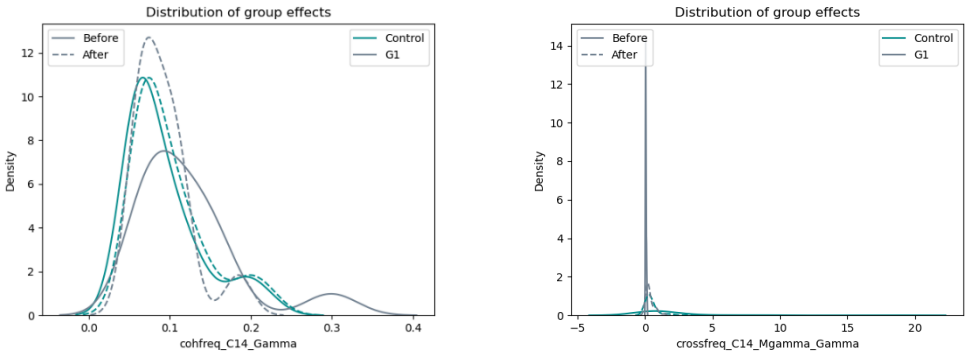


Figure 29 Comparing Pre- and Post-Group Effects: Analyzing Distribution Patterns.

The same method was applied to the distribution of group effects. A representative sample of these graphs for one neural gICA Component is presented below see Figure **29**

For a complete set of graphs, please refer to Annex 4, which is located at the end of this document.

3.3.8 Statistical analysis of harmonized features

3.3.8.1 Descriptive statistics

After completing the matching and harmonization stages, the next step involves conducting an analysis using descriptive statistics, which includes generating Box plots for each stage. It is important to note that during the matching process, the groups are narrowed down to specific interest groups.

Out of the four groups that have been discussed so far (Controls, G1, and G2), the group of primary interest regarding individuals at risk of Alzheimer's is the one comprising carriers of the genetic variation PSEN1-E280A, known as the G1 group.

Therefore, the paired testing was conducted specifically between the groups that help distinguish this risk, as presented in Table 5.

Table 5 Groups and cohorts

Paired groups	Cohorts
G1 with Controls	UdeA 1, UdeA 2, SRM, CHBMP
G1 with G2	UdeA 1, UdeA 2

Table 5, 6 and 7 present the sample sizes obtained for each paired group. These sample sizes reflect the number of individuals included in the analysis, providing important information about the availability and representativeness of the data for each specific comparison.

Table 6 The sample sizes obtained for G1, and Controls plus G2 paired group.

G1 with Controls		
	ROIs	**Neural gICA Components**
G1	48	49
Controls	96	98
Total	144	147

Table 7 The sample sizes obtained for G1, and G2 paired group.

G1 with G2		
	ROIs	**Neural gICA Components**
G1	48	49
G2	54	52
Total	102	101

Figure 30 presents a comparison of the paired groups of Power Relative in Delta band neural gICA Components before and after matching, utilizing the familiar boxplot format discussed earlier. This visualization allows for a clear understanding of the changes in the distribution and features of the data following the matching process.

By examining the boxplots, we can assess how the matching procedure has affected the distribution of variables of interest within each paired group.

In (a) the box plots illustrate data without the use of neuroHarmonize. In (b) the boxplots illustrate data with the use of neuroHarmonize, and it is evident that reducing systematic differences between groups facilitates meaningful comparisons, i.e., the median is similar between groups, the distribution is better, but the presence of outliers in most of the components does not decrease.

Finally, it is observed that the distributions of the paired groups appear more similar after matching, which indicates a successful alignment of variables and a possible reduction in potential confounding factors.

Power Relative in Delta band neural gICA Components before and after matching G1 with Controls

(a) Data without neuroHarmonize

(b) Data with neuroHarmonize

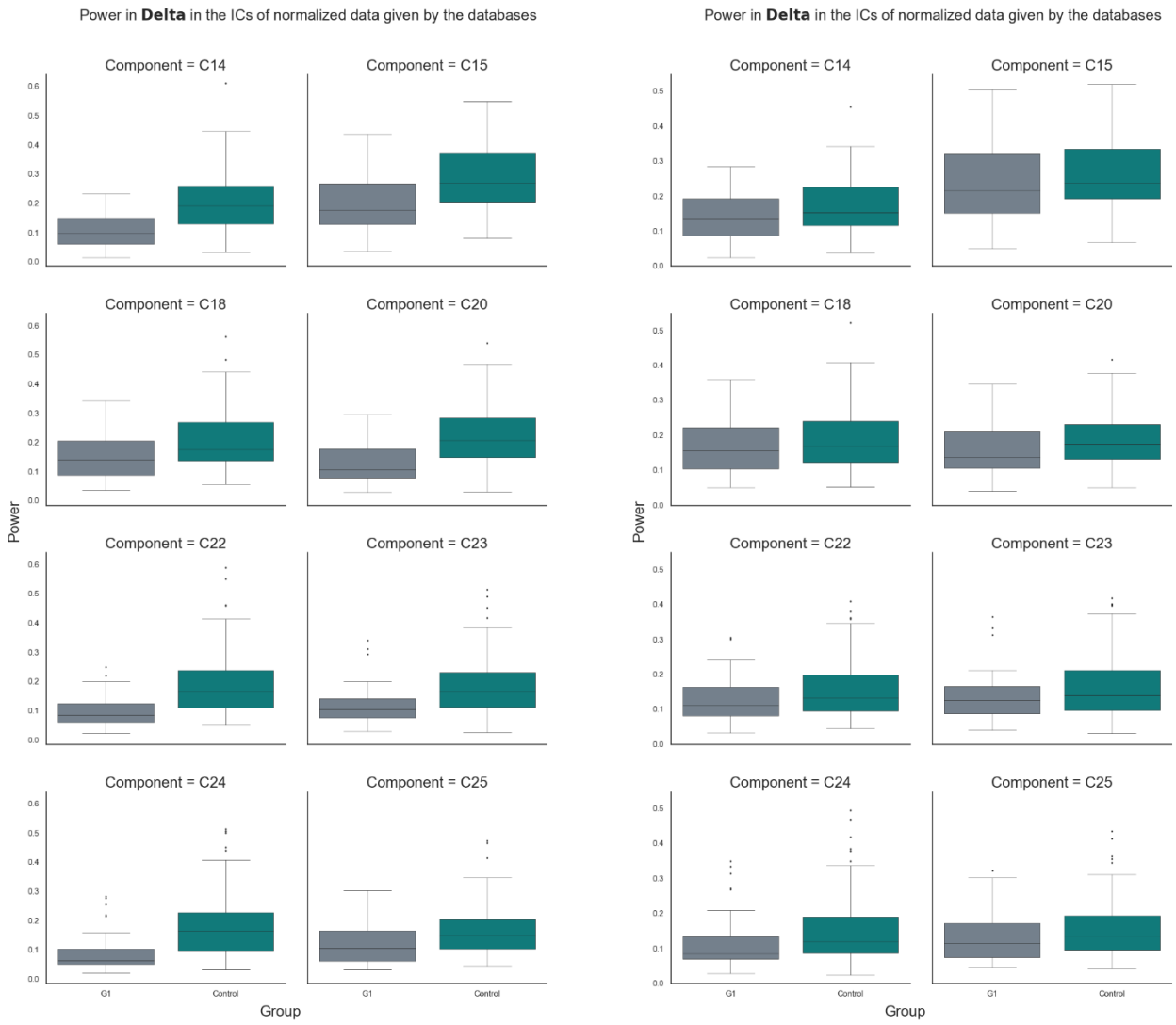


Figure 30 Power Relative in Delta band neural gICA Components before and after matching

These same graphs were generated for all the features and for the two areas of interest, namely ROIs and neural gICA Components. The results consistently demonstrate the improvements observed after using neuroHarmonize. While a selection of graphs is presented here for analysis, the remaining graphs can be found in Annex 4 for reference.

The Entropy feature has exhibited considerable variability, as evident from Figure 21 presented in section 3.3.5.2 Therefore, it is of particular interest to examine its behavior following the using neuroHarmonize. See Figure 31.

While in (a) it is not possible to observe the box of G1, in (b) the distribution, median, and length of the whiskers (representing outliers) are practically the same in both groups. This alignment and consistency in the boxplot representation after using neuroHarmonize greatly enhance the interpretability of the graph.

While analyzing the data, an unforeseen outcome emerges in relation to Relative Power, particularly within the Gamma band. Despite the unexpected nature of this result, it can be attributed directly to the decision outlined in section 3.3.7 of the neuroHarmonize methodology. This section specifies the inclusion of the Gamma band after the harmonization process, considering the unique features of Relative Power.

Remarkably, as seen in Figure 32, the obtained results demonstrate negative values for Relative Powers within the gamma band across all components.

Entropy in Gamma band neural gICA Components before and after matching G1 with Controls

(a) Data without neuroHarmonize **(b) Data with neuroHarmonize**

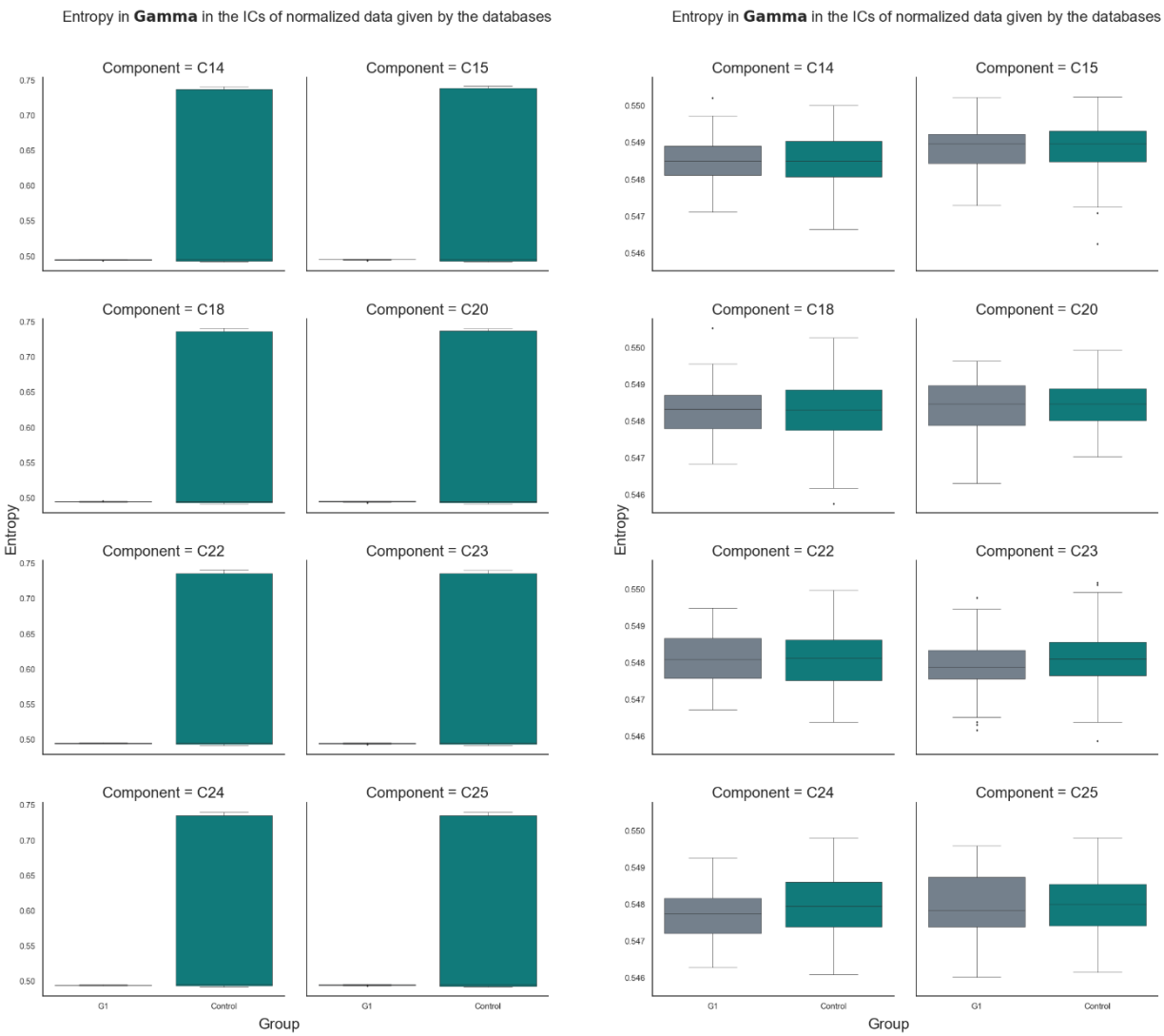


Figure 31 Entropy in Gamma band neural gICA Components before and after matching

Power Relative in Gamma band neural gICA Components before and after matching G1 with Controls

(a) Data without neuroHarmonize **(b) Data with neuroHarmonize**

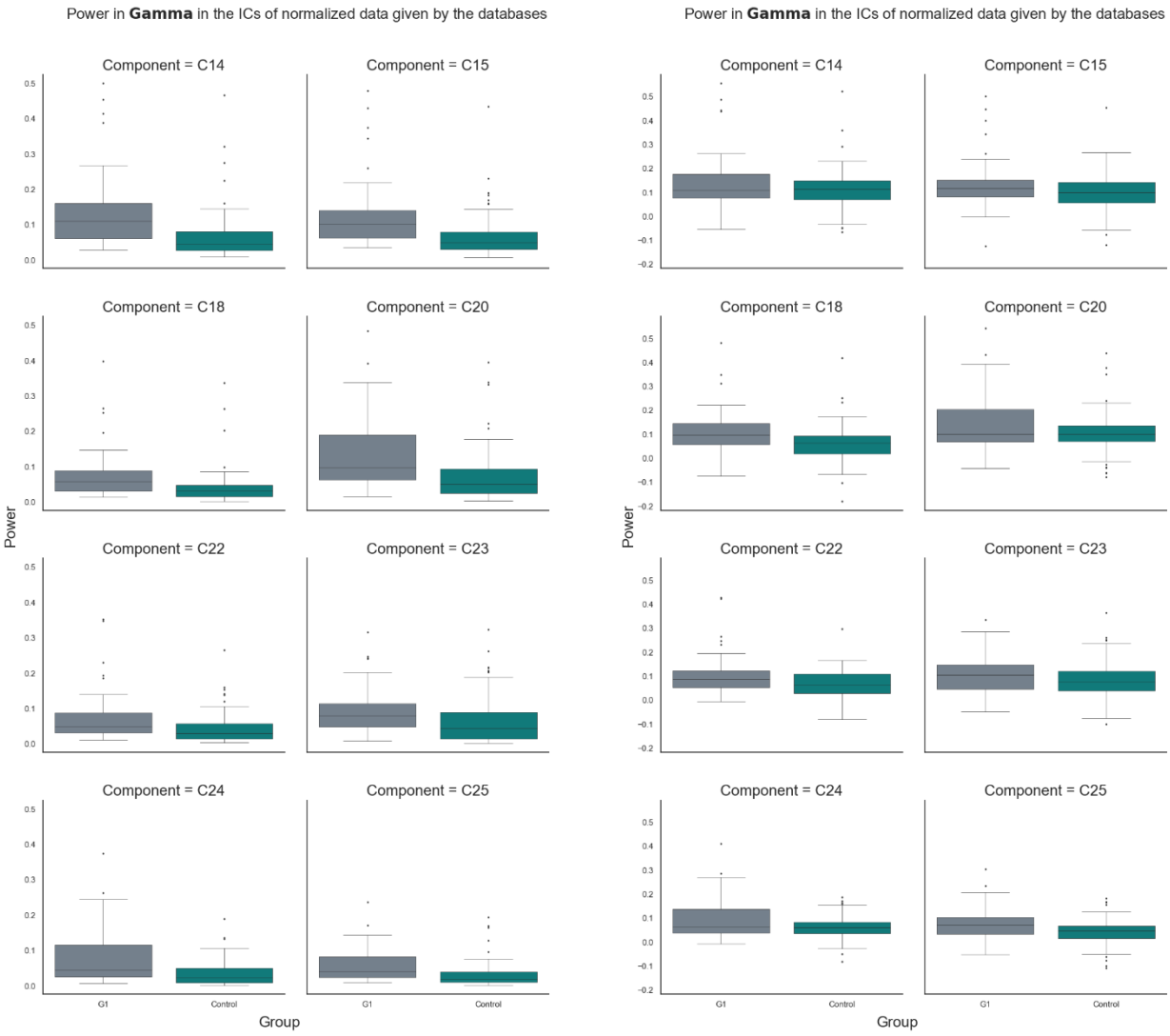


Figure 32 Power Relative in Gamma band neural gICA Components before and after matching.

3.3.8.2 Effect size

Effect sizes hold particular importance when assessing the effectiveness of harmonization techniques, offering a means to transcend significance testing and explore the tangible implications of observed distinctions. They provide standardized measurements that quantify the degree of divergence between groups (Figure 33), encapsulating the strength and orientation of relationships or discrepancies across variables, regardless of the size of the sample.

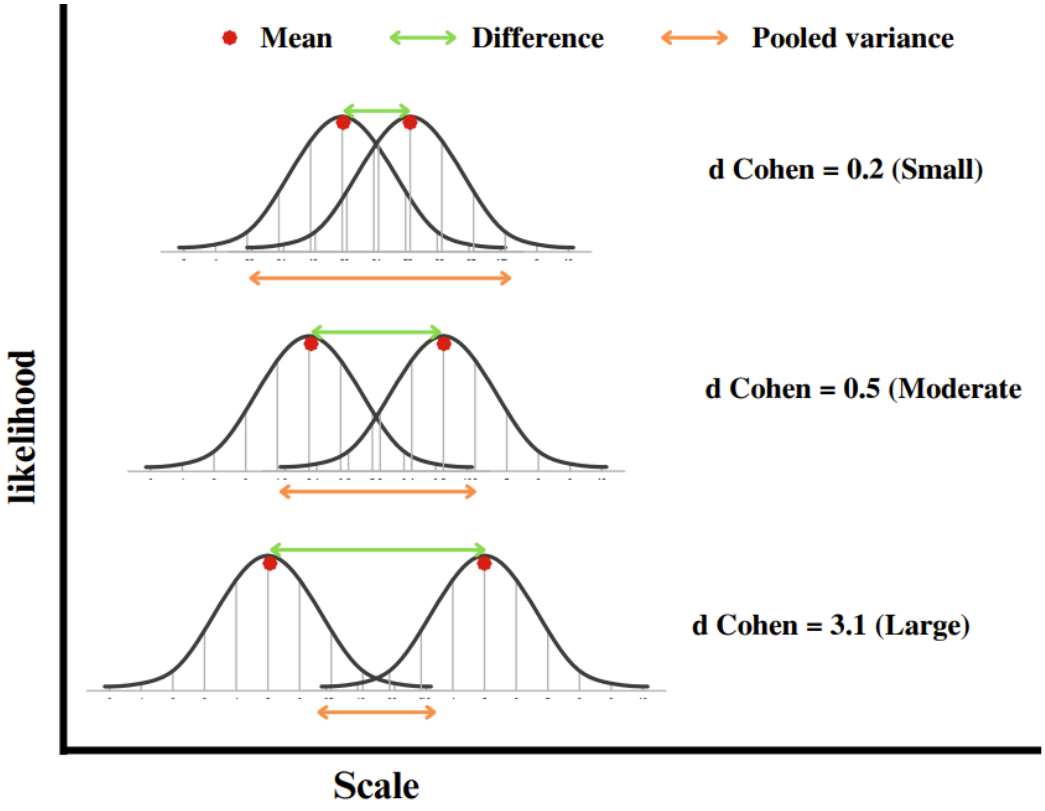


Figure 33 Cohen's Difference (d Cohen), which calculates the ratio of the difference between the mean of two groups with normal distribution (green and red) to the joint variance. Low values indicate a small effect size, high values indicate a large effect size.

From the differences between the group means and the weighted standard deviation. Cohen's d values less than 0.20 indicate the absence of an effect; values between 0.21 and 0.49 indicate a small effect; similarly, values oscillating between 0.50 and 0.70 indicate a moderate effect; finally, values greater than 0.80 indicate a large effect.

In our analysis, we employ the pingouin.compute_effsize function to calculate effect sizes. This function provides us with effect size estimates, which help us quantify the extent of differences observed within our study. It's important to note that pingouin.compute_effsize does not provide p-values, and in our specific context, the absence of p-values is not a limitation.

The reason why p-values are not relevant in this case is rooted in our focus on effect sizes for evaluating the practical significance of differences. While p-values are commonly used to determine statistical significance, they do not convey the magnitude or meaningfulness of differences. Effect sizes, on the other hand, offer a direct and interpretable measure of the strength of associations, which aligns with our objective of assessing the real-world impact of harmonization procedures.

When using the pingouin.compute_effsize library, the following Equation 18 is employed, if x and y are paired, the Cohen d_{avg} is computed:

$$d_{avg} = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{(\sigma_1^2 + \sigma_2^2)}{2}}}$$

Equation 18

Where $\bar{X}$ and $\bar{Y}$ represents the mean (average) of the values in dataset X and Y, σ_1 and σ_2 represents the variance of dataset X and Y.

The Cohen's **d** is a biased estimate of the population effect size, especially for small samples $(n < 20)$. It is often preferable to use the corrected Hedges g instead:

$$g = d \times \left(1 - \frac{3}{4(n_1 + n_2) - 9}\right)$$

The common language effect size is the proportion of pairs where x is higher than y (calculated with a brute-force approach where each observation of x is paired to each observation of y, see pingouin.wilcoxon() for more details):

$$CL = P(X > Y) + 0.5 \times P(X = Y)$$

Equation 19

P(X > Y): This represents the probability that the random variable X is greater than the random variable Y.

P(X = Y): This represents the probability that the random variable X is equal to the random variable Y.

Equation 19 consists of two terms: the first term calculates the probability of X being strictly greater than Y, and the second term considers the probability of X and

Y being equal. Since the formula is designed to evaluate the comparison between two variables, the sum of these two terms provides an estimate of the probability of X being greater than or equal to Y. The additional 0.5 in the second term is used to correct the calculation in cases where the two variables might be considered equally likely.

3.3.8.2.1 Effect size between controls across all cohorts

Table 8 presents a statistical analysis of effect size for some of the features used, providing an overview of other metrics related to data harmonization. All results are stored in Annex 5 for a more detailed exploration of this outcome.

In Table 8, it can be observed that all effect sizes among controls from different databases are large without neuroHarmonize. Notably, in Relative Power, Delta, Beta1, Beta2, Beta3, and Gamma bands stand out with effect sizes exceeding 0.8, while the lowest effect size belongs to the Theta band. Additionally, it is specified that the feature that differs the most among the 4 cohorts in controls is Gamma in the gICA component 14, and the feature that most closely resembles is Beta3 in the gICA component 23.

Table 8 Summary of the effect size by feature extraction between the control groups of the different cohorts.

Controls (UdeA1 vs UdeA2 vs SRM vs CHBMP)																
	Delta		Theta		Alpha1		Alpha2		Beta1		Beta2		Beta3		Gamma	
neuroHarmonize	Without	With	Without	With	Without	With	Without	With	Without	With	Without	With	Without	With	Without	With
Relative Power																
Max	0.96	0.10	0.32	0.07	0.45	0.10	0.55	0.04	0.82	0.11	0.84	0.08	0.88	0.10	1.04	0.36
Average	0.58	0.06	0.13	0.02	0.30	0.07	0.26	0.02	0.66	0.05	0.60	0.05	0.43	0.04	0.69	0.19
Min	0.23	0.02	0.02	0.00	0.05	0.04	0.00	0.00	0.38	0.01	0.15	0.01	0.00	0.00	0.43	0.06
Synchronization likelihood																
Max	1.63	0.10	1.61	0.09	3.38	0.11	1.33	0.04	2.48	0.05	2.88	0.04	0.03	0.02	0.98	0.08
Average	0.95	0.08	1.02	0.07	2.23	0.09	1.16	0.02	2.05	0.02	2.33	0.02	0.02	0.02	0.66	0.05
Min	0.49	0.04	0.64	0.05	1.47	0.05	0.98	0.00	1.77	0.00	2.00	0.01	0.02	0.01	0.31	0.00
Cross Frequency																
Max	0.91	0.16	0.86	0.04	0.91	0.07	0.88	0.07	0.73	0.07	0.97	0.07	1.05	0.11	1.23	0.17
Average	0.65	0.11	0.53	0.02	0.38	0.02	0.31	0.02	0.30	0.02	0.32	0.03	0.36	0.04	0.42	0.05
Min	0.00	0.08	0.10	0.00	0.01	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.02	0.00

Extreme effect sizes								
Relative Power								
	Without		With		Without		With	
Max	C14	Gamma	C25	Gamma	Min	C23	Beta3	C20 Theta
Synchronization likelihood								
	Without		With		Without		With	
Max	C14	Alpha1	C22	Alpha1	Min	C15	Gamma	C24 Gamma
Cross Frequency								
	Without		With		Without		With	
Max	C15	Gamma Mgamma	C14	Gamma Mgamma	Min	C14	Delta Mtheta	C15 Beta2 Mbeta1

After applying neuroHarmonize, the values decrease significantly, showing greater differences in the Gamma feature in gICA component 25 and greater similarity in the Theta feature in gICA component 20.

The other metrics exhibit the same behavior of reduction after neuroHarmonize, but it is specified that for SL, the features showing greater differences are Alpha1 in gICA component 14 without neuroHarmonize, and Alpha1 in gICA component 22 with neuroHarmonize. The features showing greater similarity are Gamma in gICA component 15 without neuroHarmonize, and Gamma in gICA component 24 with neuroHarmonize.

Finally, in Cross Frequency, the features showing greater differences are Gamma in gICA component 15 with the modulated Gamma band without neuroHarmonize, and Delta in gICA component 14 with the modulated Theta band with neuroHarmonize. The features showing greater similarity are Delta in gICA component 14 with the modulated Theta band without neuroHarmonize, and Beta2 in gICA component 15 with the modulated Beta1 band.

3.3.8.2.2 Effect size between paired group G1 with Controls

After understanding the behavior of the controls, the analysis is performed for the paired group results, focusing specifically on the group comparing carriers (G1) and controls.

The behavior of the effect size in Table 9 is similar to that observed in the controls Table 8. After applying neuroHarmonize, a reduction in effect size is observed for all traits. At this point, it is desirable that the effect size does not significantly decrease or increase.

Table 9 Summary of the effect size by feature extraction between the control groups of the different cohorts.

