Nov 2023

Bioengineering,
Universidad de Antioquia

ADVISOR

Ph.D. John Fredy Ochoa Gomez

STUDENT

Verónica Henao Isaza

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

Research work

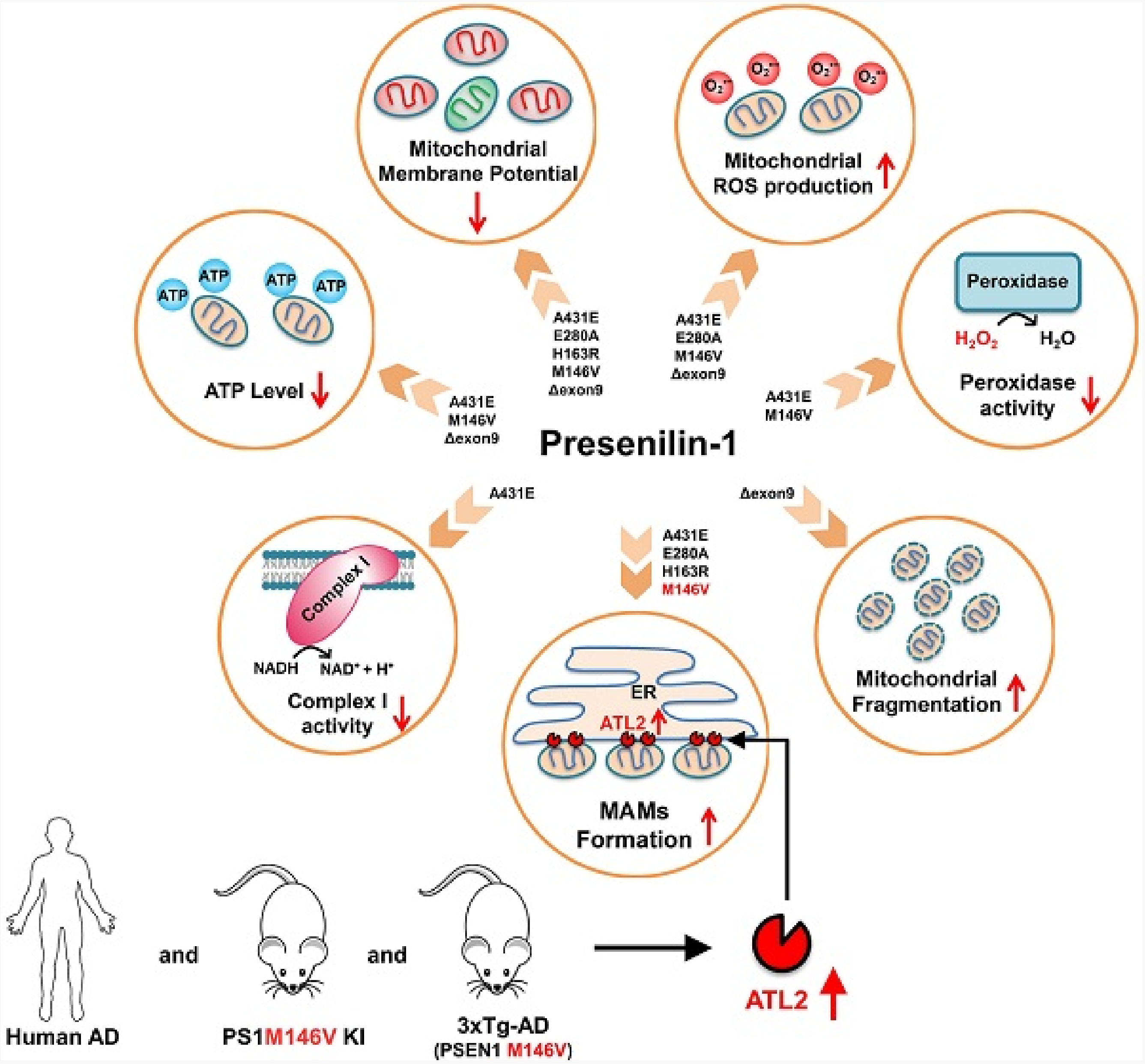
UNIVERSIDAD DE ANTIOQUIA
1803

Table of Contents

Table of Contents

The pathological genetic variants in PSEN1 are the most prevalent cause of familial Alzheimer's disease, responsible for 5-10% of early-onset AD cases.

The identification of specific EEG **biomarkers** can aid in the early diagnosis of AD and assist in the development of potential therapeutic interventions for this debilitating disease.



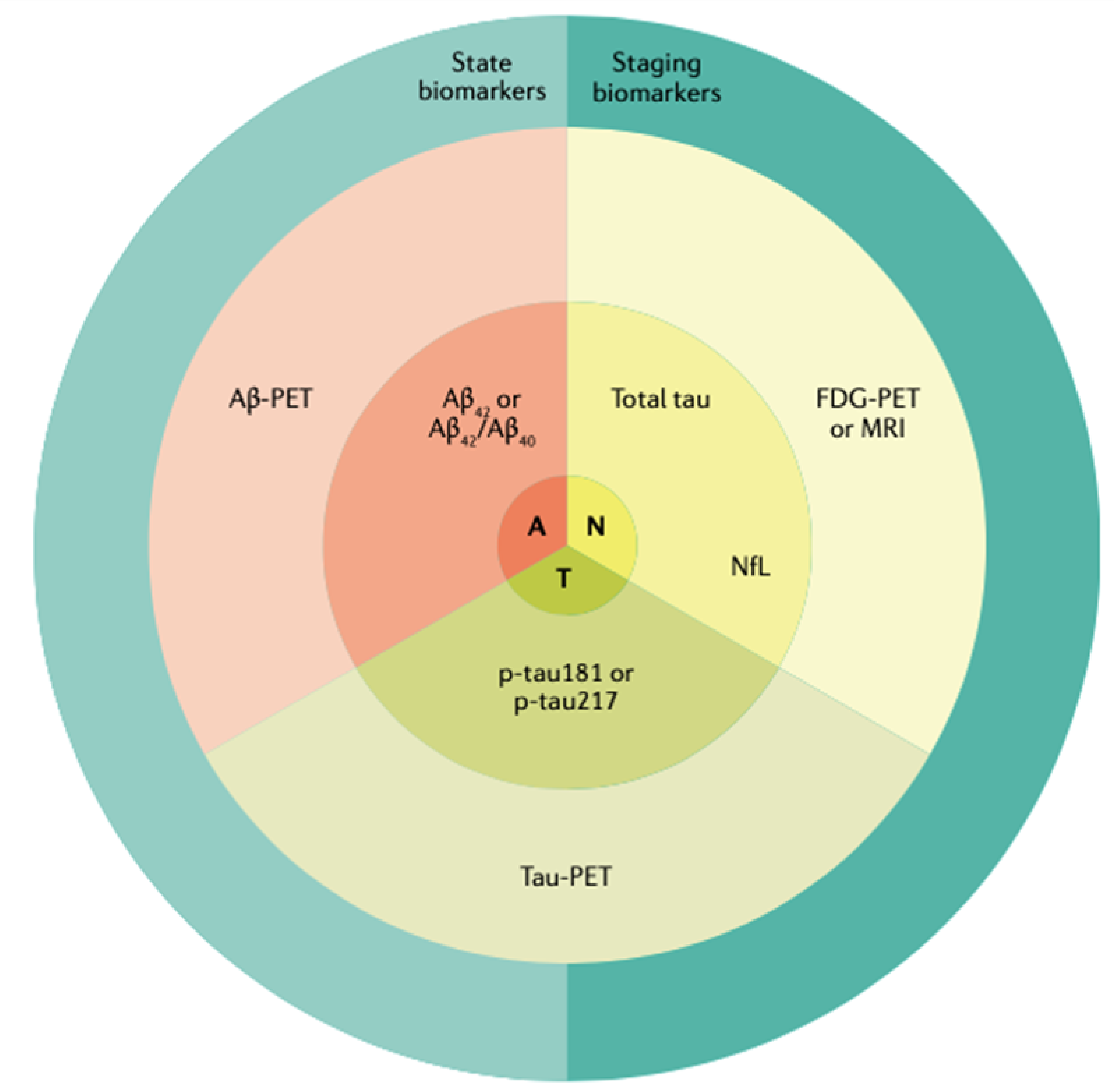
Han, J., Park, H., Maharana, C., Gwon, A.R., Park, J., Baek, S.H., Bae, H.G., Cho, Y., Kim, H.K., Sul, J.H., Lee, J., Kim, E., Kim, J., Cho, Y., Park, S., Palomera, L.F., Arumugam, T.V., Mattson, M.P., Jo, D.G. (2021). Alzheimer's disease-causing presenilin-1 mutations have deleterious effects on mitochondrial function. Theranostics, 11(18), 8855-8873. https://doi.org/10.7150/thno.59776.

gruneco
grupo neuropsicología y conducta

Biomarkers associated with A-T-N classification.

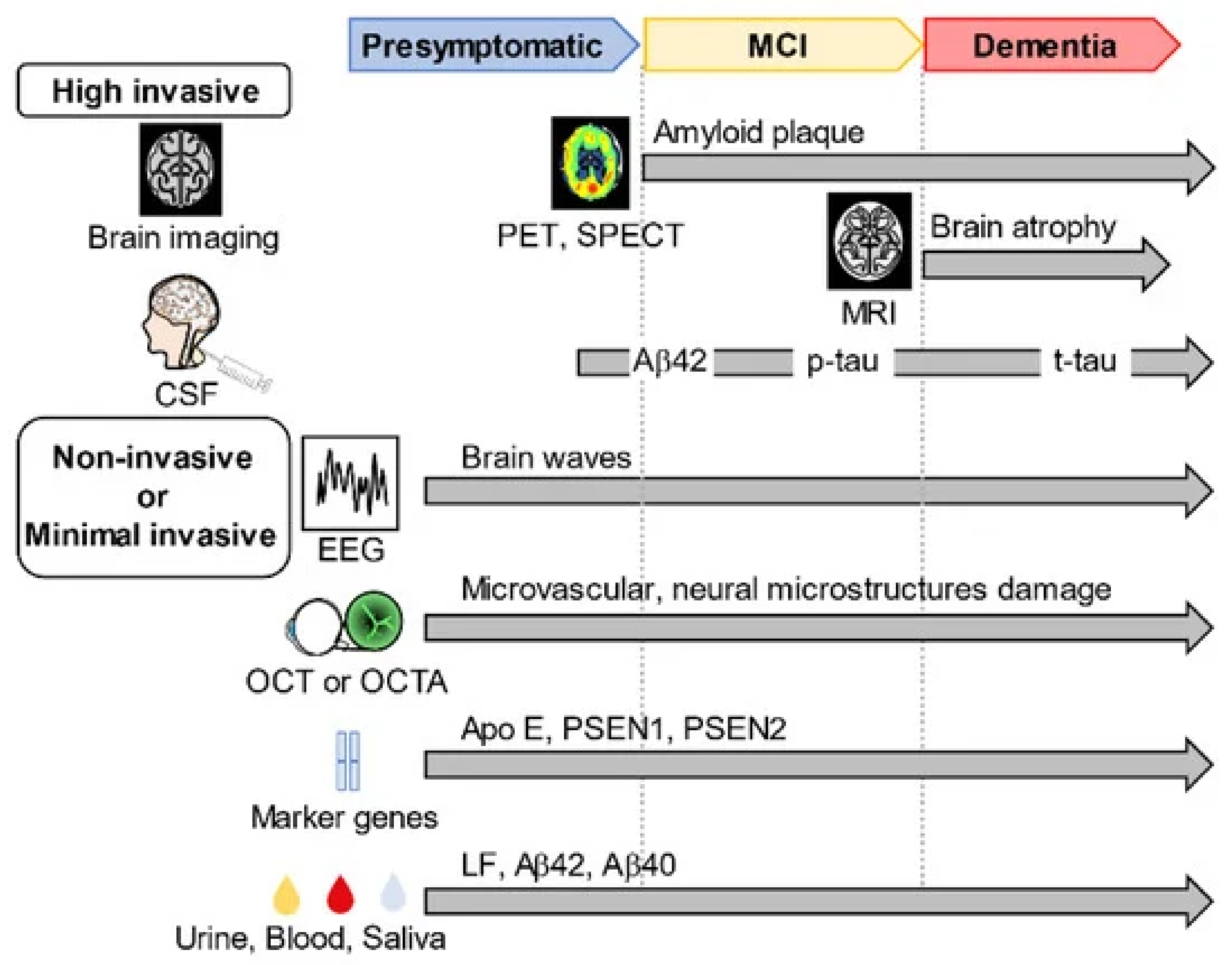
Amyloid-Tau-Neurodegeneration

Early biomarkers related to pathophysiological process of AD in pre-clinical stages.



Knopman DS, Amieva H, Petersen RC, Chételat G, Holtzman DM, Hyman BT, Nixon RA, Jones DT. Alzheimer disease. Nat Rev Dis Primers. 2021 May 13;7(1):33. doi: 10.1038/s41572-021-00269-y. PMID: 33986301; PMCID: PMC8574196.

grupo neuropsicología y conducta



Gunes, S.; Aizawa, Y.; Sugashi, T.; Sugimoto, M.; Rodrigues, P.P. Biomarkers for Alzheimer's Disease in the Current State: A Narrative Review. Int. J. Mol. Sci. 2022, 23, 4962. https://doi.org/10.3390/ijms23094962

PET:

Advantages:
High spatial resolution
Quantification of metabolic and protein markers

Limitations:
Invasive and expensive
Lower temporal resolution

$$$

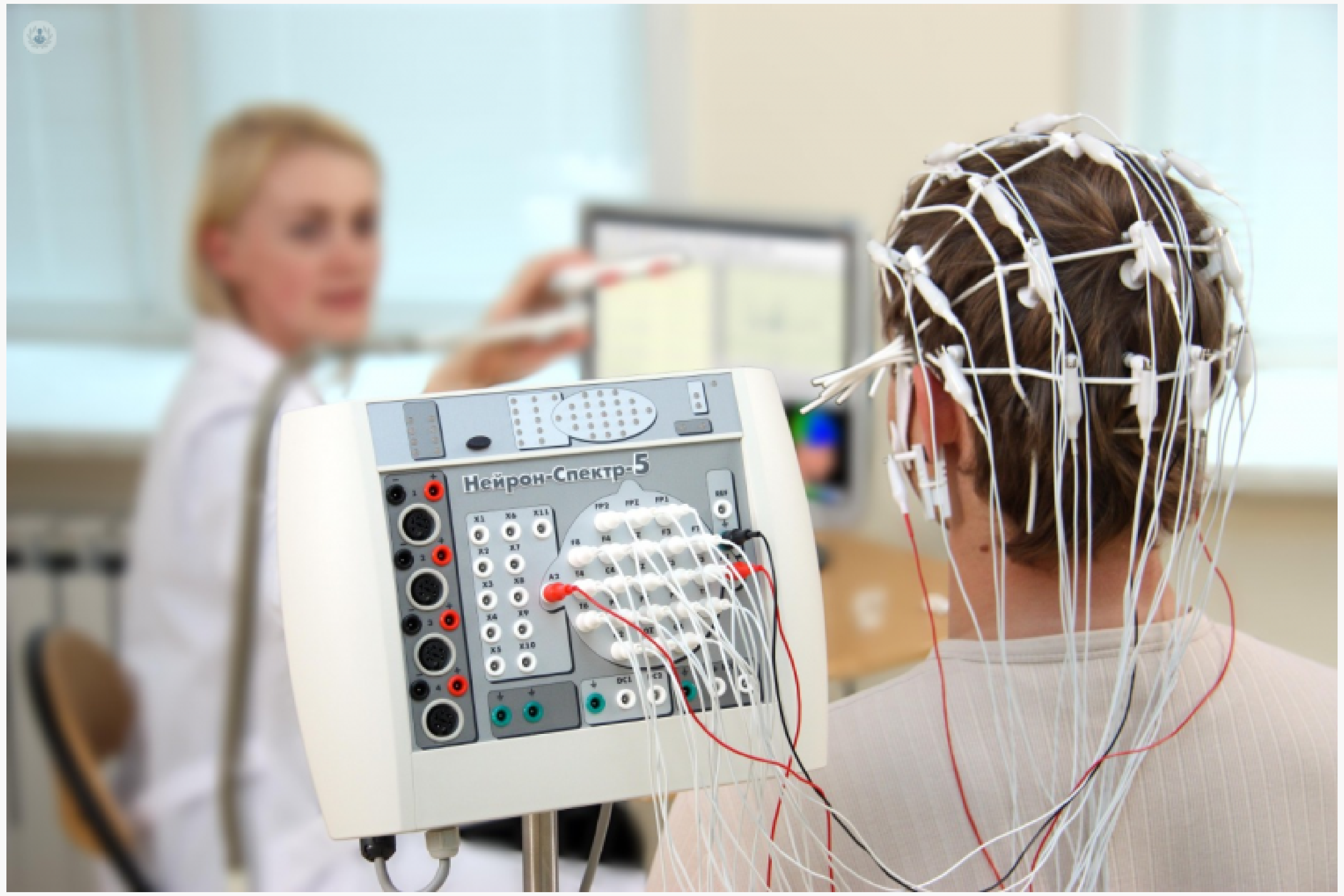
EEG:

Advantages:
High temporal resolution
Non-invasive and safe

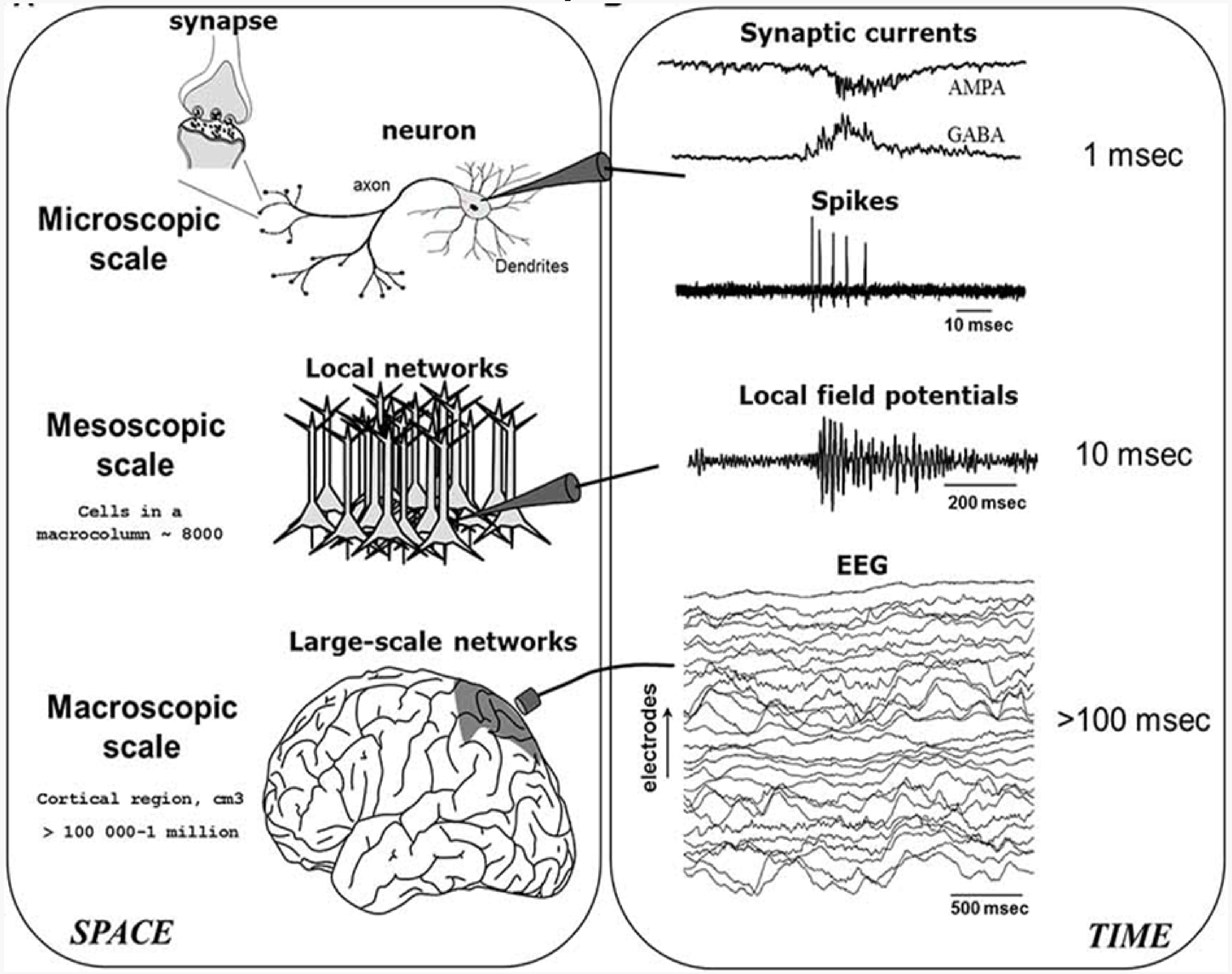
$

Limitations:
Low spatial resolution

gruneco
grupo neuropsicología y conducta



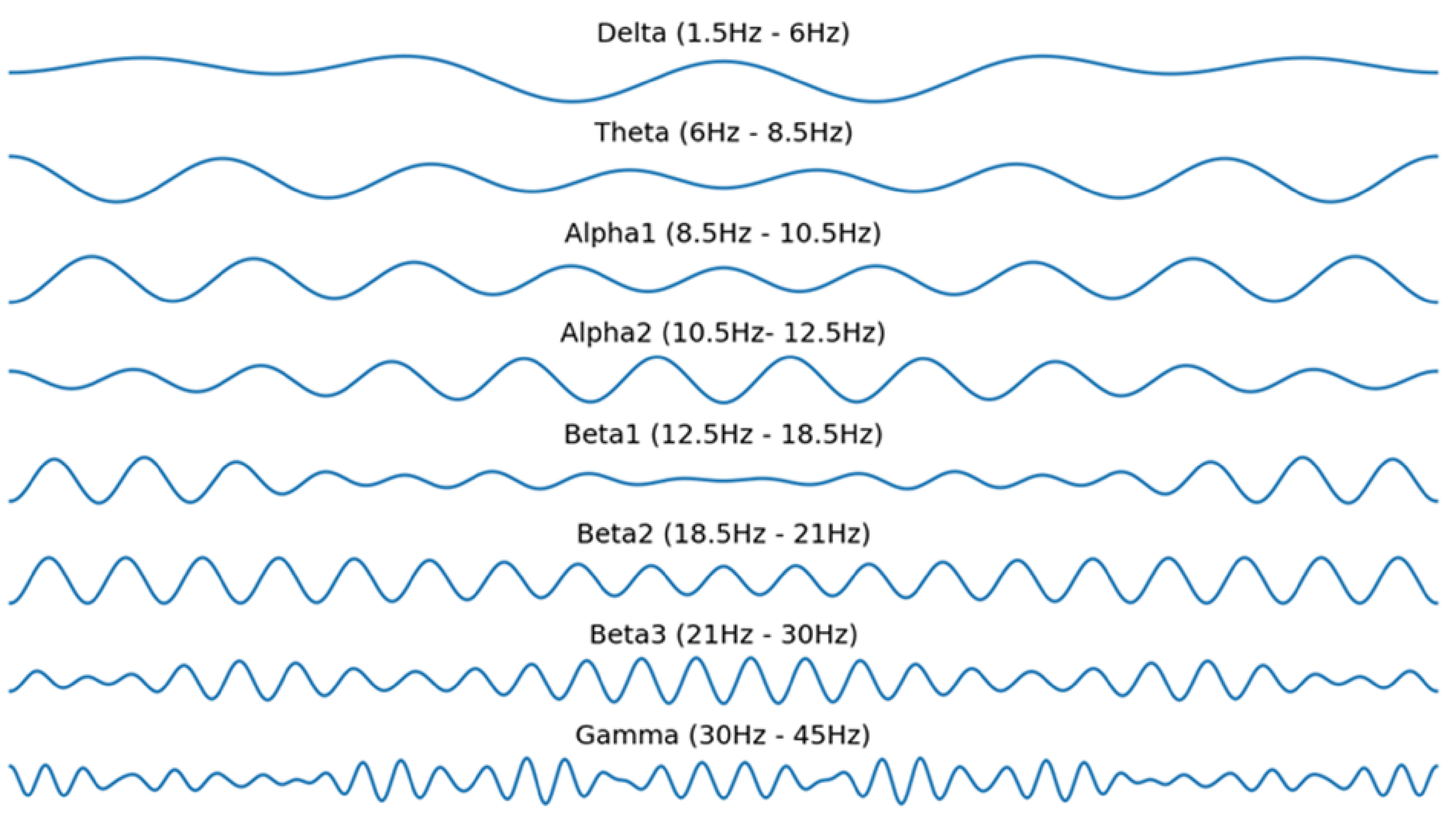
EEG is an instrument that can measure electrical brain activity.



Ros T, Baars B. J, Lanius R, et al. Tuningpathologicalbrainoscillationswithneurofeedback: a systemsneuroscienceframework. Front. Hum. Neurosci. 2014; 8:1008.

gruneco
grupo neuropsicología y conducta

These frequency bands were 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 useful reference material for other scientists.



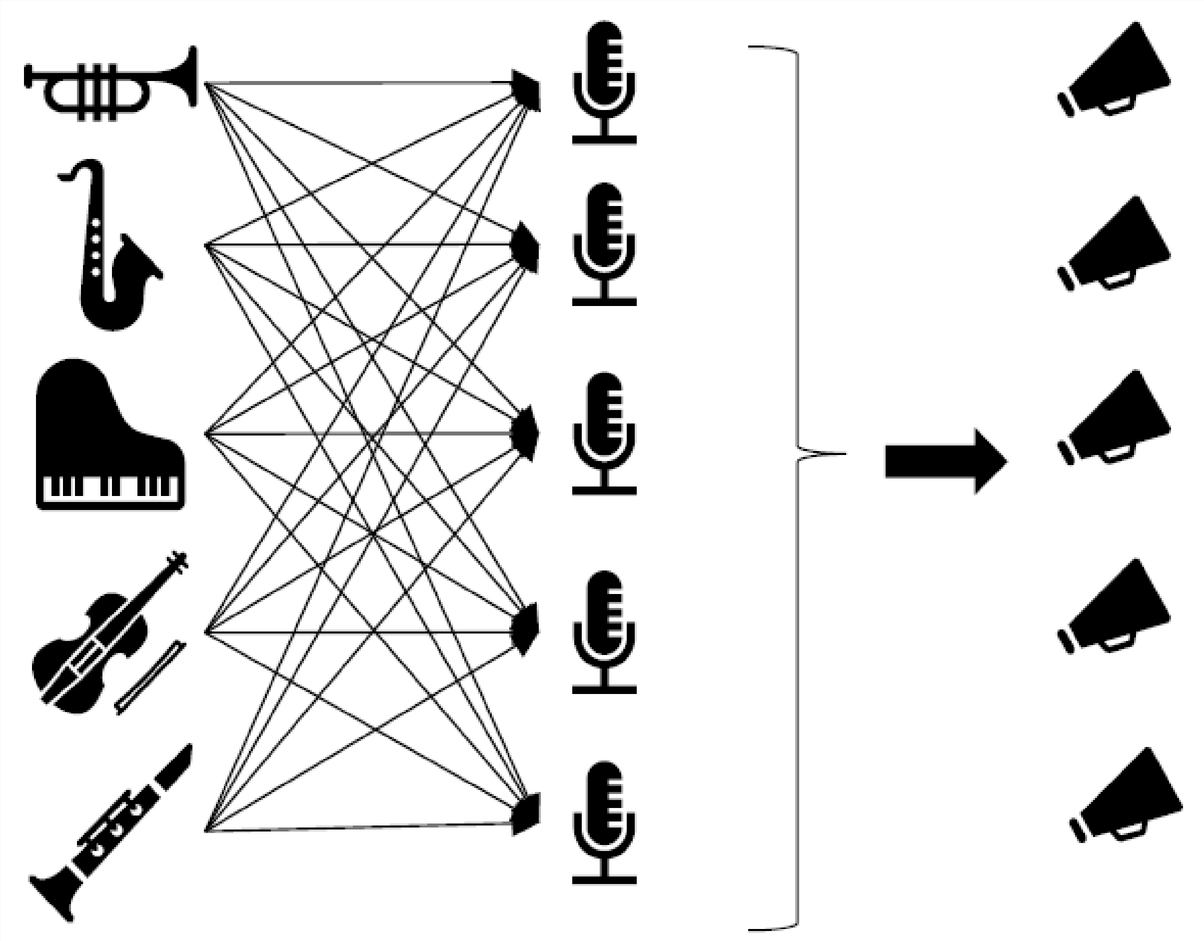
Jobert, M., Wilson, F. J., Ruigt, G. S., Brunovsky, M., Prichep, L. S., Drinkenburg, W. H., & IPEG Pharmaco-EEG Guideline Committee. (2012). Guidelines for the recording and evaluation of pharmaco-EEG data in man: the International Pharmaco-EEG Society (IPEG). Neuropsychobiology, 66(4), 201-220.