Paired group G1 vs Controls																
	Delta		Theta		Alpha1		Alpha2		Beta1		Beta2		Beta3		Gamma	
neuroHarmonize	Without	With	Without	With	Without	With	Without	With	Without	With	Without	With	Without	With	Without	With
Relative Power																
Max	1.10	0.42	0.88	0.48	0.54	0.35	0.25	0.09	1.12	0.43	1.14	0.43	1.34	0.55	0.83	0.65
Average	0.83	0.33	0.43	0.36	0.35	0.23	0.12	0.05	0.78	0.25	0.75	0.28	0.66	0.27	0.71	0.50
Min	0.56	0.22	0.17	0.24	0.01	0.09	0.05	0.00	0.28	0.00	1.14	0.43	0.10	0.01	0.49	0.39
Synchronization likelihood																
Max	0.75	0.23	0.82	0.16	1.05	0.15	0.96	0.13	0.89	0.12	0.90	0.09	0.76	0.10	0.77	0.23
Average	0.60	0.16	0.63	0.05	0.89	0.06	0.87	0.04	0.83	0.06	0.85	0.03	0.69	0.03	0.54	0.10
Min	0.50	0.07	0.45	0.01	0.67	0.00	0.77	0.01	0.77	0.01	0.80	0.09	0.56	0.01	0.33	0.01
Cross Frequency																
Max	1.21	0.44	0.76	0.14	0.72	0.20	0.73	0.20	1.10	0.35	1.34	0.38	1.35	0.44	1.26	0.44
Average	0.79	0.30	0.37	0.07	0.33	0.09	0.34	0.10	0.44	0.14	0.49	0.15	0.54	0.17	0.55	0.18
Min	0.00	0.18	0.09	0.01	0.02	0.01	0.00	0.00	0.01	0.01	0.03	0.02	0.01	0.01	0.00	0.00

Extreme effect sizes																
Relative Power																
	Without			**With**			**Without**			**With**						
Max	C14	Beta3		C25	Gamma		**Min**	C24	Alpha1		C23	Alpha2				
Synchronization likelihood																
	Without			**With**			**Without**			**With**						
Max	C14	Alpha1		C18	Delta		**Min**	C20	Gamma		C23	Alpha1				
Cross Frequency																
	Without			**With**			**Without**				**With**					
Max	C14	Beta3	Mbeta2	C18	Beta3	Mbeta3	**Min**	C14	Delta	Mtheta	C14	Gamma	Malpha1			

For Relative Power, the largest difference would be in the Beta 3 band for gICA component 14 without neuroHarmonize, and for Gamma in gICA component 25 with neuroHarmonize. For SL, larger differences are observed in the Alpha1 band for gICA component 14 without neuroHarmonize and for Delta in gICA component 18 with neuroHarmonize. Finally, in Cross Frequency, the largest difference would be in the Beta3 band for gICA component 14 with the modulated Beta2 band without neuroHarmonize, and for Beta3 in gICA component 18 with the modulated Beta3 band with neuroHarmonize.

Finally, Table 11 summarizes the percentage decrease in effect size for all Features, both in ROIs and components.

Table 10 Average percentages of reduction in effect size for each feature across all bands for both ROIs and neural gICA Components

G1 with Controls		
	ROIs	**Neural gICA Components**
Relative Power	29%	30%
Entropy	74%	84%
Coherence	18%	35%
Cross Frequency	21%	33%
Synchronization Likelihood	48%	67%

Table 11 shows that the most significant reduction occurred in Entropy, while the least reduction occurred in Relative Power. It also shows that the majority of the metrics experienced a decrease of more than 30%, with the exception of Coherence, Relative Power and Cross Frequency in ROIs with a decrease close to 20%.

The reduction in effect size was smaller for both neuronal gICA components and ROIs in terms of Relative Power compared to the other features, and larger for both gICA components and ROIs in terms of Entropy.

The tables containing the effect sizes and the summary statistics for all the features can be found in Annex 4. These tables provide a comprehensive overview of the impact of harmonization on the different variables analyzed in the study.

3.4 Discussion

The combination of processing techniques, harmonization, and subsequent statistical analysis has enabled focused research on risk factors for Alzheimer's disease [121]. During the development of the processing and harmonization pipeline, several steps were considered. This involved normalizing the data using Huber's normalization, implementing a matching process, and ultimately implementation the neuroHarmonize harmonization technique.

The outcomes of the descriptive analysis enable us to discern the disparities between controls with and without neuroHarmonize, as well as the alignment of the median concerning all cohorts. It also permits us to visualize the outcomes of the Relative Power feature in the gamma band, where negative values are obtained. While no negative values are observed in the other features, the alignment of the median remains consistent. These results are available for detailed examination in Annex 5.

Each step of the process yielded promising results, instilling confidence in the processing pipeline until the matching phase without neuroHarmonize. However, discussions arose regarding the interpretation of results and the most effective approach for analyzing the data, particularly after harmonization with

neuroHarmonize. The unexpected negative values observed in Relative Power lacked a consistent interpretation aligned with expectations.

It is evident that the balance of sample sizes plays a crucial role in achieving harmonization in data analysis [194]. Imbalanced sample sizes can significantly impact the harmonization process and potentially introduce biases or inaccuracies in the results. Therefore, it was essential to carefully consider and address any imbalances in sample sizes when implementing harmonization techniques using matching process.

Several potential factors during the harmonization process could lead to changes. These factors include modifications to the scale and distribution of energy in frequency bands, adjustments in reference values that impact magnitude and distribution of energy, and the estimation of parameters affecting energy distribution across frequency bands.

Additionally, with the Gamma band which was not subjected to the same adjustments as other bands and subsequently combined with them, may contribute to the occurrence of negative values. It is crucial to carefully consider these factors and their potential influence when interpreting negative values.

Nevertheless, the procedure was executed with a solid understanding of the underlying concepts, motivating the continuation of the proposed methodology to generate valuable information with machine learning.

Considering the situation, two paths need to be considered:

- Data without neuroHarmonize: The first path focuses on evaluating the pipeline from raw data input to paired data as a harmonization process capable of generating an accurate machine learning model.

- Data with neuroHarmonize: The second path focuses on evaluating the results using specialized libraries, in this case, neuroHarmonize, to achieve effective harmonization and produce an accurate machine learning model.

Furthermore, the reduction in effect size observed after harmonization implies the attenuation of systematic differences between the groups. By reducing the effect size, the potential impact of statistical differences related to the features that distinguish the two groups of interest, carriers of the PSEN1-E280A genetic variation and controls, is also reduced.

In the effect size results for controls, it was expected that the effect size would be small for all cohorts, demonstrating that controls are comparable and can be integrated as a single control group for subsequent comparison with the carrier group (G1). However, Table 8 shows that this is not the case without neuroHarmonize, as there are very large effect size values for most of the traits. Therefore, only a few features such as Relative Power in Beta3 for C23, SL in Gamma for C15, and Cross Frequency in Delta with Mtheta for C14 could be useful for integrating databases and subsequently classifying individuals at risk for

Alzheimer's disease. In this way, the differences captured by the classifier would not be associated with the cohorts, but only with the groups.

On the other hand, the path with neuroHarmonize Table 8 shows a reduction in effect size that is highly positive for all features as expected from the literature [83], [92], [191] (Therefore, the use of harmonized data that show greater similarity between controls from all cohorts may be the most favorable approach for evaluating the integration of multi-site databases. In theory, this suggests improved comparability and reduced variability between groups, as the differences initially observed were largely influenced by systematic factors that have now been addressed and mitigated by harmonization.

3.5 Conclusions

In conclusion, the five metrics investigated in this study, namely Shannon Entropy, Cross Frequency, Relative Power, Coherence, and Synchronization Likelihood in conjunction with crossover frequency, have proven their utility as valuable tools for analyzing EEG signals and extracting meaningful insights into underlying physiological processes.

Section 3.7 has emphasized the evaluation of two different approaches involving the use of specialized libraries (with neuroHarmonize) and the establishment of a well-controlled pipeline (without neuroHarmonize) to achieve harmonization and generate comparative results.

The neuroHarmonize approach has limitations in dealing with negative values that have no physiological basis. However, when analyzing healthy subjects (controls from different cohorts), it provides an alternative that successfully harmonizes multi-site databases while offering the prospect of consistent results when applying the machine learning model.

In the upcoming Chapter 4, we will present and discuss the machine learning model developed for the classification of Alzheimer's disease (AD), building on the two paths outlined in this Chapter 3. Our focus will be on the methodology used and the results obtained. The goal is to further evaluate and reflect on the model's performance and implications in order to gain a deeper understanding of its strengths, limitations, and potential impact on AD research and clinical practice.

Chapter 4

Machine Learning model

4.1 Introduction

The development of accurate and reliable machine learning models is crucial for identifying and classifying individuals at risk of Alzheimer's Disease (AD) based on non-invasive biomarkers. These models analyze large and complex datasets, such as EEG data, to uncover patterns and relationships that aid in accurate risk classification and enable early interventions and targeted treatments for better patient outcomes.

Support Vector Machines (SVM) is a versatile algorithm widely used in AD risk classification [195]. It identifies an optimal hyperplane to separate different classes in the feature space and has been successfully applied to various neuroimaging and biomarker data [196].

Random Forest, an ensemble learning method, combines multiple decision trees to improve model generalization and handle high-dimensional data. Its ability to capture complex interactions among features makes it popular in AD research [197], [198].

Neural networks, particularly deep learning models like Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), have shown promise in

AD risk classification. These models automatically learn hierarchical representations and intricate patterns from raw data, making them suitable for neuroimaging, genetic, and clinical data analysis [199], [200].

Gradient Boosting Machines (GBM) combine weak learners to create a strong predictive model, iteratively improving performance. GBM has been applied to AD risk classification with different data types and has demonstrated good predictive performance [201].

Logistic regression, a simple and interpretable linear model, estimates class probabilities based on input features and is widely used in AD research. It often incorporates feature selection techniques to identify relevant biomarkers for risk classification [202].

These machine learning models provide valuable tools for accurately classifying individuals at risk of Alzheimer's disease and hold great potential for improving diagnosis and treatment outcomes.

The choice of model depends on various factors such as the nature of the data, the availability of labeled samples, the desired interpretability, and the specific research objectives. Some studies have focused on the use of decision trees to classify subjects at risk of Alzheimer's disease, and all present a particular combination of data [203].

Marcos et al. [204] employed decision tree algorithms to classify individuals at risk of Alzheimer's disease based on genetic markers and cognitive assessments. The

researchers achieved a classification accuracy of 85% using a combination of genetic and cognitive data [205]. They found that specific genetic variants, such as the APOE ε4 allele, played a significant role in predicting disease risk [206]. The decision tree model demonstrated the potential of using genetic and cognitive information for early detection and risk assessment of Alzheimer's disease [207].

Ketan et al. [208] discovered that Deep Learning Ludwig Classifier produces 95% accuracy while the best outcomes of the Random Forest models produce about 87%. Among many patients, it had the highest accuracy in identifying dementia.

Shahin et al. [209] proposes an upgraded machine learning algorithm named Modified Random Forest (m-RF) to individualize between normal people and people with the risk of having Alzheimer’s disease using neuroimaging features. They have achieved an accuracy of 96.43% that is far better than other algorithms like Support Vector Machine, Adaptive Boosting, K-Nearest Neighbors, etc.

In recent studies exploring EEG for AD and aging, several methodologies and findings have emerged. García-Pretelt et al. [184] focused on developing an SVM model using gICA-derived spectral features and neuroimaging data for AD classification. Their study achieved an impressive 83% classification accuracy and identified specific genetic variants, such as the PSEN1-E280A gene mutation, as important predictors of disease risk. Miltiadous et al. [210] further investigated EEG in AD and proposed methodologies, obtaining accuracy scores of 78.5% for AD detection using decision trees and 86.3% for FTD (frontotemporal dementia)

detection using random forests. Javaid et al. [211]explored the resting state and observed a strong correlation between absolute power in delta and theta bands and aging. Additionally, a correlation was found between beta absolute power and aging during a Work Memory task. The use of the decision tree method during the Work Memory task successfully distinguished the elderly group from the middle-aged group with an impressive accuracy of 87.5%. These combined studies highlight the potential of EEG and its spectral features in aiding the classification and understanding of AD, FTD, and aging-related processes.

These references collectively showcase the effectiveness of decision tree algorithms in classifying individuals at risk of Alzheimer's disease. They highlight the importance of incorporating various data modalities, including genetic markers, cognitive assessments, and neuroimaging data, for accurate classification. The results demonstrate that decision trees can effectively capture patterns and variations in data, leading to promising classification accuracies. These findings contribute to the development of reliable and robust machine learning models for early detection and risk assessment of Alzheimer's disease.

One key aspect that emerges from these references is the significance of data harmonization in the context of machine learning for Alzheimer's disease classification. Harmonization involves combining and aligning data from multiple cohorts or datasets to increase the sample size, improve statistical power, and reduce the impact of dataset-specific biases. By harmonizing cohorts, limitations

associated with small sample sizes can be overcome achieving a more comprehensive understanding of the disease.

This project increased the amount of data through harmonization in order to search development of a more robust and generalizable machine learning model with a larger and more diverse dataset than previous projects in order to improve the accuracy of the model and contribute to the identification of consistent features or biomarkers across cohorts.

4.2 Methodology

The data in this chapter is managed using dataframes, structured as illustrated in the Figure 34. Each row corresponds to a record, while the columns represent specific feature. **For the ROIs, there are a total of 391 features (columns), while for the gICA, there are a total of 547 features (columns).**

participant_id	...	crossfreq_C9_Mbeta3_Gamma	crossfreq_C9_Mgamma_Gamma
sub-G1001	...	0.032289	0.196340
sub-G1017	...	0.097949	0.843018
sub-G1002	...	0.143042	0.941100
sub-G1000	...	0.047110	0.477489
sub-G1015	...	0.031574	0.103260
...	...	...	...
sub-CBM00156	...	0.013215	0.018351
sub-CBM00147	...	0.075288	0.072200
sub-CBM00202	...	0.016286	0.036942
sub-CBM00283	...	0.018948	0.039089
sub-CBM00284	...	0.035080	0.016786

Figure 34 Processed Dataframe Containing Model Input Information

The methodology for developing the machine learning model for Alzheimer's disease risk classification involved several steps. The first step, preprocessing, aimed to evaluate the available features and ensure data quality and consistency. After preprocessing and applying MatchIt, there were two data sets to evaluate. The first one (Table 11) consists of two groups: Carriers of the PSEN1-E280A gene (G1) and controls plus G2 (healthy), and the second (Table 12) includes two groups; carriers of the PSEN1-E280A gene (G1) with their respective control group (G2). It was noted that if any cohort was missing neuropsychological or demographic information, that specific information (neuropsychological or demographic columns) would be excluded from the model for all cohorts.

Table 11 Description of total subjects according to MacthIt for the first selected record.

Healthy: Control Group + G2 Group

	Age			Sex
	n	mean	std	F/M
Group				
Healthy	98	29.52	6.19	52/46
G1	49	30.18	5.50	29/20
Total	147			

Table 12 Description of total subjects according to MacthIt for the second selected record

	Age			Sex
	n	mean	std	F/M
Group				
G1	49	30.18	5.50	29/20
G2	49	30.63	5.52	29/20
Total	98			

Once the model and the best combination of the input parameters were selected, the feature dataset which included Relative Power, Shannon Entropy, Coherence, Cross Frequency, and Synchronization Likelihood, was partitioned into training and testing sets. This partitioning is crucial for evaluating the performance of the model. The training set is used to train the machine learning model on a subset of the data, allowing it to learn patterns and relationships between the input features and the target variable. The testing set is then used to assess the model's performance on unseen data, providing an estimate of how well the model can generalize to new data.

To assess the importance of each feature, a training graph is generated. The graph evaluates the precision of the model as each feature is added to the training one by one. The process starts with including one feature and progressively adds more features until all the features in the dataset are included. The precision of the model is calculated at each step, allowing the identification of the model with the highest precision.

By analyzing the training graph, is determined which feature set contributes the most to improving the precision of the model. These features are considered more informative and play a significant role in accurately classifying individuals at risk of Alzheimer's disease.

Once the decision tree model with the highest precision has been selected, it is used to generate a confusion matrix and evaluate the performance of the model in classifying individuals at risk of Alzheimer's disease.

As outlined in the preceding section 3.7, the methodology operates along two distinct paths.

Figure 35 represents the first path that focuses on evaluating the pipeline from raw data input to paired data as a harmonization process capable of generating an accurate machine learning model, that is, without neuroHarmonize.

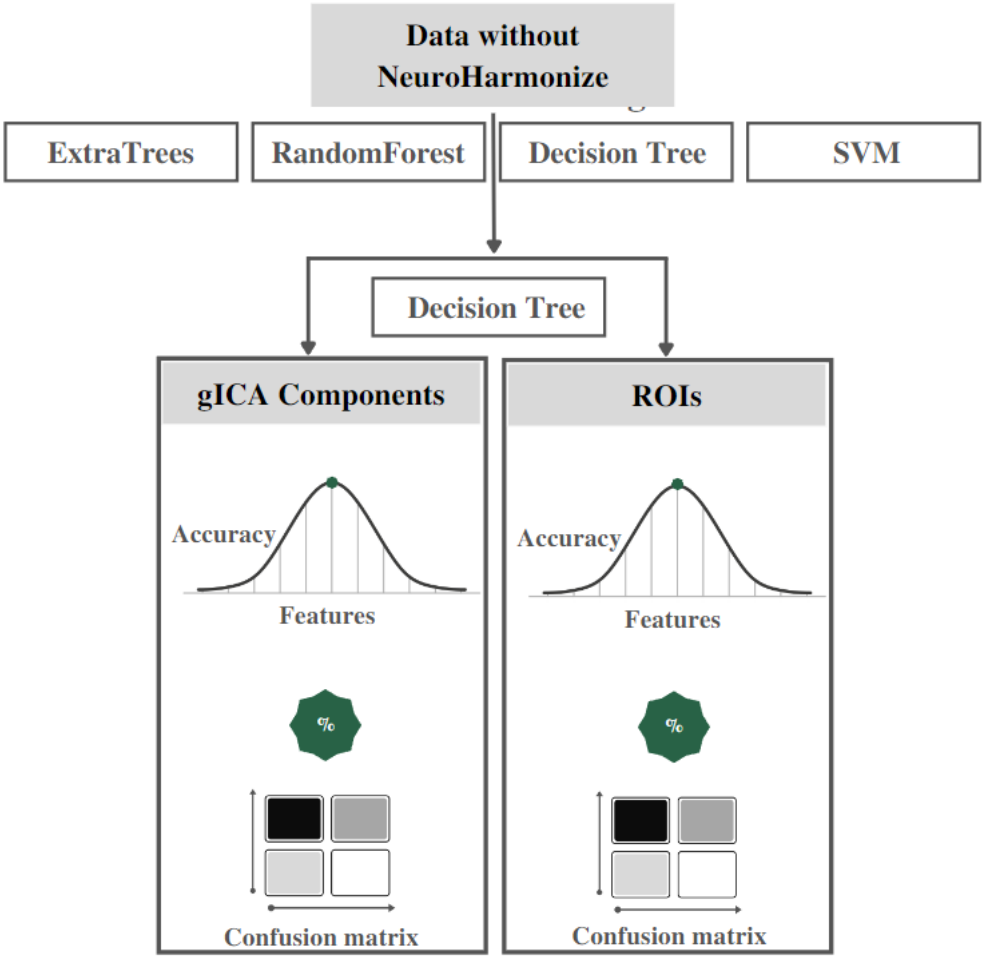


Figure 35 The first path focuses on evaluating the pipeline from raw data input to paired data.

Figure 36 shows a graphic representation of the methodology described above and represents the second path focuses on evaluating the results using specialized libraries, in this case, neuroHarmonize, to achieve effective harmonization and produce an accurate machine learning model.

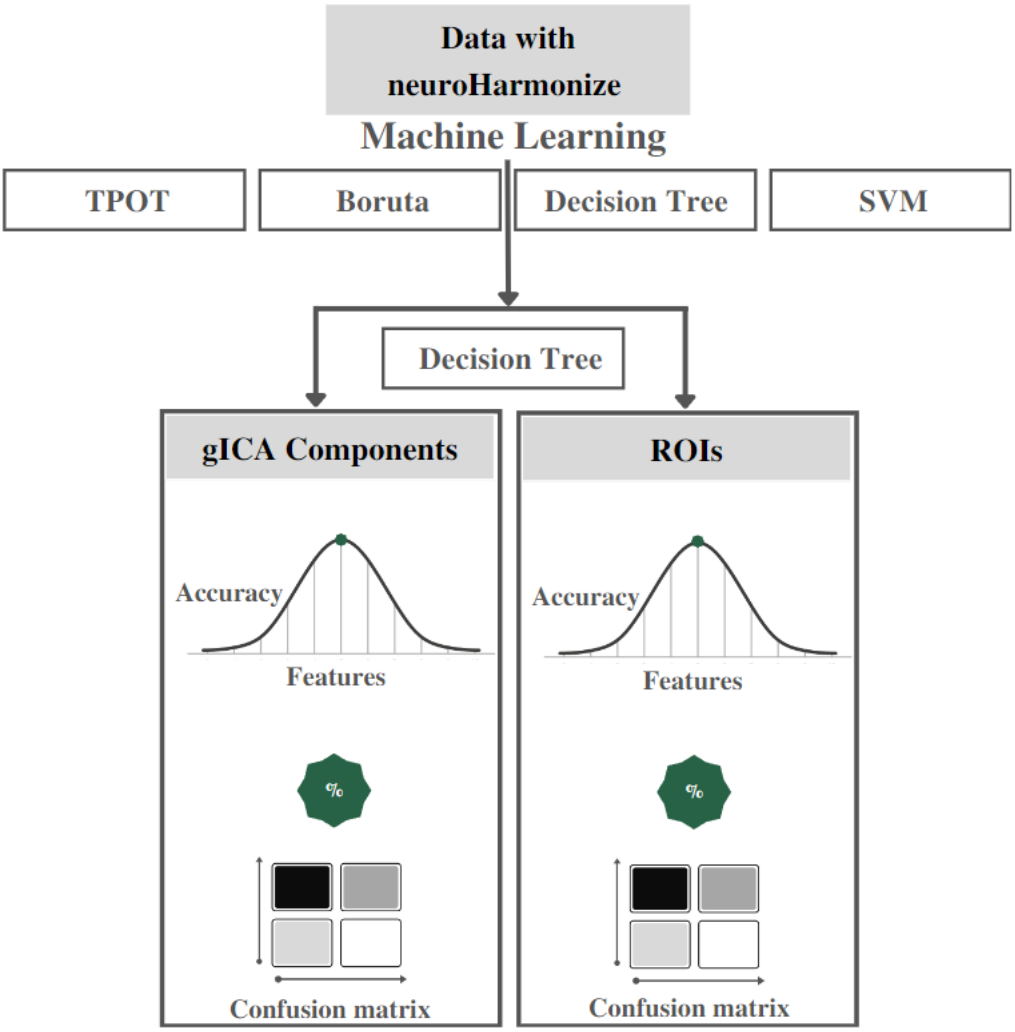


Figure 36 The second path focuses on evaluating the results using specialized libraries.

4.3 Model selection

In machine learning model selection, it is important to consider the potential issues of overfitting, sensitivity to small data variations, and the need for proper feature engineering and data preprocessing. Experimentation with different algorithms and

evaluation techniques is essential to determine the best approach for your specific problem and data.

TPOT is a powerful tool that automates the process of building machine learning pipelines, including preprocessing, feature selection, and model selection. It uses genetic programming to search for the best combination of these components and aims to optimize the performance of tree-based models.

Boruta is a feature selection algorithm that helps identify the most relevant features in a dataset for machine learning models. It is particularly useful when dealing with high-dimensional data or when there are many potential predictors. Both algorithms were tested and TPOT identified decision trees as the most suitable algorithm for the binary classification task.

Boruta and decision trees offer valuable approaches for feature selection, albeit with different features. While Boruta identifies the best features and discards the rest, decision trees provide feature importance measures without discarding any features. In this case, a decision tree algorithm was chosen directly to retain all features and facilitate the feature-to-feature curve analysis mentioned in the methodology. This approach allows for a comprehensive examination of how different features contribute to the classification process, providing insights into their individual and collective impacts on the model's performance. By leveraging the decision tree's feature importance measures, deeper understanding of the relative importance of

features can be gained and informed decisions regarding their inclusion in the final model can be made.

Lastly, considering the prevalence of SVM in classifying individuals at risk of Alzheimer's disease and the study conducted by Garcia et al. that achieved an 83% precision rate using SVM on a subset of the UdeA 1 database, this algorithm was selected for model comparison purposes. As mentioned earlier in this Chapter 4, SVM has been widely employed in Alzheimer's risk classification, making it a relevant choice to evaluate and compare against other models.

4.4 Implementation and validation of the model

4.4.1 Exploring and Loading the Data: Understanding the Dataset

Descriptive tables were generated to provide a comprehensive overview of the data. Table 13 specifically presents a description of the Relative Power in neural gICA Component 14 for the Delta band. Similar tables were created for the other frequency bands, components, and regions of interest (ROIs). The aim of these tables is to analyze and compare the Relative Power across different bands, components, and ROIs, allowing for a comprehensive understanding of the data. Annex 5 can access the complete set of tables and delve deeper into the details and insights presented in the study.

Table 13 Description of the Relative Power in neural gICA Component 14 for the Delta band

		Control	G1
power_C14_Delta	**count**	98	49
	mean	0.19	0.17
	std	0.09	0.08
	min	0.05	0.03
	25%	0.13	0.11
	50%	0.17	0.16
	75%	0.25	0.23
	max	0.51	0.32

4.4.2 Handling Incomplete Data: Removal of Inconsistent Columns and Implications for the Model

As discussed in Chapter 2, it was observed that not all cohorts had consistent availability of neuropsychological test data information for all subjects. Consequently, it became necessary to exclude these columns from the dataset used in the model. In

Table **14**, the removed columns are listed along with the corresponding amount of data present in each column for the groups Controls and G1 in the harmonized data and matching data.

Table 14 The list of removed columns for the groups Controls and G1 in the harmonized data and matching data.

Columns	Total data removed
MM_total	41
FAS_F	78

(Words that begin with the letter "F")	
FAS_S **(Words that begin with the letter "S")**	78
FAS_A **(Category of "animals")**	78

MM_total: Mini-Mental State Examination - **FAS**: verbal fluency test

By eliminating these columns, the data set was simplified to ensure consistency and reliability in the model. Removing incomplete or inconsistent data is a common practice to maintain data integrity and avoid potential bias or inaccuracy in the analysis. The decision to exclude specific columns was made to ensure the reliability and validity of the model's predictions.

4.4.3 Creating Training and Test Datasets: Data Split for Model Training and Evaluation

The data splitting was performed using a commonly used technique called "train-test split." In this technique, the dataset was divided into two portions: a training set and a test set. The division was done in a way that maintained the integrity of the dataset and preserved the relative proportions of different classes within the target variable.

To achieve this, the dataset was randomly divided, with 80% of the samples allocated to the training set and the remaining 20% allocated to the test set. This

proportion was chosen to strike a balance between having enough data for training the model and having a separate set of unseen data for evaluating its performance.

Furthermore, to ensure reproducibility of the results, a fixed random seed was set. This allowed for consistency in the data splitting process across multiple runs of the methodology.

4.4.4 Explanation of Model Cross-Validation

The cross-validation process uses the cross_val_score function from the `sklearn.model_selection` library. It first uses the pre-fitted estimator, `GS_fitted`, to estimate predictions with `random_grid`. During this process, the feature matrix `X_train` is used for both model training and evaluation, while the corresponding target vector `y_train` contains the labels of the samples in `X_train`. The cross-validation procedure divides the data into ten parts, resulting in ten iterations of training and evaluation. Each iteration involves different combinations of training and testing data.

In addition, this process makes optimal use of all available processor cores for parallel computation. The function ultimately returns a series of model performance scores that reflect the outcome of each fold in the cross-validation. These scores include various measures such as accuracy, F1 score, and more, depending on the model configuration and the nature of the problem. By averaging the scores across the ten folds, a single performance metric is derived.

4.5 Parameter selection

In this section, the parameters employed in each model for every pathway outlined in the methodology are outlined.

Detailed information on all implemented models and codes can be found in Annex 6.

4.5.1 The first path (without neuroHarmonize)

4.5.1.1 RandomizedSearchCV

The parameter configuration used for RandomizedSearchCV is as follows:

- cv: Specifies the number of cross-validation folds for evaluation.
- estimator: Sets the base estimator as a RandomForestClassifier.
- n_iter: Determines the number of parameter settings that are sampled.
- n_jobs: Utilizes all available processors for parallel computation.
- param_distributions: Specifies the range of hyperparameters to be searched, including 'bootstrap', 'criterion', 'max_depth', 'max_features', 'min_samples_leaf', 'min_samples_split', and 'n_estimators'.
- random_state: Sets the random seed for reproducibility.
- verbose: Controls the verbosity level of the output.

This configuration suggests that the best performing Random Forest classifier was found with the specified hyperparameters.

Table 15 Configuration resulting from the RandomizedSearchCV without neuroHarmonize.

Hyperparameter Optimization using RandomizedSearchCV	
Parameter	**Value**
Number of Iterations	100
Cross-Validation Folds	10
Base Estimator	RandomForestClassifier()
Number of Parallel Processes	-1
Hyperparameter Combinations Explored	
Splitting Criterion	'gini', 'entropy', 'log_loss'
Maximum Tree Depth	10, 20, 30, ..., 110, None
Maximum Features	'auto', 'sqrt'
Minimum Samples per Leaf	1, 2, 4
Minimum Samples to Split	2, 5, 10
Number of Estimators in Forest	100, 165, 231, ..., 1934, 2000
Random Seed	10
Output Verbosity	verbose=2
Optimal Model Configuration with RandomForestClassifier	
Bootstrap	False

Splitting Criterion	'log_loss'
Maximum Tree Depth	60
Minimum Samples per Leaf	2
Minimum Samples to Split	5
Number of Estimators in Forest	1541

4.5.1.2 Boruta

The specific hyperparameters used for the RandomForestClassifier estimator are as follows:

- Criterion: The criterion used to measure the quality of split points in the decision trees is based on logarithmic loss.
- Max Depth: The maximum depth of the decision trees is set to 90, which controls the complexity and depth of the trees.
- Min Samples Leaf: The minimum number of samples required to be at a leaf node is set to 2, ensuring that each leaf contains a minimum number of samples.
- Min Samples Split: The minimum number of samples required to split an internal node is set to 5, determining when a node is considered for a split.

- N Estimators: The number of trees in the random forest is set to 1000, indicating the number of decision trees that are generated and combined to make predictions.
- Random State: The random seed is set to ensure reproducibility of the results.

Table 16 Configuration resulting from the Boruta without neuroHarmonize.

Technique	**Parameters**
BorutaPy	Estimator: RandomForestClassifier
	Criterion: 'log_loss'
	Max Depth: 90
	Min Samples Leaf: 2
	Min Samples Split: 5
	Number of Estimators: 1000
	Random State: MT19937 at 0x236FCFAC340
	Verbose: 2
RandomForestClassifier	Criterion: 'log_loss'
	Max Depth: 90
	Min Samples Leaf: 2

	Min Samples Split: 5
	Number of Estimators: 1000
	Random State: MT19937 at 0x236FCFAC340

4.5.1.3 Decision tree

The optimized hyperparameters obtained through RandomizedSearchCV were employed for configuring the decision tree parameters.

4.5.1.4 Support Vector Machine (SVM)

The algorithm used in this case was a Support Vector Machine (SVM) model with the specified parameters, namely C = 0.1 and gamma = 0.001. The goal of the SVC algorithm is to find an optimal decision boundary that achieves a balance between the margin width (the separation between classes) and the accuracy of classification.

4.5.1.5 TPOT

Finally, TPOT, an automated machine learning tool, with the following parameter configuration shown Table 17. The "cv" parameter denotes the number of cross-validation folds, "generations" specifies the number of iterations for the genetic programming search, "n_jobs" determines the number of parallel jobs to run, "population size" sets the number of individuals in each generation, "random_state" ensures reproducibility, and "verbosity" controls the level of detail in the output.

Table 17 TPOT parameter configuration without neuroHarmonize.

TPOT Classifier
TPOTClassifier (cv=10, generations=5, n_jobs=-1, population_size=58, random_state=10, verbosity=3)

Table 18 Parameters for the 5 generations with using TPOT without neuroHarmonize.

Technique	Parameters
ExtraTreesClassifier	Bootstrap: True
	Criterion: 'entropy'
	Max Features: 0.75
	Min Samples Leaf: 14
	Min Samples Split: 3
	Number of Estimators: 100

4.5.2 The second path (with neuroHarmonize)

4.5.2.1 RandomizedSearchCV

The parameter configuration used for RandomizedSearchCV is as follows:

- cv: Specifies the number of cross-validation folds for evaluation.
- estimator: Sets the base estimator as a RandomForestClassifier.
- n_iter: Determines the number of parameter settings that are sampled.

- n_jobs: Utilizes all available processors for parallel computation.
- param_distributions: Specifies the range of hyperparameters to be searched, including 'bootstrap', 'criterion', 'max_depth', 'max_features', 'min_samples_leaf', 'min_samples_split', and 'n_estimators'.
- random_state: Sets the random seed for reproducibility.
- verbose: Controls the verbosity level of the output.

The final RandomForestClassifier configuration resulting from the RandomizedSearchCV is in Table 19.

Table 19 Configuration resulting from the RandomizedSearchCV with neuroHarmonize.

Hyperparameter Optimization using RandomizedSearchCV	
Parameter	**Value**
Number of Iterations	100
Cross-Validation Folds	10
Base Estimator	RandomForestClassifier()
Number of Parallel Processes	-1
Hyperparameter Combinations Explored	
Splitting Criterion	'gini', 'entropy', 'log_loss'
Maximum Tree Depth	10, 20, 30, ..., 110, None
Maximum Features	'auto', 'sqrt'

Minimum Samples per Leaf	1, 2, 4
Minimum Samples to Split	2, 5, 10
Number of Estimators in Forest	100, 165, 231, ..., 1934, 2000
Random Seed	10
Output Verbosity	verbose=2
Optimal Model Configuration with RandomForestClassifier	
Bootstrap	False
Criterion	Entropy
Maximum Tree Depth	30
Minimum Samples per Leaf	10
Number of Estimators in Forest	165

4.5.2.2 Boruta

The specific hyperparameters used for the RandomForestClassifier estimator are as follows:

- Criterion: The criterion used to measure the quality of split points in the decision trees is based on logarithmic loss.
- Max Depth: The maximum depth of the decision trees is set to 90, which controls the complexity and depth of the trees.

- Min Samples Leaf: The minimum number of samples required to be at a leaf node is set to 2, ensuring that each leaf contains a minimum number of samples.
- Min Samples Split: The minimum number of samples required to split an internal node is set to 5, determining when a node is considered for a split.
- N Estimators: The number of trees in the random forest is set to 1000, indicating the number of decision trees that are generated and combined to make predictions.
- Random State: The random seed is set to ensure reproducibility of the results.

Table 20 Configuration resulting from the Boruta with neuroHarmonize.

Technique	**Parameters**
BorutaPy	Estimator: RandomForestClassifier
	Criterion: 'log_loss'
	Max Depth: 20
	Min Samples Leaf: 2
	Min Samples Split: 5
	Number of Estimators: 1000
	Random State: MT19937 at 0x2354076B140

	Verbose: 2
RandomForestClassifier	Criterion: 'log_loss'
	Max Depth: 20
	Min Samples Leaf: 2
	Min Samples Split: 5
	Number of Estimators: 1000
	Random State: MT19937 at 0x2354076B140

4.5.2.3 Decision tree

The optimized hyperparameters obtained through RandomizedSearchCV were employed for configuring the decision tree parameters.

4.5.2.4 Support Vector Machine (SVM)

The algorithm used in this case was a Support Vector Machine (SVM) model with the specified parameters, namely C = 0.1 and gamma = 0.001. The goal of the SVC algorithm is to find an optimal decision boundary that achieves a balance between the margin width (the separation between classes) and the accuracy of classification.

4.5.2.5 TPOT

Finally, TPOT, an automated machine learning tool, with the following parameter configuration is shown Table 21. The "cv" parameter denotes the number of cross-validation folds, "generations" specifies the number of iterations for the genetic

programming search, "n_jobs" determines the number of parallel jobs to run, "population_size" sets the number of individuals in each generation, "random_state" ensures reproducibility, and "verbosity" controls the level of detail in the output.

Table 21 TPOT parameter configuration with neuroHarmonize.

TPOT Classifier
TPOTClassifier (cv=10, generations=5, n_jobs=-1, population_size=58, random_state=10, verbosity=3)

Table 22 Parameters for the 5 generations with using TPOT with neuroHarmonize.

Technique	**Parameters**
ExtraTreesClassifier	Bootstrap: True
	Criterion: 'entropy'
	Max Features: 0.75
	Min Samples Leaf: 14
	Min Samples Split: 3
	Number of Estimators: 100

Table 23 The five generations for TPOT with neuroHarmonize.

Parameters

DecisionTreeClassifier (input_matrix, criterion=gini, max_depth=9, min_samples_leaf=10, min_samples_split=14)
XGBClassifier(SGDClassifier(input_matrix, alpha=0.0, eta0=0.01, fit_intercept=False, ratio=0.0, learning_rate=invscaling, loss=log, penalty=elasticnet, power_t=0.0), GBClassifier(learning_rate=0.01, max_depth=2, min_child_weight=3, n_estimators=100, n_jobs=1, subsample=0.75, verbosity=0)
DecisionTreeClassifier (input_matrix,criterion=gini, max_depth=9, min_samples_leaf=10, min_samples_split=14)
XGBClassifier(SGDClassifier(input_matrix, alpha=0.0, eta0=0.01, fit_intercept=False, l1_ratio=0.0, learning_rate=invscaling, loss=log, penalty=elasticnet, power_t=0.0), XGBClassifier(learning_rate=0.01, max_depth=2, min_child_weight=3, n_estimators=100, n_jobs=1, subsample=0.75, verbosity=0)
RandomForestClassifier(MLPClassifier (DecisionTreeClassifier(input_matrix, criterion=gini, max_depth=7, min_samples_leaf=13,min_samples_split=4), MLPClassifier(alpha=0.1,learning_rate_init=0.1), RandomForestClassifier(bootstrap=True,criterion=entropy, max_features=0.05, min_samples_leaf=6, min_samples_split=16, n_estimators=100)
RandomForestClassifier(MLPClassifier (DecisionTreeClassifier(input_matrix, criterion=gini, max_depth=7, min_samples_leaf=13, min_samples_split=4), MLPClassifier(alpha=0.1, learning_rate_init=0.1), bootstrap=True, criterion=entropy, max_features=0.05, min_samples_leaf=6, min_samples_split=16, n_estimators=100)
DecisionTreeClassifier (input_matrix, criterion=gini, max_depth=9, min_samples_leaf=10, min_samples_split=14)
GradientBoostingClassifier(DecisionTreeClassifier(input_matrix, criterion=gini,max_depth=8,min_samples_leaf=15, min_samples_split=7),learning_rate=1.0,max_depth=9, max_features=0.95,min_samples_leaf=13, min_samples_split=5,n_estimators=100, subsample=0.85)
Generation 4
XGBClassifier(input_matrix,XGBClassifier__learning_rate=0.01, max_depth=2,min_child_weight=3, n_estimators=100, n_jobs=1, subsample=0.75, verbosity=0)

GradientBoostingClassifier(DecisionTreeClassifier(input_matrix, criterion=gini,max_depth=8,min_samples_leaf=15, min_samples_split=7), GradientBoostingClassifier(learning_rate=1.0, max_depth=9,max_features=0.95, min_samples_leaf=13,min_samples_split=5,n_estimators=100, subsample=0.85)
GradientBoostingClassifier (input_matrix, learning_rate=1.0, max_depth=2,max_features=0.65,min_samples_leaf=14,min_samples_split=5, n_estimators=100, subsample=0.85)
GradientBoostingClassifier(DecisionTreeClassifier(input_matrix, criterion=gini,max_depth=8,min_samples_leaf=15, min_samples_split=7),learning_rate=1.0,max_depth=9,max_features=0.95,min_samples_leaf=13,min_samples_split=5,n_estimators=100, subsample=0.85)

4.6 Results

Initially, a series of summary tables is provided to elucidate the elements included in the models, all adhering to the same methodology.

The number of features differs between gICA and the Regions of Interest (ROIs), considering that there are 8 gICA components and 5 ROIs. Within each component, there are 8 bands, and for each band, 5 features are evaluated Table 24 illustrates the characteristic count for each scenario.

Table 24 Specification of the number of features included in the models.

Feature Summary	
Number of features incorporated into the group independent component analysis (gICA) model	547
Number of features incorporated into the regions of interest (ROIs) model	386

4.6.1 The first path (without neuroHarmonize)

Table 25 presents a comprehensive summary of the results obtained for each model implemented using the previously mentioned methodologies. These models were specifically applied in the first path focuses on evaluating the pipeline from raw data input to paired data as a harmonization process capable of generating an accurate machine learning model.

Table 25 Comprehensive summary of the results obtained for each model without neuroHarmonize.

Healthy: Control Group + G2 Group

Groups/Models		RF-B		DT		SVM		ET-T	
		Train	Test	Train	Test	Train	Test	Train	Test
gICA	G1 vs Healthy	100%	82%	100%	80%	85%	80%	86%	86%
	G1 vs G2	87%	68%	73%	73%	50%	45%	70%	40%
ROIs	G1 vs Healthy	86%	81%	97%	87%	86%	79%	81%	79%
	G1 vs G2	83%	64%	73%	70%	62%	48%	77%	50%

RF-B: RandomForest using Boruta, **DT:** Decision Trees, **SVM:** Support vector machine, **ET-T:** ExtraTrees found with TPOP.

Below is a comprehensive description of the results obtained for each implemented model, using the dataset consisting of the G1 and control groups of neural gICA Components as an example. This particular group selection aligns with the project's objectives. However, it is important to note that all groups included in the project

underwent the same methodology for both neural gICA Components and regions of interest (ROIs). Detailed information on all implemented models can be found in Annex 6.

4.6.1.1 RandomizedSearchCV

The results obtained from applying RandomizedSearchCV to a RandomForestClassifier represent the initial step in the methodology, which involves optimizing decision trees through grid search.

The curve in Figure 37 demonstrates a 100% training accuracy, and a validation accuracy starts nearly at 70% and steadily increases until reaching an accuracy of

early at 80%. Additionally, the variability of the validation accuracy with respect to the number of samples is also evident from the curve.

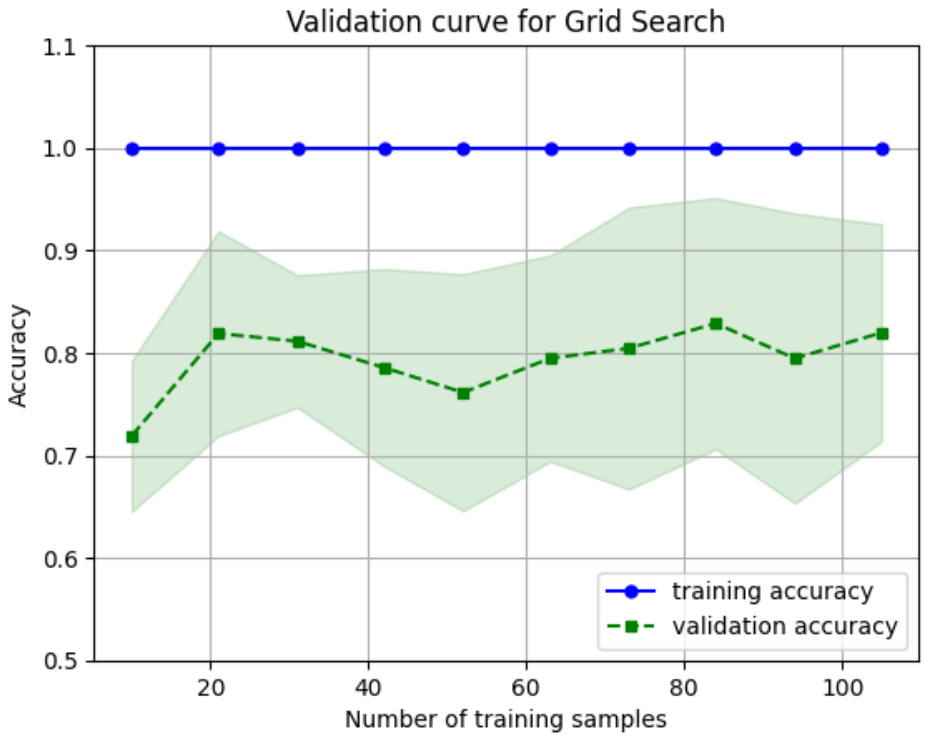


Figure 37 Validation curve for Grid Search without neuroHarmonize.

4.6.1.2 Boruta

The selection of these 19 significant features by the Boruta algorithm suggests that they have a strong influence on the predictive performance of the model. By focusing on these specific features, the model can effectively capture the relevant patterns and relationships in the data, leading to improved accuracy and performance.

To better highlight the relevant features selected by the Boruta tool, Figure 38 is presented. In this figure, the selected features are differentiated by the evaluated metric, components, bands, and modulated bands. The y-axis represents the frequency of occurrence of the discriminated element among the 19 selected features. The elements with higher significance according to the algorithm are shown in a darker color. For example, the Cross Frequency metric appears more frequently among the relevant features, but the Relative Power metric takes precedence by appearing with greater weight. Similar patterns apply to the other elements in Figure 38.

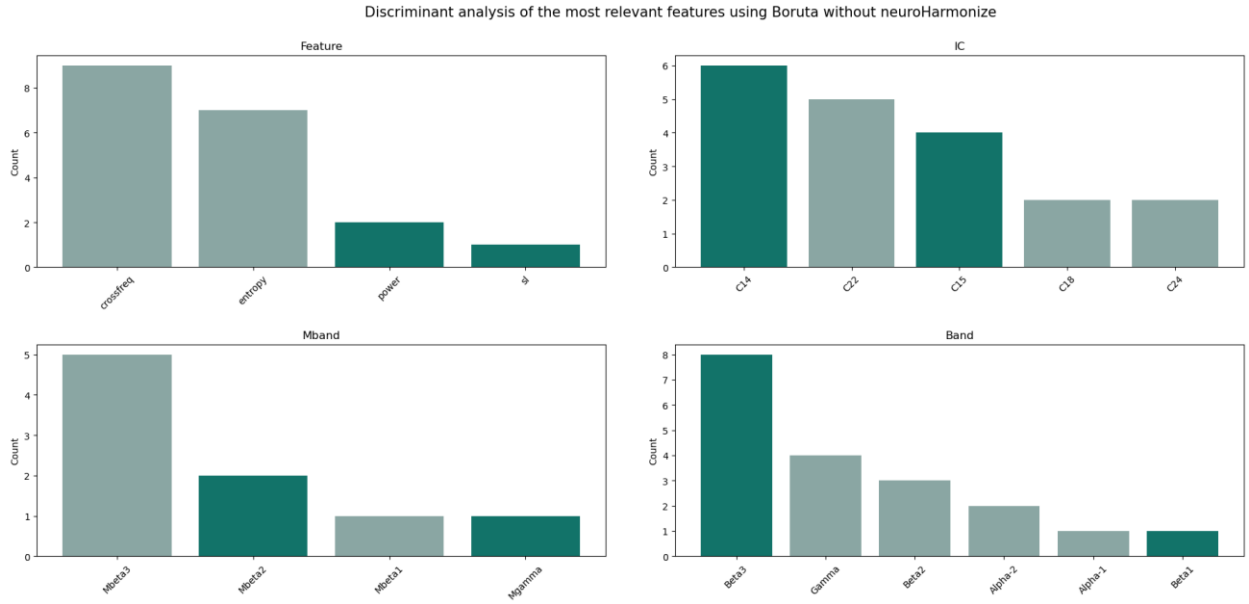


Figure 38 Discriminant analysis of the most relevant features using Boruta without neuroHarmonize.

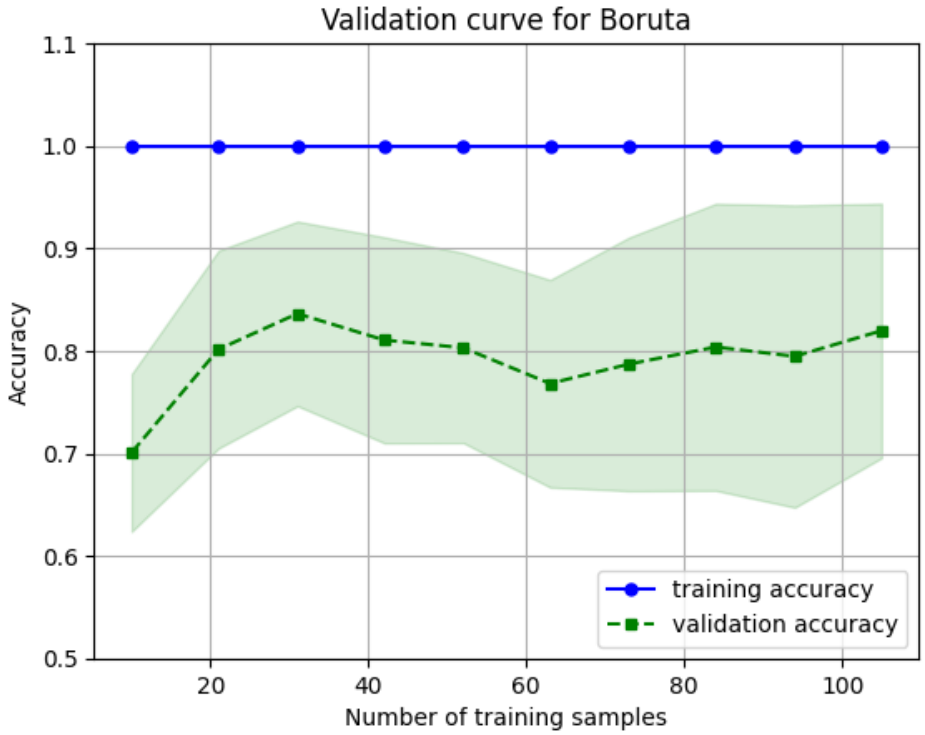


Figure 39, is the result of the input of the 19 features into the decision tree model, which was selected with the Boruta tool. The curve demonstrates a 100% training accuracy, while the validation accuracy starts nearly at 70% and it presents an increase in the accuracy between 20 and 40 samples, and steadily increases until reaching an accuracy of 82%. Additionally, the variability of the validation accuracy with respect to the number of samples is also evident from the curve.

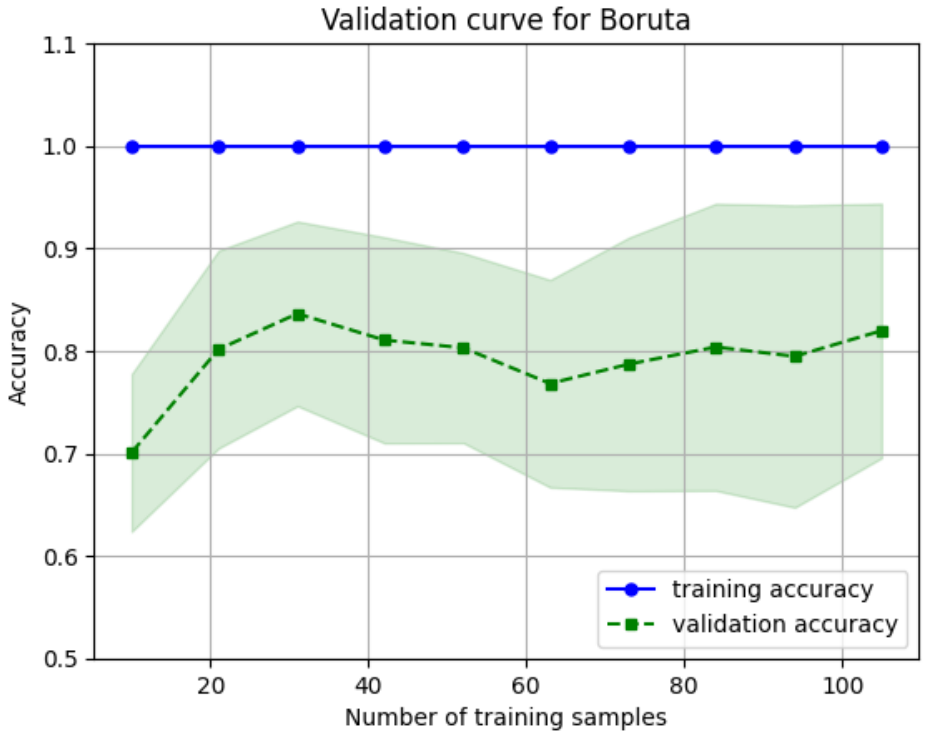


Figure 39 Validation curve for Boruta without neuroHarmonize.

4.6.1.3 Decision tree

Figure 40 visualizes the importance of different features in the classification model. The analysis reveals that Cross Frequency in the Beta3 band, particularly in component 22, holds significant importance for the classification task. Additionally, Cross Frequency in components 14 and 18, within the Gamma and Beta bands respectively, also demonstrate notable relevance. It is important to note that the graph displays only the top 10 features for better visual clarity. However, a comprehensive list of all features along with their respective relative importance can be found in the supplementary files. This comprehensive information provides

a deeper understanding of the crucial features that contribute to the accuracy and effectiveness of the classification model.