The number of arrows indicates the number of studies reporting concordant results: ↑ = 1–5 studies ↑ ↑ = 5–10 studies ↑ ↑ ↑ >10 studies. IC = inconclusive results.

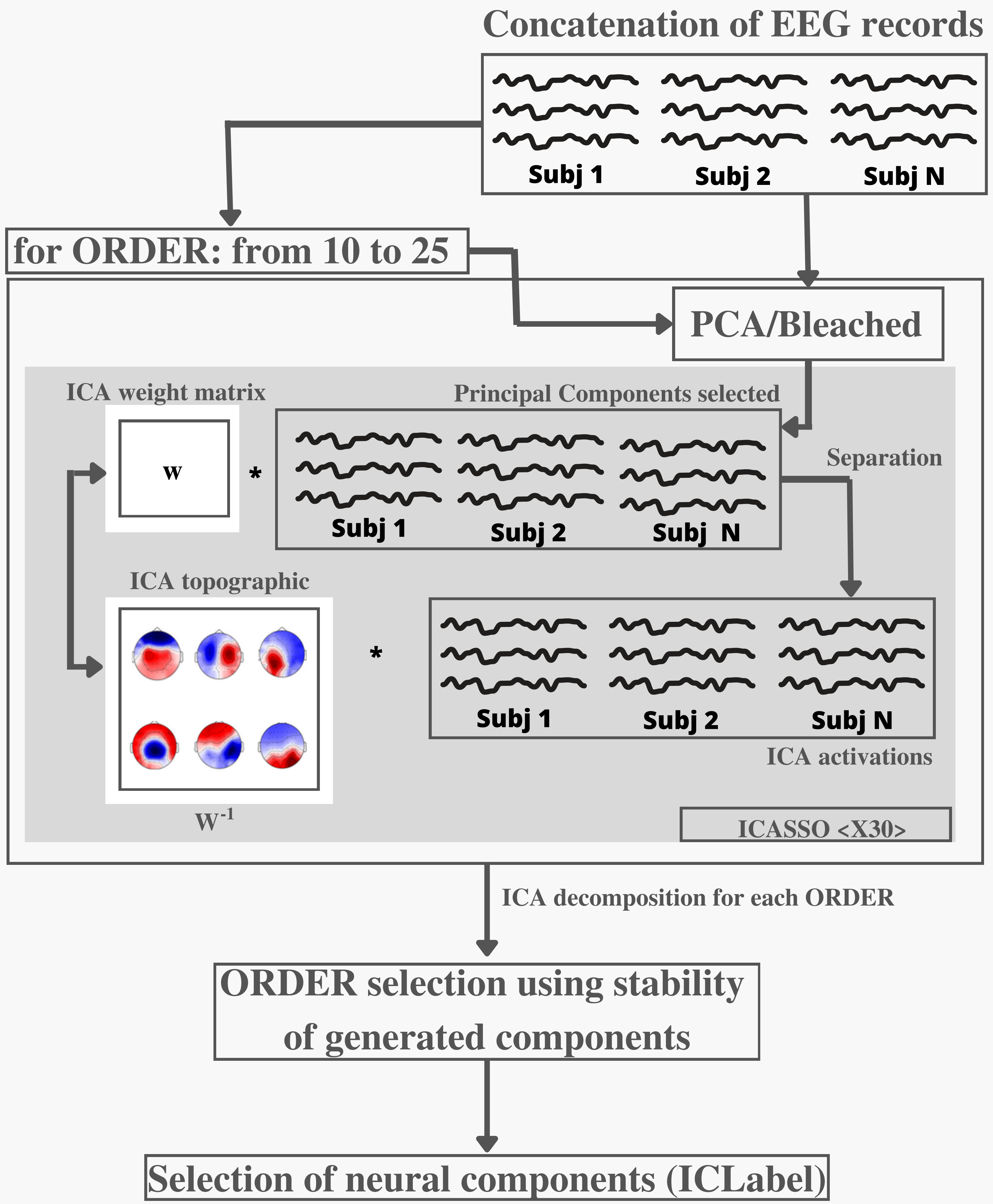
EEG band	Spectral power and phase coherence in AD	Spectral power and phase coherence in MCI
Gamma	IC	IC
Beta	↓↓	↓↓↓
Alpha	↓↓↓	↓↓↓
Theta	↑↑	↑↑
Delta	↑↑↑	↑↑

García Pretelt, F. J. (2022). Caracterización de Alzheimer temprano en poblaciones con riesgo genético mediante electroencefalografía. Bibliotecadigital.udea.edu.co. https://hdl.handle.net/10495/28940

The difference between the current work and that shown in the table is that the PSEN1-E280A populations are predisposed to Alzheimer's disease even without symptoms. It is critical to focus on these specific groups as the approach is predominantly preclinical.



1. EEG records concatenation
2. Optimal order test (10 – 25 for loop)
3. Data Whitening: PCA
4. gICA component calculation: 30 times, using ICASSO routines, evaluating max stability.



gruneco grupo neuropsicología y conducta

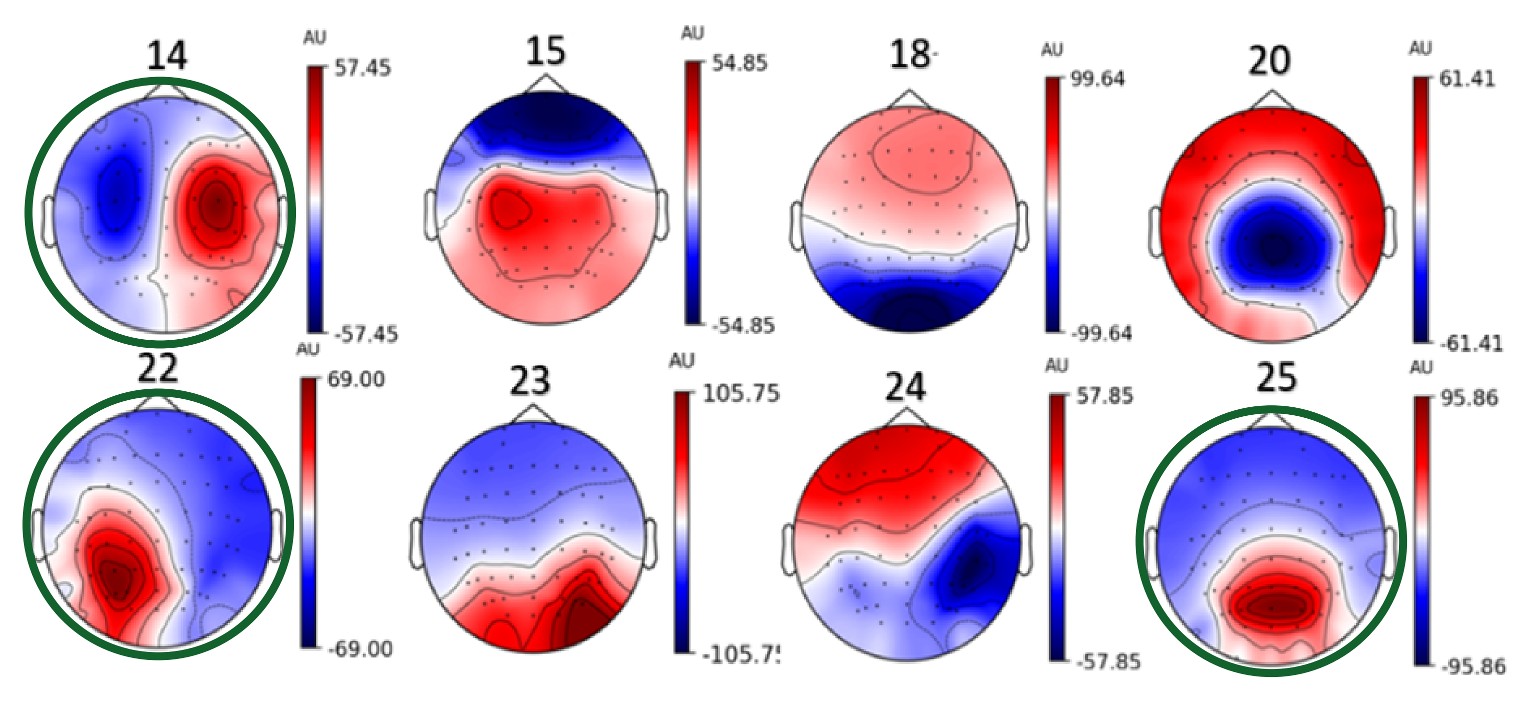
1. J. Himberg, A. Hyvärinen, y F. Esposito, «Validating the independent components of neuroimaging time series via clustering and visualization», Neuroimage, vol. 22, n.o 3, pp. 1214–1222, jul. 2004.

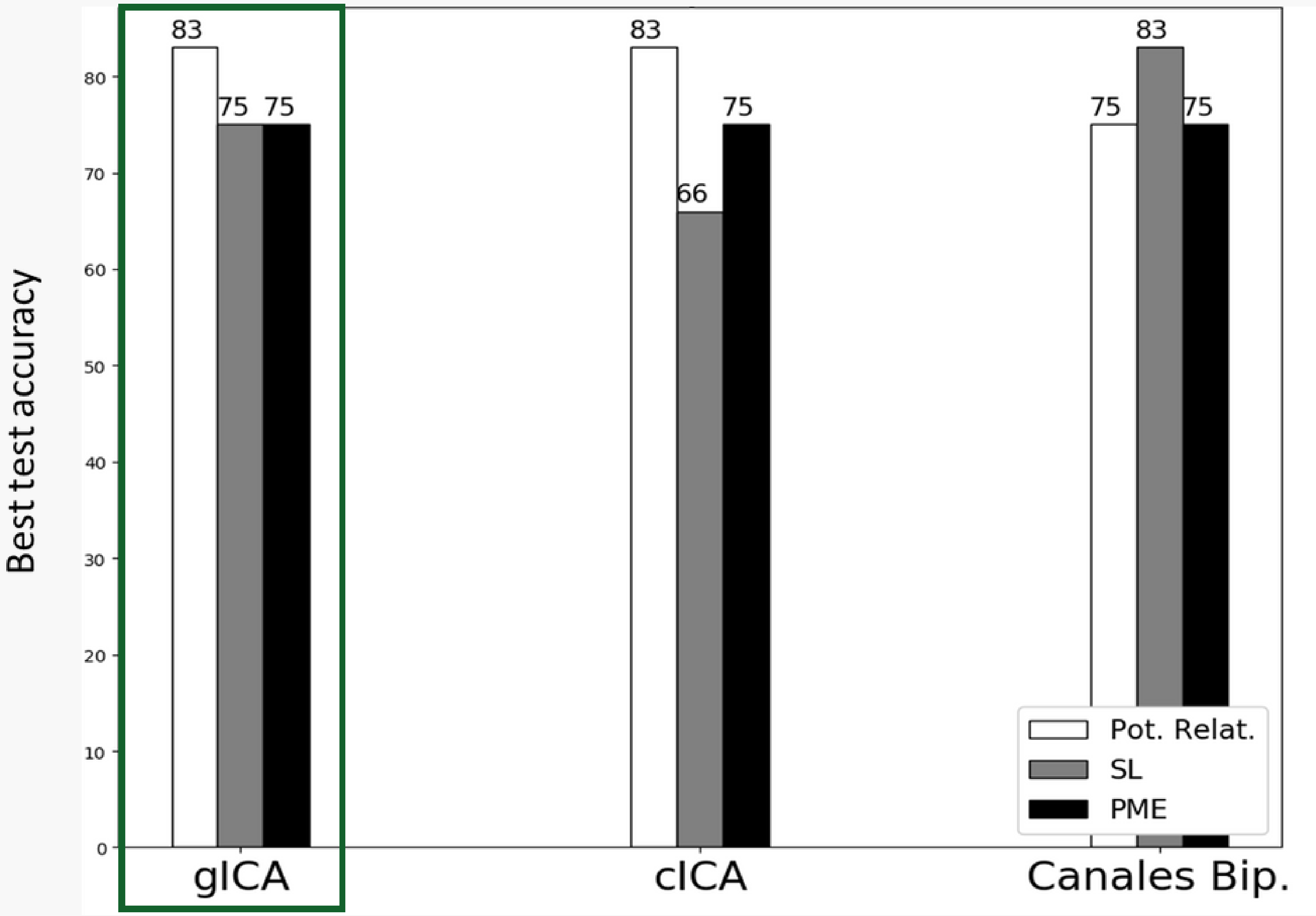
Automatic Classification of Subjects of the PSEN1-E280A Family at Risk of Developing Alzheimer's Disease Using Machine Learning and Resting State Electroencephalography

García-Pretelt FJ, Suárez-Relevo JX, Aguillon-Niño DF, Lopera-Restrepo FJ, Ochoa-Gómez JF, Tobón-Quintero CA.

	ACr	NonCr
N	27	33
Edad	32.4 ± 5.8	32.7 ± 5.8
Genero (M/F)	11/16	13/20
Escolaridad (años)	10.8 ± 2.9	13.2 ± 2.6

β
γ





García-Pretelt FJ, Suárez-Relevo JX, Aguillon-Niño DF, Lopera-Restrepo FJ, Ochoa-Gómez JF, Tobón-Quintero CA. Automatic Classification of Subjects of the PSEN1-E280A Family at Risk of Developing Alzheimer's Disease Using Machine Learning and Resting State Electroencephalography. J Alzheimers Dis. 2022;87(2):817-832. doi: 10.3233/JAD-210148. PMID: 35404271.

gruneco
grupo neuropsicología y conducta

Research Background & Motivation

Machine learning analysis reveals biomarkers for the detection of neurodegenerative diseases

Simon Lam, Muhammad Arif, Xiya Song, Mathias Uhlen, Adil Mardinoglu

doi: https://doi.org/10.1101/2022.02.15.22270625

Open Access | Article

Classifications of Neurodegenerative Disorders Using a Multiplex Blood Biomarkers-Based Machine Learning Model

by Chin-Hsien Lin [1], Shu-I Chiu [2,3], Ta-Fu Chen [1], Jyh-Shing Roger Jang [2] and Ming-Jang Chiu [1,4,5,6,*]

Evaluating the reliability of neurocognitive biomarkers of neurodegenerative diseases across countries: A machine learning approach

M. Belen Bachli [a,1], Lucas Sedeño [b,c,1], Jeremi K. Ochab [d,1], Olivier Piguet [e,f], Fiona Kumfor [e,f], Pablo Reyes [g,h], Teresa Torralva [b], María Roca [b], Juan Felipe Cardona [j], Cecilia Gonzalez Campo [b,c], Eduar Herrera [i], Andrea Slachevsky [k,l,m,n], Diana Matallana [o], Facundo Manes [b,c,e], Adolfo M. García [b,c,p], Agustín Ibáñez [b,c,e,q,r], Dante R. Chialvo [a,c]

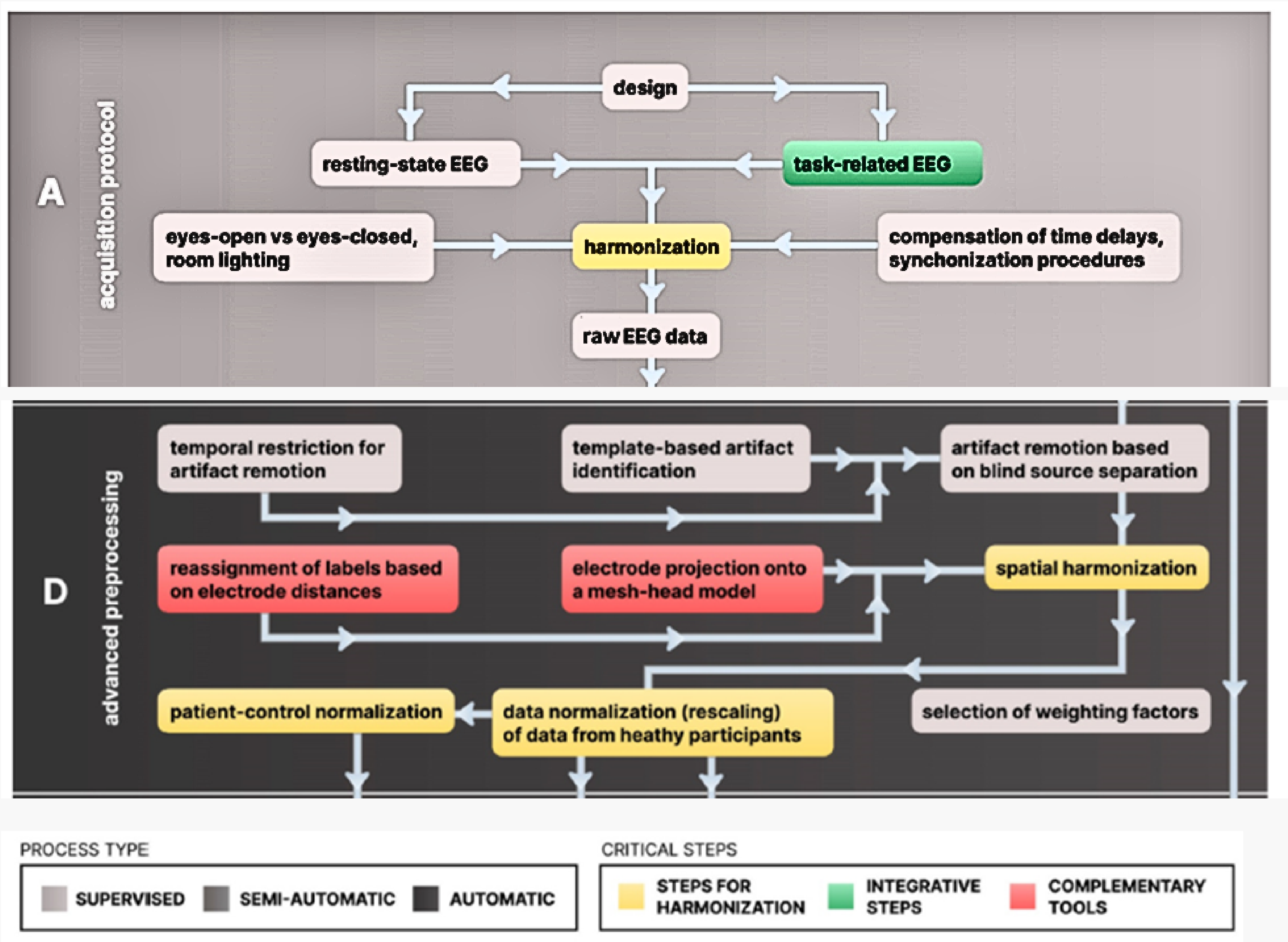
A multinomial model was constructed, which could predict neurodegenerative diseases, including AD, PD, MND, and MG, with an accuracy of 88.3%.

The developed model, which used multiple blood-based biomarkers and the Random Forest (RF) technique, achieved an accuracy of 76% in classifying AD, PD, and FTD. Furthermore, it achieved accuracies of 83% and 63% when differentiating disease severity in subgroups of the AD and PD spectra.

In Country-1, the model achieved classification rates greater than 91% for both bvFTD and AD. Moreover, it exhibited high predictive power (>0.91) when used to classify new patient cohorts from other international centers using different MRI acquisition equipment.

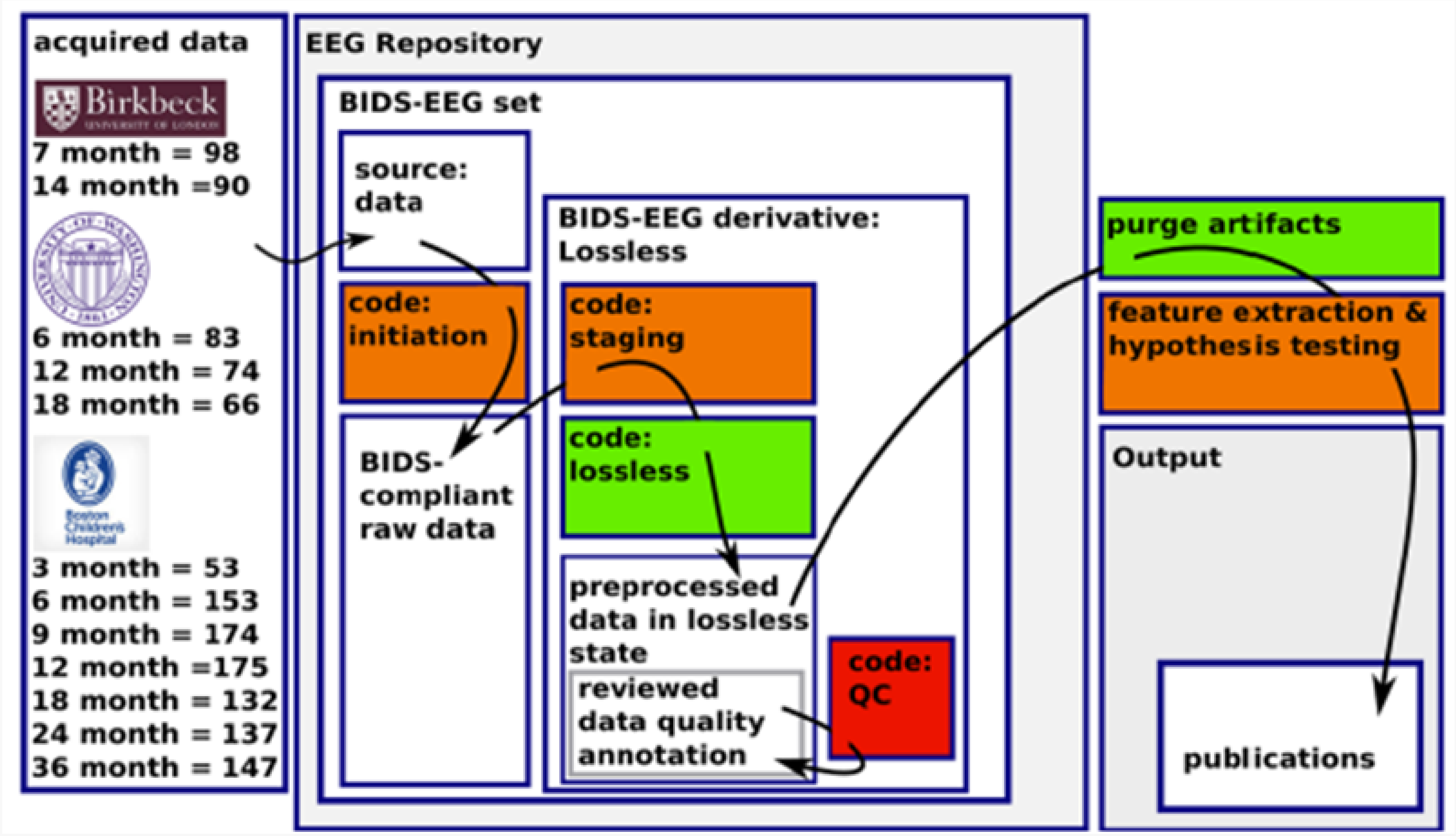
It is only applicable to symptomatic subjects; there is insufficient reference in preclinical studies.

gruneco
grupo neuropsicología y conducta



Prado, P., Birba, A., Cruzat, J., Santamaría-García, H., Parra, M., Moguilner, S., ... & Ibáñez, A. (2022). Dementia ConnEEGtome: towards multicentric harmonization of EEG connectivity in neurodegeneration. International Journal of Psychophysiology, 172, 24-38.

High-impact biomarker research is currently limited by relatively small sample sizes.



van Noordt, S., Desjardins, J. A., Huberty, S., Abou-Abbas, L., Webb, S. J., Levin, A. R., ... & Elsabbagh, M. (2020). EEG-IP: an international infant EEG data integration platform for the study of risk and resilience in autism and related conditions. Molecular Medicine, 26(1), 1-11.

gruneco
grupo neuropsicología y conducta

KEY ELEMENTS OF RESEARCH BACKGROUND & MOTIVATION

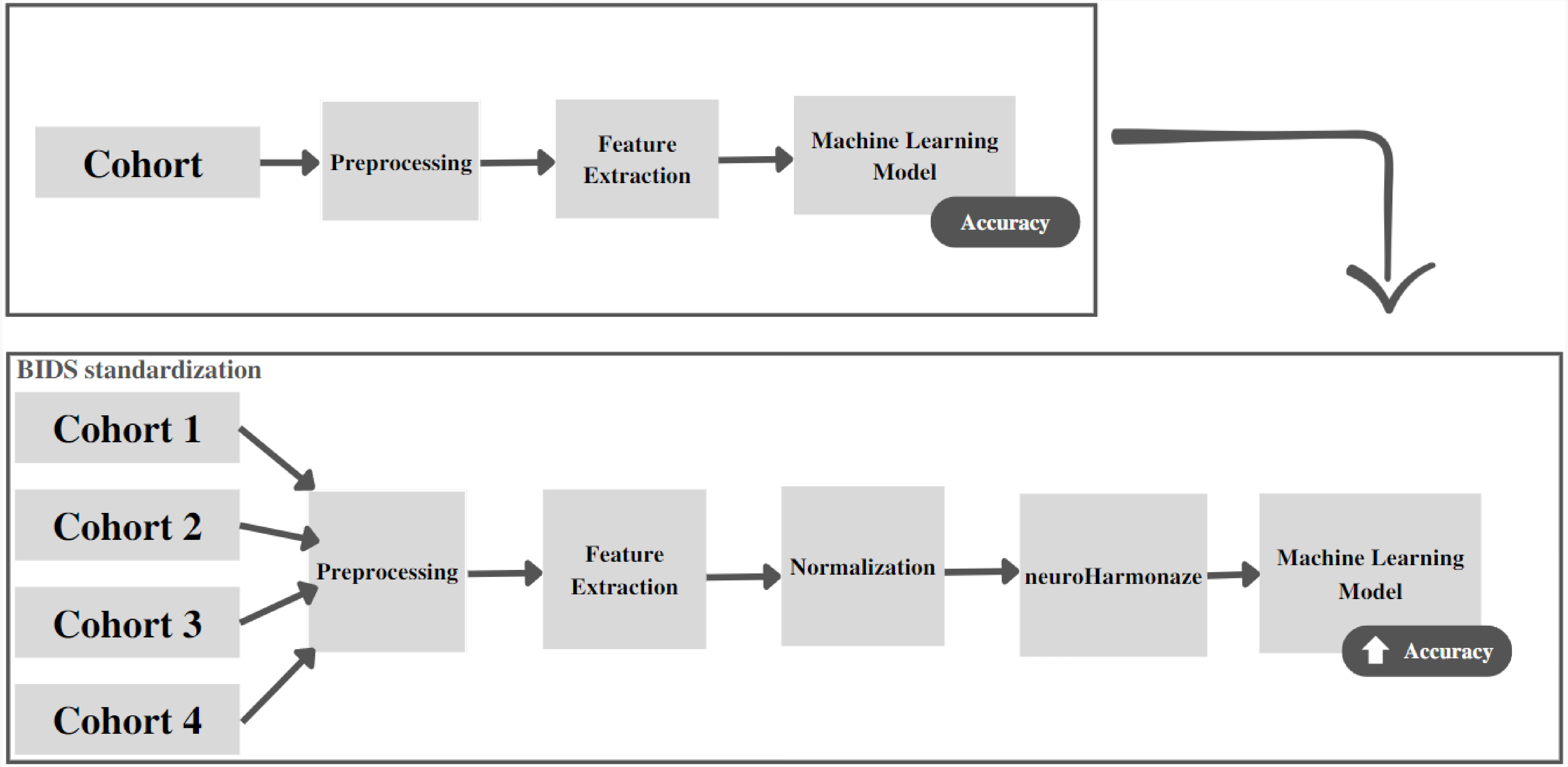
- Incorporating EEG (Electroencephalogram) usage contributes to **cost reduction.**
- Working with genetic populations provides a promising avenue for uncovering novel insights within the **preclinical** setting.
- The use of gICA could offer **stable and reproducible** insights into the spatiotemporal structure and dynamics of EEG before, during, and after experimental events.
- The reaction and dissemination of **automated processing pipelines** enable the expansion of restricted scenarios in biomarker research.
- A significant limitation lies in the scarcity of available data; however, leveraging openly **accessible databases** contributes to cost reduction.

Table of Contents

Question

"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)?"