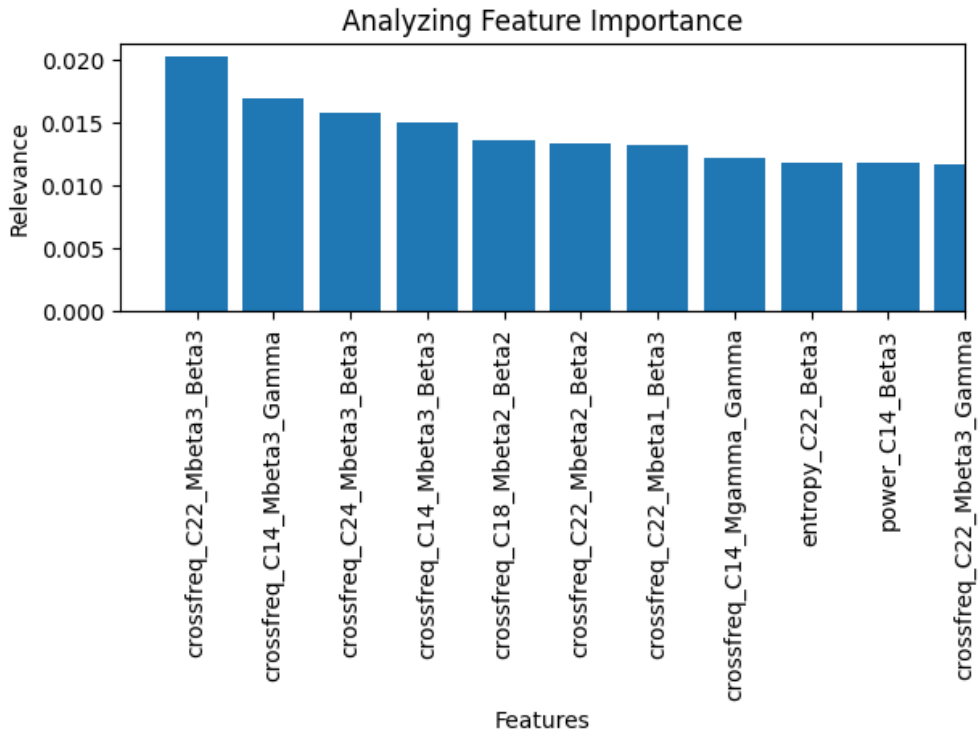


Figure 40 The importance of firsts features in the classification model without neuroHarmonize.

Figure 41 visually represents the correlation between the number of features and the accuracy of the model. The graph displays a significant increase in accuracy as the number of features increases, reaching a peak of approximately 85% within the first 40 features out of a total of 547 features. Beyond this point, accuracy slightly fluctuates but remains consistently high. This observation suggests that the initial set of features contains the most informative attributes for achieving high precision,

while additional features beyond a certain threshold do not significantly contribute to the model's performance.

Figure 41 Relationship between the number of features and the accuracy of the model without neuroHarmonize

Based on this initial finding, the evaluation focuses on the set of features that achieved the highest accuracy. In other words, the precision value is taken when it contains only the first feature, then the precision value when it includes both the first and second features, and so forth. Employing this method starting from the first graph, the outcome indicates that the optimal model is trained using only the initial 46 features.

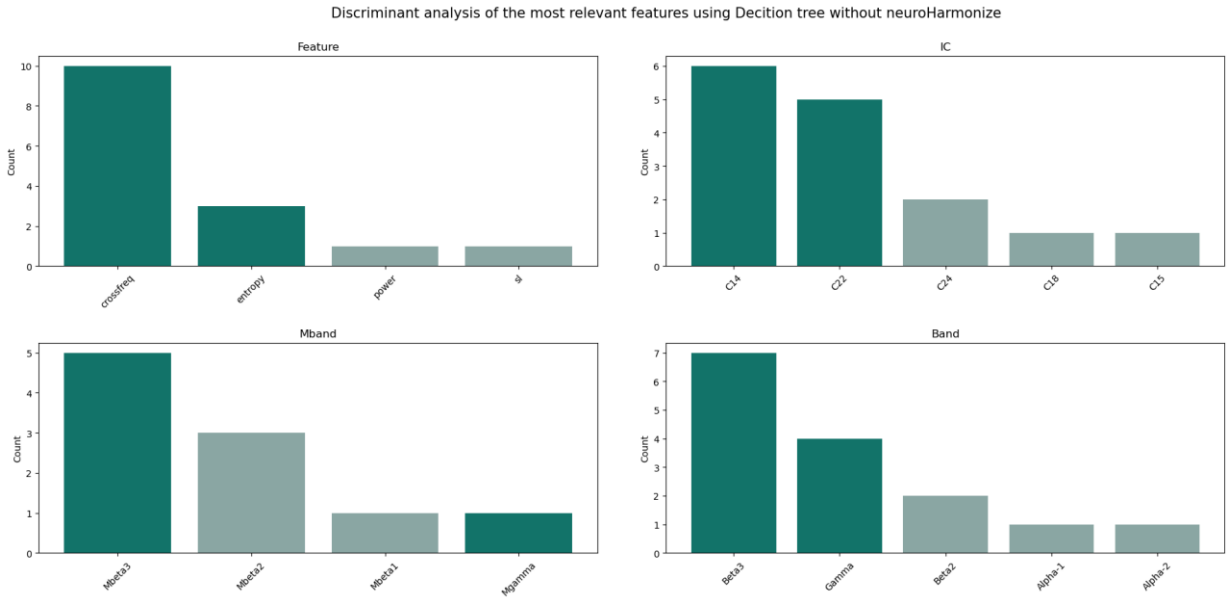


Figure 42 Discriminant analysis of the most relevant features with Decision tree without neuroHarmonize.

In Figure 42, the selected features are differentiated by the evaluated metric, components, bands, and modulated bands. The y-axis represents the frequency of occurrence of the discriminated element among the 46 selected features. The elements with higher significance according to the algorithm are shown in a darker color. For example, the Cross Frequency metric appears more frequently among the relevant features, but the Entropy metric takes precedence by appearing with greater weight. Similar patterns apply to the other elements in Figure 42.

As a result, a graph similar to Figure 41 is generated, concentrating solely on the relationship between the number of features and the corresponding increase in accuracy.

Figure 43 Relationship between the 46th best features and the accuracy of the model without neuroHarmonize.

Figure 44, the curve demonstrates a 100% training accuracy, while the validation accuracy starts nearly at 70% and steadily increases until reaching an accuracy of 80%. Additionally, the variability of the validation accuracy with respect to the number of samples is also evident from the curve.

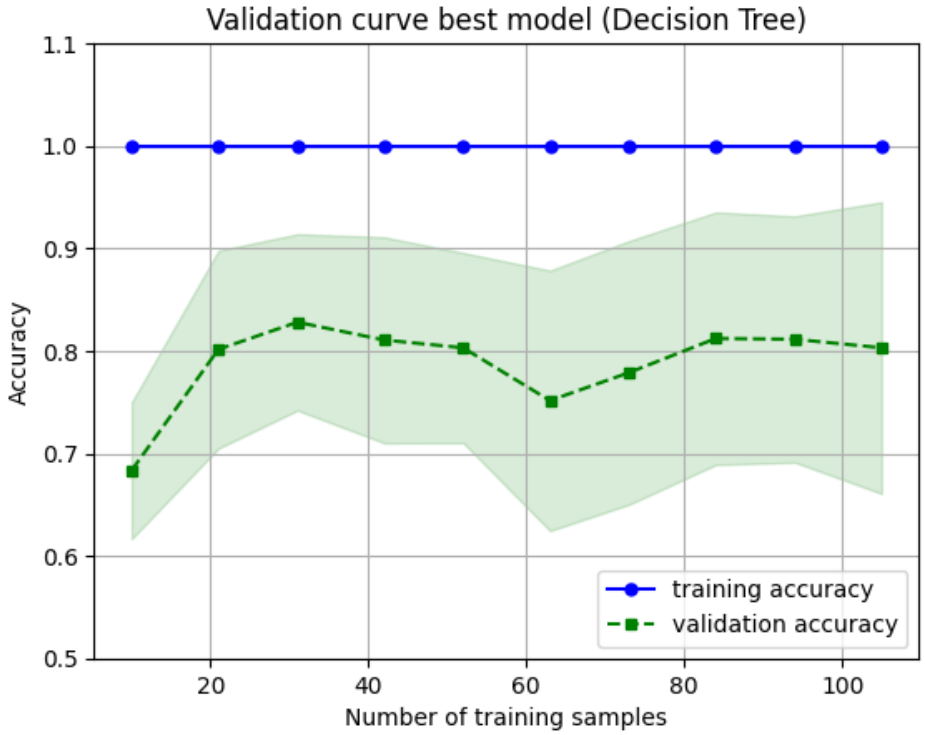


Figure 44 Validation curve for Decision Tree without neuroHarmonize.

Table 26 presents the results of the computational precision achieved by the algorithm. It demonstrates an accuracy of 97% accompanied by a standard deviation of 6%. Additionally, the precision, recall, and F1 score are reported to be 100%, indicating a high level of performance in classification.

Table 26 Results of the algorithm's computational precision for decision tree without neuroHarmonize.

```
------Computer Precision---
Precision:  0.8
Recall:  0.8
F1-score: 0.8
Accuracy: 0.86
```

```
Standard deviation: 0.11
```

Finally, a confusion matrix is generated to analyze the test set, consisting of 30 subjects. Among the G1 subjects, 8 are correctly classified, while none of them are misclassified as controls. Within the control group, 18 subjects are correctly classified, and none of them are incorrectly classified. A visual representation of the confusion matrix can be observed in Figure 45, providing a comprehensive overview of the model's performance in the test set.

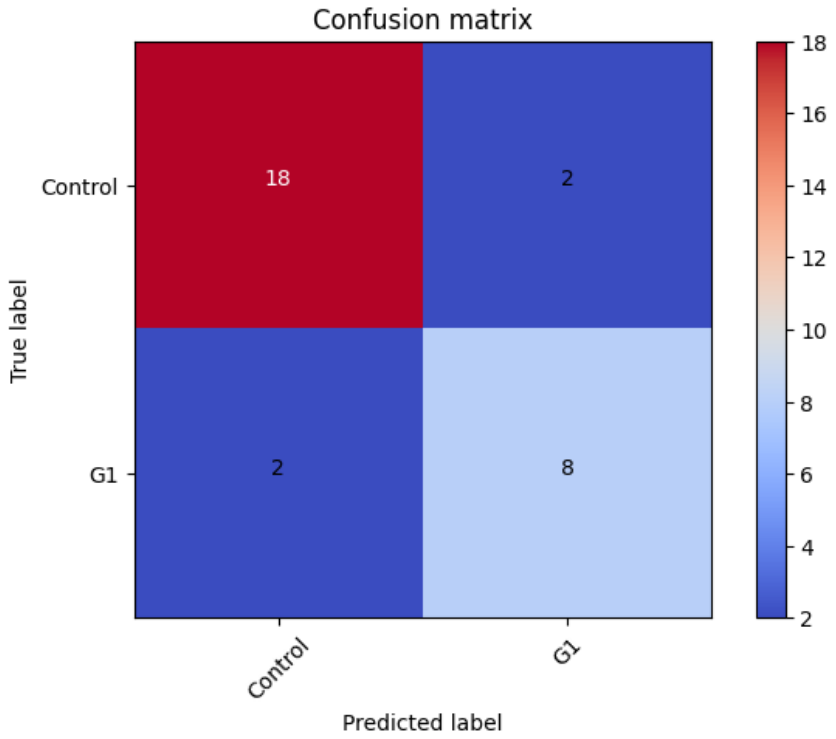


Figure 45 Confusion matrix for decision tree without neuroHarmonize.

4.6.1.4 Support Vector Machine (SVM)

Table 27 presents the results obtained from the application of the SVC algorithm. It provides information about computational precision, which refers to the accuracy

and performance of the model in classifying the data. The precision score indicates the proportion of correctly predicted instances among all instances, while other metrics such as recall and F1 score provide additional insights into the model's performance.

Table 27 Results of the algorithm's computational precision for SVM without neuroHarmonize.

```
------Computer Precision---
            ---
Precision:  0.85
Recall:  0.85
F1-score: 0.85
Accuracy: 0.80
```

4.6.1.5 TPOT

Using TPOT, Figure 46 shows the five generations of Table 18 and their respective accuracies.

The precision values represent the accuracy of the machine learning models generated by TPOT in each generation. It can be observed that the precision remains constant over all generations at a value of 86%.

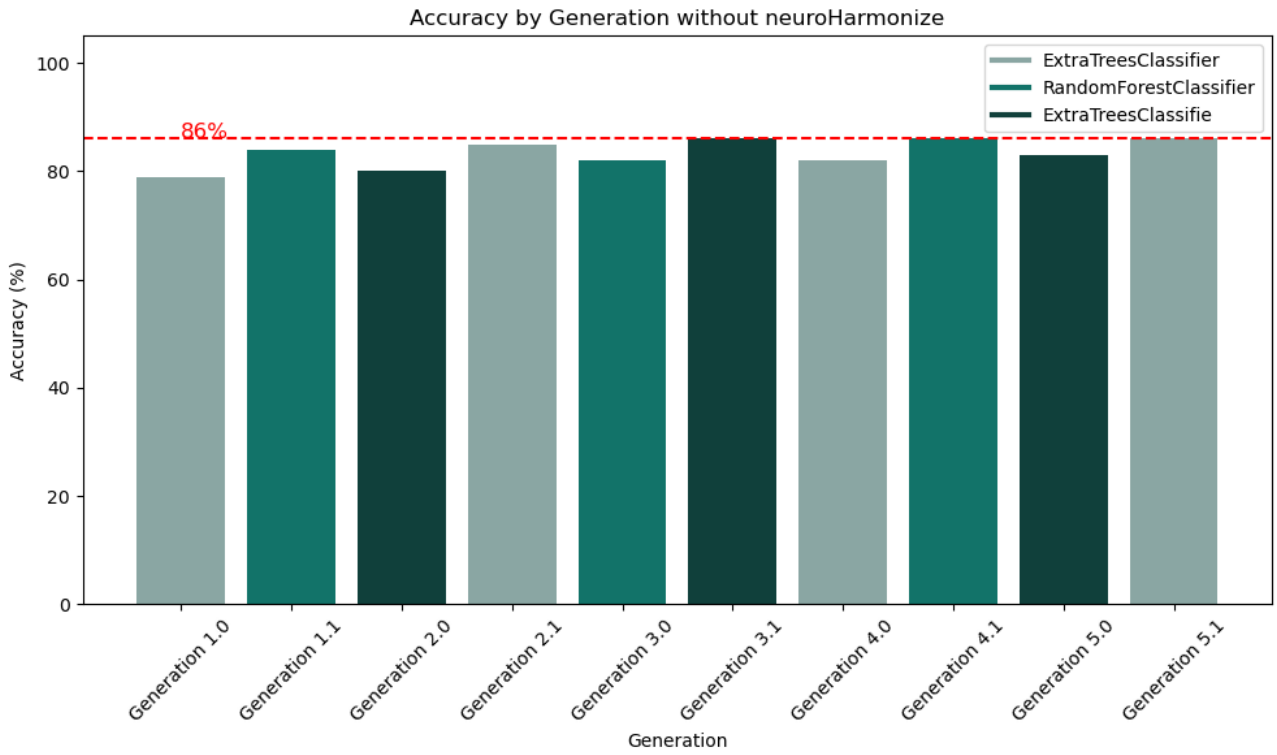


Figure 46 Five generations with ExtraTreesClassifier and the accuracy of each.

4.6.2 The second path (with neuroHarmonize)

Table 28 presents a comprehensive summary of the results obtained for each model implemented using the previously mentioned methodologies. These models were specifically applied in the second path, which emphasizes evaluating the outcomes using specialized libraries such as neuroHarmonize.

Table 28 Comprehensive summary of the results obtained for each model with neuroHarmonize.

Healthy: Control Group + G2 Group

Groups/Models		RF-B		DT		SVM		ET-T	
		Train	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
gICA	G1 vs Healthy	80%	64%	78%	70%	66%	67%	73%	66%
	G1 vs G2	75%	63%	79%	76%	60%	38%	70%	48%
ROIs	G1 vs Healthy	80%	70%	73%	70%	79%	66%	70%	68%
	G1 vs G2	83%	71%	83%	75%	62%	48%	77%	67%

RF-B: RandomForest using Boruta, **DT:** Decision Trees, **SVM:** Support vector machine, **ET-T:** ExtraTrees found with TPOP.

Below is a comprehensive description of the results obtained for each implemented model, using the dataset consisting of the G1 and control groups of neural gICA Components as an example. This group selection aligns with the project's objectives. However, it is important to note that all groups included in the project underwent the same methodology for both neural gICA Components and regions of interest (ROIs). Detailed information on all implemented models can be found in Annex 6.

4.6.2.1 RandomizedSearchCV

Figure 47, The curve demonstrates a 100% training accuracy, while the validation accuracy starts at 60% and steadily increases until reaching an accuracy of nearly

70%. Additionally, the variability of the validation accuracy with respect to the number of samples is also evident from the curve.

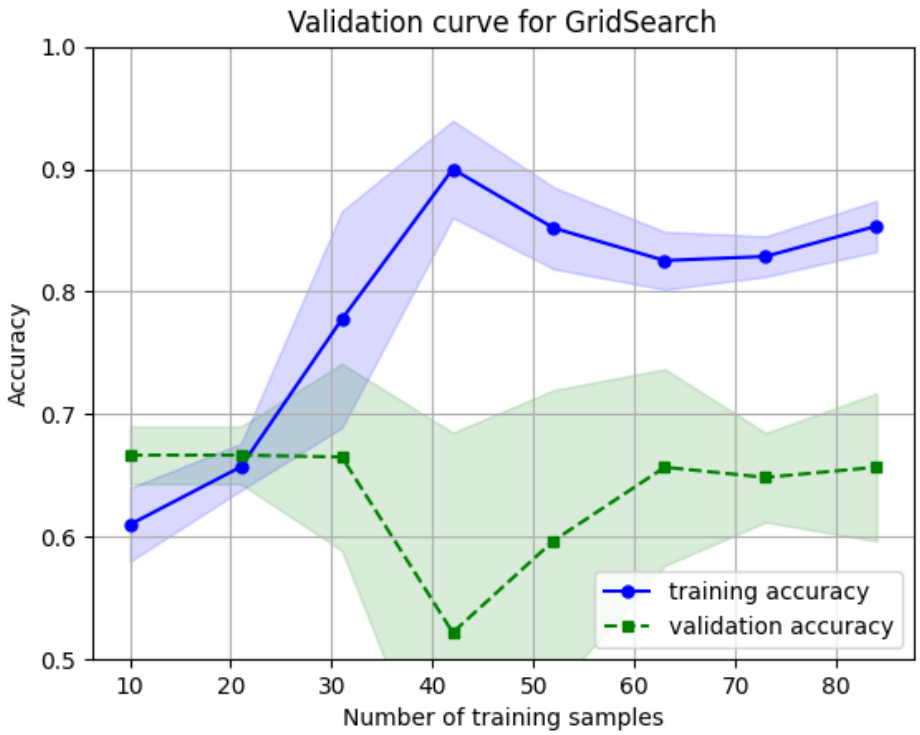


Figure 47 Validation curve for Grid Search with neuroHarmonize.

4.6.2.2 Boruta

Based on the harmonized data using neuroHarmonize, the Boruta feature selection algorithm identified three significant features for training the model.

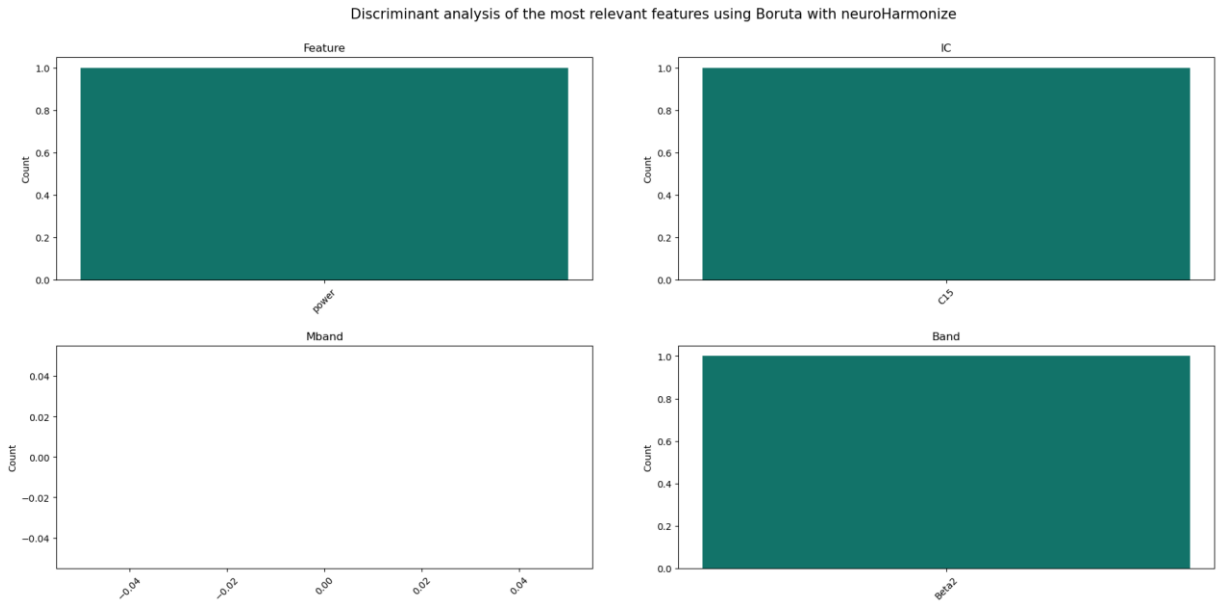


Figure 48 Discriminant analysis of the most relevant features using Boruta with neuroHarmonize.

In Figure 48, the selected features are differentiated by the evaluated metric, components, bands, and modulated bands. The y-axis represents the frequency of occurrence of the discriminated element among the 1 selected feature by Boruta.

Figure 49, the validation accuracy starts nearly at 68% and steadily increases until reaching an accuracy of 70%. Additionally, the variability of the validation accuracy with respect to the number of samples is also evident from the curve.

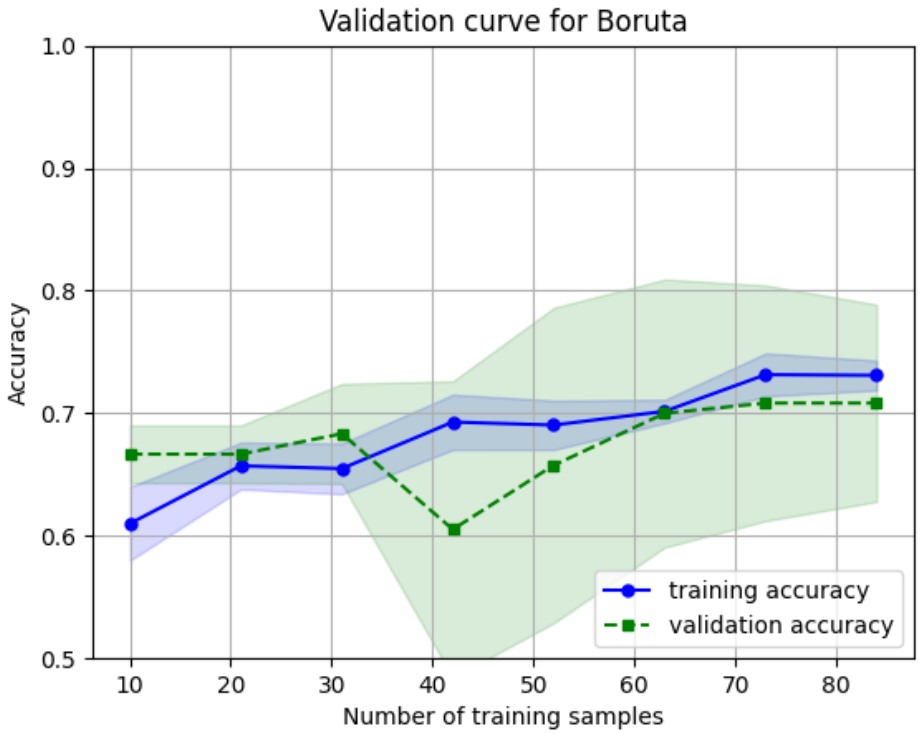


Figure 49 Validation curve for Boruta with neuroHarmonize.

4.6.2.3 Decision tree

Now, we present the results of the decision tree analysis, where the importance of all features in the database is evaluated. Each feature is initially assigned a relative importance score and enumerated in a list.

Figure 50 illustrates the importance of various features in the classification model. The x-axis represents the features, while the y-axis represents their respective importance scores. As observed, the Relative Powers in the Gamma band, specifically components 15, and 25, are among the most significant features. The three important feature is Synchronization Likelihood. For visual clarity, only the

first 10 features are displayed in the graph; however, the complete list of features with their corresponding relative importance can be found in the supplementary files.

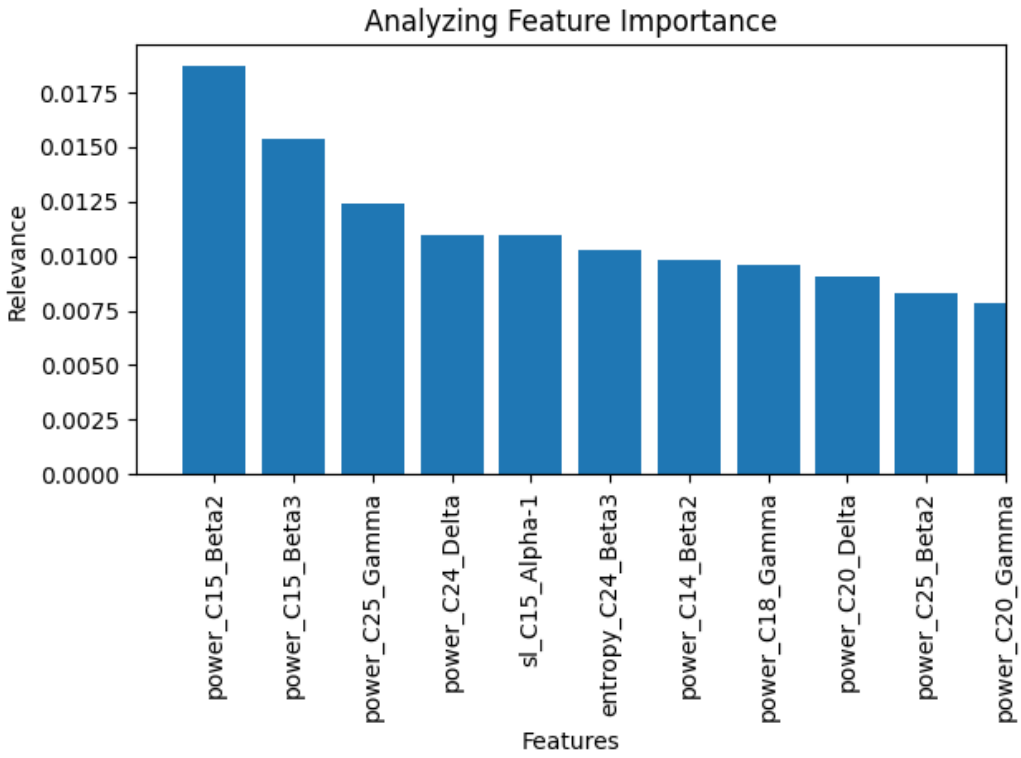


Figure 50 The importance of firsts features in the classification model with neuroHarmonize.

Once all the listed features are available, the model is trained by incrementally adding the features one by one. A training graph is then generated to visualize the relationship between the number of features entered (x-axis) and the corresponding accuracy (y-axis) achieved during each training iteration.

The graph demonstrates a significant increase in accuracy up to the first 5 features, reaching a peak of approximately 73%. This observation suggests that the most informative features for achieving high accuracy are concentrated within the initial set, and including additional features beyond a certain point does not contribute significantly to the model's performance.

Figure 51 illustrates the relationship between the number of features and the accuracy of the model.

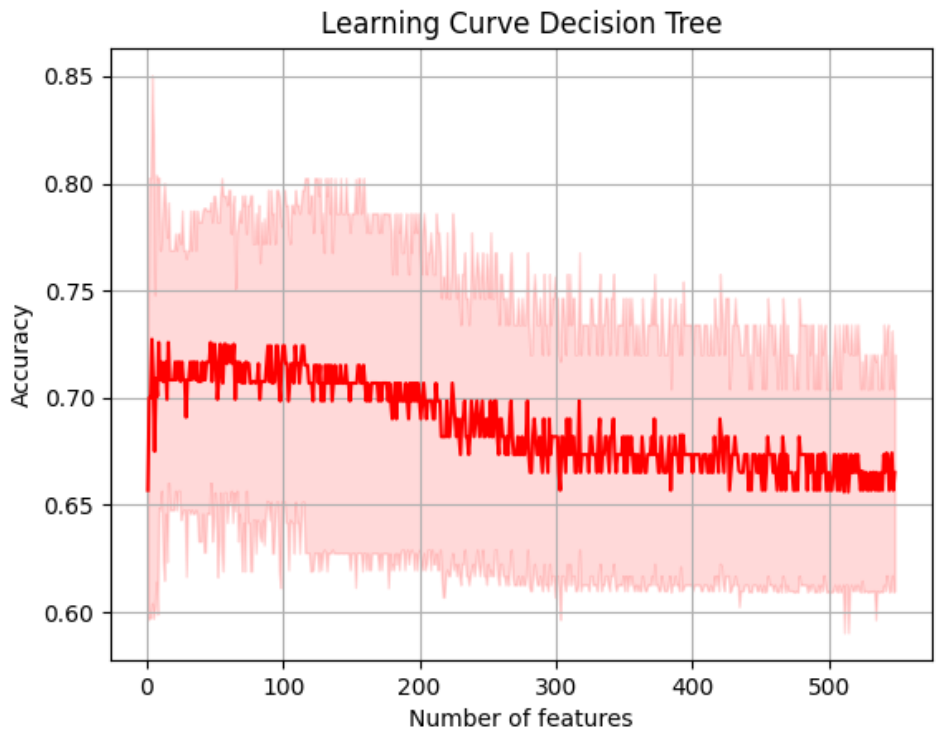


Figure 51 Relationship between the number of features and the accuracy of the model with neuroHarmonize.

Based on this initial finding, the model is subsequently trained using only the first 3 features.

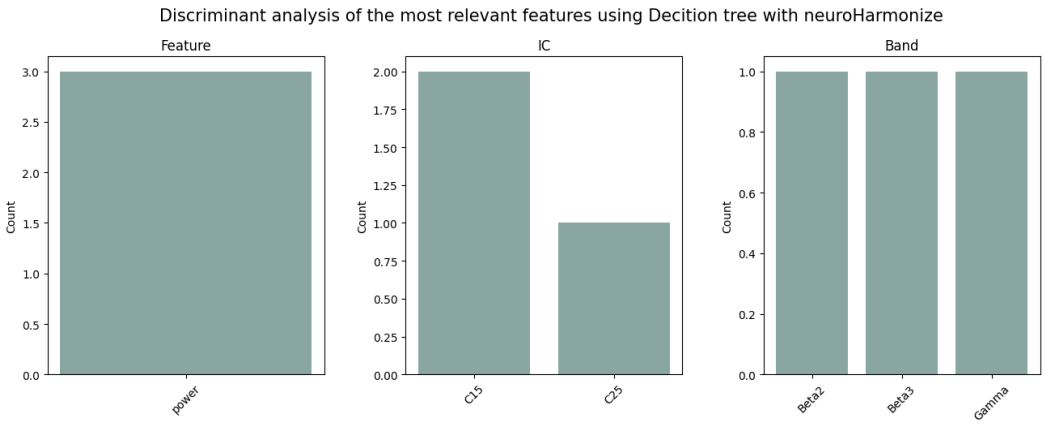


Figure 52 Discriminant analysis of the most relevant features with Decision tree with neuroHarmonize.

In Figure 52, the selected features are differentiated by the evaluated metric, components, bands, and modulated bands. The y-axis represents the frequency of occurrence of the discriminated element among the 3 selected features.

A graph like the previous Figure 51 is generated, focusing solely on the relationship between the number of features and the corresponding increase in accuracy. Notably, Figure 53 demonstrates a continuous rise in accuracy as the number of features increases, peaking at the 3rd feature reaching an accuracy of 78%. Like the learning curves, this training graph also provides insights into the variation of predictions across different training iterations.

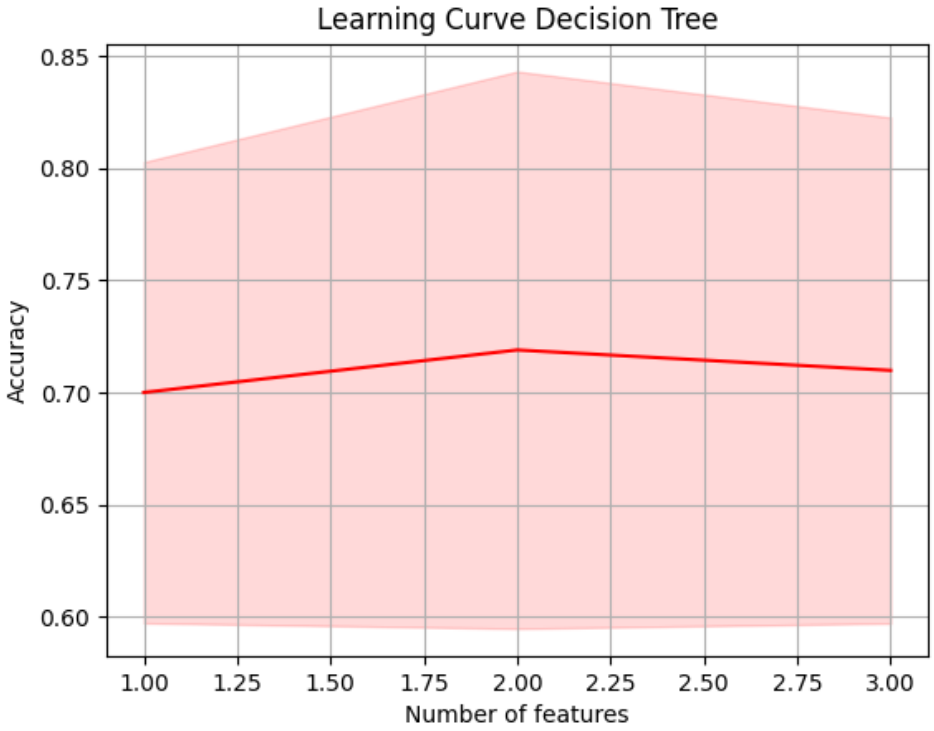


Figure 53 Relationship between the 3rd best features and the accuracy of the model with neuroHarmonize.

Figure 54, the validation accuracy starts nearly at 68% and steadily decreases until reaching an accuracy of 70%. Additionally, the variability of the validation accuracy with respect to the number of samples is also evident from the curve.

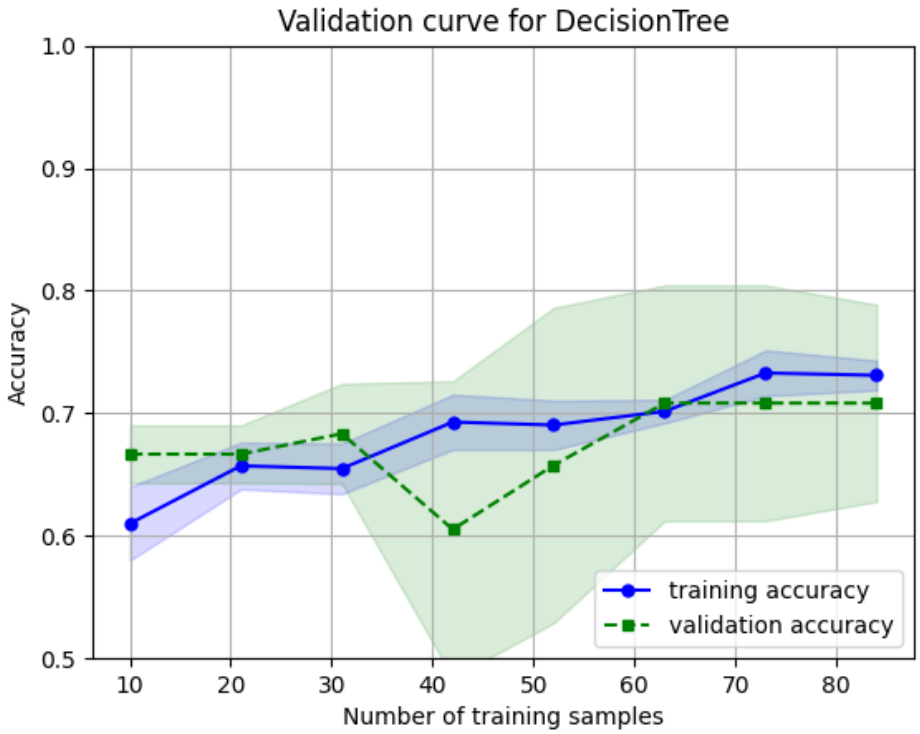


Figure 54 Validation curve for Decision Tree with neuroHarmonize.

This behavior reflects that the model is clearly overtrained, i.e., it has adapted too much to the training data and has lost its generalization capacity, despite the use of cross-validation, to improve this result it would be necessary to perform a strict feature selection.

Table 29 showcases the results of the algorithm's computational precision. It reveals an accuracy of 70%, accompanied by a standard deviation of 16%. Additionally, the precision, recall, and F1 score are reported to be 60%.

Table 29 Results of the algorithm's computational precision for decision tree with neuroHarmonize.

```
  ---Computer Precision---
Precision:  0.6
Recall:  0.6
F1-score: 0.6
Accuracy: 0.70
Standard deviation: 0.16
```

Finally, a confusion matrix is generated to analyze the test set. The test set consists of 30 subjects, out of which 4 subjects from G1 were correctly classified, and 6 subjects were erroneously classified as controls. From the healthy group, 17 subjects were correctly classified, and 3 subjects were erroneously classified. See Figure 55.

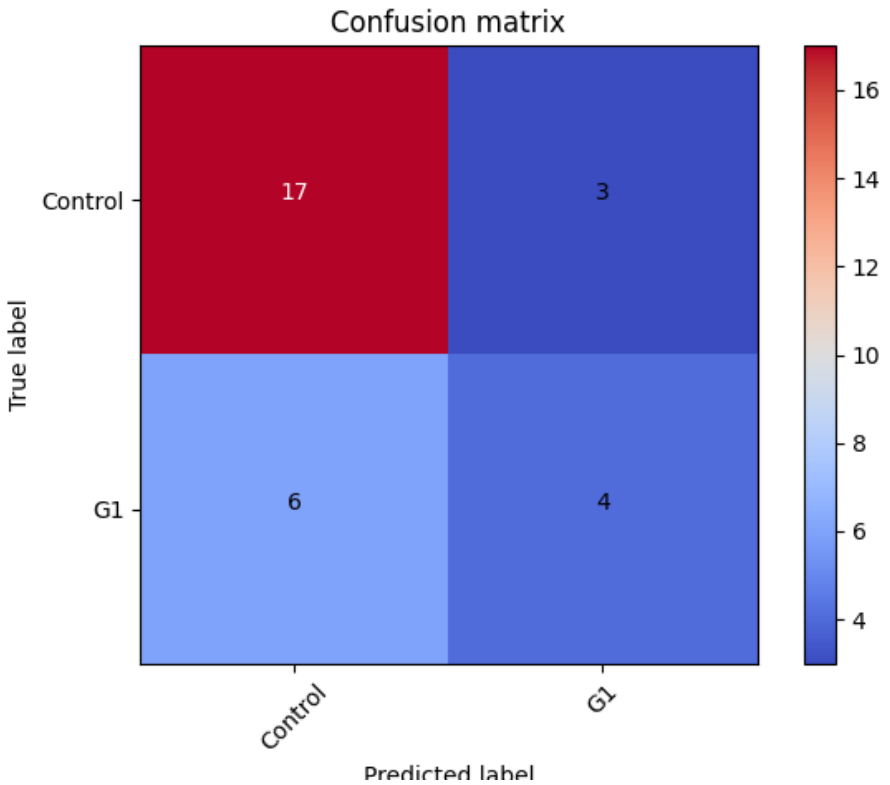


Figure 55 Confusion matrix for decision tree with neuroHarmonize.

4.6.2.4 Support Vector Machine (SVM)

Table 30 Results of the algorithm's computational precision for SVM with neuroHarmonize.

---Computer Precision--
Precision: 0.67
Recall: 1.0
F1-score: 0.8
Accuracy: 0.67

4.6.2.5 TPOT

Using TPOT, Figure 56 shows the five generations of Table 23 and their respective accuracies.

The precision values represent the accuracy of the machine learning models generated by TPOT in each generation. It can be observed that the precision remains constant over all generations at a value of 66%.

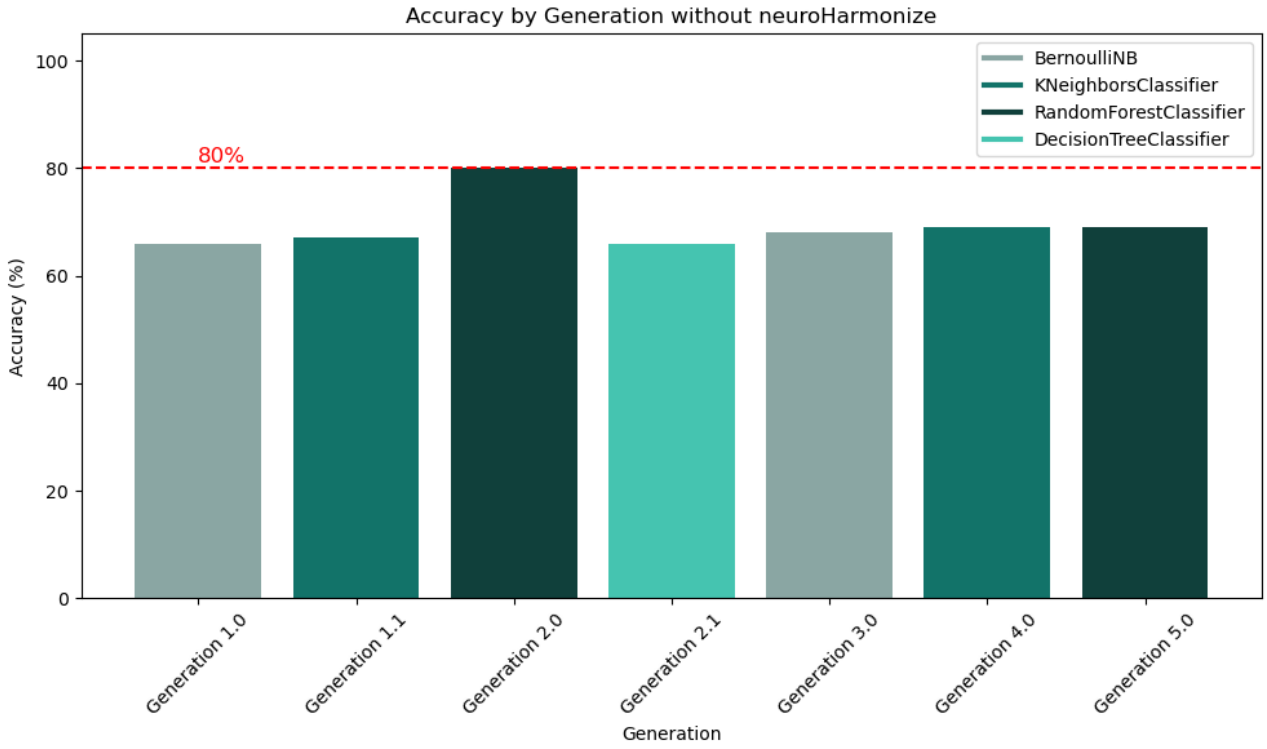


Figure 56 Five generations with different classifiers and the accuracy of each.

4.6 Discussion

The main objective of this study was to improve machine learning classification models by augmenting data to effectively categorize individuals at risk for Alzheimer's disease using non-invasive biomarkers from different databases. The results presented in this study indicate that the most optimal dataset for achieving superior classification between the two groups (G1: PSEN1-E280A gene carriers

and controls plus G2) was obtained by the second path, where the processing pipeline included the use of neuroHarmonize.

These findings are consistent with previous research demonstrating the potential of machine learning algorithms in biomarker-based AD classification. In a systematic review, Dijana et al. highlighted the importance of feature selection and model optimization in achieving accurate classification results [212]. Additionally, García-Pretelt et al. [184] used a support vector machine with an RBF kernel and achieved a test accuracy of 83% in their study, which used a smaller subset of the dataset compared to ours (27 carriers of the PSEN1-E280A gene and 33 controls from the UdeA 1 cohort). Notably, our study expanded the dataset to include 49 carriers of the PSEN1-E280A gene and 98 controls. This expansion involved additional processing steps such as channel inspection, data transformation between cohorts using an ICA matrix (58x25), Huber normalization, and MatchIt matching [184]. Consequently, we did not observe an improvement in the model's accuracy with an increase in the sample size; however, we discovered that utilizing Neuroharmonize enables the harmonization of controls across different sites.

However, the statistical analyses performed in the previous chapters lead us to speculate that the performance of the model without neuroHarmonize was able to detect differences between the groups, likely due to the fact that the controls in this case were from different cohorts. As evidenced by the effect size evaluation, these differences could introduce bias into the model.

Specifically, García-Pretelt et al. [184] found that the most relevant components were 14, 20, 22, 23, 24, and 25 for Relative Power and Cross Frequency, with which he obtained 89% accuracy in training and 83% accuracy in testing. These results are congruent and comparable to those found in this project in the flow that considers the data with neuroHarmonize, where the most relevant features are 18,20, 24, and 25 for Relative Power and Synchronization Likelihood, with which he obtained 78% accuracy in training and 65% accuracy in testing.

García-Pretelt et al. [184] also point out that the features presented had an effect size greater than 0.7 during model training, even though not all features were considered relevant in the model. Similarly, most of the features evaluated in this project without neuroHarmonize had an effect size greater than 0.8, while with neuroHarmonize they approached 0.2. Therefore, it's possible that the model with harmonized data could discriminate between the groups, but not with excessive accuracy.

Regarding the bands in García-Pretelt et al. [184] Gamma and Beta presented significant differences in the statistical part, but it was not consistent when applying the classification model. In this case, this project presents relevant results in the gamma, delta and theta bands in the classification of the model, obtaining a classification of 78% in training and 65% of accuracy in test for the data with neuroHarmonize.

In our study, the decision tree algorithm and TPOT emerged as the top-performing algorithms for both the first and second paths, consistent with other findings who successfully employed decision trees [63] and convolutional neural networks for classifying mild cognitive impairment and Alzheimer's disease [213]. Using decision trees, in the first path we used 46 features to achieve an outstanding accuracy of 85%, and in the second path we identified 5 significant features to achieve an accuracy of 65%.

The processing pipeline implemented in our study played a pivotal role in obtaining these results. The application of techniques such as gICA, Huber normalization, and MatchIt matching facilitated data structuring, enabling comparability and suitability for classifying asymptomatic individuals carrying the PSEN1-E280A genetic variant compared to the control group. These findings align with studies emphasizing the importance of data preprocessing and feature selection techniques in enhancing the accuracy of Alzheimer's disease diagnosis [81].

It is important to recognize the limitations and potential variability in the efficacy of harmonization techniques across different datasets and populations, such as the use of neuroHarmonize, which showed excellent results in integrating cohorts, but showed negative values in bands such as gamma, which could lead to inconsistencies when standardizing a protocol with this tool. Gonzalez-Escamilla et al. highlighted the difficulties associated with using machine learning models with limited MRI data and emphasized the importance of robust feature extraction methods [214].

The achieved precision of 65%, with a standard deviation of 16%, and a precision, recall, and F1 score of 60% in our study demonstrate interesting results. However, it is vital to recognize that these outcomes are derived from a specific dataset, and further validation on larger cohorts is needed. However, it is vital to recognize that these outcomes are derived from a specific dataset, and further validation on larger cohorts is needed. Moradi et al. proposed an early MRI-based Alzheimer's conversion prediction model using a machine learning framework, emphasizing the significance of model generalizability and replication [201]. Additionally, Audrey et al. emphasized the potential of machine learning models in Alzheimer's disease diagnosis using FDG-PET data, further highlighting the need for rigorous evaluation and validation [215].

Additionally, it is important to explore the effect of linear harmonization on the performance of the evaluated classification methods. By thoroughly examining the impact of linear harmonization techniques, we can gain valuable insights into their effectiveness in improving the performance and generalizability of classification models.

4.7 Conclusions

In conclusion, this study makes a significant contribution to the field by implementing a high-quality data processing pipeline for incremental data and achieving an accurate and reliable machine learning model for classifying individuals at risk for Alzheimer's disease. The results underscore the critical role

of employing appropriate data preprocessing, feature selection, and model optimization techniques to achieve high accuracy in the classification task. In particular, the use of normalization and matching techniques, as well as an increase in data volume, positively impacted data harmonization for the healthy group.

Consideration of potential limitations and variations in harmonization techniques is critical, as their effectiveness may vary across different datasets and populations. In addition, exploring the application of machine learning models to other non-invasive biomarkers holds promise for improving the accuracy and reliability of Alzheimer's disease classification.

To further improve the applicability of the model, future research should focus on replicating these results on larger cohorts. In addition, the use of generalized ICA matrices constructed from multiple databases is essential to improve the ability to effectively transform multi-site data.

Significant results include metrics that are consistent with previous studies, such as power, Synchronization Likelihood, and Cross Frequency, which should be further explored due to their consistently positive results. Similarly, consistent results were obtained for neural components 18,20, 24, and 25. Finally, the gamma, delta, theta and beta bands emerged as the primary contributors to the classification of these populations.

By addressing these issues, the field can advance our understanding of the disease and potentially contribute to the development of early and accurate diagnostic tools.

Continued efforts in this direction are critical to improving patient outcomes and facilitating timely interventions in the fight against Alzheimer's disease.

Chapter 5

General conclusions and future work

The project highlighted the importance of harmonizing EEG data and the potential of EEG-based biomarkers for early detection and screening of Alzheimer's disease. It discussed the challenges of EEG analysis, including the lack of standardized processing pipelines and organizational standards, and the need for validation methods using larger and more diverse data sets. Harmonization efforts, such as the EEG-BIDS and EEG-IP platforms, were identified as essential to integrate data and improve data quality, leading to accelerated biomarker discovery research.

Machine learning models, particularly decision trees and other algorithms, were recognized as critical tools for AD classification using non-invasive biomarkers. The project emphasized the importance of harmonizing data from multiple cohorts to increase sample size, improve statistical power, and identify consistent features or biomarkers across cohorts. The development of a robust and generalizable machine learning model using a larger and more diverse dataset was a key objective of the project.

The results obtained from the machine learning model showed promising results, with high accuracy achieved using the decision tree algorithm and TPOT. The

implemented processing techniques, including gICA, Huber normalization, and MatchIt matching, played an important role in structuring the data and improving accuracy. However, the inclusion of neuroHarmonize did not yield the expected results, indicating the need for further exploration and evaluation.

In conclusion, this project made a significant contribution to dementia research by developing a processing pipeline and machine learning model for accurately classifying individuals at risk for Alzheimer's disease. It highlights the importance of standardized pipelines, data harmonization, and the adoption of BIDS to improve accessibility and reproducibility in neuroscience research. Further validation in larger cohorts and exploration of other non-invasive biomarkers were recommended for future research. In addition, addressing imbalanced sample sizes and understanding the impact of linear harmonization on classification methods were identified as important considerations for improving the reliability and robustness of harmonization techniques in different areas of study.

References

[1] C. A. Reyes-Ortiz, S. Pacheco, C. A. Slovacek, M. Jiang, I. C. Salinas-Fernandez, and J. M. Ocampo-Chaparro, "Medical falls among older adults in Latin American cities," *Rev Salud Publica (Bogota)*, vol. 22, no. 5, Oct. 2020, doi: 10.15446/RSAP.V22N5.84883.

[2] N. Unidas and U. Nations, "Ageing in Latin America and the Caribbean: inclusion and rights of older persons," Dec. 2022, Accessed: May 13, 2023. [Online]. Available: https://repositorio.cepal.org/handle/11362/48568

[3] R. A. Cohen, M. M. Marsiske, and G. E. Smith, "Neuropsychology of aging," *Handb Clin Neurol*, vol. 167, pp. 149–180, Jan. 2019, doi: 10.1016/B978-0-12-804766-8.00010-8.

[4] A. Kumar, J. Sidhu, A. Goyal, and J. W. Tsao, "Alzheimer Disease," *StatPearls*, pp. 1–27, Jun. 2022, Accessed: May 13, 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK499922/

[5] C. R. Jack *et al.*, "NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease," *Alzheimers Dement*, vol. 14, no. 4, pp. 535–562, Apr. 2018, doi: 10.1016/J.JALZ.2018.02.018.

[6] P. Monllor *et al.*, "Electroencephalography as a Non-Invasive Biomarker of Alzheimer's Disease: A Forgotten Candidate to Substitute CSF Molecules?," *Int J Mol Sci*, vol. 22, no. 19, Oct. 2021, doi: 10.3390/IJMS221910889.

[7] R. Cassani, M. Estarellas, R. San-Martin, F. J. Fraga, and T. H. Falk, "Systematic Review on Resting-State EEG for Alzheimer's Disease Diagnosis and Progression Assessment," *Dis Markers*, vol. 2018, 2018, doi: 10.1155/2018/5174815.

[8] G. Cecchetti *et al.*, “Resting-state electroencephalographic biomarkers of Alzheimer’s disease,” *Neuroimage Clin*, vol. 31, Jan. 2021, doi: 10.1016/J.NICL.2021.102711.

[9] C. Babiloni *et al.*, “International Federation of Clinical Neurophysiology (IFCN) – EEG research workgroup: Recommendations on frequency and topographic analysis of resting state EEG rhythms. Part 1: Applications in clinical research studies,” *Clinical Neurophysiology*, vol. 131, no. 1, pp. 285–307, Jan. 2020, doi: 10.1016/J.CLINPH.2019.06.234.

[10] C. Babiloni *et al.*, “Measures of resting state EEG rhythms for clinical trials in Alzheimer’s disease: Recommendations of an expert panel,” *Alzheimer’s & Dementia*, vol. 17, no. 9, pp. 1528–1553, Sep. 2021, doi: 10.1002/ALZ.12311.

[11] A. H. H. Al-Nuaimi, E. Jammeh, L. Sun, and E. Ifeachor, “Complexity Measures for Quantifying Changes in Electroencephalogram in Alzheimer’s Disease,” *Complexity*, vol. 2018, 2018, doi: 10.1155/2018/8915079.