Hypothesis



gruneco grupo neuropsicología y conducta

Table of Contents

Objectives

"To build an ML model that allows classifying subjects at risk of ADD, using non-invasive biomarkers from multiple databases."

Identify early biomarkers of Alzheimer's in asymptomatic carriers.

1. **Build and standardize a database** with multimodal information, taking multisite databases, using tools that facilitate data storage and manipulation before and during processing.
2. **Harmonize the database** to obtain comparable relevant electrophysiological and clinical parameters among healthy subjects using biomedical data processing techniques.
3. **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.

Table of Contents

1. **Build and standardize a database**
   - Data collection
   - BIDS standardization procedure

2. **Harmonize the database**
   - PREP
   - Huber normalization
   - Feature extraction
   - Harmonization

3. **Design a machine learning model**
   - Hyperparameter optimization
   - Using TPOT and Boruta to evaluate models
   - Decision tree
   - SVM

Type of study

Cross-Sectional Analytical Observational Study.

Population

Healthy control groups and individuals with a genetic **risk for AD** carrying the PSEN1-E280A gene

Inclusion criteria

- Inclusion of datasets containing EEG measurements.
- Emphasis on data collected during periods when participants were in a resting state or had their eyes closed.
- Consideration of data from individuals diagnosed with Alzheimer's and data from healthy individuals (without Alzheimer's), facilitating meaningful comparisons.
- Inclusion of data containing results from neuropsychological assessments.

Exclusion criteria

- Exclusion of datasets collected from portable equipment or with fewer than 58 electrodes.
- Exclusion of datasets from private sources without explicit consent from the creators.
- Exclusion of datasets that have been preprocessed.
- Exclusion of datasets that do not provide information on the acquisition protocol and the tests conducted on participants.

Table of Contents

V Research Results

Build and standardize a database

Search query

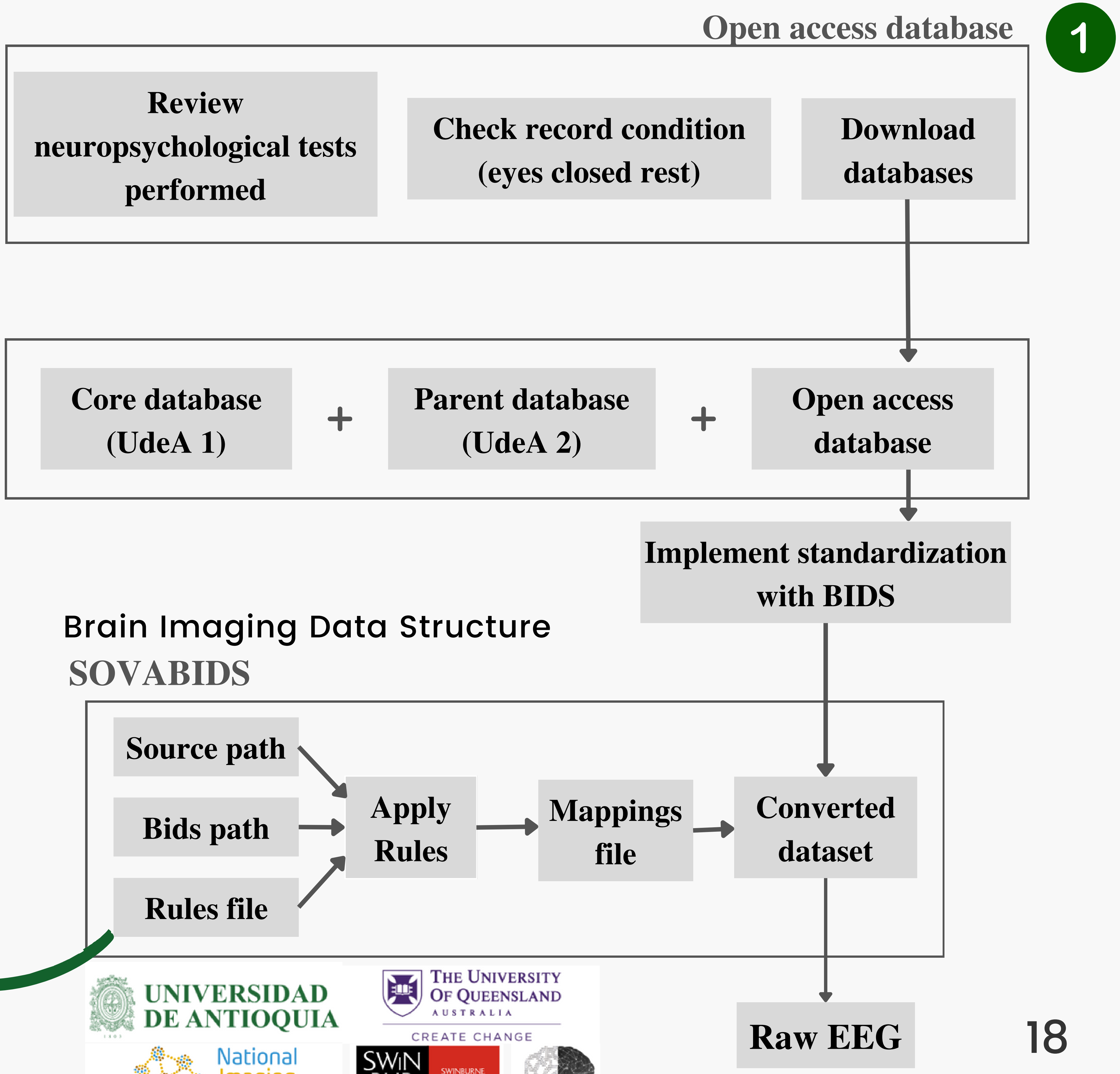
(EEG AND (CLOSE EYES OR RESTING STATE)) OR (ALZHEIMER OR HEALTHY SUBJECTS) AND NEUROPSYCHOLOGICAL TEST

Rules File

```
entities:
   task : rest
   session : S1

dataset_description :
   Name : MyDataset
   Authors :
      - Alice
      - Bob

sidecar :
   EEGReference : FCz
   PowerLineFrequency : 50

non-bids:
   eeg_extension : .cnt
   path_analysis:
      pattern : _data/%entities.subject%.cnt
```

Mappings File

```
- IO:
   source: data\P001.cnt
   target: BIDS\sub-P001\ses-S1\eeg\sub-P001_ses-S1_task-rest_eeg.vhdr
entities:
   session: 'S1'
   subject: 'P001'
   task: 'rest'
sidecar:
   EEGReference: FCz
   PowerLineFrequency: 50
- IO:
   source: data\P002.cnt
   target: BIDS\sub-P002\ses-S1\eeg\sub-P002_ses-S1_task-rest_eeg.vhdr
entities:
   session: 'S1'
   subject: 'P002'
   task: 'rest'
sidecar:
   EEGReference: FCz
   PowerLineFrequency: 50
```



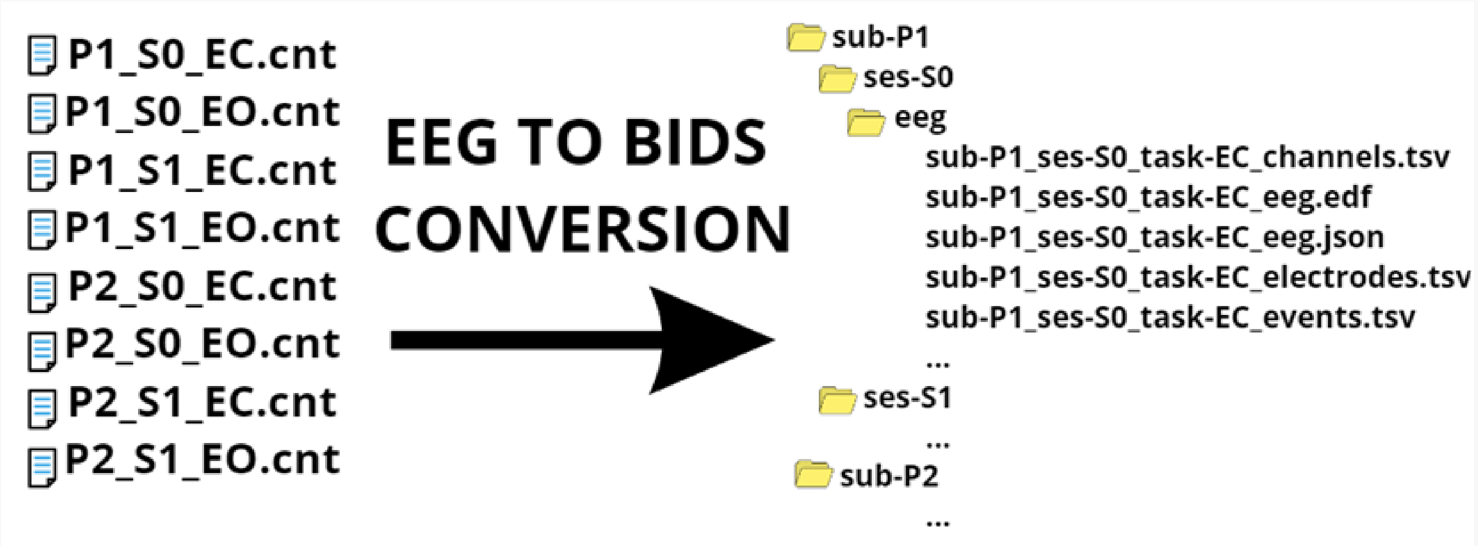
gruneco
grupo neuropsicología y conducta

UNIVERSIDAD DE ANTIOQUIA
THE UNIVERSITY OF QUEENSLAND AUSTRALIA
CREATE CHANGE
National Imaging Facility
SWINBURNE UNIVERSITY OF TECHNOLOGY
NeuroCo

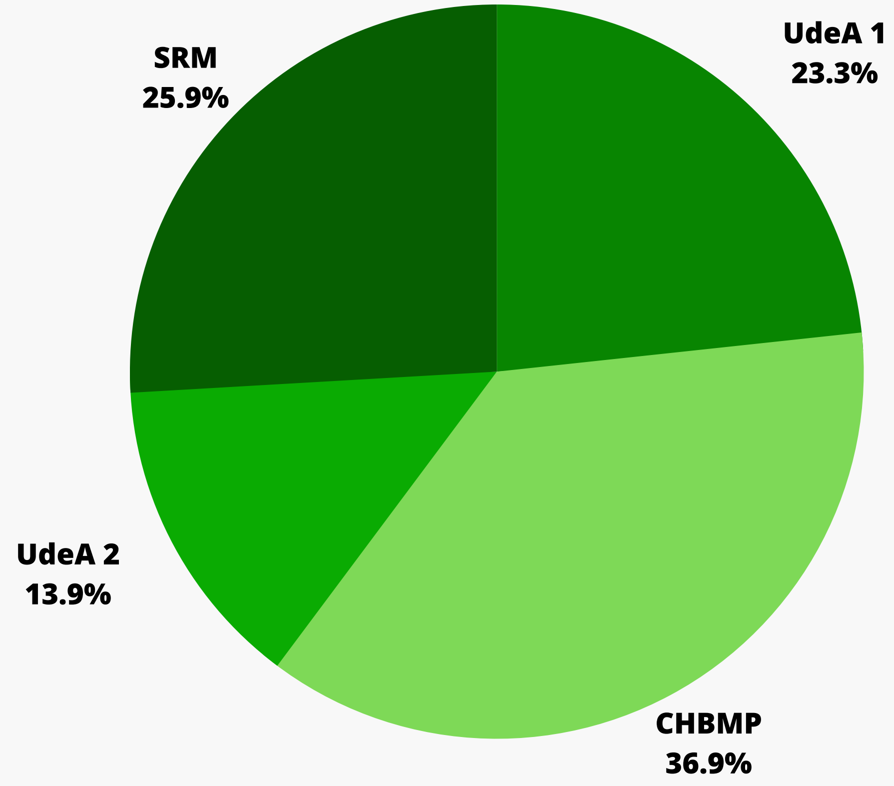
1

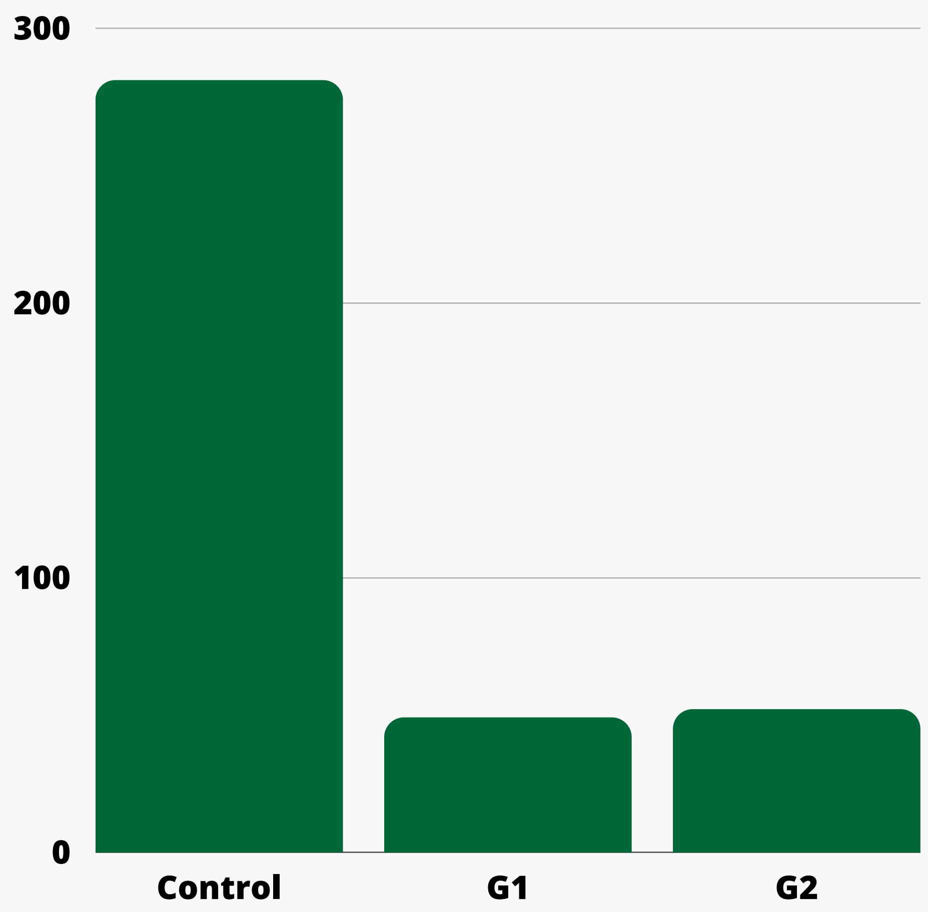
V Research Results

Build and standardize a database



		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			





gruneco
grupo neuropsicología y conducta

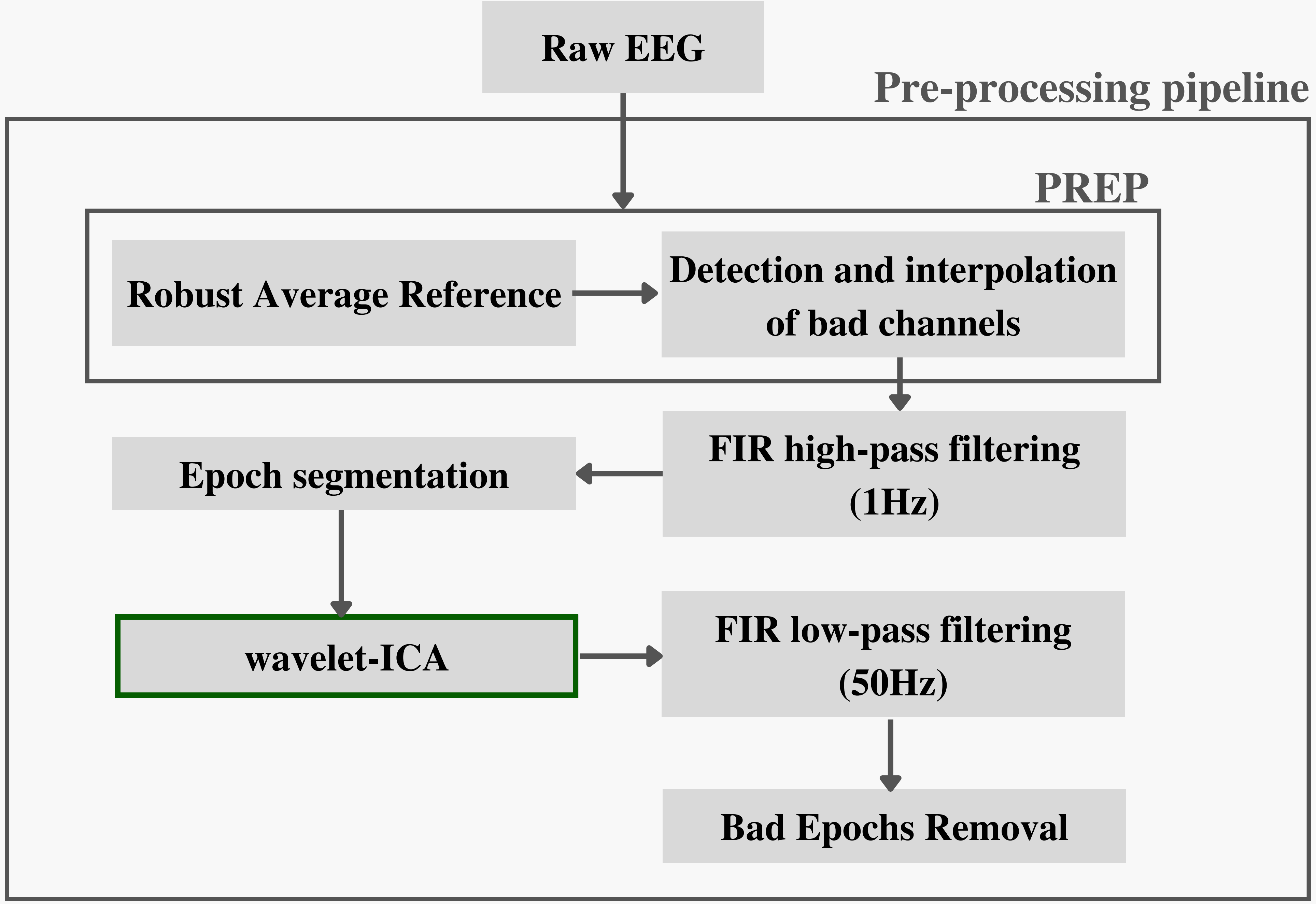
KEY ELEMENTS OF BUILD AND STANDARDIZE A DATABASE

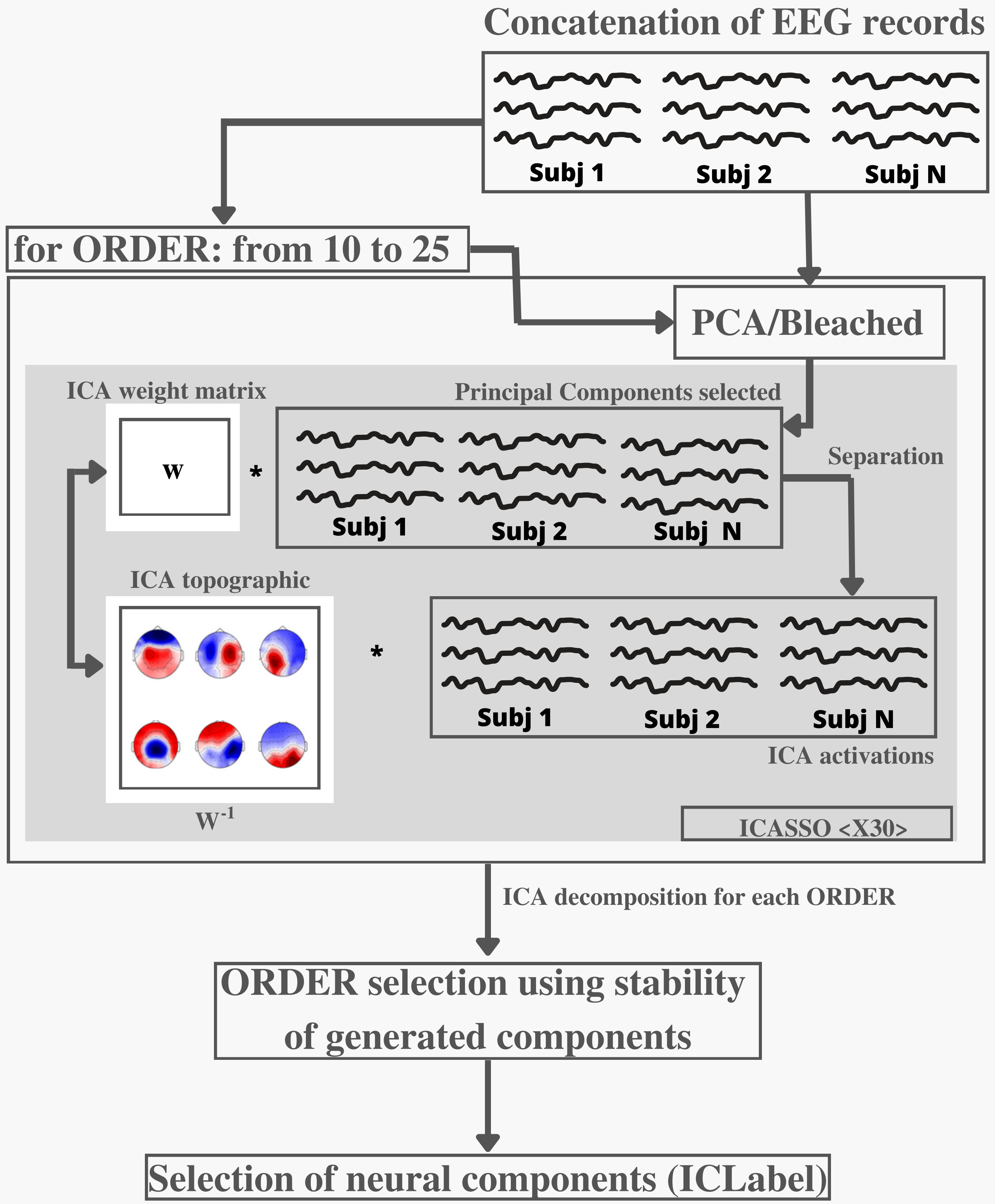
- Enhancing comparability challenges among independent recordings through the implementation of **BIDS in EEG data** yields a harmonized, scalable, and integrated data state.
- .A total of 457 records were obtained from four cohorts—two internal and two external—with 33% proprietary data and 67% external data. While maintaining overall control data, the majority comes from **external sources**, emphasizing their significant contribution.

V Research Results

Harmonize the database

2





gruneco
grupo neuropsicología y conducta

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.

V Research Results

Harmonize the database

Bad Epochs Removal

Record-specific constant

FIR filtering low-pass (20Hz) a copy of the EEG signal

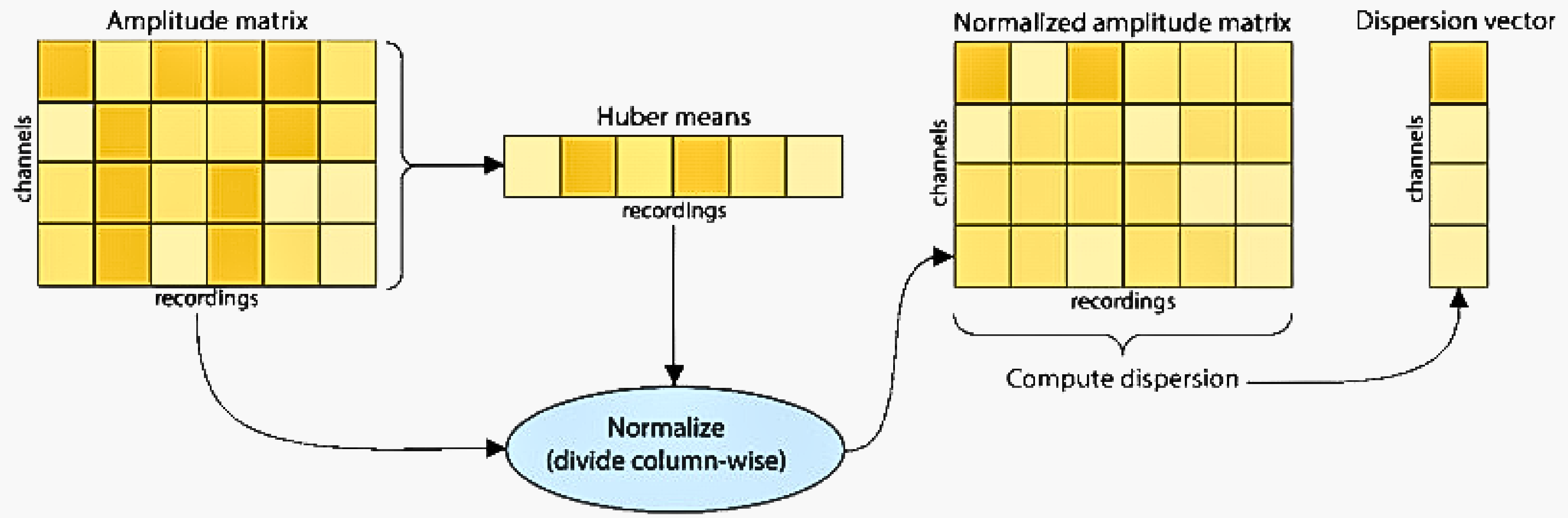
Robust standard deviation of each channel

Huber Mean

Signal

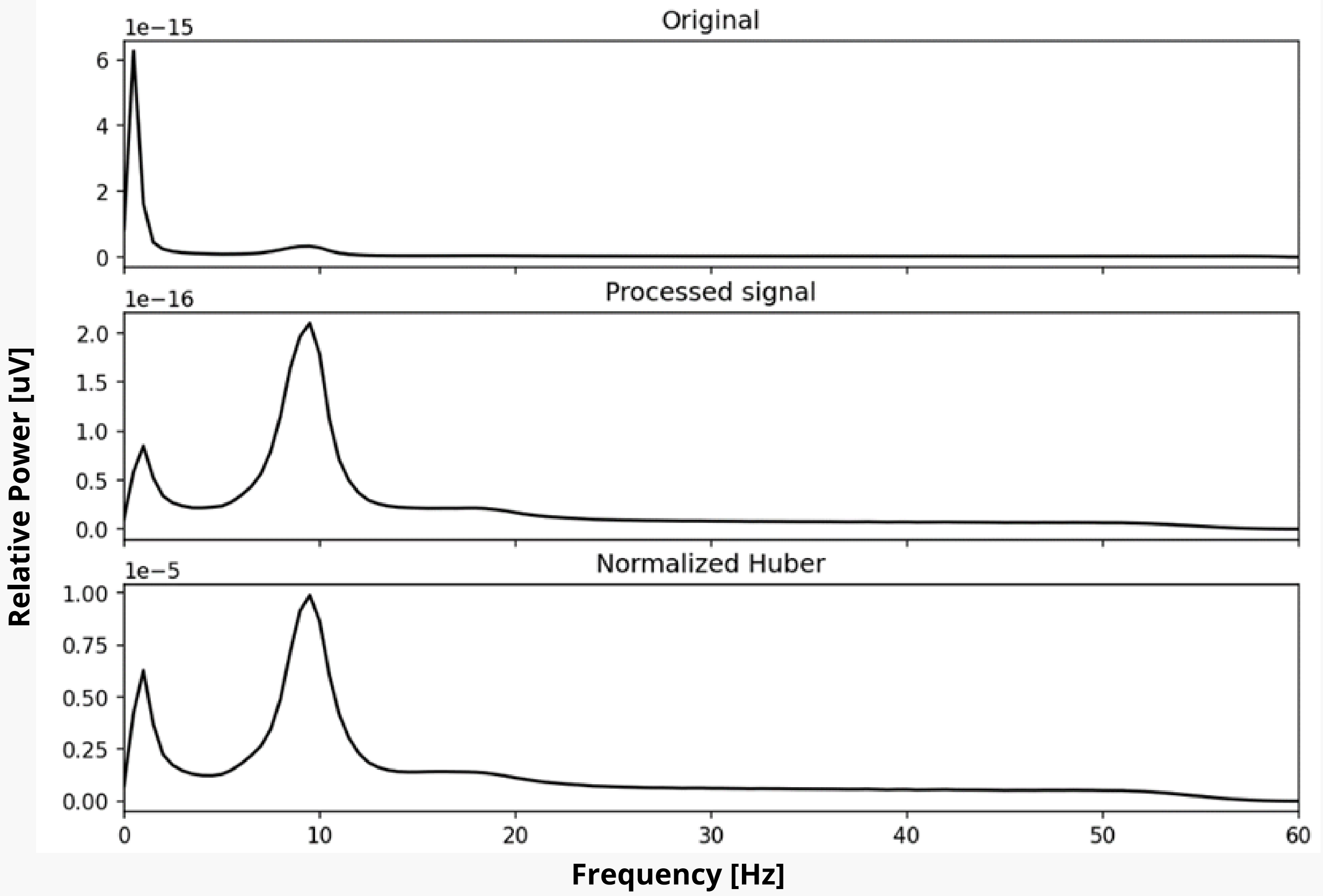
constant

EEG signal divided by the record-specific constant



Bigdely-Shamlo, N., Touryan, J., Ojeda, A., Kothe, C., Mullen, T., & Robbins, K. (2020). Automated EEG mega-analysis I: Spectral and amplitude characteristics across studies. NeuroImage, 207, 116361.