[12] M. Rashid *et al.*, “Current Status, Challenges, and Possible Solutions of EEG-Based Brain-Computer Interface: A Comprehensive Review,” *Front Neurorobot*, vol. 14, p. 25, Jun. 2020, doi: 10.3389/FNBOT.2020.00025.

[13] L. Pan, H. Zheng, and T. Li, “Effects of the indoor environment on EEG and thermal comfort assessment in males,” *Build Environ*, vol. 228, p. 109761, Jan. 2023, doi: 10.1016/J.BUILDENV.2022.109761.

[14] A. de Cheveigné, “ZapLine: A simple and effective method to remove power line artifacts,” *Neuroimage*, vol. 207, p. 116356, Feb. 2020, doi: 10.1016/J.NEUROIMAGE.2019.116356.

[15] M. K. Islam, A. Rastegarnia, and Z. Yang, “Methods for artifact detection and removal from scalp EEG: A review,” *Neurophysiologie Clinique/Clinical Neurophysiology*, vol. 46, no. 4–5, pp. 287–305, Nov. 2016, doi: 10.1016/J.NEUCLI.2016.07.002.

[16] A. Pedroni, A. Bahreini, and N. Langer, “Automagic: Standardized preprocessing of big EEG data,” *Neuroimage*, vol. 200, pp. 460–473, Oct. 2019, doi: 10.1016/J.NEUROIMAGE.2019.06.046.

[17] A. R. Javed *et al.*, “Artificial Intelligence for Cognitive Health Assessment: State-of-the-Art, Open Challenges and Future Directions,” *Cognitive Computation 2023*, vol. 1, pp. 1–46, Jun. 2023, doi: 10.1007/S12559-023-10153-4.

[18] H. Hu, R. Kumar Das, A. Martin, T. Zurales, D. Dowling, and A. Khan, “A Survey on EEG Data Analysis Software,” *Sci 2023, Vol. 5, Page 23*, vol. 5, no. 2, p. 23, Jun. 2023, doi: 10.3390/SCI5020023.

[19] J. T. Fuller *et al.*, “Biological and Cognitive Markers of Presenilin1 E280A Autosomal Dominant Alzheimer’s Disease: A Comprehensive Review of the Colombian Kindred,” *J Prev Alzheimers Dis*, vol. 6, no. 2, p. 112, 2019, doi: 10.14283/JPAD.2019.6.

[20] “Normalization and Standardization of Data.” https://akashkunwar.hashnode.dev/understanding-the-differences-between-normalization-and-standardization-of-data (accessed Aug. 21, 2023).

[21] J. Caruana, “Harmonization,” *Global Encyclopedia of Public Administration, Public Policy, and Governance*, pp. 1–9, 2016, doi: 10.1007/978-3-319-31816-5_2281-1.

[22] P. Prado *et al.*, “Dementia ConnEEGtome: Towards multicentric harmonization of EEG connectivity in neurodegeneration,” *International Journal of Psychophysiology*, vol. 172, pp. 24–38, Feb. 2022, doi: 10.1016/J.IJPSYCHO.2021.12.008.

[23] L. M. Alexander *et al.*, “Data Descriptor: An open resource for transdiagnostic research in pediatric mental health and learning disorders,” *Sci Data*, vol. 4, Dec. 2017, doi: 10.1038/SDATA.2017.181.

[24] “OpenNeuro.” https://openneuro.org/ (accessed May 13, 2023).

[25] N. Bigdely-Shamlo, J. Touryan, A. Ojeda, C. Kothe, T. Mullen, and K. Robbins, “Automated EEG mega-analysis I: Spectral and amplitude characteristics across studies,” *Neuroimage*, vol. 207, p. 116361, Feb. 2020, doi: 10.1016/J.NEUROIMAGE.2019.116361.

[26] K. J. Gorgolewski *et al.*, “The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments,” *Scientific Data 2016 3:1*, vol. 3, no. 1, pp. 1–9, Jun. 2016, doi: 10.1038/sdata.2016.44.

[27] C. R. Pernet *et al.*, “EEG-BIDS, an extension to the brain imaging data structure for electroencephalography,” *Scientific Data*, vol. 6, no. 1. Nature Research, pp. 1–5, Dec. 01, 2019. doi: 10.1038/s41597-019-0104-8.

[28] M. C. Biagioni and J. E. Galvin, “Using biomarkers to improve detection of Alzheimer’s disease,” *Neurodegener Dis Manag*, vol. 1, no. 2, p. 127, Apr. 2011, doi: 10.2217/NMT.11.11.

[29] T. M. Rutkowski, M. S. Abe, T. Komendzinski, H. Sugimoto, S. Narebski, and M. Otake-Matsuura, “Machine learning approach for early onset dementia neurobiomarker using EEG network topology features,” *Front Hum Neurosci*, vol. 17, p. 1155194, Jun. 2023, doi: 10.3389/FNHUM.2023.1155194/BIBTEX.

[30] A. Khan and S. Zubair, “Development of a three tiered cognitive hybrid machine learning algorithm for effective diagnosis of Alzheimer’s disease,” *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 10, pp. 8000–8018, Nov. 2022, doi: 10.1016/J.JKSUCI.2022.07.016.

[31] A. Chaddad, Y. Wu, R. Kateb, and A. Bouridane, “Electroencephalography Signal Processing: A Comprehensive Review and Analysis of Methods and Techniques,” *Sensors 2023, Vol. 23, Page 6434*, vol. 23, no. 14, p. 6434, Jul. 2023, doi: 10.3390/S23146434.

[32] L. Del Fabro, E. Bondi, F. Serio, E. Maggioni, A. D’Agostino, and P. Brambilla, “Machine learning methods to predict outcomes of pharmacological treatment in psychosis,” *Translational Psychiatry 2023 13:1*, vol. 13, no. 1, pp. 1–15, Mar. 2023, doi: 10.1038/s41398-023-02371-z.

[33] S. Iannaccone, E. Houdayer, A. Spina, G. Nocera, and F. Alemanno, “Quantitative EEG for early differential diagnosis of dementia with Lewy bodies,” *Front Psychol*, vol. 14, Apr. 2023, doi: 10.3389/FPSYG.2023.1150540.

[34] S. Asadzadeh, T. Yousefi Rezaii, S. Beheshti, A. Delpak, and S. Meshgini, “A systematic review of EEG source localization techniques and their applications on diagnosis of brain abnormalities,” *J Neurosci Methods*, vol. 339, p. 108740, Jun. 2020, doi: 10.1016/J.JNEUMETH.2020.108740.

[35] C. T. Briels *et al.*, “Reproducibility of EEG functional connectivity in Alzheimer’s disease,” *Alzheimers Res Ther*, vol. 12, no. 1, pp. 1–14, Jun. 2020, doi: 10.1186/S13195-020-00632-3/TABLES/3.

[36] E. R. , S. J. H. , J. T. M. , S. S. A. , & H. A. J. Kandel, *Principles of Neural Science*. 2013.

[37] H. Berger, “Über das elektroenkephalogramm des menschen,” *Arch Psychiatr Nervenkr*, vol. 1, p. 87, 1929.

[38] M. E. Raichle and A. Z. Snyder, “A default mode of brain function: a brief history of an evolving idea,” *Neuroimage*, vol. 37, no. 4, pp. 1083–1090, Oct. 2007, doi: 10.1016/J.NEUROIMAGE.2007.02.041.

[39] D. A. Fair *et al.*, “Development of distinct control networks through segregation and integration,” *Proc Natl Acad Sci U S A*, vol. 104, no. 33, pp. 13507–13512, Aug. 2007, doi: 10.1073/PNAS.0705843104/SUPPL_FILE/05843TABLE3.PDF.

[40] M. Jobert, F. J. Wilson, G. S. F. Ruigt, M. Brunovsky, L. S. Prichep, and W. H. I. M. Drinkenburg, “Guidelines for the recording and evaluation of pharmaco-EEG data in man: the International Pharmaco-EEG Society (IPEG),” *Neuropsychobiology*, vol. 66, no. 4, pp. 201–220, 2012, doi: 10.1159/000343478.

[41] G. Bernardi, M. Betta, E. Ricciardi, P. Pietrini, G. Tononi, and F. Siclari, “Regional Delta Waves In Human Rapid Eye Movement Sleep,” *Journal of Neuroscience*, vol. 39, no. 14, pp. 2686–2697, Apr. 2019, doi: 10.1523/JNEUROSCI.2298-18.2019.

[42] M. B. Bin Heyat, D. Lai, F. Akhtar, M. A. Bin Hayat, and S. Azad, “Short time frequency analysis of theta activity for the diagnosis of bruxism on EEG sleep record,” *Studies in Computational Intelligence*, vol. 875, pp. 63–83, 2020, doi: 10.1007/978-3-030-35252-3_4/FIGURES/12.

[43] P. Danjou *et al.*, “Electrophysiological assessment methodology of sensory processing dysfunction in schizophrenia and dementia of the Alzheimer type,” *Neurosci Biobehav Rev*, vol. 97, pp. 70–84, Feb. 2019, doi: 10.1016/J.NEUBIOREV.2018.09.004.

[44] K. Corace, R. Baysarowich, M. Willows, A. Baddeley, N. Schubert, and V. Knott, “Resting State EEG Activity Related to Impulsivity in People with Prescription Opioid Use Disorder,” *Psychiatry Res Neuroimaging*, vol. 321, p. 111447, Apr. 2022, doi: 10.1016/J.PSCYCHRESNS.2022.111447.

[45] X. Ma, L. Song, B. Hong, Y. Li, and Y. Li, “Relationships between EEG and thermal comfort of elderly adults in outdoor open spaces,” *Build Environ*, vol. 235, p. 110212, May 2023, doi: 10.1016/J.BUILDENV.2023.110212.

[46] J. A. C. Saeid Sanei, “EEG Signal Processing and Machine Learning - Saeid Sanei, Jonathon A. Chambers - Google Libros,” 2022. https://books.google.com.co/books?hl=es&lr=&id=yt9BEAAAQBAJ&oi=fnd&pg=PA17&dq=Gamma,+Although+the+amplitudes+of+these+rhythms+are+very+small+and+their+occurrence+is+rare,+the+detection+of+these+rhythms+can+be+used+to+confirm+certain+brain+diseases.+&ots=usglQxZpbK&sig=CLn8utyYpKkKTkB4n1c5RDsgs64&redir_esc=y#v=onepage&q&f=false (accessed May 14, 2023).

[47] F. W. Pfrieger, “Neurodegenerative Diseases and Cholesterol: Seeing the Field Through the Players,” *Front Aging Neurosci*, vol. 13, Nov. 2021, doi: 10.3389/FNAGI.2021.766587.

[48] OME, “Global action plan on the public health response to dementia 2017 - 2025,” *Geneva: World Health Organization*, p. 27, 2017, Accessed: May 14, 2023. [Online]. Available: http://apps.who.int/bookorders.

[49] M. Ayaz, A. Nawaz, F. Naz, F. Ullah, A. Sadiq, and Z. U. Islam, “Phytochemicals-based Therapeutics against Alzheimer’s Disease: An Update,” *Curr Top Med Chem*, vol. 22, no. 22, pp. 1811–1820, Sep. 2022, doi: 10.2174/1568026622666220815104305.

[50] J. Andrade-Guerrero *et al.*, “Alzheimer’s Disease: An Updated Overview of Its Genetics,” *International Journal of Molecular Sciences 2023, Vol. 24, Page 3754*, vol. 24, no. 4, p. 3754, Feb. 2023, doi: 10.3390/IJMS24043754.

[51] X. X. Zhang, Y. Tian, Z. T. Wang, Y. H. Ma, L. Tan, and J. T. Yu, “The Epidemiology of Alzheimer’s Disease Modifiable Risk Factors and Prevention,” *The Journal of Prevention of Alzheimer’s Disease 2021 8:3*, vol. 8, no. 3, pp. 313–321, Apr. 2021, doi: 10.14283/JPAD.2021.15.

[52] J. Alber *et al.*, “Developing retinal biomarkers for the earliest stages of Alzheimer’s disease: What we know, what we don’t, and how to move forward,” *Alzheimer’s & Dementia*, vol. 16, no. 1, pp. 229–243, Jan. 2020, doi: 10.1002/ALZ.12006.

[53] K. A. Jellinger, “Recent update on the heterogeneity of the Alzheimer’s disease spectrum,” *J Neural Transm*, vol. 129, no. 1, pp. 1–24, Jan. 2022, doi: 10.1007/S00702-021-02449-2/FIGURES/3.

[54] E. Giacobini, A. C. Cuello, and A. Fisher, “Reimagining cholinergic therapy for Alzheimer’s disease,” *Brain*, vol. 145, no. 7, pp. 2250–2275, Jul. 2022, doi: 10.1093/BRAIN/AWAC096.

[55] Z. R. Chen, J. B. Huang, S. L. Yang, and F. F. Hong, “Role of Cholinergic Signaling in Alzheimer’s Disease,” *Molecules 2022, Vol. 27, Page 1816*, vol. 27, no. 6, p. 1816, Mar. 2022, doi: 10.3390/MOLECULES27061816.

[56] C. Ramos *et al.*, “Substance Use-Related Cognitive Decline in Families with Autosomal Dominant Alzheimer’s Disease: A Cohort Study,” *Journal of Alzheimer’s Disease*, vol. 85, no. 4, pp. 1423–1439, Jan. 2022, doi: 10.3233/JAD-215169.

[57] G. B. Frisoni *et al.*, “The probabilistic model of Alzheimer disease: the amyloid hypothesis revised,” *Nature Reviews Neuroscience 2021 23:1*, vol. 23, no. 1, pp. 53–66, Nov. 2021, doi: 10.1038/s41583-021-00533-w.

[58] F. Maestú, W. de Haan, M. A. Busche, and J. DeFelipe, “Neuronal excitation/inhibition imbalance: core element of a translational perspective on Alzheimer pathophysiology,” *Ageing Res Rev*, vol. 69, p. 101372, Aug. 2021, doi: 10.1016/J.ARR.2021.101372.

[59] D. S. Knopman *et al.*, “The National Institute on Aging and the Alzheimer’s Association Research Framework for Alzheimer’s disease: Perspectives from the Research Roundtable,” *Alzheimer’s & Dementia*, vol. 14, no. 4, pp. 563–575, Apr. 2018, doi: 10.1016/J.JALZ.2018.03.002.

[60] R. Khoury and E. Ghossoub, “Diagnostic biomarkers of Alzheimer’s disease: A state-of-the-art review,” *Biomark Neuropsychiatry*, vol. 1, p. 100005, Dec. 2019, doi: 10.1016/J.BIONPS.2019.100005.

[61] S. Gunes, Y. Aizawa, T. Sugashi, M. Sugimoto, and P. P. Rodrigues, “Biomarkers for Alzheimer’s Disease in the Current State: A Narrative Review,” *International Journal of Molecular Sciences 2022, Vol. 23, Page 4962*, vol. 23, no. 9, p. 4962, Apr. 2022, doi: 10.3390/IJMS23094962.

[62] A. J. Atkinson *et al.*, “Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework,” *Clin Pharmacol Ther*, vol. 69, no. 3, pp. 89–95, Jan. 2001, doi: 10.1067/MCP.2001.113989.

[63] B. Jiao *et al.*, “Neural biomarker diagnosis and prediction to mild cognitive impairment and Alzheimer’s disease using EEG technology,” *Alzheimers Res Ther*, vol. 15, no. 1, pp. 1–14, Dec. 2023, doi: 10.1186/S13195-023-01181-1/FIGURES/6.

[64] N. Chedid, J. Tabbal, A. Kabbara, S. Allouch, and M. Hassan, “The development of an automated machine learning pipeline for the detection of Alzheimer’s Disease,” *Scientific Reports 2022 12:1*, vol. 12, no. 1, pp. 1–7, Oct. 2022, doi: 10.1038/s41598-022-22979-3.

[65] A. A. Horvath *et al.*, “Subclinical epileptiform activity accelerates the progression of Alzheimer’s disease: A long-term EEG study,” *Clinical Neurophysiology*, vol. 132, no. 8, pp. 1982–1989, Aug. 2021, doi: 10.1016/J.CLINPH.2021.03.050.

[66] M. K. Jónsdóttir, J. Harrison, and K. I. Hannesdóttir, “The ambivalence toward neuropsychology in dementia research, diagnosis, and drug development: Myths and misconceptions,” *Alzheimer’s & Dementia*, vol. 19, no. 5, pp. 2175–2181, May 2023, doi: 10.1002/ALZ.12909.

[67] C. H. Ciolek and S. Y. Lee, “Cognitive Issues in the Older Adult,” *Guccione’s Geriatric Physical Therapy*, pp. 425–452, Jan. 2020, doi: 10.1016/B978-0-323-60912-8.00019-1.

[68] J. Hobson, “The Montreal Cognitive Assessment (MoCA),” *Occup Med (Chic Ill)*, vol. 65, no. 9, pp. 764–765, Dec. 2015, doi: 10.1093/OCCMED/KQV078.

[69] V. Korten *et al.*, “Prevalence of HIV-associated neurocognitive disorder (HAND) in Turkey and assessment of Addenbrooke’s Cognitive

Examination Revised (ACE-R) test as a screening tool," *HIV Med*, vol. 22, no. 1, pp. 60–66, Jan. 2021, doi: 10.1111/HIV.12957.

[70] V. Singhal, "Clinical Approach to Acute Decline in Sensorium," *Indian J Crit Care Med*, vol. 23, no. Suppl 2, p. S120, 2019, doi: 10.5005/JP-JOURNALS-10071-23188.

[71] M. F. Folstein, S. E. Folstein, and P. R. McHugh, "'Mini-mental state'. A practical method for grading the cognitive state of patients for the clinician," *J Psychiatr Res*, vol. 12, no. 3, pp. 189–198, Nov. 1975, doi: 10.1016/0022-3956(75)90026-6.

[72] S. García-Herranz, M. C. Díaz-Mardomingo, C. Venero, and H. Peraita, "Accuracy of verbal fluency tests in the discrimination of mild cognitive impairment and probable Alzheimer's disease in older Spanish monolingual individuals," *https://doi.org/10.1080/13825585.2019.1698710*, 2019, doi: 10.1080/13825585.2019.1698710.

[73] M. K. Yeung and J. Lin, "Probing depression, schizophrenia, and other psychiatric disorders using fNIRS and the verbal fluency test: A systematic review and meta-analysis," *J Psychiatr Res*, vol. 140, pp. 416–435, Aug. 2021, doi: 10.1016/J.JPSYCHIRES.2021.06.015.

[74] L. M. Wright, M. De Marco, and A. Venneri, "Current Understanding of Verbal Fluency in Alzheimer's Disease: Evidence to Date," *Psychol Res Behav Manag*, vol. Volume 16, pp. 1691–1705, May 2023, doi: 10.2147/PRBM.S284645.

[75] J. Melin *et al.*, "Traceability and comparability through crosswalks with the NeuroMET Memory Metric," *Scientific Reports 2023 13:1*, vol. 13, no. 1, pp. 1–12, Mar. 2023, doi: 10.1038/s41598-023-32208-0.

[76] D. Kondziella *et al.*, "European Academy of Neurology guideline on the diagnosis of coma and other disorders of consciousness," *Eur J Neurol*, vol. 27, no. 5, pp. 741–756, May 2020, doi: 10.1111/ENE.14151.

[77] W. Duan, X. Chen, Y. J. Wang, W. Zhao, H. Yuan, and X. Lei, "Reproducibility of power spectrum, functional connectivity and network construction in resting-state EEG," *J Neurosci Methods*, vol. 348, p. 108985, Jan. 2021, doi: 10.1016/J.JNEUMETH.2020.108985.

[78] J. R. Almeida, L. B. Silva, I. Bos, P. J. Visser, and J. L. Oliveira, “A methodology for cohort harmonisation in multicentre clinical research,” *Inform Med Unlocked*, vol. 27, p. 100760, Jan. 2021, doi: 10.1016/J.IMU.2021.100760.

[79] P. A. Valdes-Sosa *et al.*, “The Cuban Human Brain Mapping Project, a young and middle age population-based EEG, MRI, and cognition dataset,” *Scientific Data 2021 8:1*, vol. 8, no. 1, pp. 1–12, Feb. 2021, doi: 10.1038/s41597-021-00829-7.

[80] T. E. Cope, R. S. Weil, E. Düzel, B. C. Dickerson, and J. B. Rowe, “Advances in neuroimaging to support translational medicine in dementia,” *J Neurol Neurosurg Psychiatry*, vol. 92, no. 3, pp. 263–270, Mar. 2021, doi: 10.1136/JNNP-2019-322402.

[81] P. Prado *et al.*, “Dementia ConnEEGtome: Towards multicentric harmonization of EEG connectivity in neurodegeneration,” *International Journal of Psychophysiology*, vol. 172, pp. 24–38, Feb. 2022, doi: 10.1016/J.IJPSYCHO.2021.12.008.

[82] S. Moguilner *et al.*, “Multi-feature computational framework for combined signatures of dementia in underrepresented settings,” *J Neural Eng*, vol. 19, no. 4, p. 046048, Aug. 2022, doi: 10.1088/1741-2552/AC87D0.

[83] F. Hu *et al.*, “Image harmonization: A review of statistical and deep learning methods for removing batch effects and evaluation metrics for effective harmonization,” *Neuroimage*, vol. 274, p. 120125, Jul. 2023, doi: 10.1016/J.NEUROIMAGE.2023.120125.

[84] M. Bento, I. Fantini, J. Park, L. Rittner, and R. Frayne, “Deep Learning in Large and Multi-Site Structural Brain MR Imaging Datasets,” *Front Neuroinform*, vol. 15, p. 82, Jan. 2022, doi: 10.3389/FNINF.2021.805669/BIBTEX.

[85] M. Li *et al.*, “Harmonized-Multinational qEEG norms (HarMNqEEG),” *Neuroimage*, vol. 256, p. 119190, Aug. 2022, doi: 10.1016/J.NEUROIMAGE.2022.119190.

[86] G. Chiarion, L. Sparacino, Y. Antonacci, L. Faes, and L. Mesin, “Connectivity Analysis in EEG Data: A Tutorial Review of the State of the

Art and Emerging Trends," *Bioengineering 2023, Vol. 10, Page 372*, vol. 10, no. 3, p. 372, Mar. 2023, doi: 10.3390/BIOENGINEERING10030372.

[87] Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T. H. Falk, and J. Faubert, "Deep learning-based electroencephalography analysis: a systematic review," *J Neural Eng*, vol. 16, no. 5, p. 051001, Aug. 2019, doi: 10.1088/1741-2552/AB260C.

[88] A. S. Ballesteros, P. Prado, A. Ibanez, J. A. M. Perez, and S. Moguilner, "A pipeline for large-scale assessments of dementia EEG connectivity across multicentric settings," 2023, doi: 10.31219/OSF.IO/H2WGV.

[89] Y. Cherapanamjeri, S. Mohanty, and M. Yau, "List decodable mean estimation in nearly linear time," *Proceedings - Annual IEEE Symposium on Foundations of Computer Science, FOCS*, vol. 2020-November, pp. 141–148, Nov. 2020, doi: 10.1109/FOCS46700.2020.00022.

[90] A. Jaramillo-Jimenez *et al.*, "Spectral features of resting-state EEG in Parkinson's Disease: A multicenter study using functional data analysis," *Clinical Neurophysiology*, vol. 151, pp. 28–40, Jul. 2023, doi: 10.1016/j.clinph.2023.03.363.

[91] J. C. Beer *et al.*, "Longitudinal ComBat: A method for harmonizing longitudinal multi-scanner imaging data," *Neuroimage*, vol. 220, p. 117129, Oct. 2020, doi: 10.1016/J.NEUROIMAGE.2020.117129.

[92] W. E. Johnson, C. Li, and A. Rabinovic, "Adjusting batch effects in microarray expression data using empirical Bayes methods," *Biostatistics*, vol. 8, no. 1, pp. 118–127, Jan. 2007, doi: 10.1093/BIOSTATISTICS/KXJ037.

[93] J. P. Fortin *et al.*, "Harmonization of multi-site diffusion tensor imaging data," *Neuroimage*, vol. 161, pp. 149–170, Nov. 2017, doi: 10.1016/J.NEUROIMAGE.2017.08.047.

[94] J. P. Fortin *et al.*, "Harmonization of cortical thickness measurements across scanners and sites," *Neuroimage*, vol. 167, pp. 104–120, Feb. 2018, doi: 10.1016/J.NEUROIMAGE.2017.11.024.

[95] R. Pomponio *et al.*, "Harmonization of large multi-site imaging datasets: Application to 10,232 MRIs for the analysis of imaging patterns of structural

brain change throughout the lifespan," *bioRxiv*, p. 784363, Sep. 2019, doi: 10.1101/784363.

[96] J. A. Desjardins, S. van Noordt, S. Huberty, S. J. Segalowitz, and M. Elsabbagh, "EEG Integrated Platform Lossless (EEG-IP-L) pre-processing pipeline for objective signal quality assessment incorporating data annotation and blind source separation," *J Neurosci Methods*, vol. 347, p. 108961, Jan. 2021, doi: 10.1016/J.JNEUMETH.2020.108961.

[97] V. P. Kumaravel, E. Farella, E. Parise, and M. Buiatti, "NEAR: An artifact removal pipeline for human newborn EEG data," *Dev Cogn Neurosci*, vol. 54, p. 101068, Apr. 2022, doi: 10.1016/J.DCN.2022.101068.

[98] K. Kingphai and Y. Moshfeghi, "On EEG Preprocessing Role in Deep Learning Effectiveness for Mental Workload Classification," *Communications in Computer and Information Science*, vol. 1493 CCIS, pp. 81–98, 2021, doi: 10.1007/978-3-030-91408-0_6/TABLES/3.

[99] J. van Driel, C. N. L. Olivers, and J. J. Fahrenfort, "High-pass filtering artifacts in multivariate classification of neural time series data," *J Neurosci Methods*, vol. 352, p. 109080, Mar. 2021, doi: 10.1016/J.JNEUMETH.2021.109080.

[100] W. Peng, "EEG preprocessing and denoising," *EEG Signal Processing and Feature Extraction*, pp. 71–87, Jan. 2019, doi: 10.1007/978-981-13-9113-2_5/COVER.

[101] S. Pattisapu and S. Ray, "Stimulus-induced narrow-band gamma oscillations in humans can be recorded using open-hardware low-cost EEG amplifier," *PLoS One*, vol. 18, no. 1, p. e0279881, Jan. 2023, doi: 10.1371/JOURNAL.PONE.0279881.

[102] M. C. Guerrero, J. S. Parada, and H. E. Espitia, "EEG signal analysis using classification techniques: Logistic regression, artificial neural networks, support vector machines, and convolutional neural networks," *Heliyon*, vol. 7, no. 6, p. e07258, Jun. 2021, doi: 10.1016/J.HELIYON.2021.E07258.

[103] S. Guan *et al.*, "The Profiles of Non-stationarity and Non-linearity in the Time Series of Resting-State Brain Networks," *Front Neurosci*, vol. 14, p. 493, Jun. 2020, doi: 10.3389/FNINS.2020.00493/BIBTEX.

[104] T. Popov *et al.*, “Test–retest reliability of resting-state EEG in young and older adults,” *Psychophysiology*, vol. 00, p. e14268, Mar. 2023, doi: 10.1111/PSYP.14268.

[105] M. Grobbelaar *et al.*, “A Survey on Denoising Techniques of Electroencephalogram Signals Using Wavelet Transform,” *Signals 2022, Vol. 3, Pages 577-586*, vol. 3, no. 3, pp. 577–586, Aug. 2022, doi: 10.3390/SIGNALS3030035.

[106] A. Echtioui, W. Zouch, M. Ghorbel, M. Ben Slima, A. Ben Hamida, and C. Mhiri, “Automated EEG Artifact Detection Using Independent Component Analysis,” *2020 International Conference on Advanced Technologies for Signal and Image Processing, ATSIP 2020*, Sep. 2020, doi: 10.1109/ATSIP49331.2020.9231574.

[107] K. Yasoda, R. S. Ponmagal, K. S. Bhuvaneshwari, and K. Venkatachalam, “Automatic detection and classification of EEG artifacts using fuzzy kernel SVM and wavelet ICA (WICA),” *Soft comput*, vol. 24, no. 21, pp. 16011–16019, Nov. 2020, doi: 10.1007/S00500-020-04920-W/TABLES/3.

[108] S. A. Khoshnevis and R. Sankar, “Applications of Higher Order Statistics in Electroencephalography Signal Processing: A Comprehensive Survey,” *IEEE Rev Biomed Eng*, vol. 13, pp. 169–183, 2020, doi: 10.1109/RBME.2019.2951328.

[109] W. Deng, Y. Liu, J. Hu, and J. Guo, “The small sample size problem of ICA: A comparative study and analysis,” *Pattern Recognit*, vol. 45, no. 12, pp. 4438–4450, Dec. 2012, doi: 10.1016/J.PATCOG.2012.06.010.

[110] N. Mammone, F. La Foresta, and F. C. Morabito, “Automatic artifact rejection from multichannel scalp EEG by wavelet ICA,” *IEEE Sens J*, vol. 12, no. 3, pp. 533–542, 2012, doi: 10.1109/JSEN.2011.2115236.

[111] X. Jiang, G. Bin Bian, and Z. Tian, “Removal of Artifacts from EEG Signals: A Review,” *Sensors (Basel)*, vol. 19, no. 5, Mar. 2019, doi: 10.3390/S19050987.

[112] M. P. S. Chawla, “PCA and ICA processing methods for removal of artifacts and noise in electrocardiograms: A survey and comparison,” *Appl Soft*

Comput, vol. 11, no. 2, pp. 2216–2226, Mar. 2011, doi: 10.1016/J.ASOC.2010.08.001.

[113] L. Zhang, Z. Li, F. Zhang, R. Gu, W. Peng, and L. Hu, "Demystifying signal processing techniques to extract task- related EEG responses for psychologists," *Brain Science Advances*, vol. 6, no. 3, pp. 171–188, Sep. 2020, doi: 10.26599/BSA.2020.9050018.

[114] C. Chen, Z. Mei, and Z. Huang, "A clinical EEG research platform to support progressive model construction," *Proceedings - 2022 15th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics, CISP-BMEI 2022*, 2022, doi: 10.1109/CISP-BMEI56279.2022.9979832.

[115] K. Li *et al.*, "Feature Extraction and Identification of Alzheimer's Disease based on Latent Factor of Multi-Channel EEG," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 1557–1567, 2021, doi: 10.1109/TNSRE.2021.3101240.

[116] M. Ouchani, S. Gharibzadeh, M. Jamshidi, and M. Amini, "A Review of Methods of Diagnosis and Complexity Analysis of Alzheimer's Disease Using EEG Signals," *Biomed Res Int*, vol. 2021, 2021, doi: 10.1155/2021/5425569.

[117] R. Grech *et al.*, "Review on solving the inverse problem in EEG source analysis," *J Neuroeng Rehabil*, vol. 5, p. 25, 2008, doi: 10.1186/1743-0003-5-25.

[118] M. Gavaret, L. Maillard, and J. Jung, "High-resolution EEG (HR-EEG) and magnetoencephalography (MEG)," *Neurophysiologie Clinique/Clinical Neurophysiology*, vol. 45, no. 1, pp. 105–111, Mar. 2015, doi: 10.1016/J.NEUCLI.2014.11.011.

[119] F. Hasanzadeh, M. Mohebbi, and R. Rostami, "Graph theory analysis of directed functional brain networks in major depressive disorder based on EEG signal," *J Neural Eng*, vol. 17, no. 2, p. 026010, Mar. 2020, doi: 10.1088/1741-2552/AB7613.

[120] S. Abdulla, M. Diykh, R. L. Laft, K. Saleh, and R. C. Deo, "Sleep EEG signal analysis based on correlation graph similarity coupled with an ensemble

extreme machine learning algorithm," *Expert Syst Appl*, vol. 138, p. 112790, Dec. 2019, doi: 10.1016/J.ESWA.2019.07.007.

[121] P. M. Rossini *et al.*, "Early diagnosis of Alzheimer's disease: the role of biomarkers including advanced EEG signal analysis. Report from the IFCN-sponsored panel of experts," *Clinical Neurophysiology*, vol. 131, no. 6, pp. 1287–1310, Jun. 2020, doi: 10.1016/J.CLINPH.2020.03.003.

[122] L. Gabard-Durnam, W. Xie, J. Yedukondalu, and L. Dev Sharma, "Circulant Singular Spectrum Analysis and Discrete Wavelet Transform for Automated Removal of EOG Artifacts from EEG Signals," *Sensors 2023, Vol. 23, Page 1235*, vol. 23, no. 3, p. 1235, Jan. 2023, doi: 10.3390/S23031235.

[123] K. D. Tzimourta *et al.*, "Analysis of electroencephalographic signals complexity regarding Alzheimer's Disease," *Computers & Electrical Engineering*, vol. 76, pp. 198–212, Jun. 2019, doi: 10.1016/J.COMPELECENG.2019.03.018.

[124] N. I. Abbasi, R. Bose, A. Bezerianos, N. V. Thakor, and A. Dragomir, "EEG-Based Classification of Olfactory Response to Pleasant Stimuli," *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, pp. 5160–5163, Jul. 2019, doi: 10.1109/EMBC.2019.8857673.

[125] M. Mosayebi-Samani, A. Jamil, R. Salvador, G. Ruffini, J. Haueisen, and M. A. Nitsche, "The impact of individual electrical fields and anatomical factors on the neurophysiological outcomes of tDCS: A TMS-MEP and MRI study," *Brain Stimul*, vol. 14, no. 2, pp. 316–326, Mar. 2021, doi: 10.1016/J.BRS.2021.01.016.

[126] B. Y. Angulo-Ruiz, V. Muñoz, E. I. Rodríguez-Martínez, and C. M. Gómez, "Absolute and relative variability changes of the resting state brain rhythms from childhood and adolescence to young adulthood," *Neurosci Lett*, vol. 749, Apr. 2021, doi: 10.1016/J.NEULET.2021.135747.

[127] J. V. Candy, "Multichannel Spectral Estimation: An Approach to Estimating/Analyzing Vibrational Systems," Jan. 2020, doi: 10.2172/1592017.

[128] H. Bokil, P. Andrews, J. E. Kulkarni, S. Mehta, and P. P. Mitra, “Chronux: A platform for analyzing neural signals,” *J Neurosci Methods*, vol. 192, no. 1, pp. 146–151, Sep. 2010, doi: 10.1016/J.JNEUMETH.2010.06.020.

[129] H. Fort, “Entropy as missing information: from Shannon’s information theory to Jaynes’ maximum entropy principle,” *Forecasting with Maximum Entropy*, pp. 1-1-1–26, Nov. 2022, doi: 10.1088/978-0-7503-3931-5CH1.

[130] L. C. Amarantidis and D. Abásolo, “Interpretation of Entropy Algorithms in the Context of Biomedical Signal Analysis and Their Application to EEG Analysis in Epilepsy,” *Entropy 2019, Vol. 21, Page 840*, vol. 21, no. 9, p. 840, Aug. 2019, doi: 10.3390/E21090840.

[131] C. Pappalettera, F. Miraglia, M. Cotelli, P. M. Rossini, and F. Vecchio, “Analysis of complexity in the EEG activity of Parkinson’s disease patients by means of approximate entropy,” *Geroscience*, vol. 44, no. 3, pp. 1599–1607, Jun. 2022, doi: 10.1007/S11357-022-00552-0/FIGURES/2.

[132] A. C. Hull and J. B. Morton, “Activity-State Entropy: A novel brain entropy measure based on spatial patterns of activity,” *J Neurosci Methods*, vol. 393, p. 109868, Jun. 2023, doi: 10.1016/J.JNEUMETH.2023.109868.

[133] H. Ahmadieh and F. Ghassemi, “Assessing the Effects of Alzheimer Disease on EEG Signals Using the Entropy Measure: A Meta-analysis,” *Basic Clin Neurosci*, vol. 13, no. 2, p. 153, Mar. 2022, doi: 10.32598/BCN.2021.1144.3.

[134] J. A. Thiele, A. Richter, and K. Hilger, “Multimodal Brain Signal Complexity Predicts Human Intelligence,” *eNeuro*, vol. 10, no. 2, Feb. 2023, doi: 10.1523/ENEURO.0345-22.2022.

[135] O. Sporns, G. Tononi, and G. M. Edelman, “Connectivity and complexity: the relationship between neuroanatomy and brain dynamics,” *Neural Networks*, vol. 13, no. 8–9, pp. 909–922, Nov. 2000, doi: 10.1016/S0893-6080(00)00053-8.

[136] C. H. Chang, T. Furukawa, T. Asahina, K. Shimba, K. Kotani, and Y. Jimbo, “Coupling of in vitro Neocortical-Hippocampal Coculture Bursts Induces Different Spike Rhythms in Individual Networks,” *Front Neurosci*, vol. 16, p. 663, May 2022, doi: 10.3389/FNINS.2022.873664/BIBTEX.

[137] S. L. Yan, X. L. Yang, H. Yang, and Z. K. Sun, “Decreased coherence in the model of the dorsal visual pathway associated with Alzheimer’s disease,” *Scientific Reports 2023 13:1*, vol. 13, no. 1, pp. 1–13, Mar. 2023, doi: 10.1038/s41598-023-30535-w.

[138] H. Ono *et al.*, “Dynamic cortical and tractography atlases of proactive and reactive alpha and high-gamma activities,” *Brain Commun*, vol. 5, no. 2, Mar. 2023, doi: 10.1093/BRAINCOMMS/FCAD111.

[139] L. Rürup *et al.*, “Altered gamma and theta oscillations during multistable perception in schizophrenia,” *International Journal of Psychophysiology*, vol. 155, pp. 127–139, Sep. 2020, doi: 10.1016/J.IJPSYCHO.2020.06.002.

[140] W. A. Huang *et al.*, “Transcranial alternating current stimulation entrains alpha oscillations by preferential phase synchronization of fast-spiking cortical neurons to stimulation waveform,” *Nature Communications 2021 12:1*, vol. 12, no. 1, pp. 1–20, May 2021, doi: 10.1038/s41467-021-23021-2.

[141] Y. Zhong *et al.*, “A review on pathology, mechanism, and therapy for cerebellum and tremor in Parkinson’s disease,” *npj Parkinson’s Disease 2022 8:1*, vol. 8, no. 1, pp. 1–9, Jun. 2022, doi: 10.1038/s41531-022-00347-2.

[142] F. S. Racz, A. Czoch, Z. Kaposzta, O. Stylianou, P. Mukli, and A. Eke, “Multiple-Resampling Cross-Spectral Analysis: An Unbiased Tool for Estimating Fractal Connectivity With an Application to Neurophysiological Signals,” *Front Physiol*, vol. 13, p. 132, Mar. 2022, doi: 10.3389/FPHYS.2022.817239/BIBTEX.

[143] J. F. Morici, N. V. Weisstaub, and C. L. Zold, “Hippocampal-medial prefrontal cortex network dynamics predict performance during retrieval in a context-guided object memory task,” *Proc Natl Acad Sci U S A*, vol. 119, no. 20, p. e2203024119, May 2022, doi: 10.1073/PNAS.2203024119/SUPPL_FILE/PNAS.2203024119.SAPP.PDF.

[144] J. M. Cassidy, A. Wodeyar, R. Srinivasan, and S. C. Cramer, “Coherent neural oscillations inform early stroke motor recovery,” *Hum Brain Mapp*, vol. 42, no. 17, pp. 5636–5647, Dec. 2021, doi: 10.1002/HBM.25643.

[145] Y. Qin, T. Menara, D. S. Bassett, and F. Pasqualetti, “Phase-amplitude coupling in neuronal oscillator networks,” *Phys Rev Res*, vol. 3, no. 2, p. 023218, Jun. 2021, doi: 10.1103/PHYSREVRESEARCH.3.023218/FIGURES/5/MEDIUM.

[146] J. Riddle, A. McFerren, and F. Frohlich, “Causal role of cross-frequency coupling in distinct components of cognitive control,” *Prog Neurobiol*, vol. 202, p. 102033, Jul. 2021, doi: 10.1016/J.PNEUROBIO.2021.102033.

[147] T. H. Falk, F. J. Fraga, L. Trambaiolli, and R. Anghinah, “EEG amplitude modulation analysis for semi-automated diagnosis of Alzheimer’s disease,” *EURASIP J Adv Signal Process*, vol. 2012, no. 1, pp. 1–9, Aug. 2012, doi: 10.1186/1687-6180-2012-192/COMMENTS.

[148] B. Boashash, “Time frequency signal analysis and processing : a comprehensive reference,” 2015.

[149] Z. M. Smith, B. Delgutte, and A. J. Oxenham, “Chimaeric sounds reveal dichotomies in auditory perception,” *Nature*, vol. 416, no. 6876, p. 87, Mar. 2002, doi: 10.1038/416087A.

[150] M. Wischnewski, I. Alekseichuk, and A. Opitz, “Neurocognitive, physiological, and biophysical effects of transcranial alternating current stimulation,” *Trends Cogn Sci*, vol. 27, no. 2, pp. 189–205, Feb. 2023, doi: 10.1016/J.TICS.2022.11.013.

[151] E. Weiss, M. Kann, and Q. Wang, “Neuromodulation of Neural Oscillations in Health and Disease,” *Biology 2023, Vol. 12, Page 371*, vol. 12, no. 3, p. 371, Feb. 2023, doi: 10.3390/BIOLOGY12030371.

[152] B. Lega, J. Burke, J. Jacobs, and M. J. Kahana, “Slow-Theta-to-Gamma Phase–Amplitude Coupling in Human Hippocampus Supports the Formation of New Episodic Memories,” *Cerebral Cortex*, vol. 26, no. 1, pp. 268–278, Jan. 2016, doi: 10.1093/CERCOR/BHU232.

[153] J. Feingold, D. J. Gibson, B. Depasquale, and A. M. Graybiel, “Bursts of beta oscillation differentiate postperformance activity in the striatum and motor cortex of monkeys performing movement tasks,” *Proc Natl Acad Sci U S A*, vol. 112, no. 44, pp. 13687–13692, Nov. 2015, doi: 10.1073/PNAS.1517629112/SUPPL_FILE/PNAS.201517629SI.PDF.

[154] D. Klepl, F. He, W. Min, D. Blackburn, and P. Sarrigiannis, “Bispectrum-based Cross-frequency Functional Connectivity: Classification of Alzheimer’s disease,” *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, vol. 2022-July, pp. 305–308, 2022, doi: 10.1109/EMBC48229.2022.9871366.

[155] T. Montez, K. Linkenkaer-Hansen, B. W. van Dijk, and C. J. Stam, “Synchronization likelihood with explicit time-frequency priors,” *Neuroimage*, vol. 33, no. 4, pp. 1117–1125, Dec. 2006, doi: 10.1016/J.NEUROIMAGE.2006.06.066.

[156] H. Yu, J. Liu, L. Cai, J. Wang, Y. Cao, and C. Hao, “Functional brain networks in healthy subjects under acupuncture stimulation: An EEG study based on nonlinear synchronization likelihood analysis,” *Physica A: Statistical Mechanics and its Applications*, vol. 468, pp. 566–577, Feb. 2017, doi: 10.1016/J.PHYSA.2016.10.068.

[157] I. V. Stuldreher, N. Thammasan, J. B. F. Van Erp, and A. M. Brouwer, “Physiological synchrony in EEG, electrodermal activity and heart rate reflects shared selective auditory attention,” *J Neural Eng*, vol. 17, no. 4, p. 046028, Aug. 2020, doi: 10.1088/1741-2552/ABA87D.

[158] T. Wu, X. Zhang, and Z. Liu, “Understanding the mechanisms of brain functions from the angle of synchronization and complex network,” *Frontiers of Physics 2022 17:3*, vol. 17, no. 3, pp. 1–23, Apr. 2022, doi: 10.1007/S11467-022-1161-6.

[159] D. Posthuma, E. J. C. De Geus, E. J. C. M. Mulder, D. J. A. Smit, D. I. Boomsma, and C. J. Stam, “Genetic components of functional connectivity in the brain: The heritability of synchronization likelihood,” *Hum Brain Mapp*, vol. 26, no. 3, p. 191, Nov. 2005, doi: 10.1002/HBM.20156.

[160] Y. K. Kim, E. Park, A. Lee, C. H. Im, and Y. H. Kim, “Changes in network connectivity during motor imagery and execution,” *PLoS One*, vol. 13, no. 1, Jan. 2018, doi: 10.1371/JOURNAL.PONE.0190715.

[161] S. Nobukawa, T. Yamanishi, S. Kasakawa, H. Nishimura, M. Kikuchi, and T. Takahashi, “Classification Methods Based on Complexity and Synchronization of Electroencephalography Signals in Alzheimer’s

Disease," *Front Psychiatry*, vol. 11, p. 255, Apr. 2020, doi: 10.3389/FPSYT.2020.00255/BIBTEX.

[162] L. Billeci, A. Badolato, L. Bachi, and A. Tonacci, "Machine Learning for the Classification of Alzheimer's Disease and Its Prodromal Stage Using Brain Diffusion Tensor Imaging Data: A Systematic Review," *Processes 2020, Vol. 8, Page 1071*, vol. 8, no. 9, p. 1071, Sep. 2020, doi: 10.3390/PR8091071.

[163] J. Ray, L. Wijesekera, and S. Cirstea, "Machine learning and clinical neurophysiology," *J Neurol*, vol. 269, no. 12, pp. 6678–6684, Dec. 2022, doi: 10.1007/S00415-022-11283-9/TABLES/1.

[164] R. S. Olson, O. Edu, and J. H. Moore, "TPOT: A Tree-based Pipeline Optimization Tool for Automating Machine Learning," vol. 64. PMLR, pp. 66–74, Dec. 04, 2016. Accessed: May 14, 2023. [Online]. Available: https://proceedings.mlr.press/v64/olson_tpot_2016.html

[165] N. Pilnenskiy and I. Smetannikov, "Feature Selection Algorithms as One of the Python Data Analytical Tools," *Future Internet 2020, Vol. 12, Page 54*, vol. 12, no. 3, p. 54, Mar. 2020, doi: 10.3390/FI12030054.

[166] Olson Randal, "Using TPOT - TPOT," *University of Pennsylvania*, 2021. http://epistasislab.github.io/tpot/using/ (accessed May 14, 2023).

[167] D'Agostino Andrea, "Feature Selection with Boruta in Python | by Andrea D'Agostino | Towards Data Science," 2021. https://towardsdatascience.com/feature-selection-with-boruta-in-python-676e3877e596 (accessed May 14, 2023).

[168] R. J. Urbanowicz, M. Meeker, W. La Cava, R. S. Olson, and J. H. Moore, "Relief-based feature selection: Introduction and review," *J Biomed Inform*, vol. 85, pp. 189–203, Sep. 2018, doi: 10.1016/J.JBI.2018.07.014.

[169] M. A. Azhar and P. A. Thomas, "Comparative Review of Feature Selection and Classification modeling," *2019 6th IEEE International Conference on Advances in Computing, Communication and Control, ICAC3 2019*, Dec. 2019, doi: 10.1109/ICAC347590.2019.9036816.

[170] C. J. Markiewicz *et al.*, "The openneuro resource for sharing of neuroscience data," *Elife*, vol. 10, Oct. 2021, doi: 10.7554/ELIFE.71774.

[171] R. Rajora, A. Kumar, S. Malhotra, and A. Sharma, “Data security breaches and mitigating methods in the healthcare system: A review,” *Proceedings - 2022 International Conference on Computational Modelling, Simulation and Optimization, ICCMSO 2022*, pp. 325–330, 2022, doi: 10.1109/ICCMSO58359.2022.00070.

[172] B. Maharathi *et al.*, “Multi-modal data integration platform combining clinical and preclinical models of post subarachnoid hemorrhage epilepsy,” *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, vol. 2022-July, pp. 3459–3463, 2022, doi: 10.1109/EMBC48229.2022.9871864.

[173] H. Altaheri *et al.*, “Deep learning techniques for classification of electroencephalogram (EEG) motor imagery (MI) signals: a review,” *Neural Computing and Applications 2021*, pp. 1–42, Aug. 2021, doi: 10.1007/S00521-021-06352-5.

[174] G. Niso *et al.*, “Open and reproducible neuroimaging: From study inception to publication,” *Neuroimage*, vol. 263, p. 119623, Nov. 2022, doi: 10.1016/J.NEUROIMAGE.2022.119623.

[175] J. Bosch-Bayard, L. Galan, E. Aubert Vazquez, T. Virues Alba, and P. A. Valdes-Sosa, “Resting State Healthy EEG: The First Wave of the Cuban Normative Database,” *Front Neurosci*, vol. 14, p. 555119, Dec. 2020, doi: 10.3389/FNINS.2020.555119.

[176] S. Van Noordt *et al.*, “EEG-IP: An international infant EEG data integration platform for the study of risk and resilience in autism and related conditions,” *Molecular Medicine*, vol. 26, no. 1, pp. 1–11, May 2020, doi: 10.1186/S10020-020-00149-3/FIGURES/5.

[177] L. J. Gabard-Durnam, A. S. M. Leal, C. L. Wilkinson, and A. R. Levin, “The harvard automated processing pipeline for electroencephalography (HAPPE): Standardized processing software for developmental and high-artifact data,” *Front Neurosci*, vol. 12, p. 97, Feb. 2018, doi: 10.3389/FNINS.2018.00097/BIBTEX.

[178] B. Daniel, L. Tim, S. Øyvind, and B. Jochen, “EEG-derived brain graphs are reliable measures for exploring exercise-induced changes in brain

networks," *Scientific Reports 2021 11:1*, vol. 11, no. 1, pp. 1–13, Oct. 2021, doi: 10.1038/s41598-021-00371-x.

[179] Y.-J. Mantilla-Ramos, "sovabids v0.3.1-alpha+21.gf251ba5.dirty documentation," 2021. https://sovabids.readthedocs.io/en/latest/ (accessed May 15, 2023).

[180] G. Niso *et al.*, "Good scientific practice in EEG and MEG research: Progress and perspectives," *Neuroimage*, vol. 257, p. 119056, Aug. 2022, doi: 10.1016/J.NEUROIMAGE.2022.119056.

[181] L. Waller *et al.*, "ENIGMA HALFpipe: Interactive, reproducible, and efficient analysis for resting-state and task-based fMRI data," *Hum Brain Mapp*, vol. 43, no. 9, pp. 2727–2742, Jun. 2022, doi: 10.1002/HBM.25829.

[182] J. Suarez-Revelo, J. Ochoa-Gomez, and J. Duque-Grajales, "Improving test-retest reliability of quantitative electroencephalography using different preprocessing approaches," *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, vol. 2016-October, pp. 961–964, Oct. 2016, doi: 10.1109/EMBC.2016.7590861.

[183] J. X. Suárez-Revelo, J. F. Ochoa-Gómez, and C. A. Tobón-Quintero, "Validation of EEG Pre-processing Pipeline by Test-Retest Reliability," *Communications in Computer and Information Science*, vol. 916, pp. 290–299, 2018, doi: 10.1007/978-3-030-00353-1_26/FIGURES/3.

[184] F. J. García-Pretelt, J. X. Suárez-Relevo, D. F. Aguillon-Niño, F. J. Lopera-Restrepo, J. F. Ochoa-Gómez, and C. A. Tobón-Quintero, "Automatic Classification of Subjects of the PSEN1-E280A Family at Risk of Developing Alzheimer's Disease Using Machine Learning and Resting State Electroencephalography," *Journal of Alzheimer's Disease*, vol. 87, no. 2, pp. 817–832, Jan. 2022, doi: 10.3233/JAD-210148.

[185] J. F. Ochoa *et al.*, "Successful Object Encoding Induces Increased Directed Connectivity in Presymptomatic Early-Onset Alzheimer's Disease," *Journal of Alzheimer's Disease*, vol. 55, no. 3, p. 1195, 2017, doi: 10.3233/JAD-160803.

[186] J. F. Ochoa *et al.*, “Precuneus Failures in Subjects of the PSEN1 E280A Family at Risk of Developing Alzheimer’s Disease Detected Using Quantitative Electroencephalography,” *J Alzheimers Dis*, vol. 58, no. 4, pp. 1229–1244, 2017, doi: 10.3233/JAD-161291.

[187] M. L. Manaog and L. Parisi, “M-ar-K-PCA and M-ar-K-FastICA: Robust Feature Extraction for Classification of Non-Gaussian and Entropic Data,” Jun. 2022, doi: 10.21203/RS.3.RS-1560908/V1.

[188] M. Wang, X. Cui, T. Wang, T. Jiang, F. Gao, and J. Cao, “Eye blink artifact detection based on multi-dimensional EEG feature fusion and optimization,” *Biomed Signal Process Control*, vol. 83, p. 104657, May 2023, doi: 10.1016/J.BSPC.2023.104657.

[189] D. E. Ho *et al.*, “Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference,” *Political Analysis*, vol. 15, no. 3, pp. 199–236, Jun. 2007, doi: 10.1093/PAN/MPL013.

[190] J. Raffo, Raffo, and Julio, “MATCHIT: Stata module to match two datasets based on similar text patterns,” May 2020, Accessed: May 15, 2023. [Online]. Available: https://EconPapers.repec.org/RePEc:boc:bocode:s457992

[191] M. Reynolds, T. Chaudhary, M. E. Torbati, D. L. Tudorascu, K. Batmanghelich, and the A. D. N. Initiative, “ComBat Harmonization: Empirical Bayes versus Fully Bayes Approaches,” *bioRxiv*, p. 2022.07.13.499561, Jul. 2022, doi: 10.1101/2022.07.13.499561.

[192] P. A. Valdes-Sosa *et al.*, “The Cuban Human Brain Mapping Project, a young and middle age population-based EEG, MRI, and cognition dataset,” *Scientific Data 2021 8:1*, vol. 8, no. 1, pp. 1–12, Feb. 2021, doi: 10.1038/s41597-021-00829-7.