Data Normalization

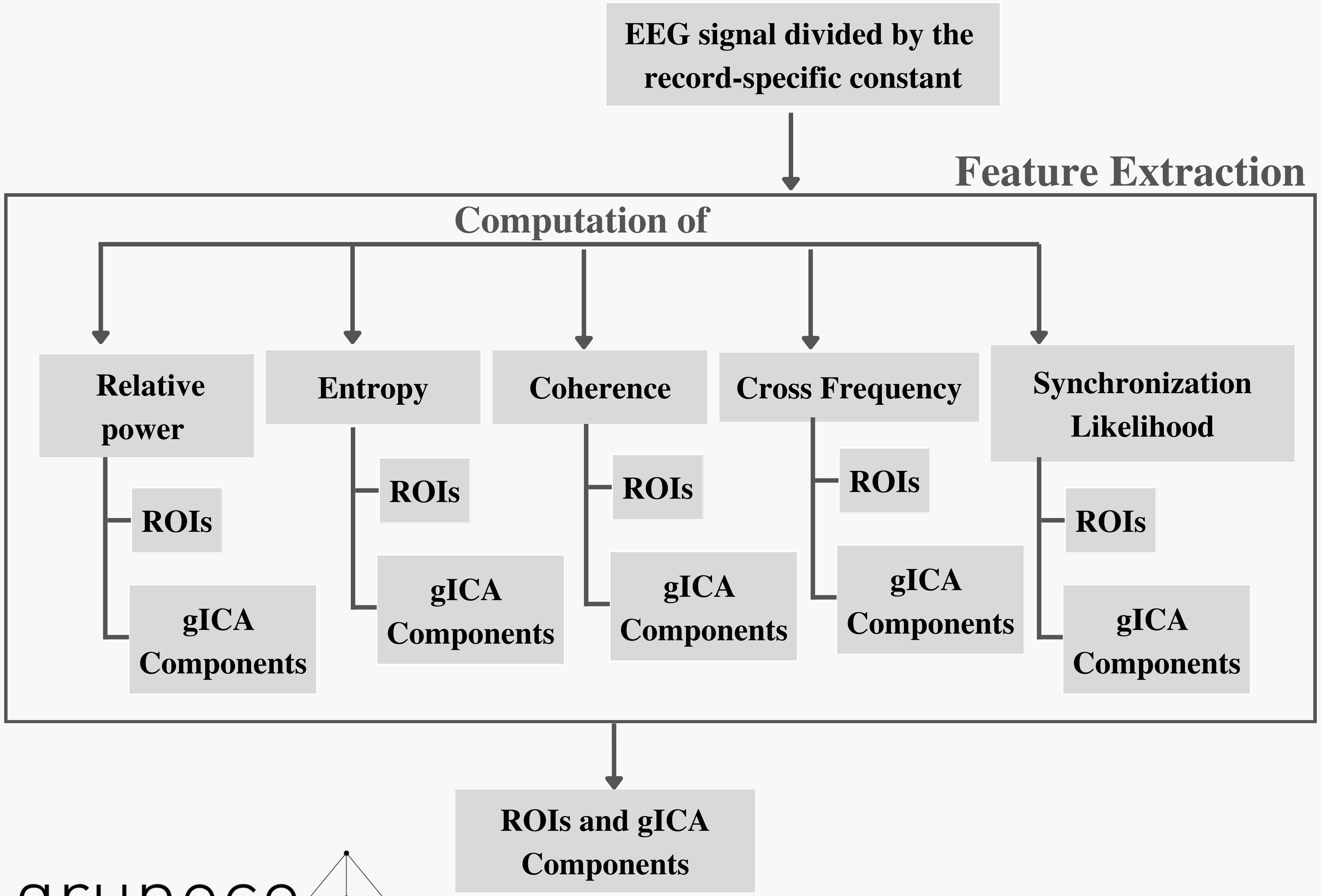


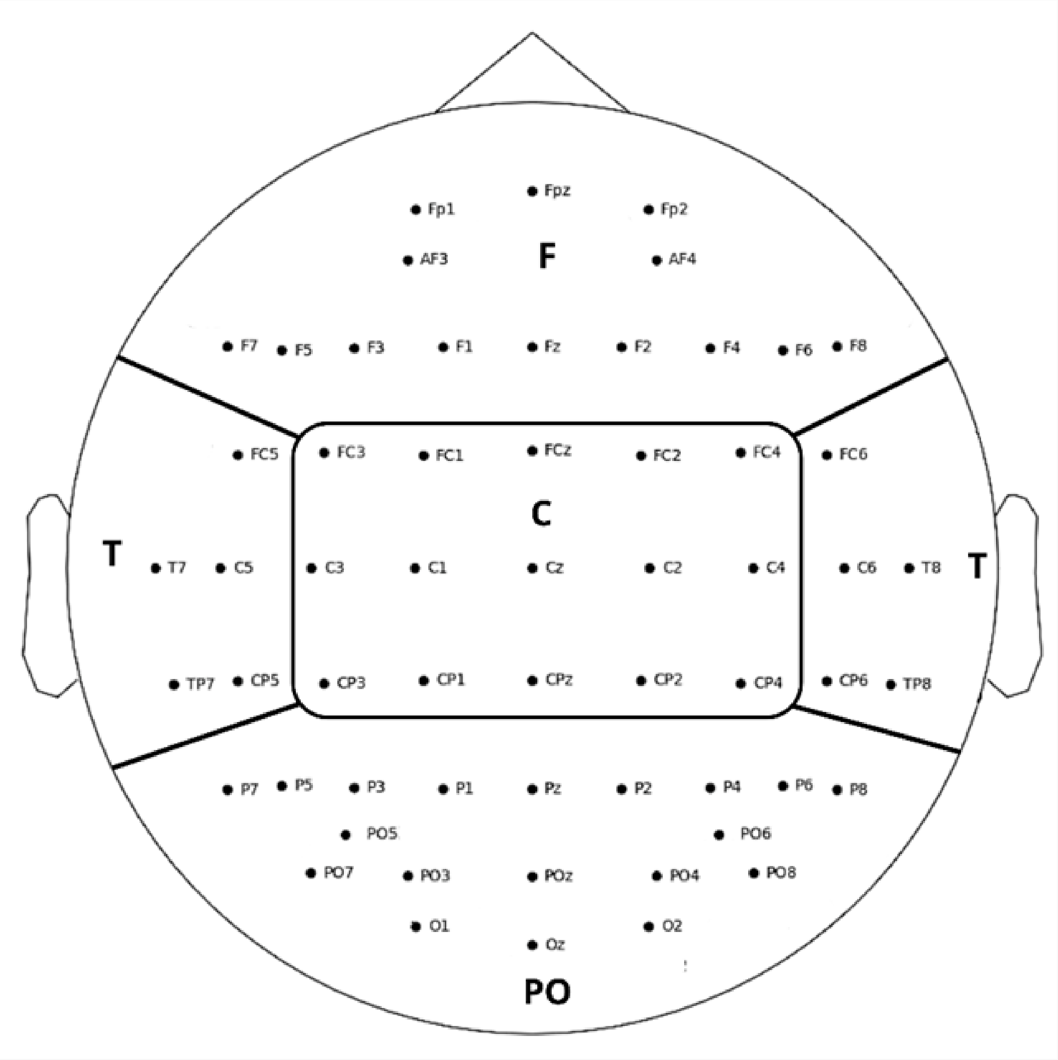
"Data normalization is a process employed in data analysis and statistics to transform variables onto a consistent scale or comparable range. Its primary aim is to mitigate scale effects and ensure that variables exhibit a similar distribution, simplifying data comparison and analysis."

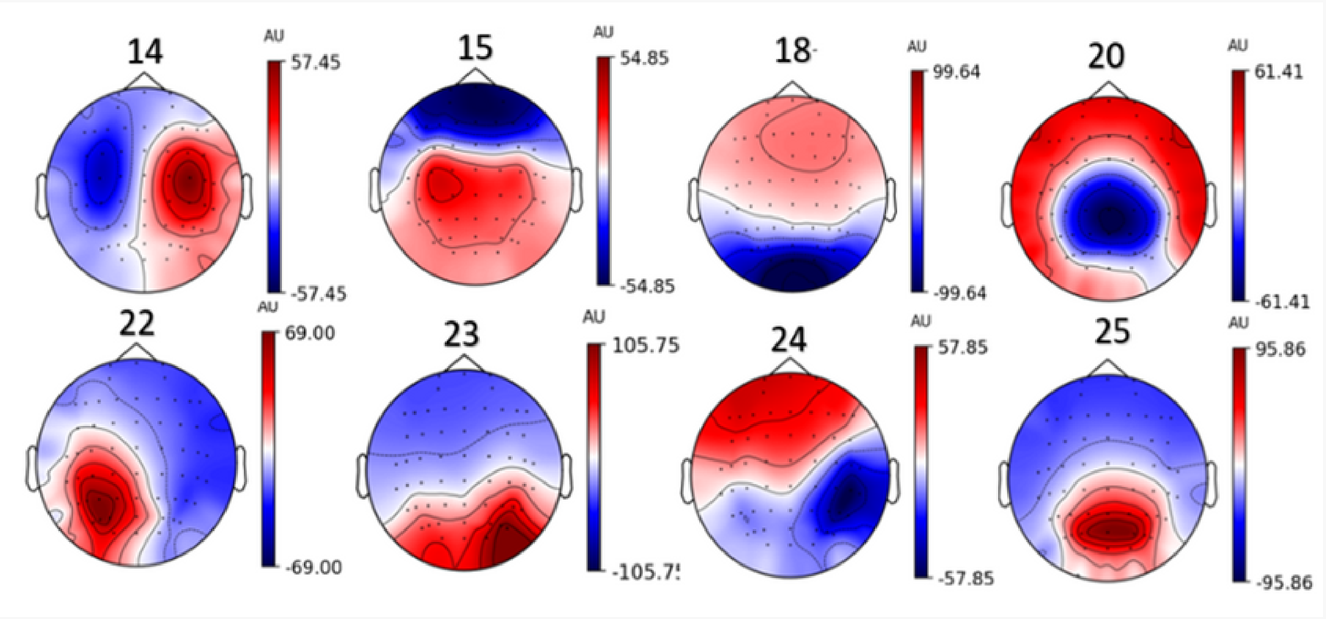
gruneco
grupo neuropsicología y conducta

V Research Results

Harmonize the database

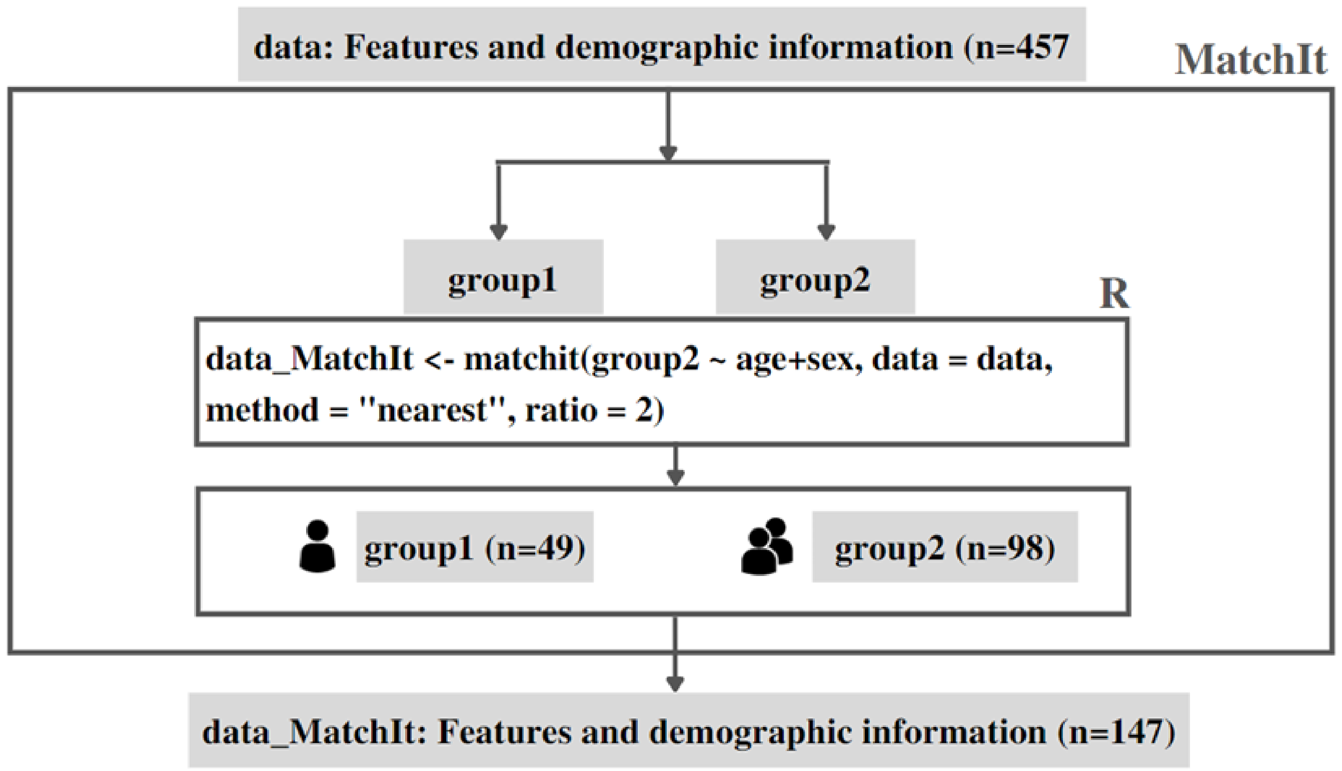






gruneco
grupo neuropsicología y conducta

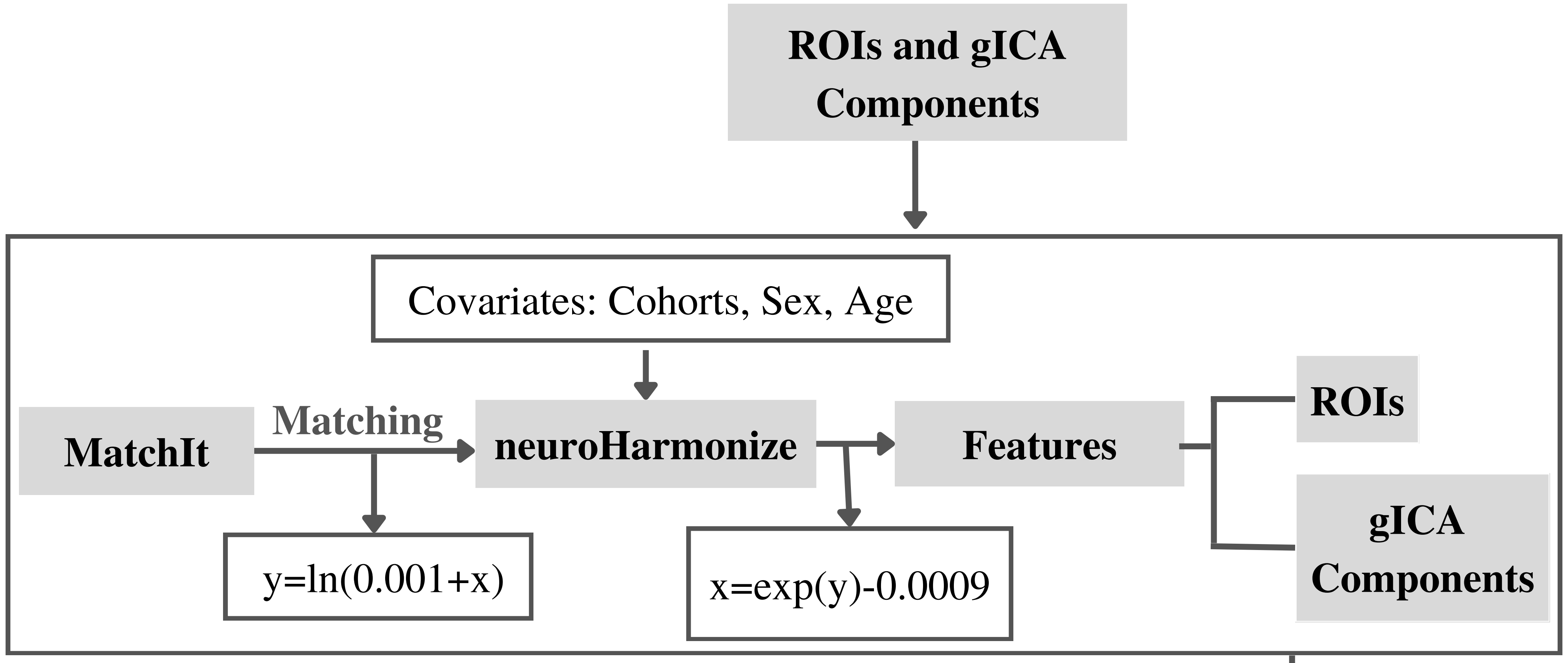
V Research Results
Harmonize the database



		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			

Healthy: Control Group + G2 Group

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



The neuroHarmonize tool, leveraging the combat method previously employed in harmonizing cohorts with genetic data and medical images, corrects site effects while intentionally preserving the effects of covariates.

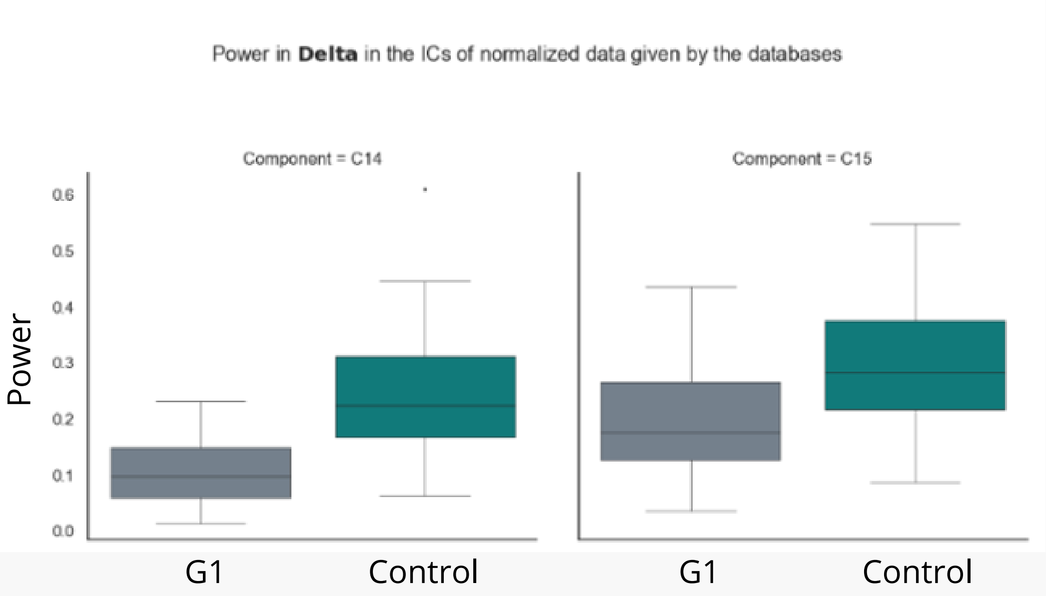
gruneco
grupo neuropsicología y conducta

V Research Results
Harmonize the database

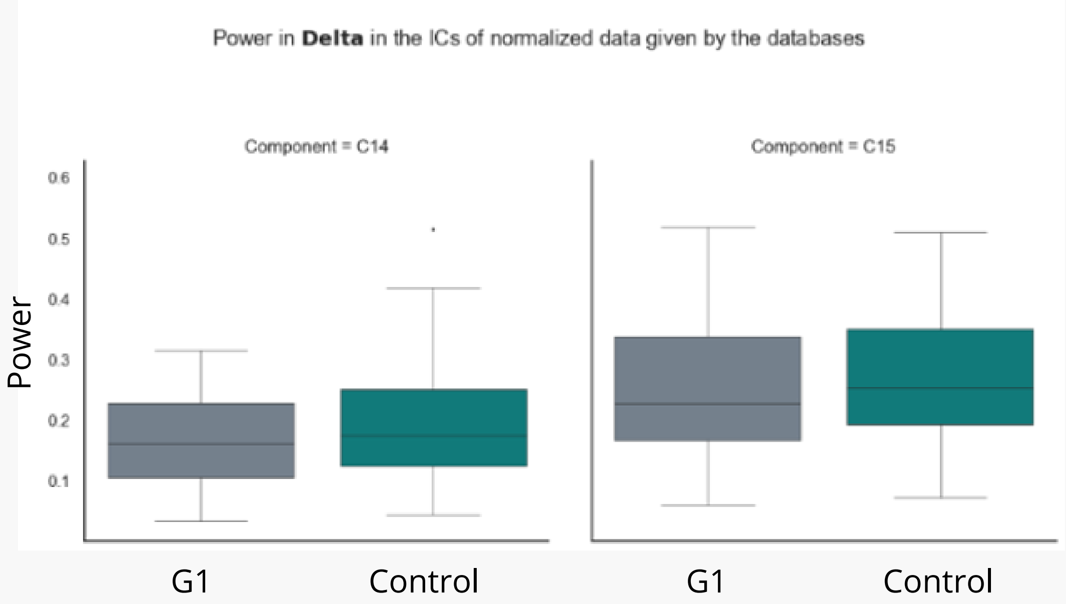
Table 9 Summary of the effect size by feature extraction between the control groups of the different cohorts.

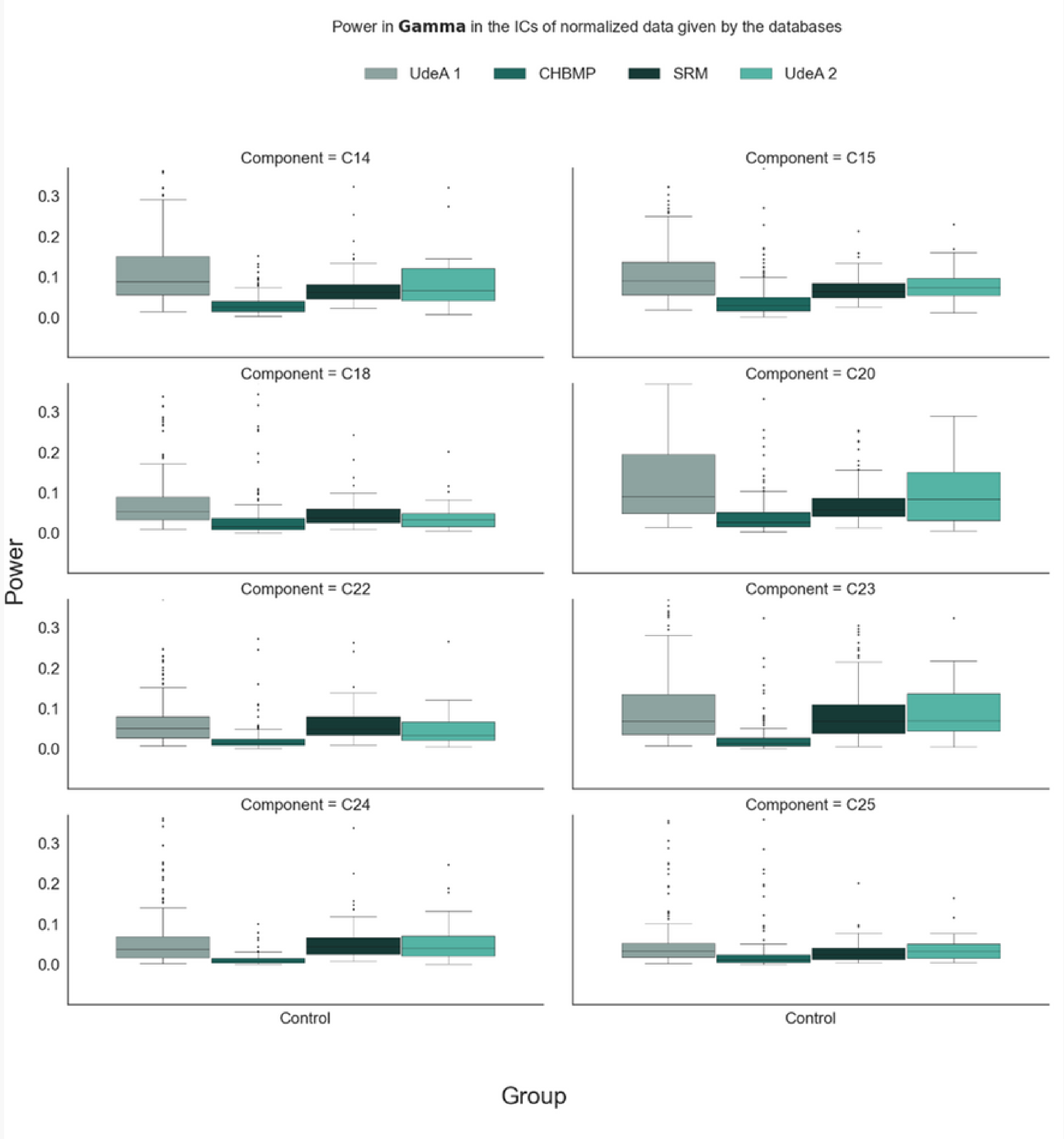
Controls (UdeA1 vs UdeA2 vs SRM vs CHBMP)														
	Delta		Theta		Alpha1		Alpha2		Beta1		Beta2		Gamma	
neuroHarmonize	Without	With	Without	With	Without	With	Without	With	Without	With	Without	With	Without	With
Relative Power														
Max	0.92	0.07	0.59	0.05	0.69	0.09	0.49	0.06	1.07	0.10	0.93	0.08	1.13	0.32
Average	0.57	0.05	0.16	0.02	0.45	0.06	0.24	0.05	0.83	0.06	0.60	0.05	0.69	0.17
Min	0.25	0.01	0.01	0.00	0.18	0.05	0.05	0.02	0.45	0.01	0.04	0.08	0.32	0.05

(a) Data without neuroHarmonize



(b) Data with neuroHarmonize





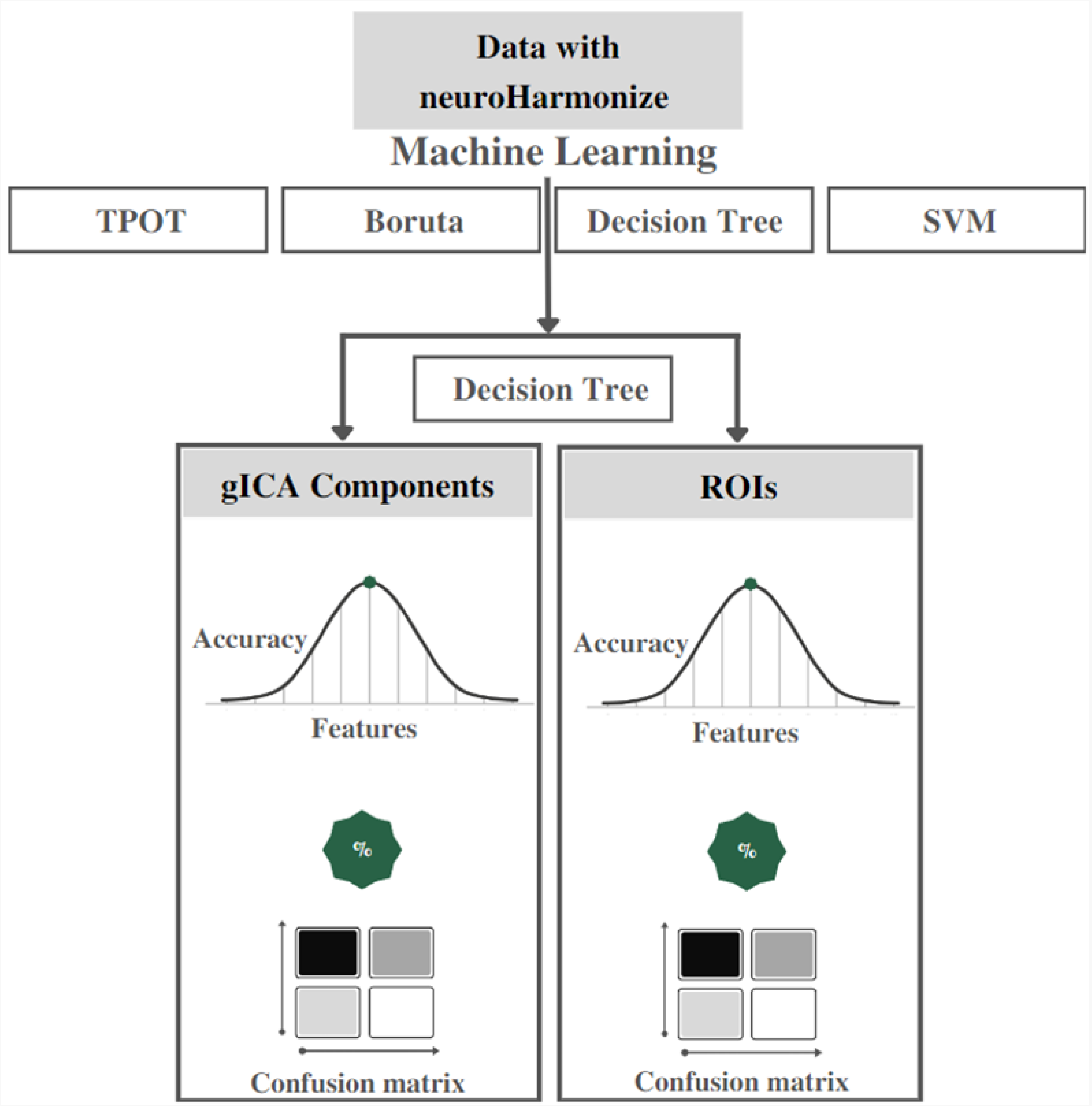
grupo neuropsicología y conducta

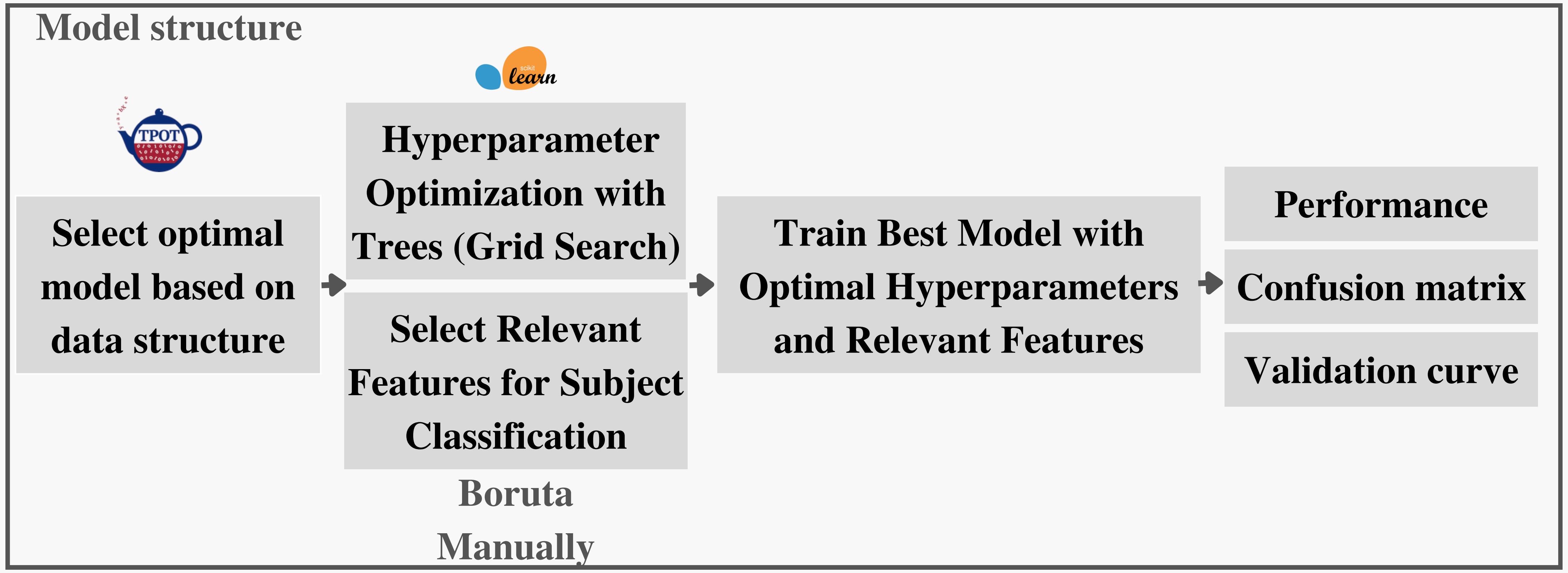
KEY ELEMENTS OF HARMONIZE THE DATABASE

- The **automated** PREP **preprocessing** pipeline incorporates a previously introduced gICA, as discussed in one of the background references.
- The normalization stage aims to **reduce variability** across EEG.
- Processing involves intercepting the minimum 58 channels present in all databases to implement **consistent analysis** across cohorts.
- Combat corrects site effects while deliberately preserving the **effects of covariates.**
- Following harmonization, the **reduction in effect size** observed among healthy groups across the four cohorts suggests increased comparability at the individual level

V Research Results

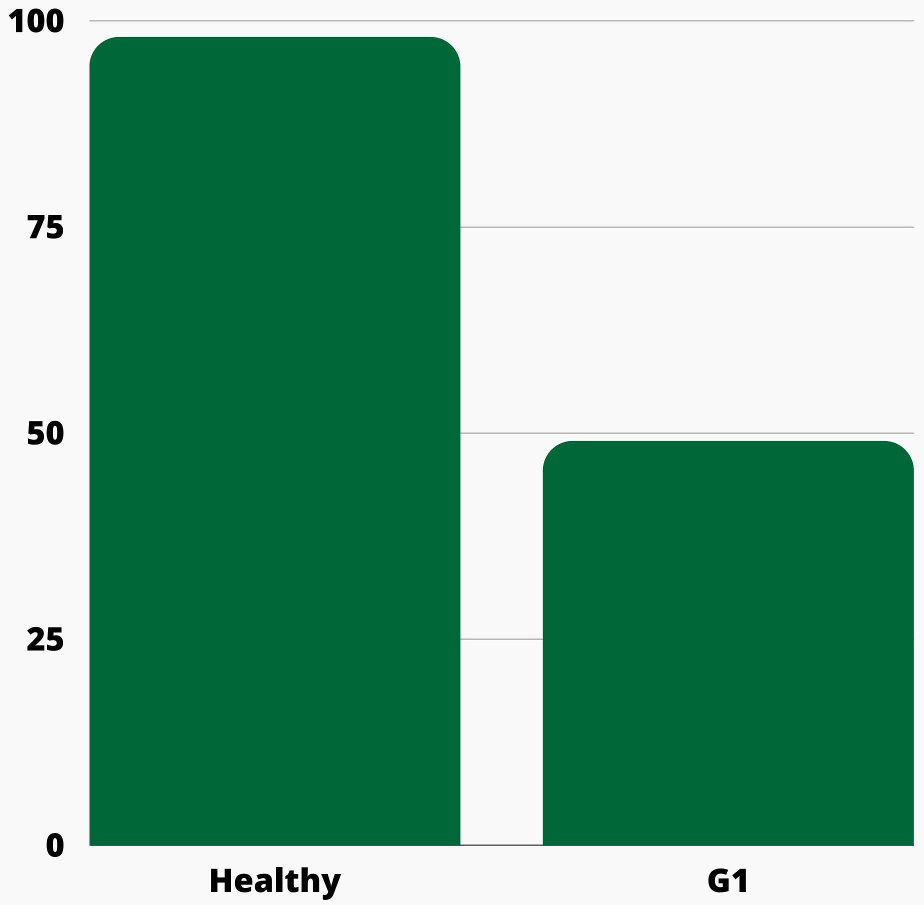
Design a machine learning model





		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			

Healthy: Control Group + G2 Group



V Research Results

Design a machine learning model

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

Participant_id	Feature_Component_Band
sub_Group001	Value

```
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
           ...  ...                       ...                       ...
```

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.

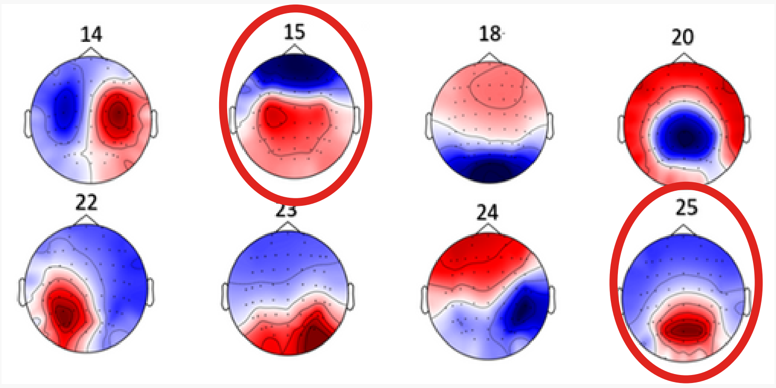
gruneco
grupo neuropsicología y conducta