[193] Zapata Luisa, “Desarrollo de aplicación de servicios web basado en estándares de informática médica para el preprocesamiento y visualización de registros EEG,” 2022, Accessed: May 20, 2023. [Online]. Available: https://bibliotecadigital.udea.edu.co/handle/10495/30073

[194] N. K. Dinsdale, M. Jenkinson, and A. I. L. Namburete, “Deep learning-based unlearning of dataset bias for MRI harmonisation and confound removal,”

Neuroimage, vol. 228, p. 117689, Mar. 2021, doi: 10.1016/J.NEUROIMAGE.2020.117689.

[195] K. D. Tzimourta *et al.*, "Machine Learning Algorithms and Statistical Approaches for Alzheimer's Disease Analysis Based on Resting-State EEG Recordings: A Systematic Review," *https://doi.org/10.1142/S0129065721300023*, vol. 31, no. 5, Feb. 2021, doi: 10.1142/S0129065721300023.

[196] G. Chen *et al.*, "Classification of Alzheimer disease, mild cognitive impairment, and normal cognitive status with large-scale network analysis based on resting-state functional MR imaging," *Radiology*, vol. 259, no. 1, pp. 213–221, Apr. 2011, doi: 10.1148/RADIOL.10100734.

[197] S. F. Eskildsen, P. Coupé, D. García-Lorenzo, V. Fonov, J. C. Pruessner, and D. L. Collins, "Prediction of Alzheimer's disease in subjects with mild cognitive impairment from the ADNI cohort using patterns of cortical thinning," *Neuroimage*, vol. 65, pp. 511–521, Jan. 2013, doi: 10.1016/j.neuroimage.2012.09.058.

[198] M. Velazquez, Y. Lee, and for the A. D. N. Initiative, "Random forest model for feature-based Alzheimer's disease conversion prediction from early mild cognitive impairment subjects," *PLoS One*, vol. 16, no. 4, Apr. 2021, doi: 10.1371/JOURNAL.PONE.0244773.

[199] T. Jo, K. Nho, and A. J. Saykin, "Deep Learning in Alzheimer's Disease: Diagnostic Classification and Prognostic Prediction Using Neuroimaging Data," *Front Aging Neurosci*, vol. 11, p. 220, Aug. 2019, doi: 10.3389/FNAGI.2019.00220/FULL.

[200] G. W. Cha, H. J. Moon, and Y. C. Kim, "Comparison of Random Forest and Gradient Boosting Machine Models for Predicting Demolition Waste Based on Small Datasets and Categorical Variables," *International Journal of Environmental Research and Public Health 2021, Vol. 18, Page 8530*, vol. 18, no. 16, p. 8530, Aug. 2021, doi: 10.3390/IJERPH18168530.

[201] E. Moradi, A. Pepe, C. Gaser, H. Huttunen, and J. Tohka, "Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects," *Neuroimage*, vol. 104, pp. 398–412, Jan. 2015, doi: 10.1016/J.NEUROIMAGE.2014.10.002.

[202] J. Schrouff *et al.*, “PRoNTo: pattern recognition for neuroimaging toolbox,” *Neuroinformatics*, vol. 11, no. 3, pp. 319–337, Jul. 2013, doi: 10.1007/S12021-013-9178-1.

[203] M. Jin and W. Deng, “Predication of different stages of Alzheimer’s disease using neighborhood component analysis and ensemble decision tree,” *J Neurosci Methods*, vol. 302, pp. 35–41, May 2018, doi: 10.1016/J.JNEUMETH.2018.02.014.

[204] M. V. F. Silva, C. D. M. G. Loures, L. C. V. Alves, L. C. De Souza, K. B. G. Borges, and M. D. G. Carvalho, “Alzheimer’s disease: risk factors and potentially protective measures,” *J Biomed Sci*, vol. 26, no. 1, May 2019, doi: 10.1186/S12929-019-0524-Y.

[205] A. Biffi *et al.*, “Genetic variation and neuroimaging measures in Alzheimer disease,” *Arch Neurol*, vol. 67, no. 6, pp. 677–685, Jun. 2010, doi: 10.1001/ARCHNEUROL.2010.108.

[206] H. Stocker, T. Möllers, L. Perna, and H. Brenner, “The genetic risk of Alzheimer’s disease beyond APOE ε4: systematic review of Alzheimer’s genetic risk scores,” *Translational Psychiatry 2018 8:1*, vol. 8, no. 1, pp. 1–9, Aug. 2018, doi: 10.1038/s41398-018-0221-8.

[207] J. S. Yokoyama *et al.*, “Decision tree analysis of genetic risk for clinically heterogeneous Alzheimer’s disease,” *BMC Neurol*, vol. 15, no. 1, Mar. 2015, doi: 10.1186/S12883-015-0304-6.

[208] K. Gupta, N. Jiwani, and P. Whig, “An Efficient Way of Identifying Alzheimer’s Disease Using Deep Learning Techniques,” *Lecture Notes in Networks and Systems*, vol. 479, pp. 455–465, 2023, doi: 10.1007/978-981-19-3148-2_38/COVER.

[209] M. S. Ali, M. K. Islam, J. Haque, A. A. Das, D. S. Duranta, and M. A. Islam, “Alzheimer’s Disease Detection Using m-Random Forest Algorithm with Optimum Features Extraction,” *2021 1st International Conference on Artificial Intelligence and Data Analytics, CAIDA 2021*, pp. 1–6, Apr. 2021, doi: 10.1109/CAIDA51941.2021.9425212.

[210] A. Miltiadous *et al.*, “Alzheimer’s Disease and Frontotemporal Dementia: A Robust Classification Method of EEG Signals and a Comparison of

Validation Methods," *Diagnostics 2021, Vol. 11, Page 1437*, vol. 11, no. 8, p. 1437, Aug. 2021, doi: 10.3390/DIAGNOSTICS11081437.

[211] H. Javaid, R. Manor, E. Kumarnsit, and S. Chatpun, "Decision Tree in Working Memory Task Effectively Characterizes EEG Signals in Healthy Aging Adults," *IRBM*, vol. 43, no. 6, pp. 705–714, Dec. 2022, doi: 10.1016/J.IRBM.2021.12.001.

[212] D. Oreski, S. Oreski, and B. Klicek, "Effects of dataset characteristics on the performance of feature selection techniques," *Appl Soft Comput*, vol. 52, pp. 109–119, Mar. 2017, doi: 10.1016/J.ASOC.2016.12.023.

[213] S. Fouladi, A. A. Safaei, N. Mammone, F. Ghaderi, and M. J. Ebadi, "Efficient Deep Neural Networks for Classification of Alzheimer's Disease and Mild Cognitive Impairment from Scalp EEG Recordings," *Cognitive Computation 2022 14:4*, vol. 14, no. 4, pp. 1247–1268, Jun. 2022, doi: 10.1007/S12559-022-10033-3.

[214] S. Sharma and P. K. Mandal, "A Comprehensive Report on Machine Learning-based Early Detection of Alzheimer's Disease using Multi-modal Neuroimaging Data," *ACM Comput Surv*, vol. 55, no. 2, Mar. 2023, doi: 10.1145/3492865.

[215] A. Katako *et al.*, "Machine learning identified an Alzheimer's disease-related FDG-PET pattern which is also expressed in Lewy body dementia and Parkinson's disease dementia," *Sci Rep*, vol. 8, no. 1, Dec. 2018, doi: 10.1038/S41598-018-31653-6.

[216] Henao Isaza V, Mantilla-Ramos Y, Cadavid Castro V, and Zapata Saldarriaga L, "GRUNECO/eeg_harmonization," 2021. https://github.com/GRUNECO/eeg_harmonization (accessed May 15, 2023).

Annexes

Annex 1: Procedure BIDS

To obtain the BIDS format, the open-source tool sovabids was used, sovabids is a python package for automating eeg2bids conversion [179], and sovabids can be used through:

1. Its python API
2. Its CLI entry points
3. Its JSON-RPC entry points (needs a server running the backend)
4. Its minimal web-app GUI

To understand the structure of the tool, the step called sovabids can be taken from Figure 57 of the methodology.

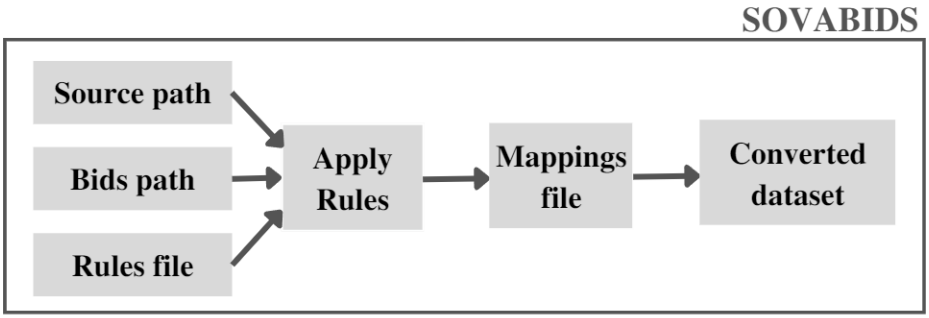


Figure 57 Sovabids methodology

The basic architecture is:

1. A source path with the original dataset.
2. A bids path that will be the output path of the conversion.
3. A rules file that configures how the conversion is done from the general perspective.
4. A mapping file that encodes how the conversion is performed to each individual file of the dataset.
5. Download and relocate the database.

Note: All subjects must be in the same folder

6. Evaluate the availability and distribution of the information and the format (Verify if it is in BIDS format or if you have the necessary information to convert to BIDS).

Test BIDS

Allows you to identify if a database is in BIDS format. To perform this test, go to the following GitHub repository:

Sovaharmony unpublished package [216], and in the misc folder, run the pybids_test.py file.

1. If the database is in BIDS format, the result of running the code would be:

BIDS format OK

2. If the database is not in BIDS format, the result of running the code would be:

Try to convert of Dataset in BIDS format using "conversion_bids"

Conversion

1. The first thing to do is create a rules file.

Take as a base a .yml file like the one presented in the Github repository and edit each parameter as explained in the file's documentation:

- task: refers to the type of task or condition.
- Name: name of the database
- Authors: Database authors
- PowerLineFrequency: The spectrum graph is made to show the frequency by visual inspection.
- EEGReference: Reference channel
- Channels: Channel information
- eeg_extension: EEG file extension
- pattern: structure of the EEG file name

If the file has a complex structure due to the number of tasks, conditions, sessions, etc. You need to create an additional parameter Figure 58:

- fields: specifies each of the elements that differentiate the records within the database.

Example:

```
path_analysis:
  pattern : database/%ignore%_%entities.subject%_%entities.session%_%entities.task%.cnt
  #pattern : database/(.+)_(.+)_(.+)_(.+).cnt
  #fields :
  #  - ignore
  #  - entities.subject
  #  - entities.session
  #  - entities.task
```

Figure 58 Pattern Rules File Example

For the design and creation of the rules file, there are useful tools such as https://regex101.com/ which is a regular expression tester with syntax highlighting, explanation, cheat sheet for PHP/PCRE, Python, GO, JavaScript, Java, C#/.NET, Rust.

2. Open the file conversion_bids.py edit the variables; **source_path, bids_path** and **rules_path.**
3. Run the file conversion_bids.py

****For a step-by-step implementation of sovabids see Annex 1 or the direct source of sovathe bids****

Rules

The databases taken from websites already have the BIDS structure implemented, so it was only necessary to apply the standard to the initial databases.

The conversion process was performed separately for each database, and the resulting rule files are displayed below.

Figure 59 show database description along with the baseline and reference electrode of the assembly. To identify the before channels, a channel standard was published for BIDS which includes a restricted keyword list for field type channels, shared with MEG and iEEG modalities [27]. In this case, the VEO and HEO types were marked. Additionally, the characteristics of the non-BIDS database, such as the data extension and file name, are specified. with the "Entities" in the file name represent relevant elements such as session, subject group, homework, etc. Currently, there are tools available to automatically generate regex as you type.

```
dataset_description:                        # Configuring the dataset_description.json file
  Name : UdeA 1                           # Name of the dataset, set up as a fixed string
  Authors:                                  # Here I put the personnel involved in the acquisition of the dataset
    - Gruneco

sidecar:                                    # Configuring the sidecar eeg file
  PowerLineFrequency : 60                   # Noted from the visual inspection of the eeg spectrum
  EEGReference : Mastoide derecha           # As mentioned in https://www.nature.com/articles/sdata2018308

channels:                                   # Configuring the channels tsv
  type :                                    # This property allow us to overwrite channel types inferred by MNE
    VEO : VEOG                              # Here the syntax is <channel name> : <channel type according to bids notation>
    HEO : HEOG                              # Here we set the type of F3, it was already correctly inferred by mne but it is
                                            #included to illustrate retyping of various channels.
non-bids:                                   # Additional configuration not belonging specifically to any of the previous objects
  eeg_extension : .cnt                      # Sets which extension to read as an eeg file
  path_analysis:                            # Some bids properties can be inferred from the path of the source files
    # For example here we extract from the path the "subject" child of the "entities" object
    pattern : Codificado_EEG_V0_V1_V2_V3\/(.+)\/(.+_.{3})_(.+)_(.+).cnt
    fields :
      - entities.session
      - entities.subject
      - ignore
      - entities.task
```

Figure 59 Rules File Example of UdeA 1

The second database (Figure 60) that was converted to BIDS has a file structure that is quite like the first. The authors, baseline, reference electrode, and non-EEG channels have been preserved. However, in this case, the file names represent tasks independently.

It is important to note that both methods of creating the rules file are valid, and they generate standardized and compatible files with respect to the groups, tasks, and file types that will be generated and used later in the processing pipeline.

```
dataset_description:                        # Configuring the dataset_description.json file
  Name : UdeA 2                           # Name of the dataset, set up as a fixed string
  Authors:                                  # Here I put the personnel involved in the acquisition of the dataset
    - Gruneco

sidecar:                                    # Configuring the sidecar eeg file
  PowerLineFrequency : 60                   # Noted from the visual inspection of the eeg spectrum
  EEGReference : Mastoide derecha           # As mentioned in https://www.nature.com/articles/sdata2018308

channels:                                   # Configuring the channels tsv
  type :                                    # This property allow us to overwrite channel types inferred by MNE
    VEO : VEOG                              # Here the syntax is <channel name> : <channel type according to bids notation>
    HEO : HEOG                              # Here we set the type of F3, it was already correctly inferred by mne but it is
                                            #included to illustrate retyping of various channels.
non-bids:                                   # Additional configuration not belonging specifically to any of the previous objects
  eeg_extension : .cnt                      # Sets which extension to read as an eeg file
  path_analysis:                            # Some bids properties can be inferred from the path of the source files
    pattern : NEW/%entities.subject%_%ignore%.cnt # ALZCE001_RES, DFT003_EEG

entities:                                   # Configuring the file name structure of bids
  task : resting                            # Setting the task of all files to a fixed string
```

Figure 60 Rules File Example of UdeA 2

Annex 2: Optimizing procedure of the Processing Pipeline

Installation of sova package

1. Open a command console and run the following command lines, see Figure 61.

```
# Installation package sovaflow
pip install git+https: // gitfront.io/r/GRUNECO/xiGpXFpQvM2T/sovaflow.git
# Installation package sovareject
pip install git+https: //gitfront.io/r/yjmantilla/5c5817890b14af2e5ae8ae9ba3f14522c337522e/sovareject.git
# Installation package sovachronux
pip install git+https: //gitfront.io/r/GRUNECO/5wAdEXRh7oTf/sovachronux.git
# Installation package eeg_harmonization
git clone git+https: https://github.com/GRUNECO/eeg_harmonization.git
```

Figure 61 command for installation of packages

2. Enter the location of the installed package and execute the command line that allows you to install the libraries necessary for the execution of the code, see Figure 62.

```
cd eeg_harmonization
pip install -r requirements-install.txt
```

Figure 62 command necessary for the execution of the installation code.

3. When aiming to synchronize databases obtained from various repositories, collected by different devices, with varying sampling frequencies and channels, it is essential to utilize the Sovaharmony package for processing. Alternatively, if only the processing aspect is required, the sovapipeline package can be used. Nevertheless, it is highly recommended to employ the Sovaharmony package as it provides comprehensive pre-processing and post-processing functionalities.

Table 31 package list

Resource	Description	Location on GitHub
Sovachronux (private)	It is inspired by the MATLAB Chronux tool that allows loading, visualization and analysis of neurobiological time series data such as EEG. This includes spectral analysis with the multitaper technique.	https://github.com/ GRUNECO/ sovachronux/ tree/ main/ sovachronux
Sovawica (private)	It is inspired by thresh, a MATLAB wavelet function by M. Misiti, Y. Misiti, G. Oppenheim, J.M. Poggi 12-Mar-96.	https://github.com/ GRUNECO/ sovawica
Sovareject (private)	Artifact rejection routine with parameterized and automatic thresholds.	https://github.com/ GRUNECO/ sovareject
Sovaview (private)	Software for EEG signal visualization	https://github.com/ GRUNECO/ sovaview
Sovapipeline (private)	EEG signal preprocessing at rest. It integrates the stage of	https://github.com /GRUNECO/ sovapipeline

	PREP, wICA and the rejection of times.	
Sovaharmony (public)	Integrates sova packages for use in multiple cohorts	https://github.com/ GRUNECO/ eeg_harmonization/ tree/ main /sovaharmony
neuroharmonaze.py (public)	Matching and implement the neuroHarmonize library	https://github.com/ GRUNECO/ eeg_harmonization/ blob/ main/ misc/ neuroharmonaze.py
Data_analysis_ ML_Harmonization_ Proyect (public)	Descriptive analysis and model implementation	https://github.com/ GRUNECO/ Data_analysis_ML_Harmonization_Proyect

Verify the installation of the libraries and the versions of the previously installed packages, see Figure 63.

```
pip freeze
```

Figure 63 Output installed packages in requirements format.

4. Run preprocessing routine, found in the eeg_harmonization repository, in the Python file **preprocessing.py.**
5. Run preprocessing routine, found in the eeg_harmonization repository, in the Python file **postprocessing.py.**

6. Use the "Data_analysis_ML_Harmonization_Proyect" repository to generate the necessary DataFrames and graphics.
7. Perform matching and implementation of neuroHarmonize.
8. Use "Data_analysis_ML_Harmonization_Proyect" again for graph generation.
9. Run **ML_models_G1_ic. ipynb** in "Data_analysis_ML_Harmonization_Proyect" for machine learning model deployment

Annex 3: Feature Extraction

Link to Feature Extraction

Annex 4: Harmonization of extracted features

Link to Harmonization of extracted features

Annex 5: Statistical analysis of harmonized features

Link to result of sovaHarmony (without neuroHarmonize)

Link to result of neuroHarmonize

Link to result of effect size tables

Annex 6: Implementation and validation of the model

Link to result of the implementation and validation of the model

Methodology for Learning Curve Analysis

To assess the performance and generalization ability of the optimized model obtained from the grid search, a learning curve analysis was conducted. This analysis helps determine how the model's accuracy varies as the number of training samples increases.

The learning curve was constructed using the following steps:

1. Learning Curve Generation: The learning_curve function was utilized to generate the learning curve. It takes the following inputs:

- The optimized model.
- The training dataset (X_train and y_train).
- The train_sizes parameter, which defines the proportion of the training dataset to use. In this case, it was set to start with 10% of the training data and gradually increase up to 100% in 10 equal steps.
- Cross-validation (cv) was performed with a value of 10, which splits the data into 10 folds.
- The learning_curve function was executed in parallel using all available processors (n_jobs=-1) to expedite the process.

2. Calculation of Mean and Standard Deviation: The mean and standard deviation of the training and validation scores were calculated across the different training set sizes. These scores represent the accuracy of the model.

3. Plotting the Learning Curve: A line plot was created to visualize the learning curve. The x-axis represents the number of training samples, while the y-axis represents accuracy. The training accuracy was plotted in blue with markers, and the validation accuracy was plotted in green with dashed lines and markers.

- The area between the mean + standard deviation and mean - standard deviation for both training and validation accuracy was filled to represent the variance.

By following this methodology, the learning curve analysis provided insights into the model's performance with varying training set sizes. It helped assess the model's ability to generalize well and detect any overfitting or underfitting issues.

Optimizing Decision Trees using Grid Search: Fine-tuning Hyperparameters for Improved Performance

To optimize the performance of decision trees, a systematic approach called grid search was employed. Grid search involves evaluating the model's performance by systematically searching through a predefined set of hyperparameters to find the optimal combination that yields the best results.

The following hyperparameters were considered for optimization:

1. Number of estimators: A range of values from 100 to 2000, with a step size of 30, was explored.

2. Maximum number of features: Two options were considered: 'auto' and 'sqrt'.
3. Maximum depth: A range of values from 10 to 110, with a step size of 11, was examined. Additionally, a value of None was included to allow for unlimited depth.
4. Minimum samples required to split an internal node: Three values were tested: 2, 5, and 10.
5. Minimum samples required to be a leaf node: Three values were evaluated: 1, 2, and 4.
6. Bootstrap sampling: Two options were considered: True and False.
7. Criterion for splitting: Three criteria were assessed: 'gini', 'entropy', and 'log_loss'.

A random grid was constructed using these hyperparameters, encompassing various combinations for testing.

To perform the grid search, a random forest classifier was employed as the base estimator. The random search algorithm, RandomizedSearchCV, was utilized to explore the hyperparameter space. This algorithm conducts a randomized search by sampling a specified number of combinations from the grid and evaluating their performance using cross-validation.

During the search, the algorithm was configured to perform 100 iterations and use 10-fold cross-validation. It was set to run in parallel using all available processors

(n_jobs=-1) to expedite the process. The random_state was fixed to 10 for result reproducibility.

Finally, the model was fitted on the training data (X_train and y_train) using the optimized hyperparameters obtained from the grid search.

By implementing this methodology, the decision tree model's hyperparameters were fine-tuned to enhance its performance, ultimately leading to improved predictive capabilities. Methodology for Feature Selection using Boruta.

This was tried methodology aimed to select the relevant features from the dataset based on their importance in the classification process.

1. Initializing the Boruta Feature Selector:

- The BorutaPy class was utilized to perform the feature selection.
- The verbose parameter was set to 2 to display detailed information during the selection process.
- The estimator parameter was set to the best-selected model obtained from previous steps.
- The max_iter parameter was set to 100, which specifies the maximum number of iterations to run.
- The random_state parameter was set to 10 to ensure reproducibility.

2. Fitting the Feature Selector:

- The feature selector (feat_selector) was fit to the training data (X_train, y_train) using the fit () method.

- During the fitting process, Boruta evaluated the importance of each feature by comparing it with randomized versions of the dataset.

3. The best-selected model (best_selected) was fitted to the transformed feature set (X_transform) and the target variable (y_train).

By following this methodology, the Boruta feature selection technique was applied to identify the most relevant features for classification. The selected features were then used to train a model and evaluate its performance using classification metrics and cross-validation.

Methodology for Feature Selection using Decision Trees

In this phase, the focus was on analyzing the importance of features using a decision tree-based approach. By understanding the significance of different features, we can gain insights into their contribution to the overall predictive power of the model.

A decision tree model was trained using the training dataset, with the target variable being the class labels. The model was trained to determine the importance of each feature in the classification process.

1. Feature Importance Scores:

- The feature importance scores were calculated using the trained decision tree model.
- These scores provide a quantitative measure of how much each feature contributes to the model's predictive performance.

2. Ranking and Visualization:

- The features were ranked based on their importance scores in descending order.
- Each feature was associated with a corresponding score, representing its relative importance.
- The top-ranked features, with the highest scores, are the most influential in the classification task.

The analysis of feature importance helps identify the most influential features, enabling us to focus on the key variables that contribute significantly to the model's predictive performance. By understanding the relative importance of features, we can make informed decisions regarding feature selection and potentially improve the effectiveness of our predictive models.

Methodology for SVM (Grid Search)

The following methodology describes the steps involved in optimizing the SVM model using grid search.

1. Defining the Parameter Grid:

- The parameter grid consists of different combinations of hyperparameters that will be evaluated during the search process.
- The 'C' parameter controls the regularization strength, with a range of values defined using a logarithmic scale.

- The 'gamma' parameter controls the kernel coefficient, with a range of values defined using a logarithmic scale, along with 'Auto' and 'scale' options.
- The 'kernel' parameter specifies the type of kernel function to be used, with options for 'rbf' (Radial basis function) and 'poly' (Polynomial) kernels.

2. Model and Grid Search Setup:

- An SVM classifier (SVC) is instantiated.
- GridSearchCV is used to perform grid search.
- The SVM classifier and the parameter grid are provided as inputs to GridSearchCV.
- The 'n_jobs' parameter allows for parallel processing to speed up the grid search.
- Cross-validation is performed with 10 folds using the 'cv' parameter.

3. Grid Search Execution:

- The SVM model is trained and evaluated for each combination of hyperparameters in the grid.
- The performance of each model is assessed using cross-validation.

4. Selection of Best Model:

- The best performing SVM model is identified based on the evaluation results.
- The best estimator, representing the SVM model with the optimal hyperparameters, is obtained.

By systematically searching through different combinations of hyperparameters, the grid search approach helps identify the SVM model with the best performance. The chosen model, determined by the evaluation results, represents the optimized SVM model for the given dataset and task.

Methodology for TPOT

To streamline the model optimization process, the TPOTClassifier is employed. TPOT utilizes genetic programming to automatically search and select the best combination of machine learning algorithms and their hyperparameters. The following methodology outlines the steps involved in using TPOT for model optimization.

1. Setting Parameters:

- The number of generations determines the number of iterations TPOT will go through to evolve the best model.
- The population size determines the number of individuals (candidate models) in each generation.

- Cross-validation (CV) is performed with the specified number of folds (cv) to evaluate the fitness of each candidate model.
- The random_state parameter ensures reproducibility of results.
- Verbosity level (verbosity) controls the amount of information displayed during optimization.
- The n_jobs parameter allows parallel processing for faster execution.

2. Optimization:

- The TPOTClassifier is applied to the dataset for model optimization.
- TPOT automatically evolves a population of candidate models using genetic programming.
- Each candidate model undergoes evaluation through the specified number of CV folds to assess its performance.

3. Selection of Best Model:

- TPOT identifies the best-performing model based on a fitness metric, such as accuracy or ROC-AUC.
- The best model is selected as the output of the optimization process.

The utilization of TPOT simplifies the process of model optimization by automating the search for the best combination of algorithms and hyperparameters. This methodology allows for efficient exploration of the model space and enables the

identification of highly performing models without the need for manual trial and error.

Complementary material

To access the supplementary material associated with this study, please follow the link provided below:

Link to Supplementary Material

The supplementary material provides additional information and resources that complement the findings and methodology presented in this research. It includes detailed tables, figures, code scripts, and any other supporting materials that can further enhance the understanding and reproducibility of the study.

By accessing the supplementary material, readers can gain deeper insights into the experimental procedures, additional analysis, and extended results that may not be included in the main manuscript. It is recommended to explore the supplementary material to obtain a comprehensive understanding of the study's findings.

Should you have any difficulties accessing or downloading the supplementary material, please contact the corresponding author or the research team for further assistance.