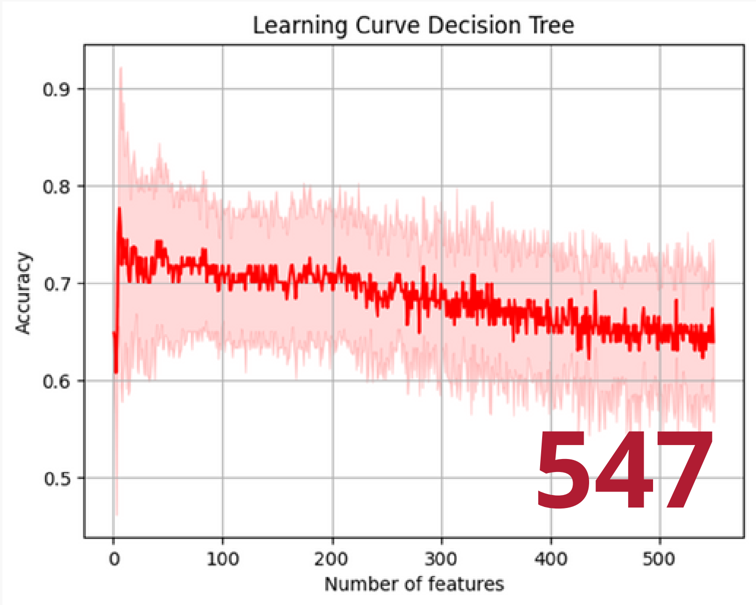
V Research Results

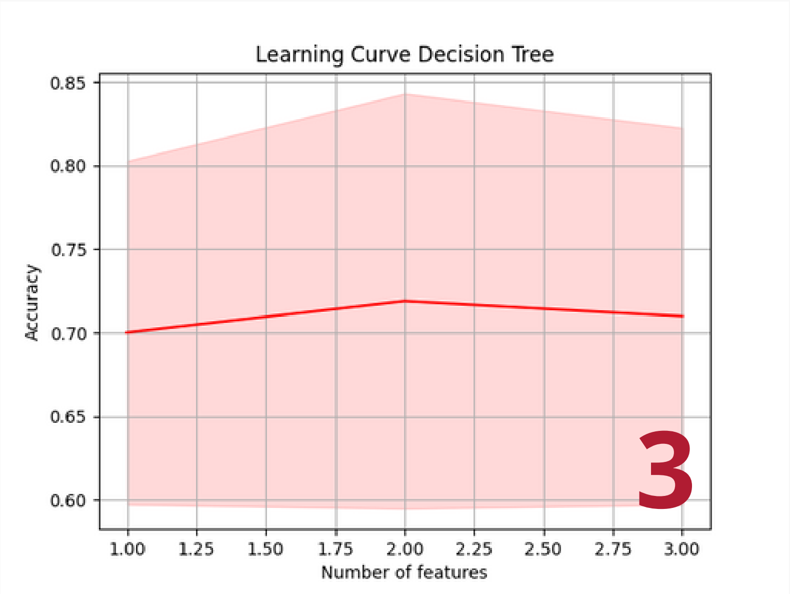
Design a machine learning model

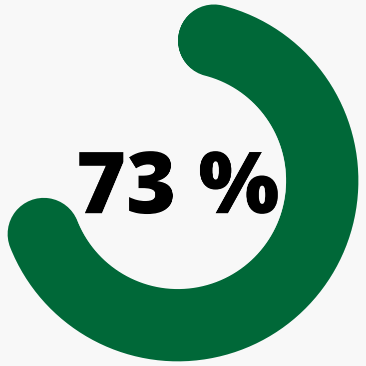
3

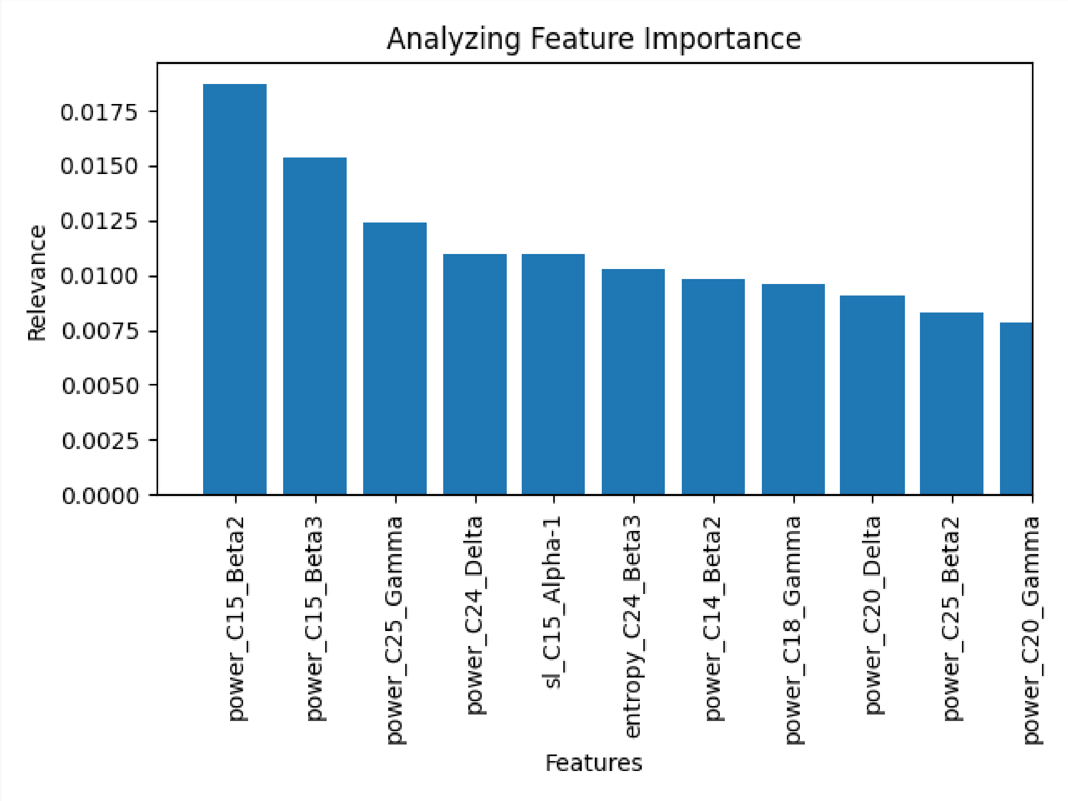
58x25

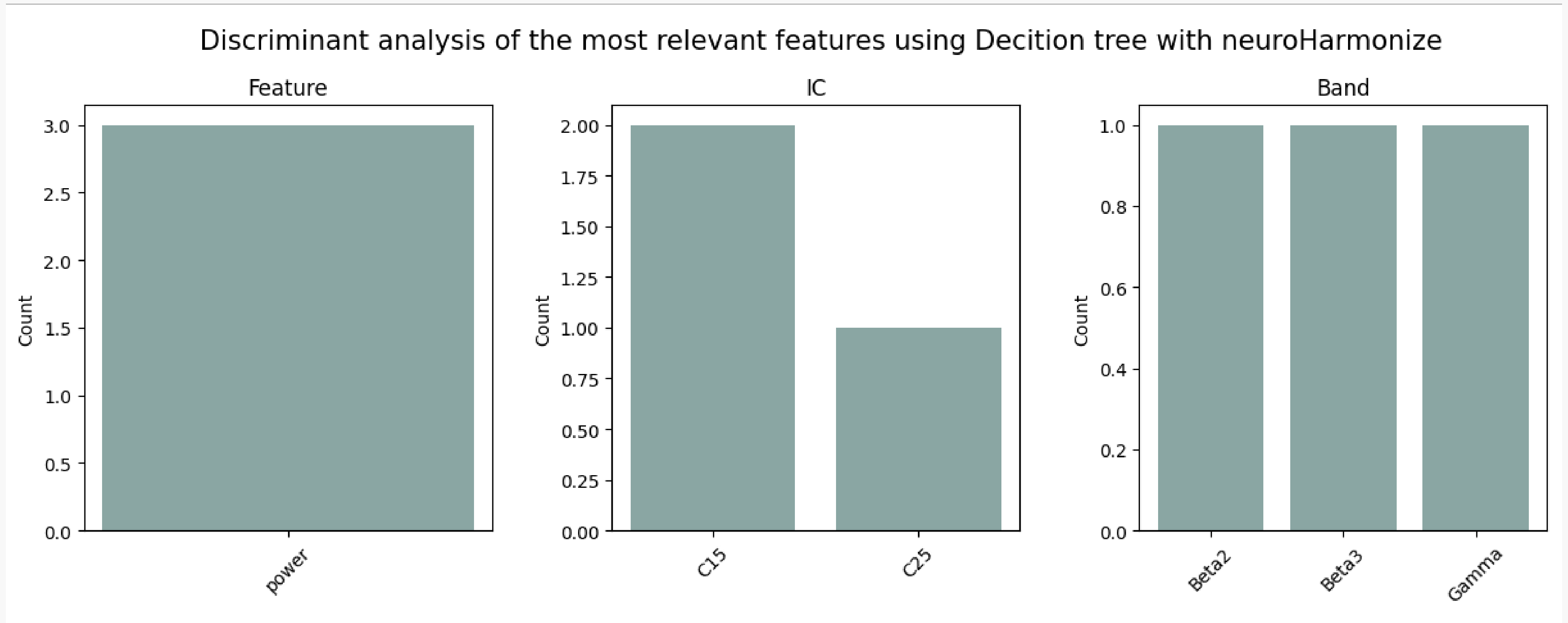










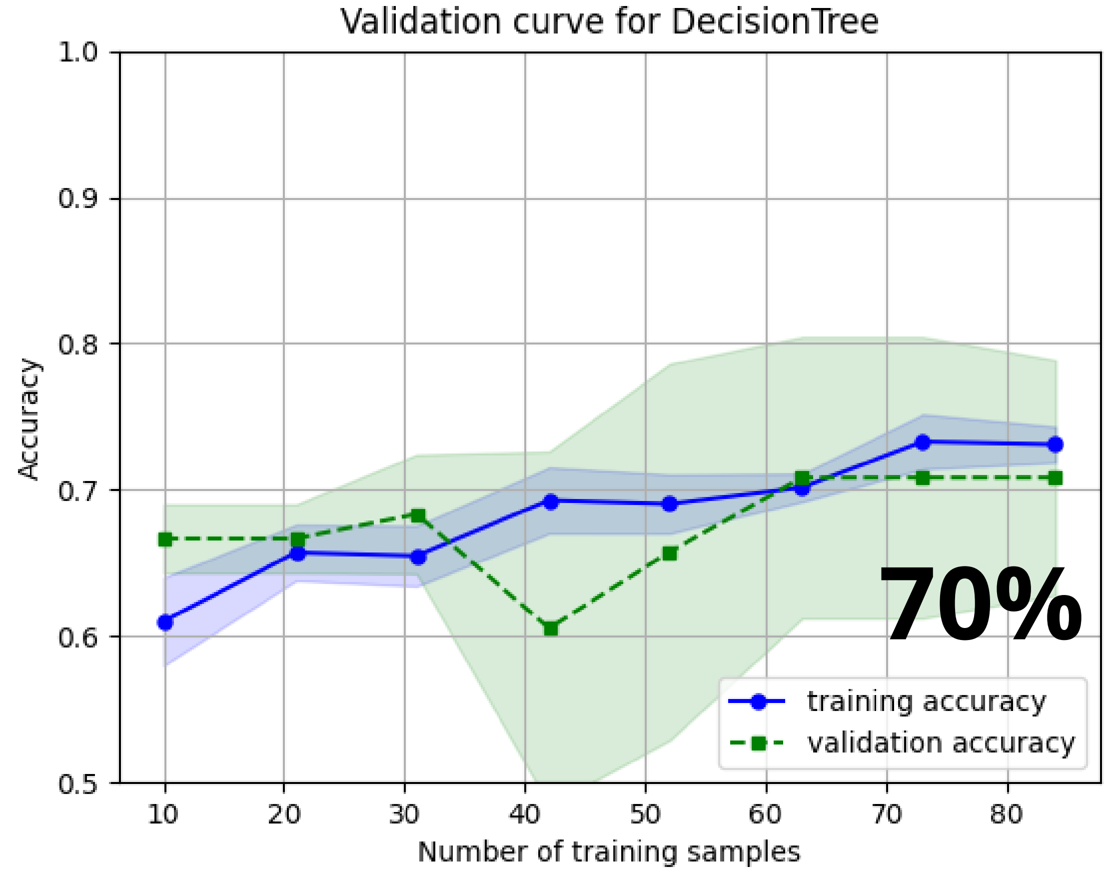


V Research Results

Design a machine learning model

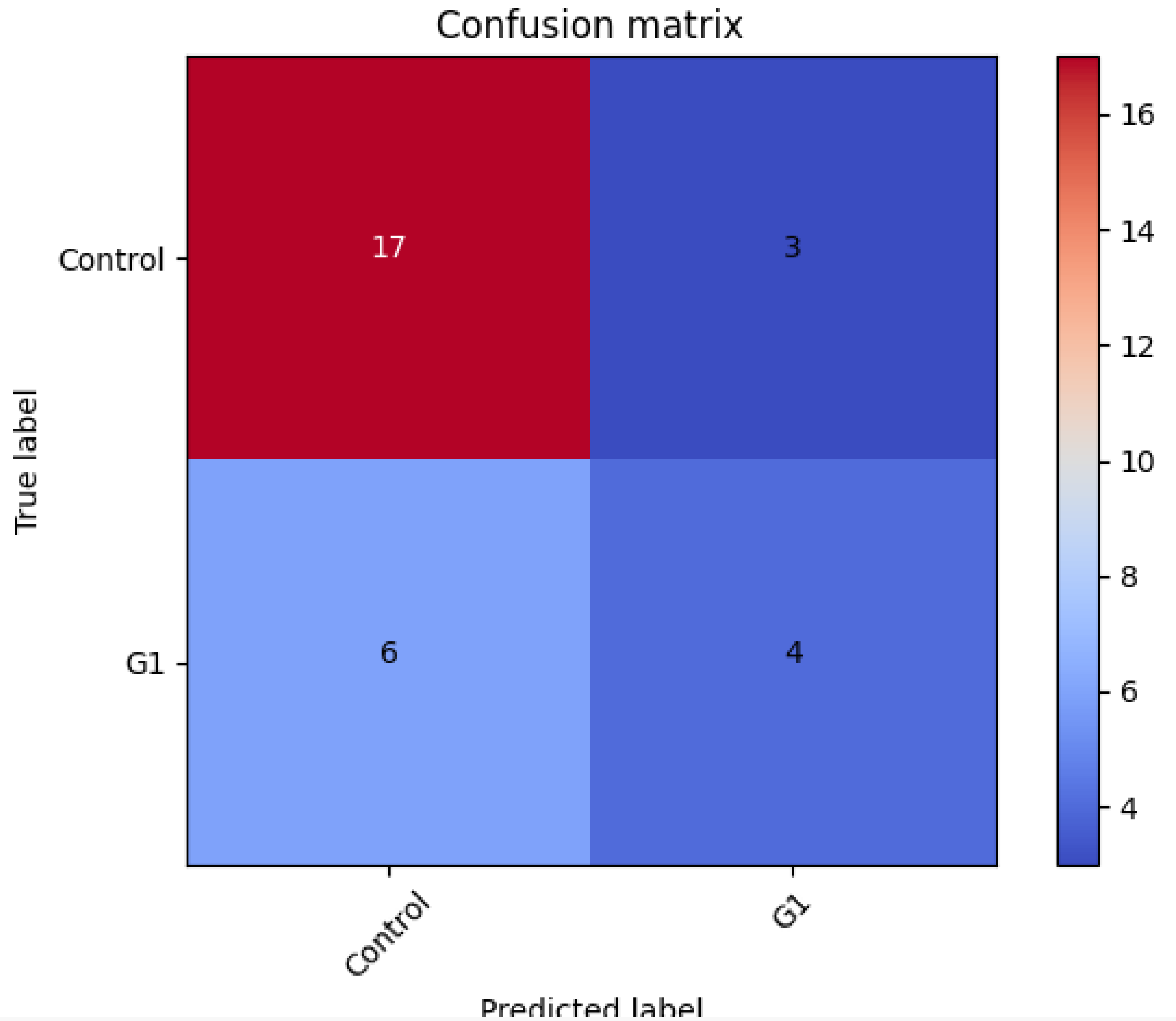
Test 20%

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



---Computer Precision---

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

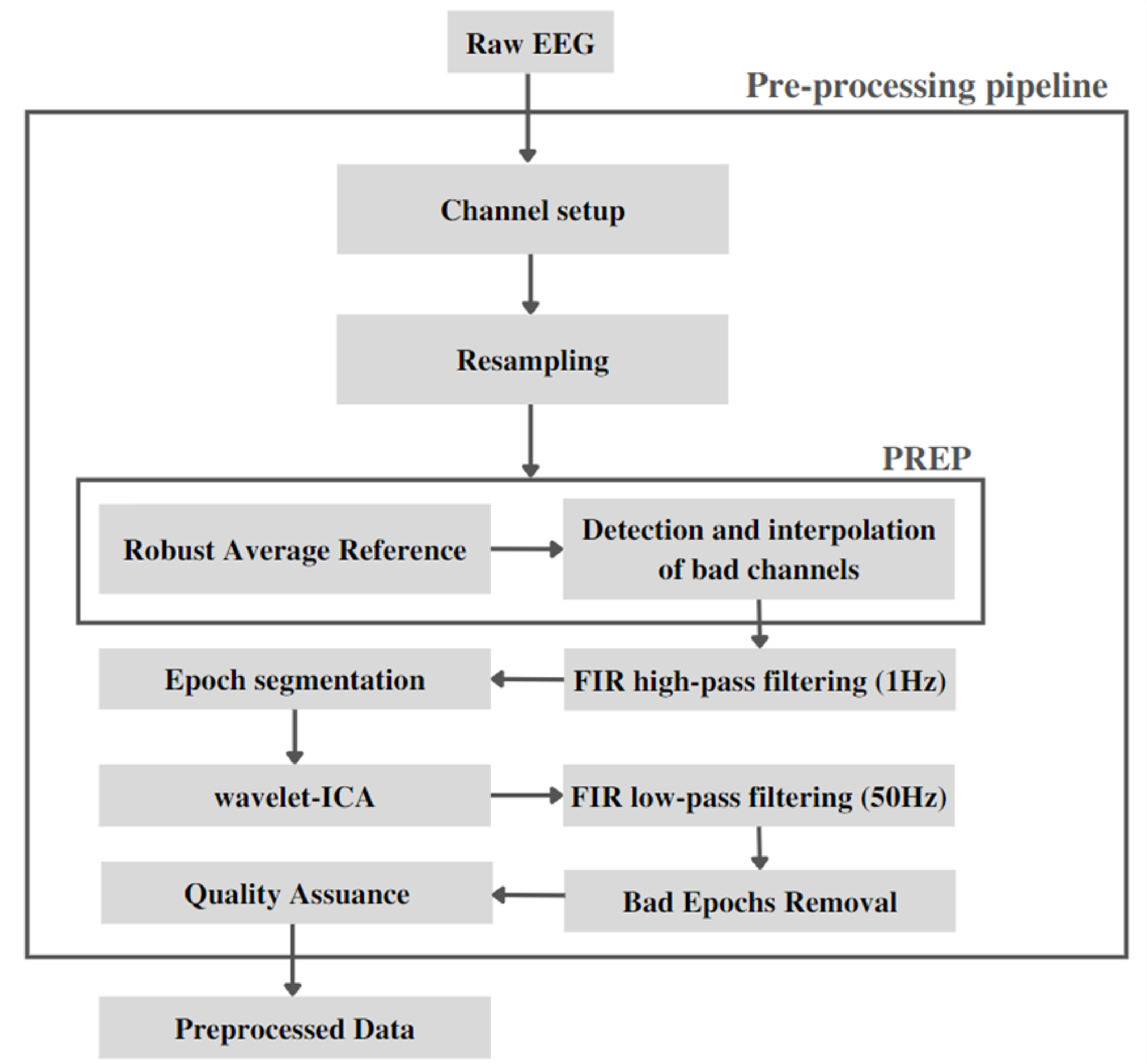


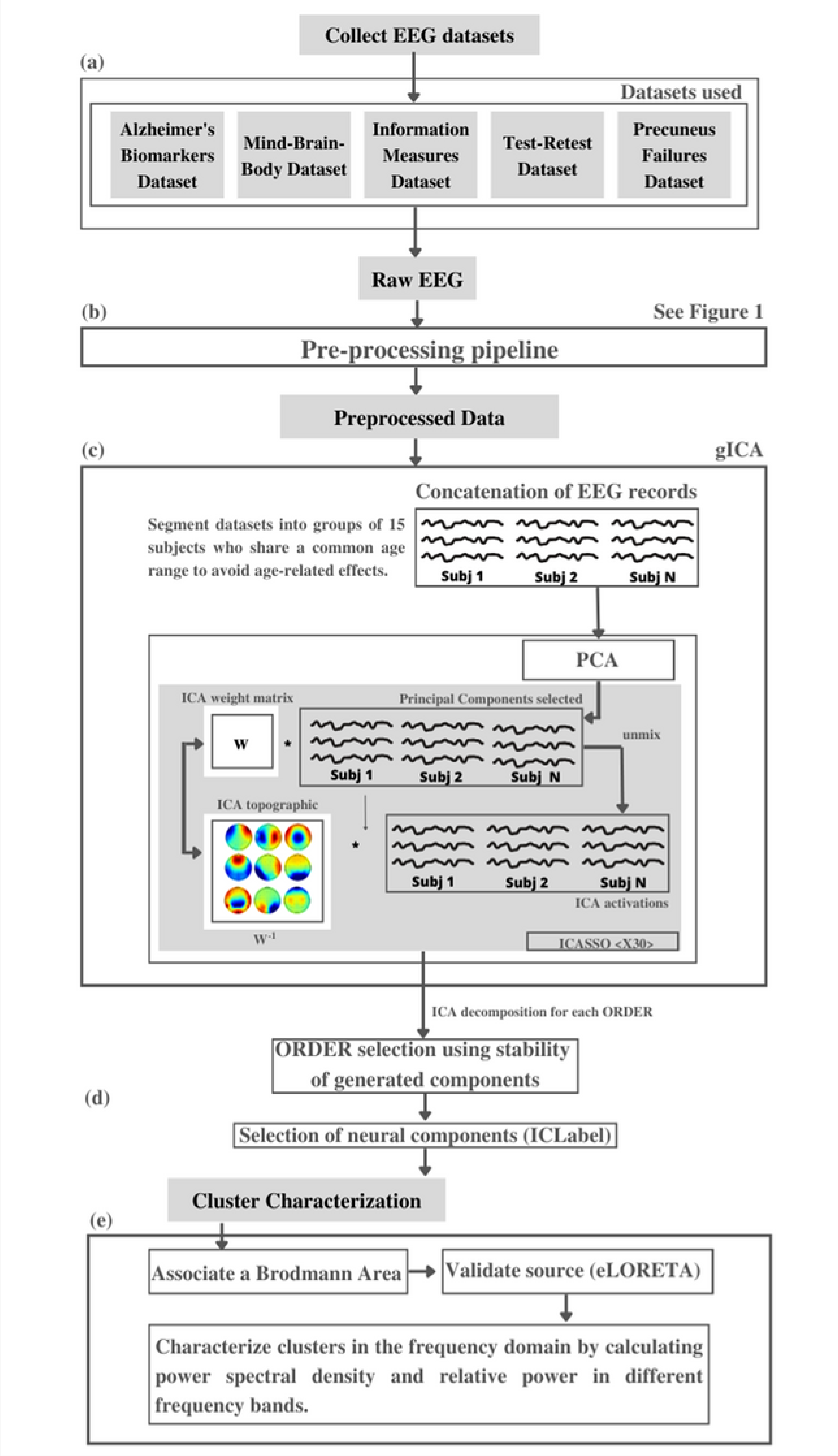
KEY ELEMENTS OF DESIGN A MACHINE LEARNING MODEL

- The **model incorporates 147 subjects** with 547 features for components and 386 features for regions.
- Three features were sufficient to achieve a **73% accuracy** in the model, including the relative power metric, components 15 and 25, and the Beta and Gamma bands.
- The validation curve exhibited a **accuracy of 70%**, while the confusion matrix effectively classified 17 controls out of 20 and 4 carriers out of 10, utilizing 20% of the data for each group.

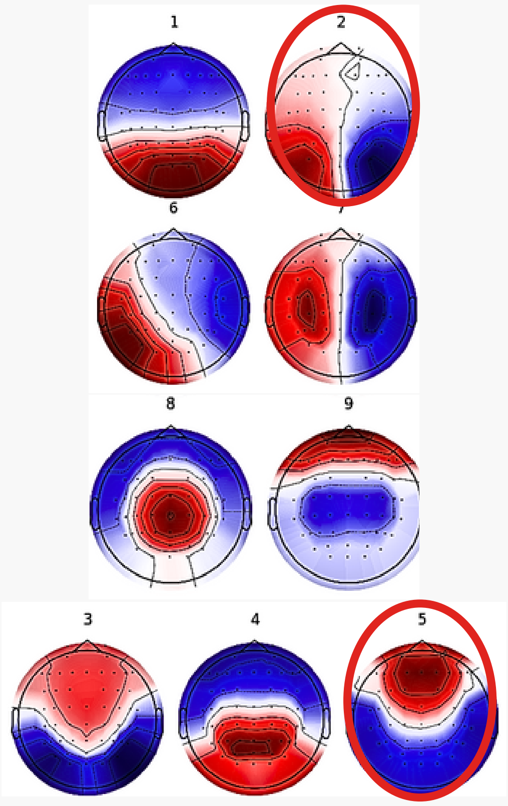
V Research Results

Design a machine learning model

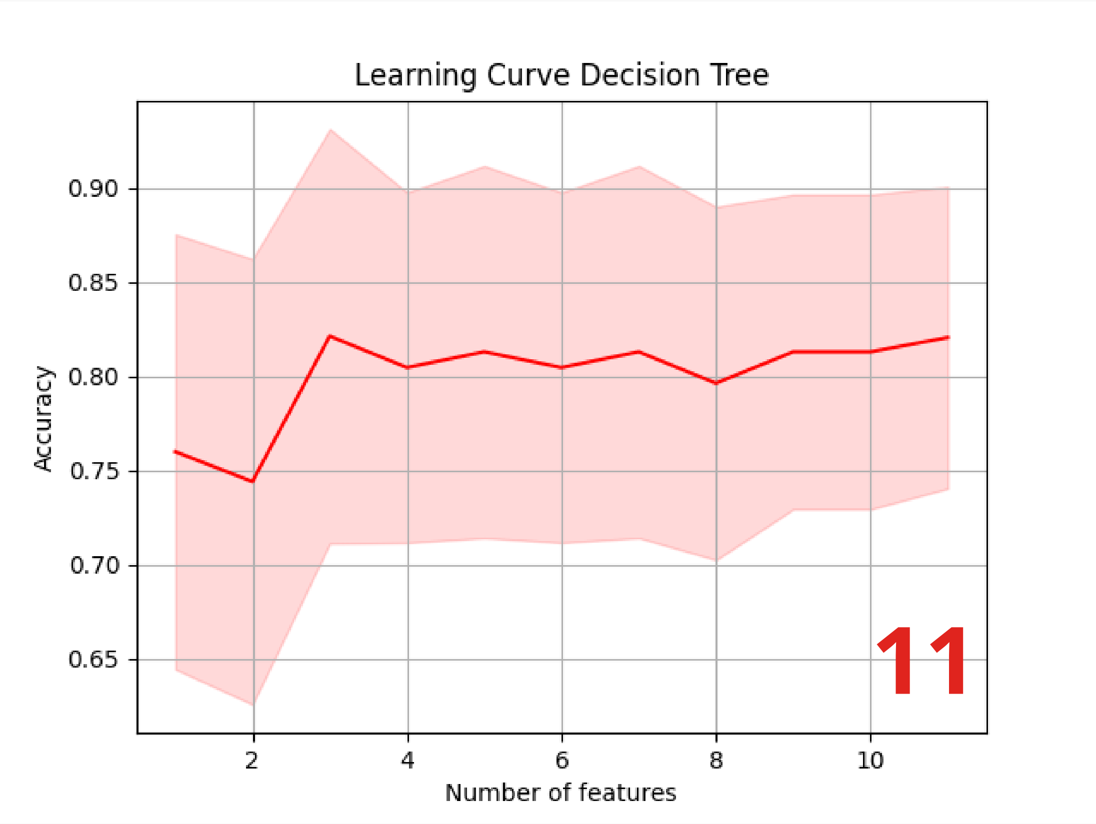




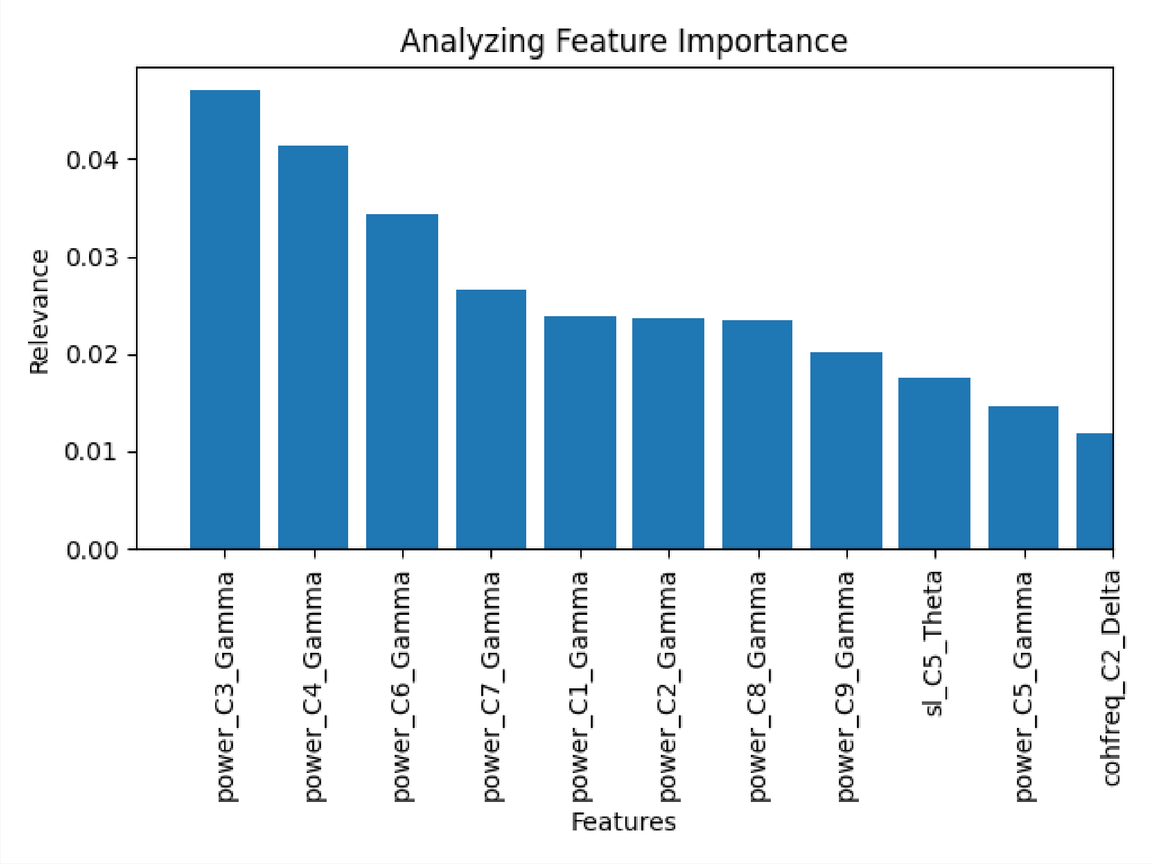
54x10



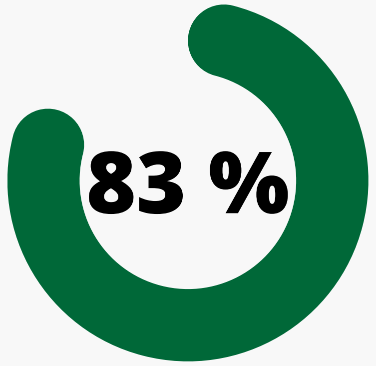
1
2
6
7
8
9
3
4
5



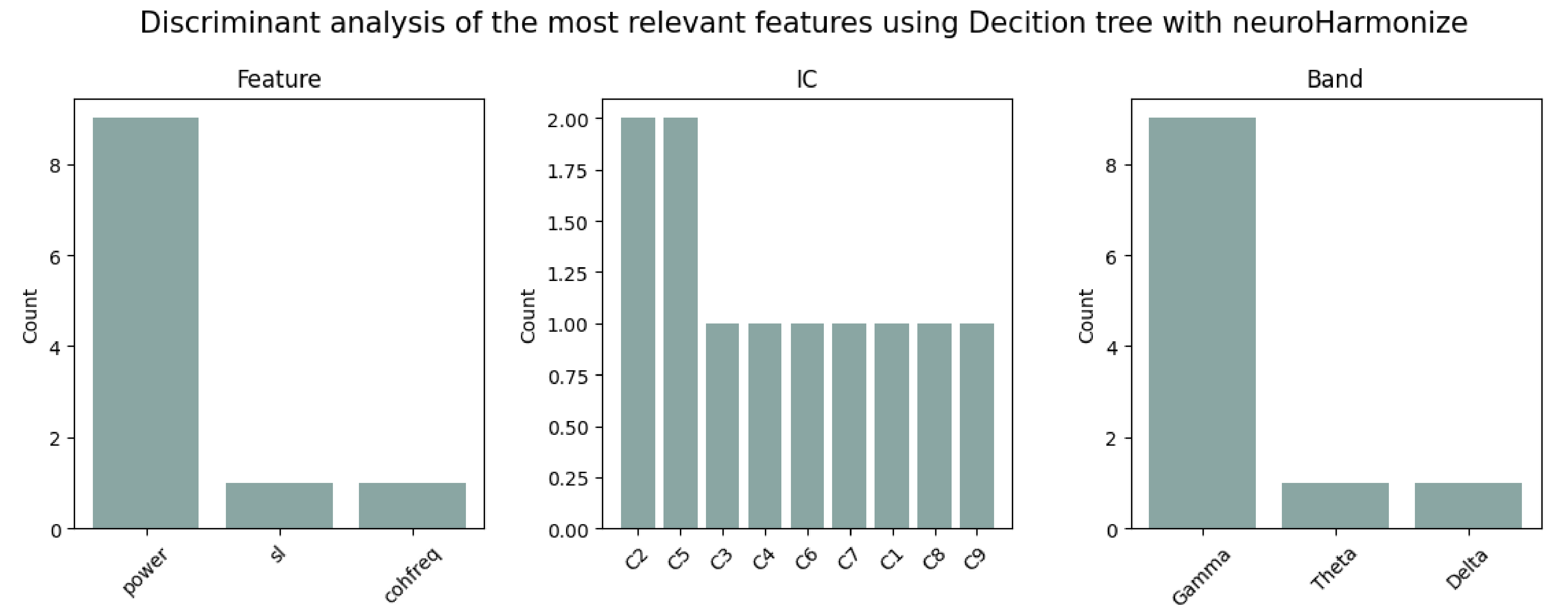
Learning Curve Decision Tree
Accuracy
0.90
0.85
0.80
0.75
0.70
0.65
2
4
6
8
10
Number of features
11



Analyzing Feature Importance
Relevance
0.04
0.03
0.02
0.01
0.00
power_C3_Gamma
power_C4_Gamma
power_C6_Gamma
power_C7_Gamma
power_C1_Gamma
power_C2_Gamma
power_C8_Gamma
power_C9_Gamma
sl_C5_Theta
power_C5_Gamma
cohfreq_C2_Delta
Features



83 %



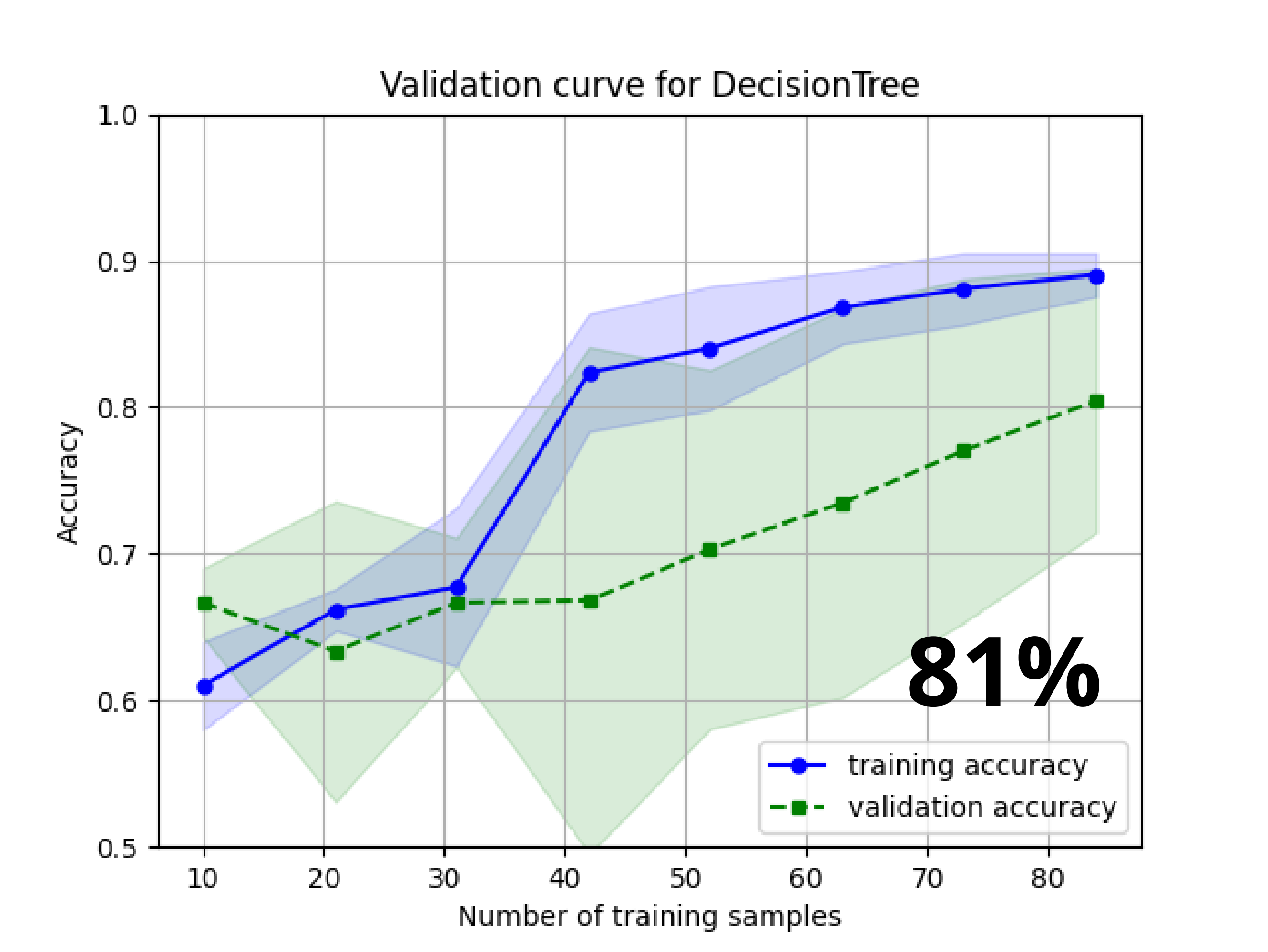
Discriminant analysis of the most relevant features using Decition tree with neuroHarmonize
Feature
Count
0
2
4
6
8
power
sl
cohfreq
IC
Count
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
C2
C5
C3
C4
C6
C7
C1
C8
C9
Band
Count
0
2
4
6
8
Gamma
Theta
Delta

V Discussions

Reproducible Neuronal Components found

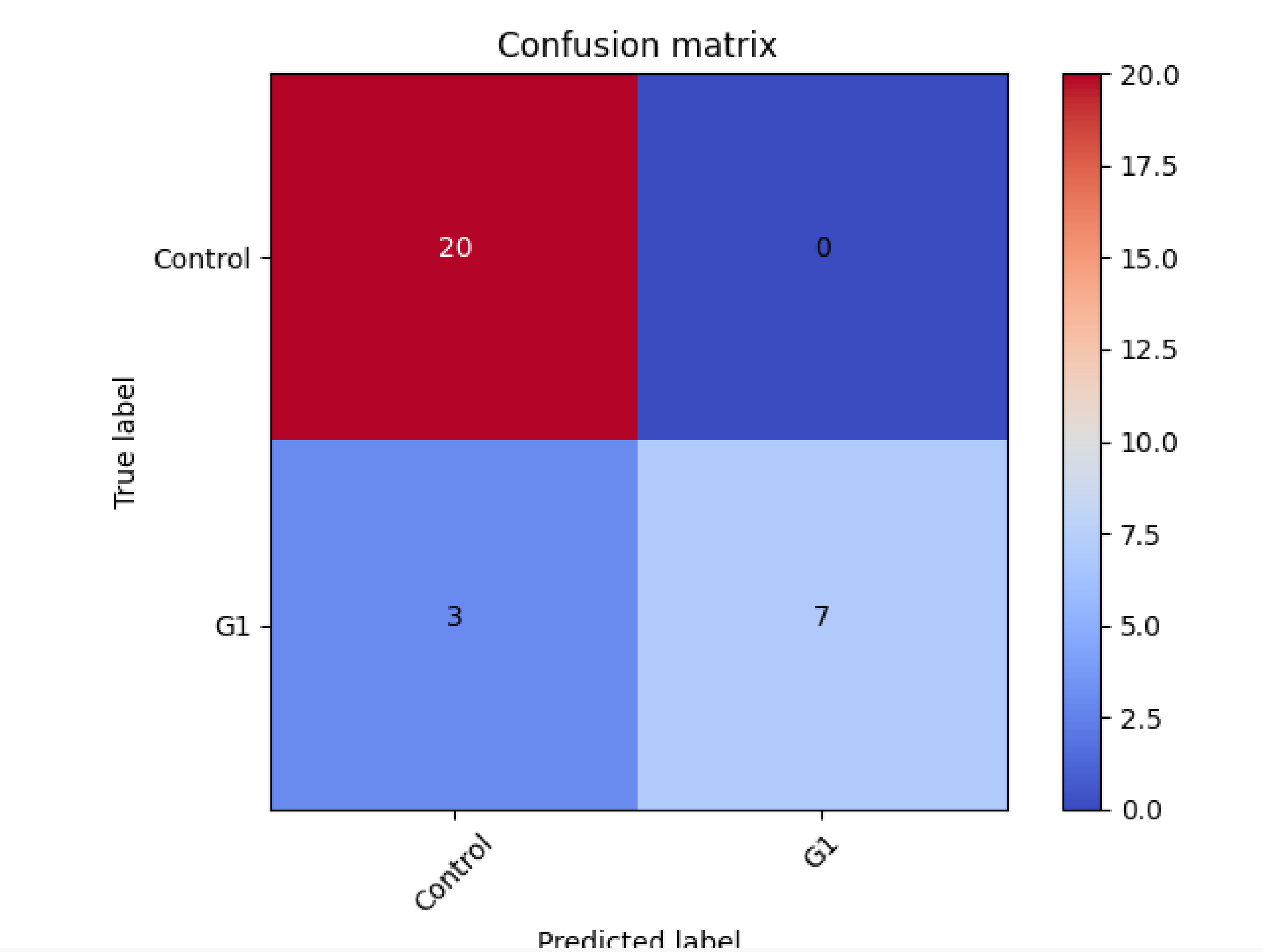
Test 20%

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



---Computer Precision---

Precision: 1.0
Recall: 0.6
F1-score: 0.75
Accuracy: 0.81
Standard deviation: 0.08



KEY ELEMENTS OF DESIGN A MACHINE LEARNING MODEL

- A **new gICA** was utilized, created using **5 databases** but applying the matrix to the same 147 matched subjects from the 4 cohorts of the project.
- Eleven features were sufficient to achieve a **83% accuracy** in the model, including the relative power metric, sl, coherence, components 2 and 5, and the Delta, Theta and Gamma bands.
- The validation curve exhibited a **accuracy of 81%**, while the confusion matrix effectively classified 20 controls out of 20 and 7 carriers out of 10, utilizing 20% of the data for each group.

Table of Contents

Build and standardize a database

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.

Harmonize the database

Each step of the process yielded promising results, instilling confidence in the **processing pipeline** until the matching phase. 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.

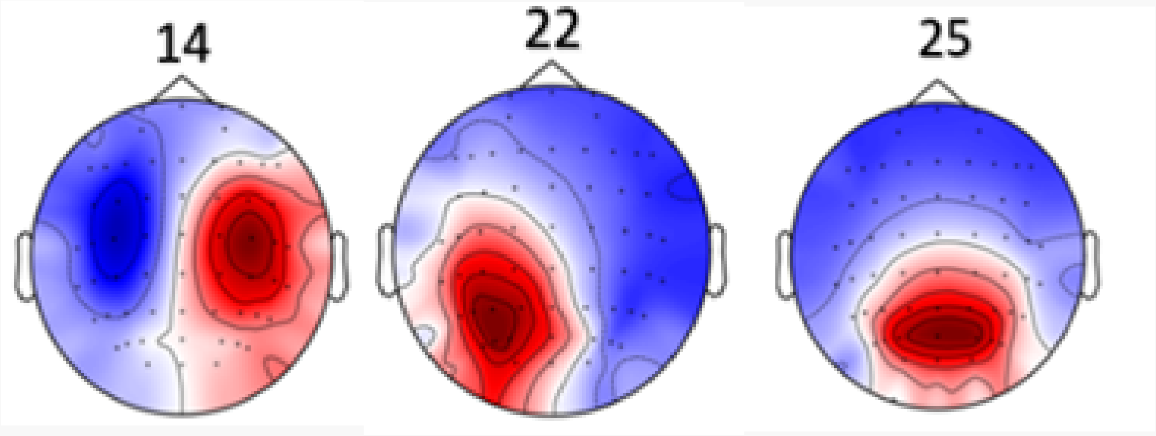
Design a machine learning model

SVM - gICA (58X25) - 1 Cohort

83 %

β

γ



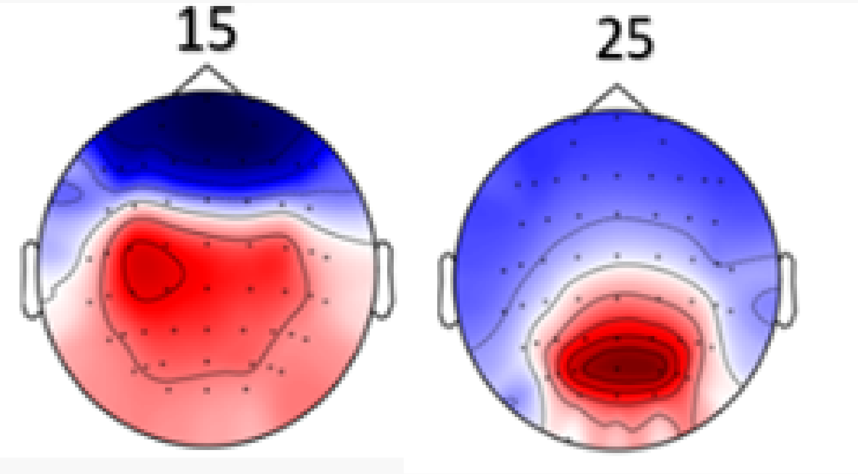
Relative Power and Cross Frequency

DT - gICA (58X25) - 4 Cohort

70 %

β

γ



Relative Power

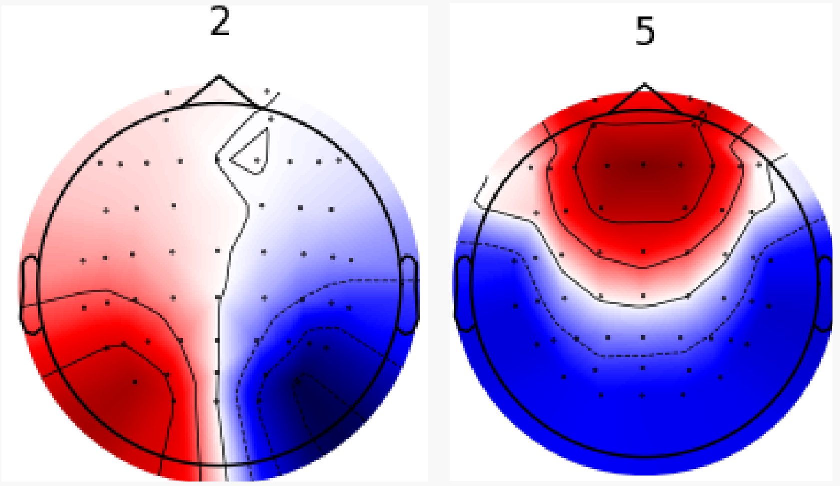
DT - gICA (54X10) - 4 Cohort

81 %

γ

θ

Δ



Relative Power and Synchronization Likelihood

V Conclusions

1. Incorporating EEG usage significantly contributes to **cost reduction**, underscoring the economic benefits of leveraging this technology in research and clinical applications.

2. Exploring genetic populations offers a promising avenue for uncovering novel insights in the **preclinical** context, highlighting the potential of genetics in advancing our understanding of various conditions.

3. The use of gICA proves valuable in offering **stable and reproducible** insights into the spatiotemporal structure and dynamics of EEG, demonstrating its efficacy in harmonizing data before, during, and after experimental events.

4. Leveraging openly **accessible databases** contributes to cost reduction and research efficiency, showcasing the benefits of tapping into existing datasets for scientific inquiry.

Table of Contents

V Limitations

1. Results should be generalized cautiously as at the moment they are applicable only to the population with the **PSEN1-E280A variant**.

2. The study faces a constraint in the availability of **comprehensive datasets**, which hinders the depth and breadth of the analysis and may impact the generalizability of findings.

3. The findings obtained using setups with 58 or 54 electrodes should be tested on smaller and more **portable** setups to assess their applicability in different settings.

Table of Contents

1. To further enhance the model's applicability, future research should focus on replicating these results in **larger cohorts**.

2. Addressing **sample size imbalances** and understanding the impact of linear harmonization methods on classification techniques are highlighted as crucial considerations to strengthen the reliability and robustness of harmonization techniques across various research areas.

3. A potential avenue for future work could involve the evaluation of generalized group Independent Component Analysis (gICA) matrices that **aggregate data from multiple databases** originating from different regions and populations. This approach could serve as a valuable tool for implementing harmonized processing pipelines and expanding research possibilities.

Table of Contents

VI Contributions

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

ALZHEIMER'S ASSOCIATION
AAIC 23

David Aguillon[1,2,3]; Alejandro Guerrero[2,3]; Daniel Vasquez[1]; Valeria Cadavid[2]; Verónica Henao[2]; Ximena Suárez[2]; Alberto Jaramillo-Jimenez[1]; Isabel Marquez[1]; Francisco Lopera[1,2]; David Pineda[1,2]; Carlos Tobón[1,2], John Ochoa[2].

1. Grupo de Neurociencias de Antioquia, Facultad de Medicina, Universidad de Antioquia
2. Grupo Neuropsicología y Conducta, Facultad de Medicina, Universidad de Antioquia
3. Semillero de Investigación Sinapsis, Facultad de Medicina, Universidad de Antioquia

Alzheimer's Association International Conference

Dr. David Aguillón

Dr. John Ochoa

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

Veronica Henao Isaza[1], Valeria Cadavid Castro[1], Luisa Zapata Saldarriaga[1], Yorguin Mantilla-Ramos[1], Carlos Tobón Quintero[1], Jazmin Suarez-Revelo[1], and John Ochoa Gómez[1]

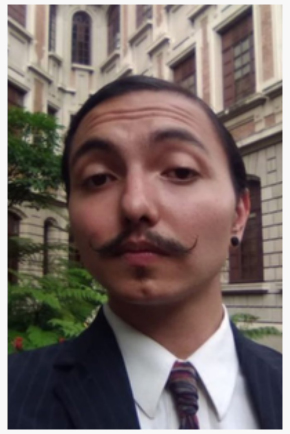
DOI: 10.22541/au.168570191.12788016/v1
Under review: Biomedical Signal Processing and Control, Q1 , SJR 2022

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

Alberto Jaramillo-Jimenez[a,b,c,d,e,f,*], Diego A. Tovar-Rios[a,b,g,h], Johann Alexis Ospina[i], Yorguin-Jose Mantilla-Ramos[d,f], Daniel Loaiza-López[d,f], Verónica Henao Isaza[d,f], Luisa María Zapata Saldarriaga[d,f], Valeria Cadavid Castro[d,f], Jazmin Ximena Suarez-Revelo[c,d], Yamile Bocanegra[c,d], Francisco Lopera[c], David Antonio Pineda-Salazar[c,d], Carlos Andrés Tobón Quintero[c,d,j], John Fredy Ochoa-Gomez[d], Miguel Germán Borda[a,b,k], Dag Aarsland[a,b,l], Laura Bonanni[m], Kolbjørn Brønnick[a,b]

[a] *Centre for Age-Related Medicine (SESAM), Stavanger University Hospital. Stavanger, Norway*
[b] *Faculty of Health Sciences, University of Stavanger. Stavanger, Norway*
[c] *Grupo de Neurociencias de Antioquia, Universidad de Antioquia, School of Medicine. Medellín, Colombia*
[d] *Grupo Neuropsicología y Conducta, Universidad de Antioquia, School of Medicine. Medellín, Colombia*
[e] *Semillero de Investigación SINAPSIS, Universidad de Antioquia, School of Medicine. Medellín, Colombia*
[f] *Semillero de Investigación NeuroCo, Universidad de Antioquia, School of Medicine & School of Engenieering. Medellín, Colombia*
[g] *Universidad del Valle, Grupo de Investigación en Estadística Aplicada - INFERIR, Faculty of Engineering, Santiago de Cali, Colombia*
[h] *Universidad del Valle, Prevención y Control de la Enfermedad Crónica - PRECEC, Faculty of Health, Santiago de Cali, Colombia*
[i] *Facultad de Ciencias Básicas, Universidad Autónoma de Occidente, Santiago de Cali, Colombia*
[j] *Área Investigación e Innovación, Hospital Alma Mater de Antioquia. Medellín, Colombia*
[k] *Semillero de Neurociencias y Envejecimiento, Pontificia Universidad Javeriana, Ageing Institute, Medical School. Bogotá, Colombia*
[l] *Department of Old Age Psychiatry, Institute of Psychiatry, Psychology, and Neuroscience, King's College London. London, UK*
[m] *Department of Medicine and Aging Sciences, G. d'Annunzio University. Chieti, Italy*

DOI: 10.1016/j.clinph.2023.03.363. PMID: 37146531
Clinical Neurophysiology, Q1, SJR 2022



Dr. Alberto Jaramillo-Jimenez

NeuroCo
Neurociencias Computacionales
Universidad de Antioquia

gruneco
grupo neuropsicología y conducta

VI Relevant publications and works

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

John Fredy Ochoa-Gómez[1*†],
Yorguin-José Mantilla-Ramos[1,3†], Verónica Henao Isaza[1,3],
Carlos Andrés Tobón[1], Francisco Lopera[2],
David Aguillón[2], Jazmín Ximena Suárez[1]

DOI: 10.1101/2023.11.14.566952
Pre print biorxiv

Dr. John Ochoa

gruneco
grupo neuropsicología y conducta

VI Relevant publications and works

Research internship

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.

SAPIENZA
UNIVERSITÀ DI ROMA

Nov 2023

Bioengineering,
Universidad de Antioquia

ADVISOR

Ph.D. John Fredy Ochoa
Gomez

STUDENT

Verónica Henao
Isaza

Thank you for listening!

Reference

About me

Table of Contents

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.

There is evidence that the pathophysiological processes in AD begin decades before the manifestation of clinical symptoms.

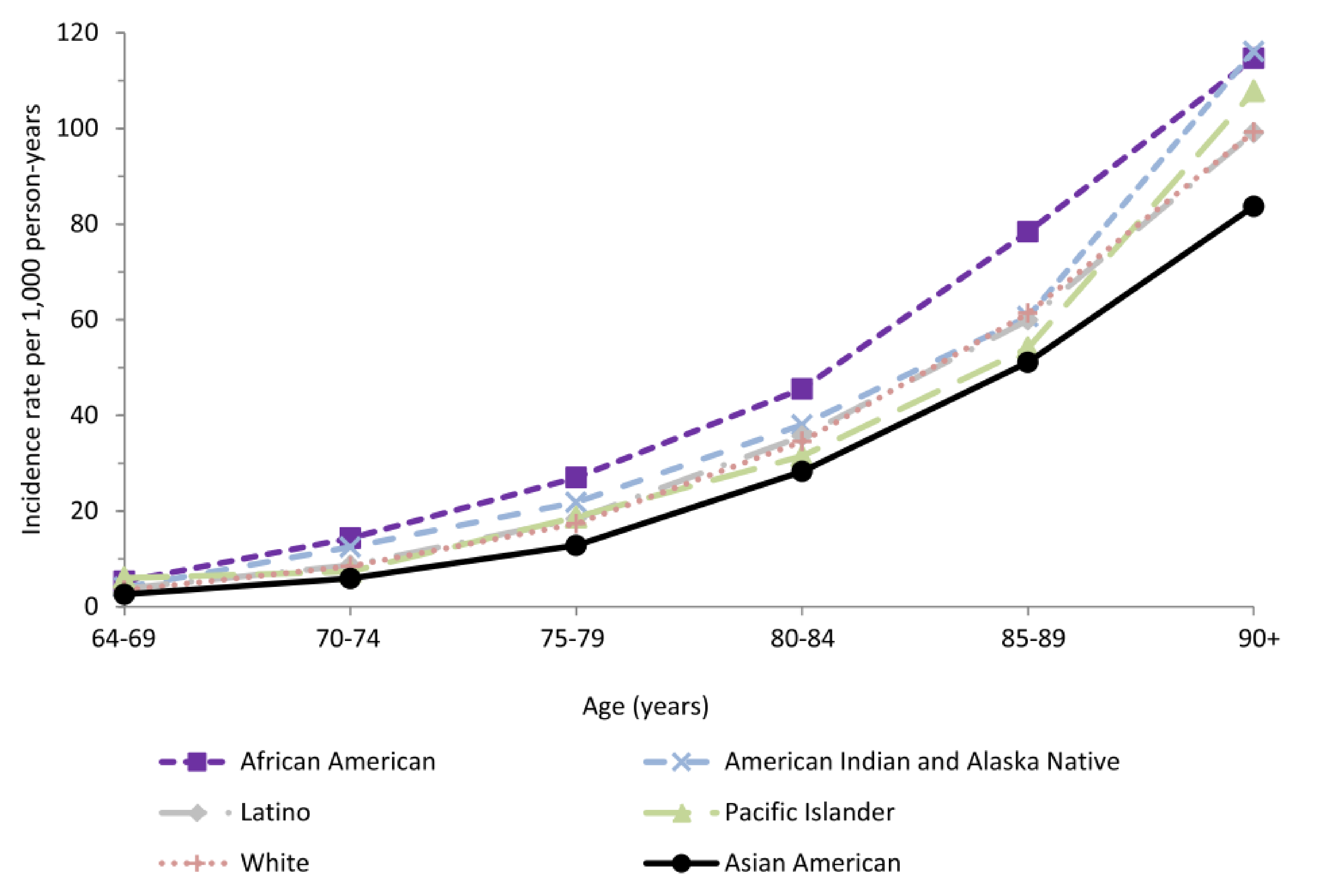


Fig. 2. Dementia incidence rates per 1000 person-years by age and race-ethnicity, 2000–2013.

Mayeda, E. R., Glymour, M. M., Quesenberry, C. P., & Whitmer, R. A. (2016). Inequalities in dementia incidence between six racial and ethnic groups over 14 years. Alzheimer's & Dementia, 12(3), 216-224.

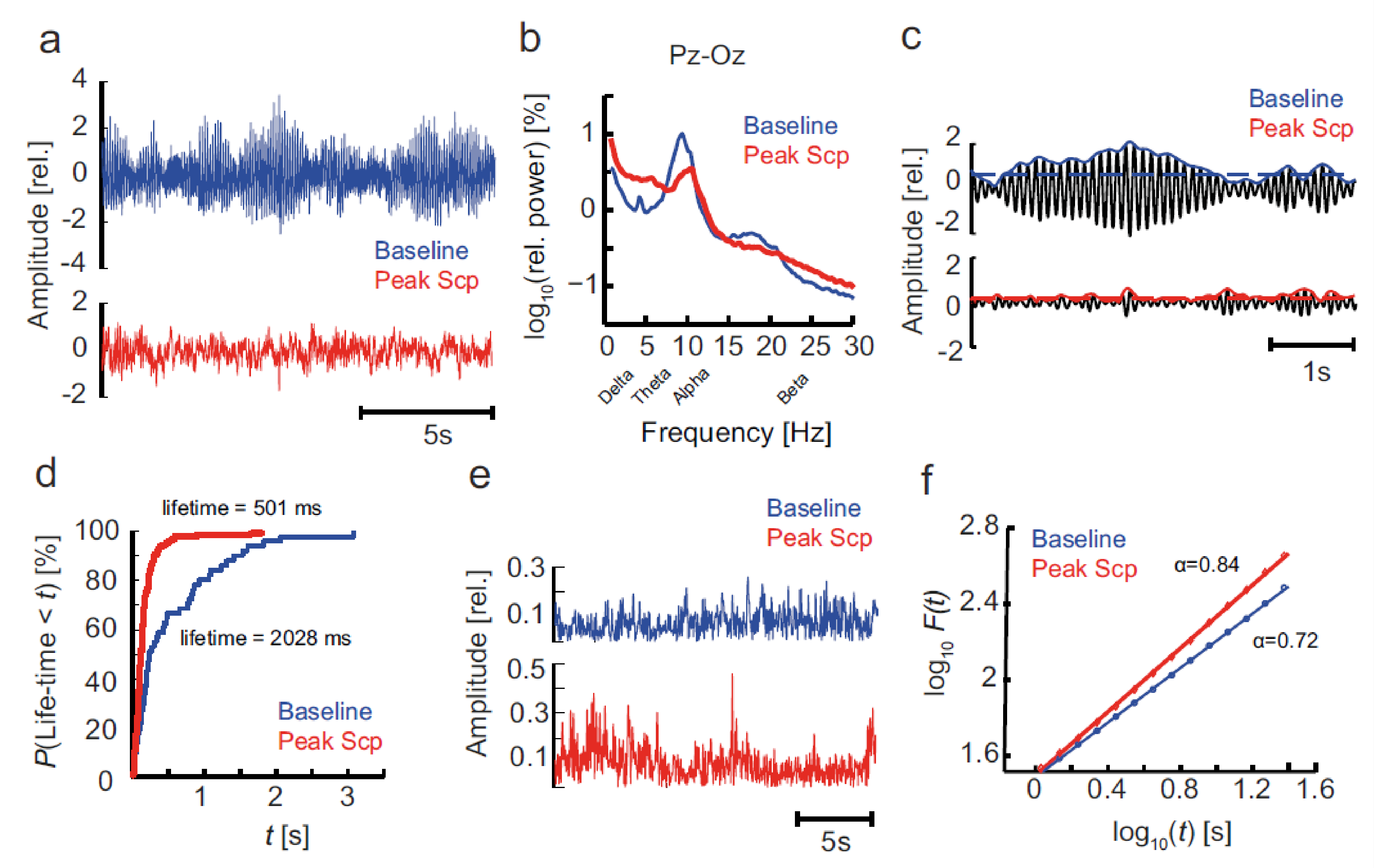
gruneco
grupo neuropsicología y conducta

complement 1

VI Definitions

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"

Atkinson, C., & Bingham, S. A. (2001). Mammographic breast density as a biomarker of effects of isoflavones on the female breast. Breast Cancer Research, 4(1), 1-4.



Simpraga, S., Alvarez-Jimenez, R., Mansvelder, H. D., Van Gerven, J. M., Groeneveld, G. J., Poil, S. S., & Linkenkaer-Hansen, K. (2017). EEG machine learning for accurate detection of cholinergic intervention and Alzheimer's disease. Scientific reports, 7(1), 5775.

gruneco
grupo neuropsicología y conducta

complement 2

Neuropsychological tests

The use of approaches based on resting-state EEG and neuropsychological test could be beneficial in neurology or even primary care.

MMSE

A set of 11 questions that doctors and other healthcare professionals commonly use to check for cognitive impairment (problems with thinking, communication, understanding and memory).

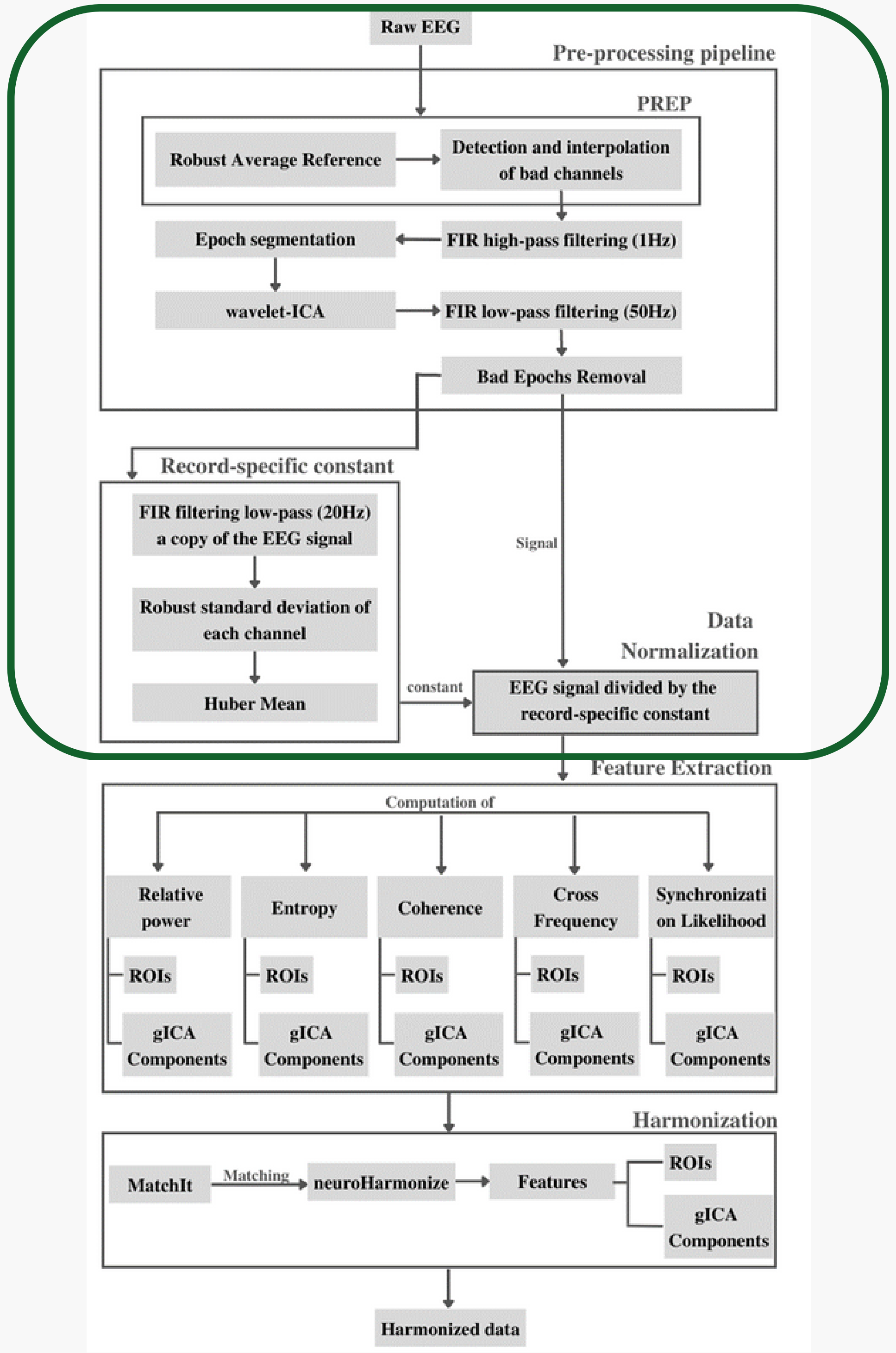
MoCA

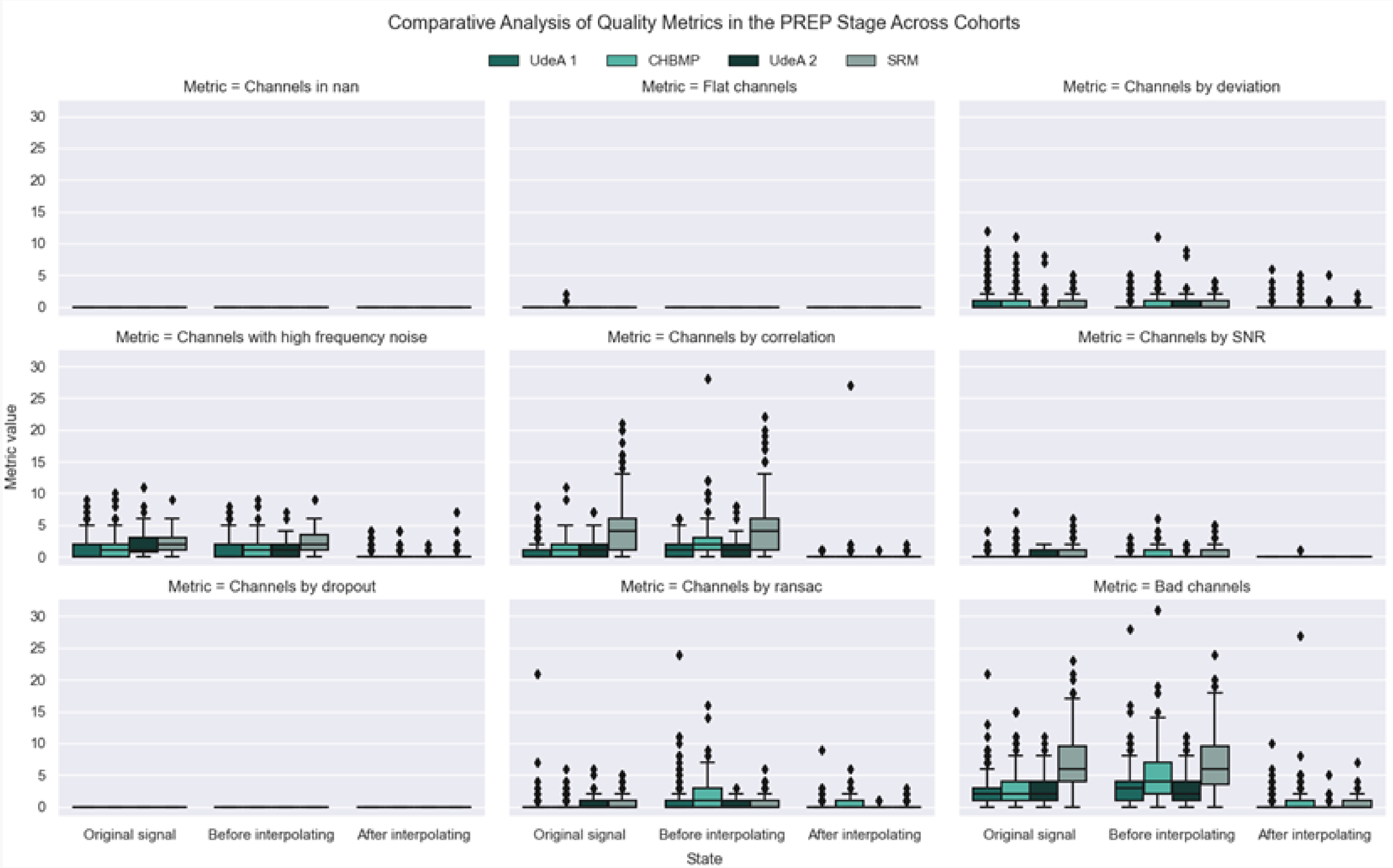
Montreal Cognitive Assessment: The sensitivity of MoCA for detecting MCI is 90%, compared to 18% for the MMSE.

VFT

It is an individual application test that is used to carry out a rapid global evaluation of language and executive functions.

IV Research Results
Harmonize the database

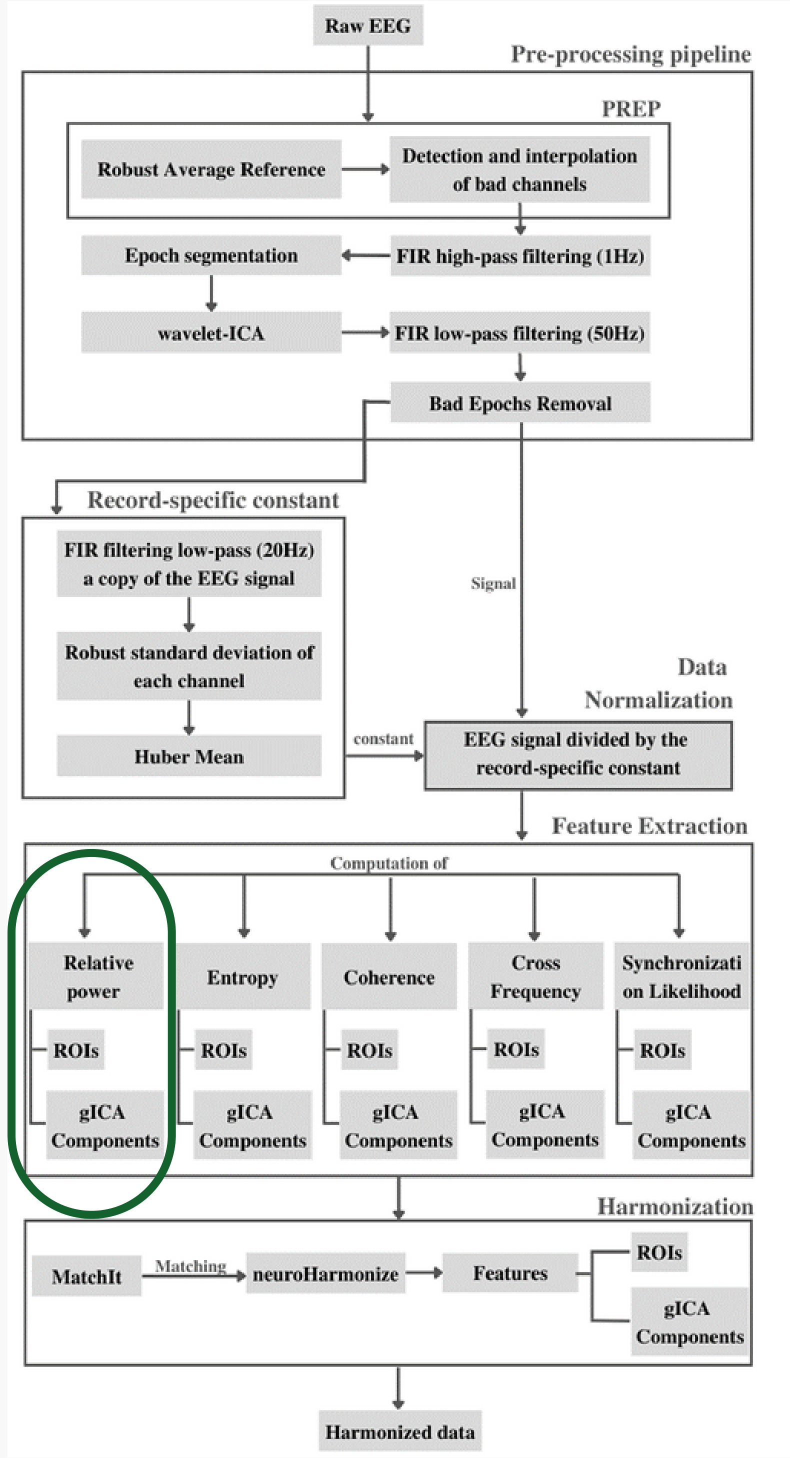




grupo neuropsicología y conducta

complement 4

IV Research Results
Harmonize the database



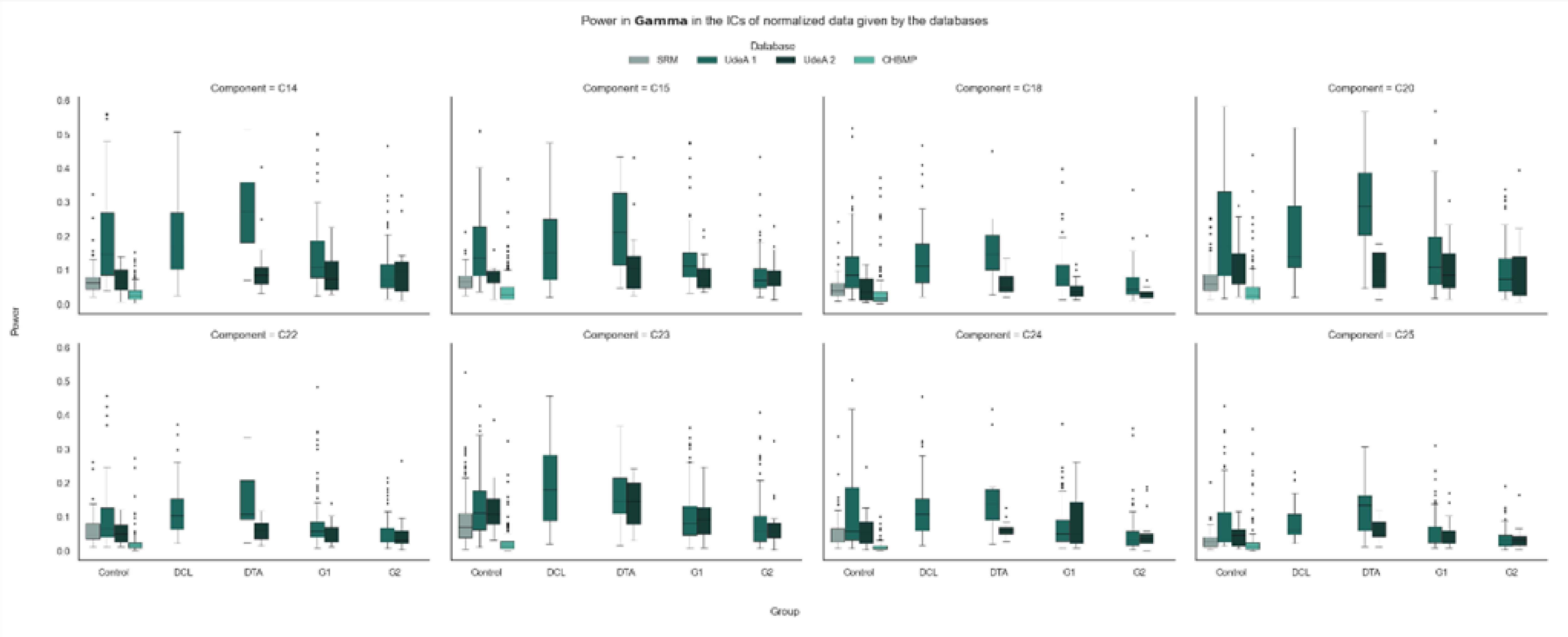
Multitaper Method (MTM)

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

K = 2TW - 1

u = taper or windowing function

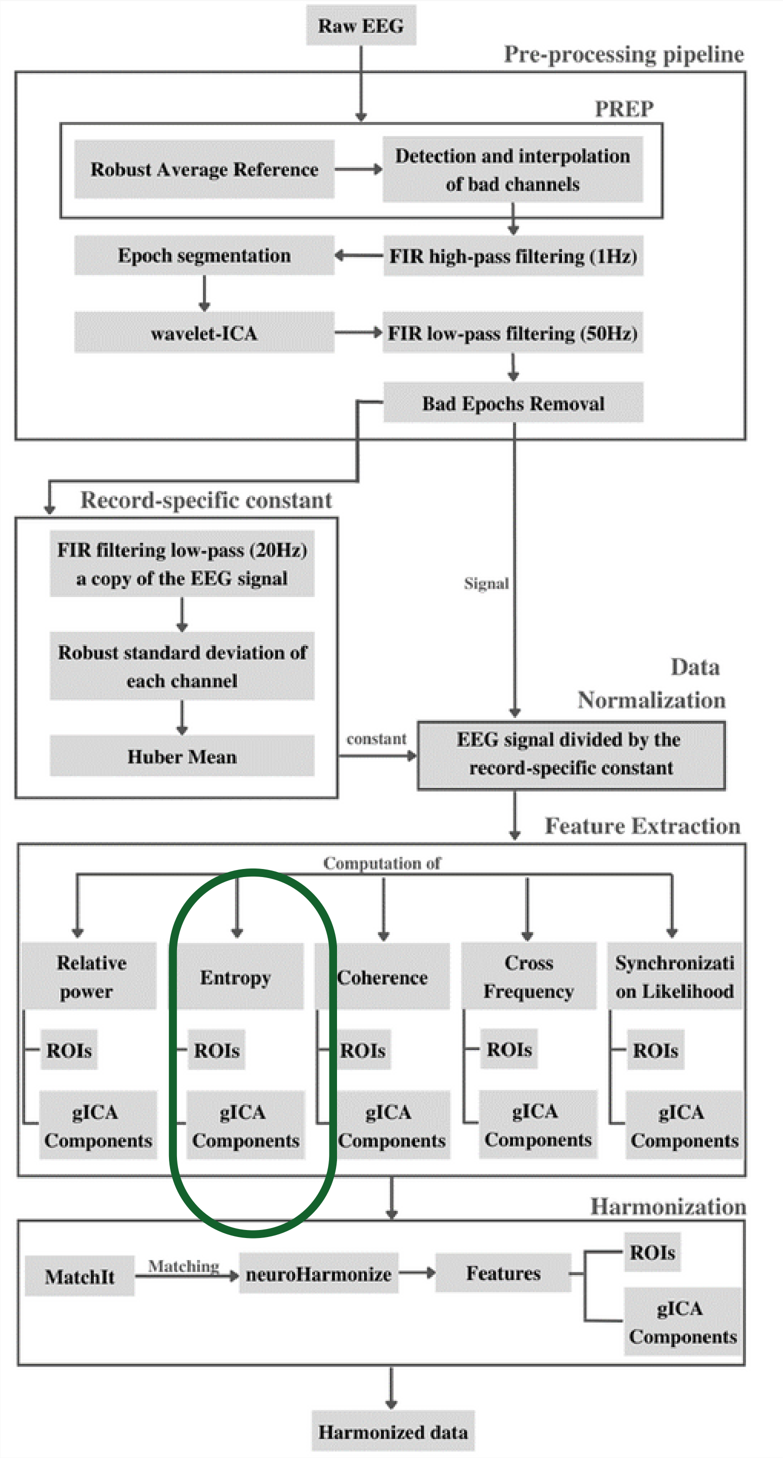
Bokil, H., Andrews, P., Kulkarni, J. E., Mehta, S., & Mitra, P. P. (2010). Chronux: a platform for analyzing neural signals. Journal of neuroscience methods, 192(1), 146-151.



grupo neuropsicología y conducta

IV Research Results

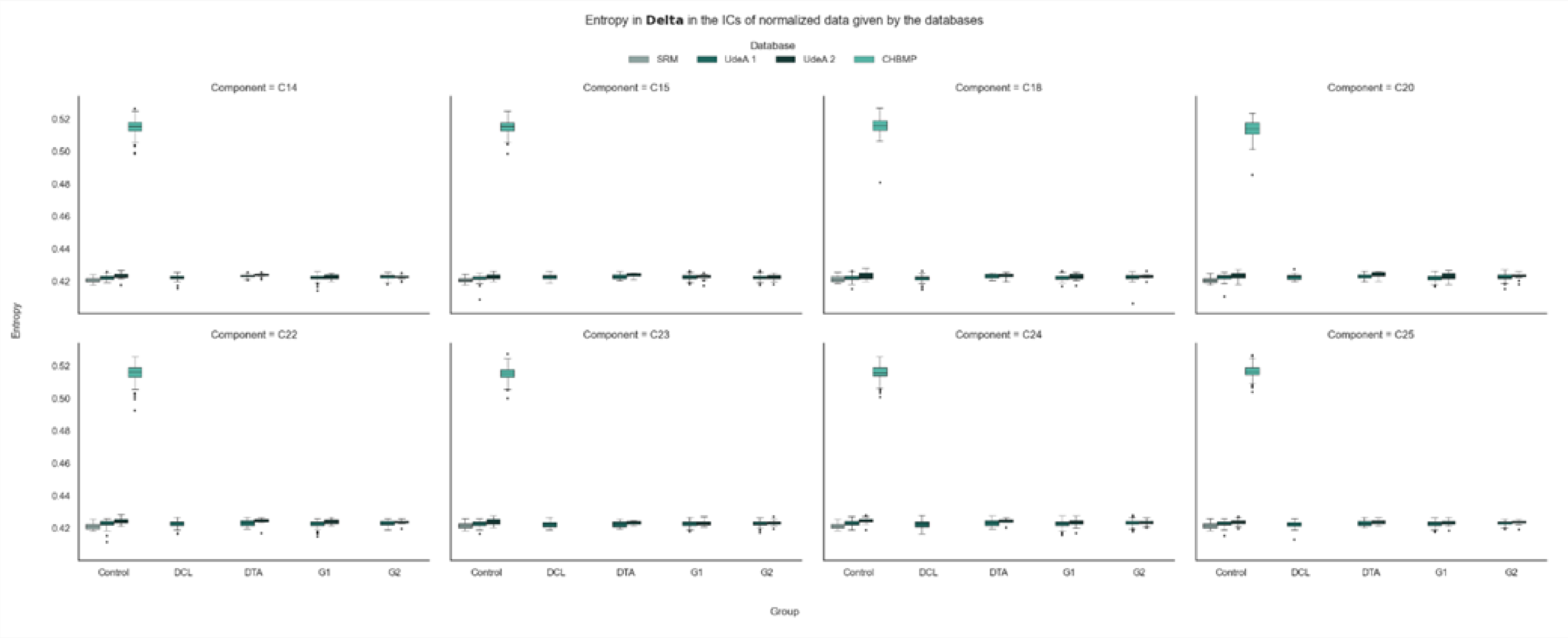
Harmonize the database



Shannon entropy

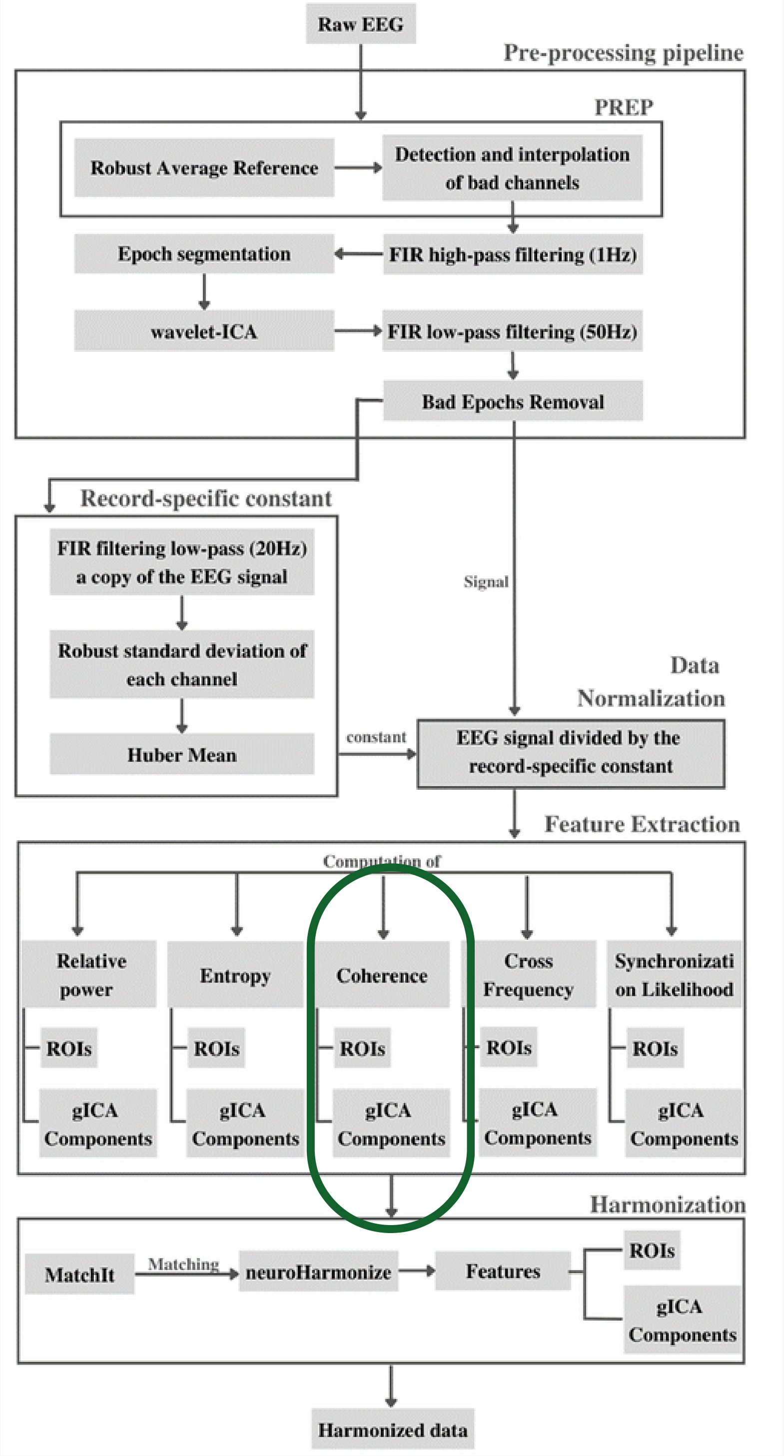
probabilities of the event(i)

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



complement 6

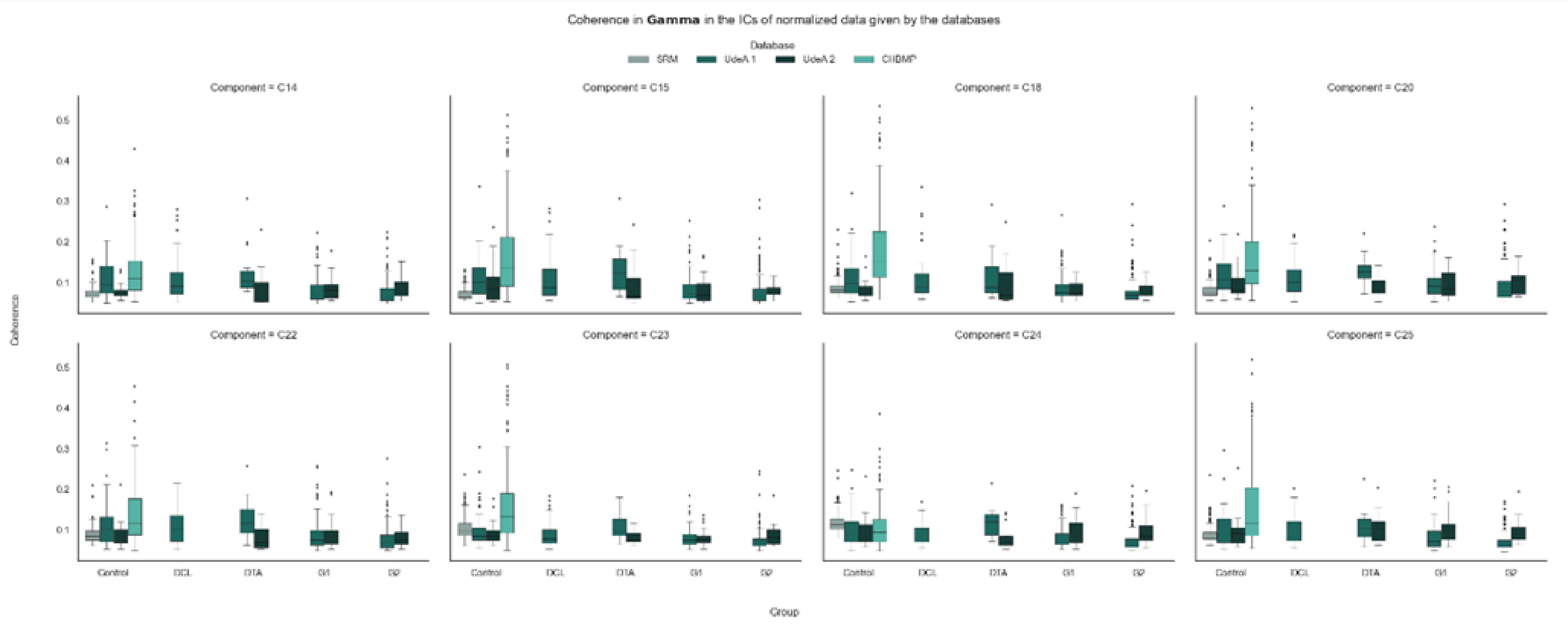
IV Research Results
Harmonize the database



Coherence

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

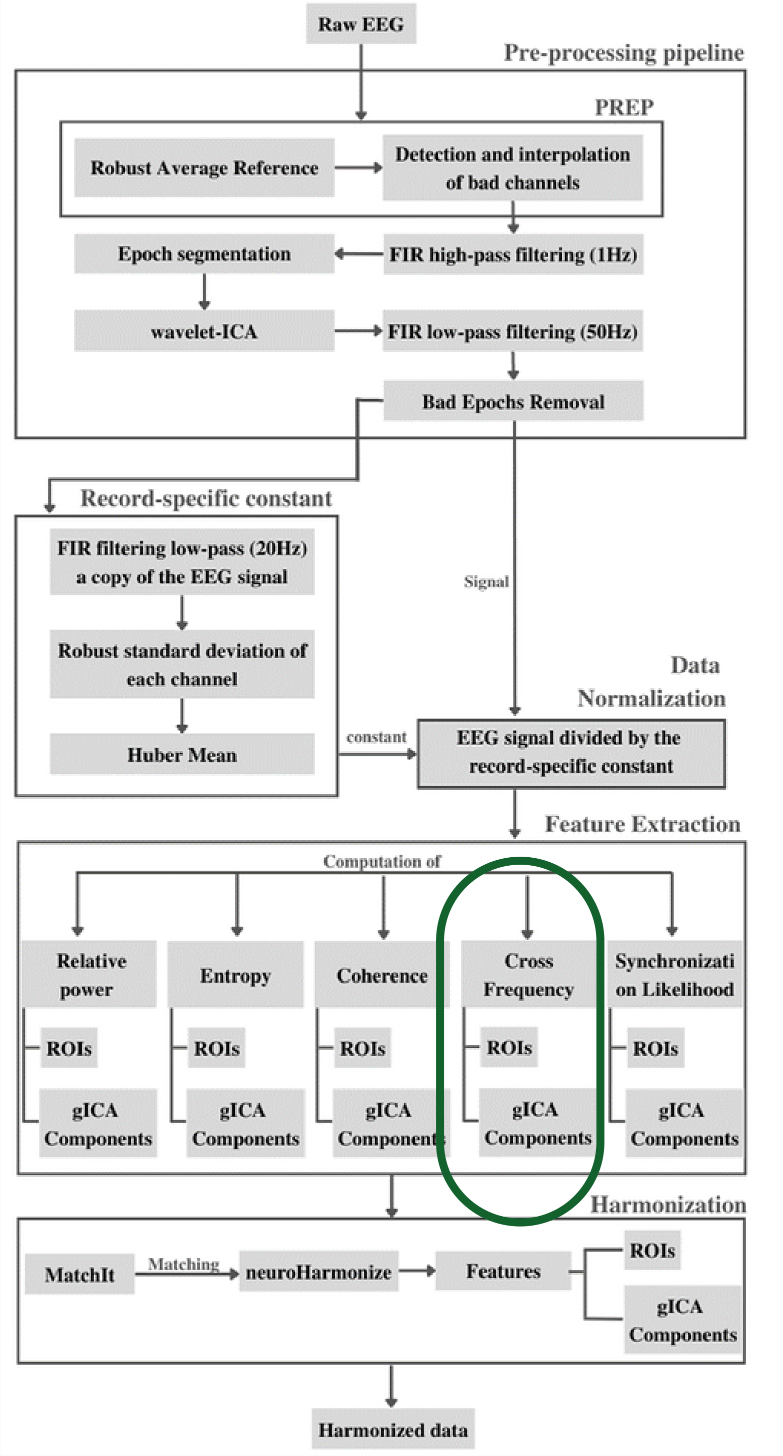
$$C_{xy} = \frac{|P_{xy}|^2}{P_{xx} * P_{yy}}$$



grupo neuropsicología y conducta

IV Research Results

Harmonize the database



Cross Frequency

Hilbert transform H{*}

Fourier transform F{*}

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

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

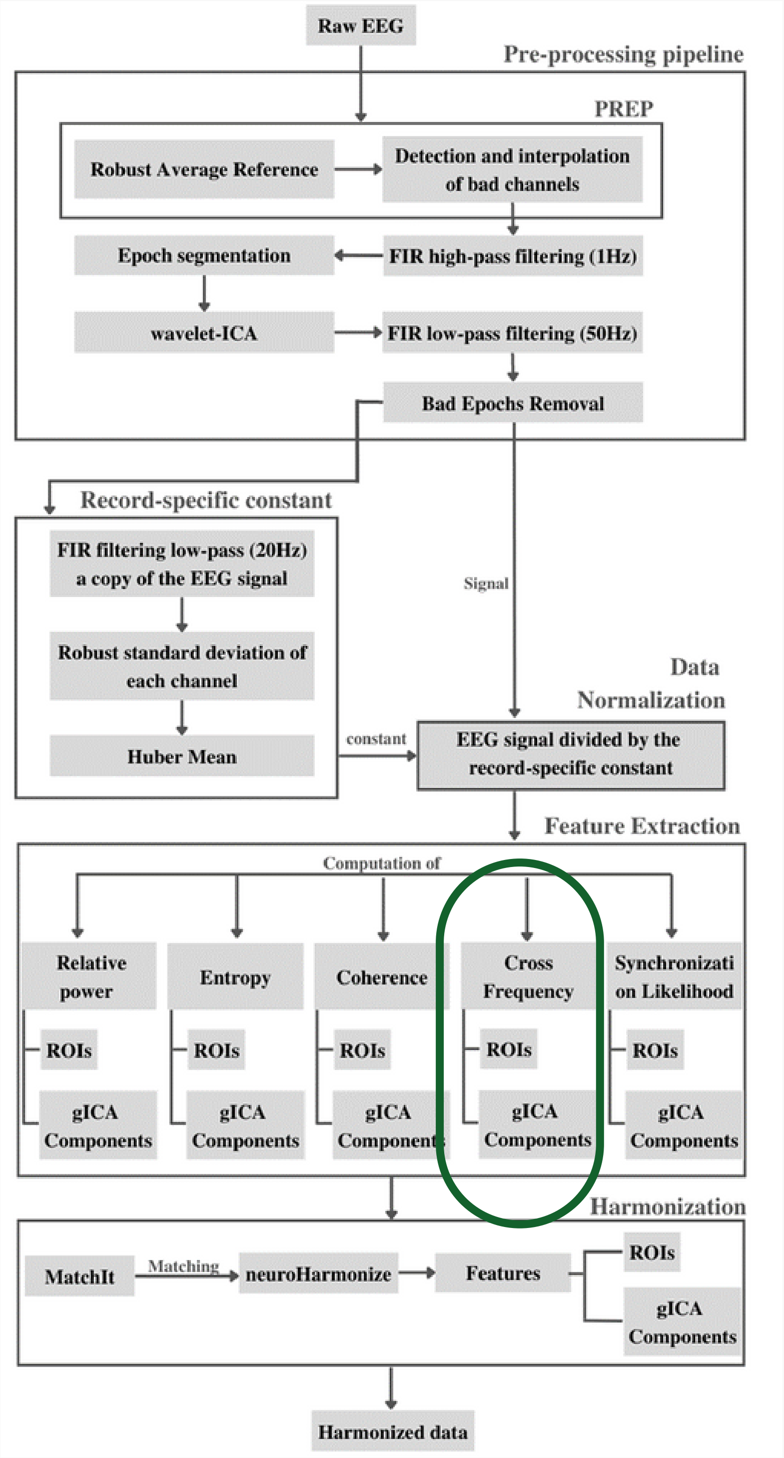
$$\varepsilon_i(m; f) = |F\{e_i(m,n)\}|$$

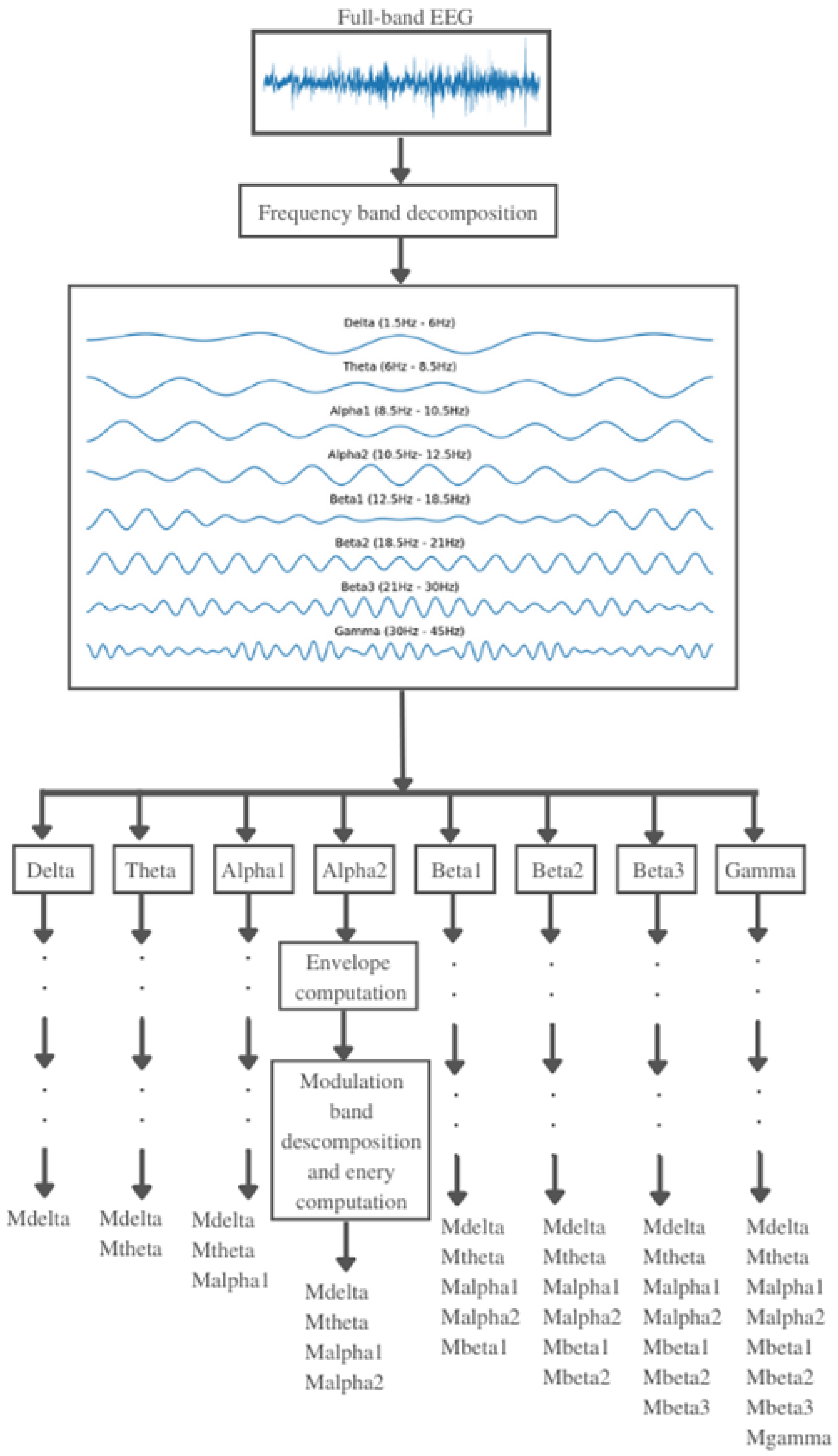
$$PME_{i,j} = \frac{\bar{\varepsilon}_{i,j}}{\sum_{i=1}^{k}\sum_{j=1}^{k-1}\bar{\varepsilon}_{i,j}} x100\%$$

gruneco
grupo neuropsicología y conducta

IV Research Results

Harmonize the database

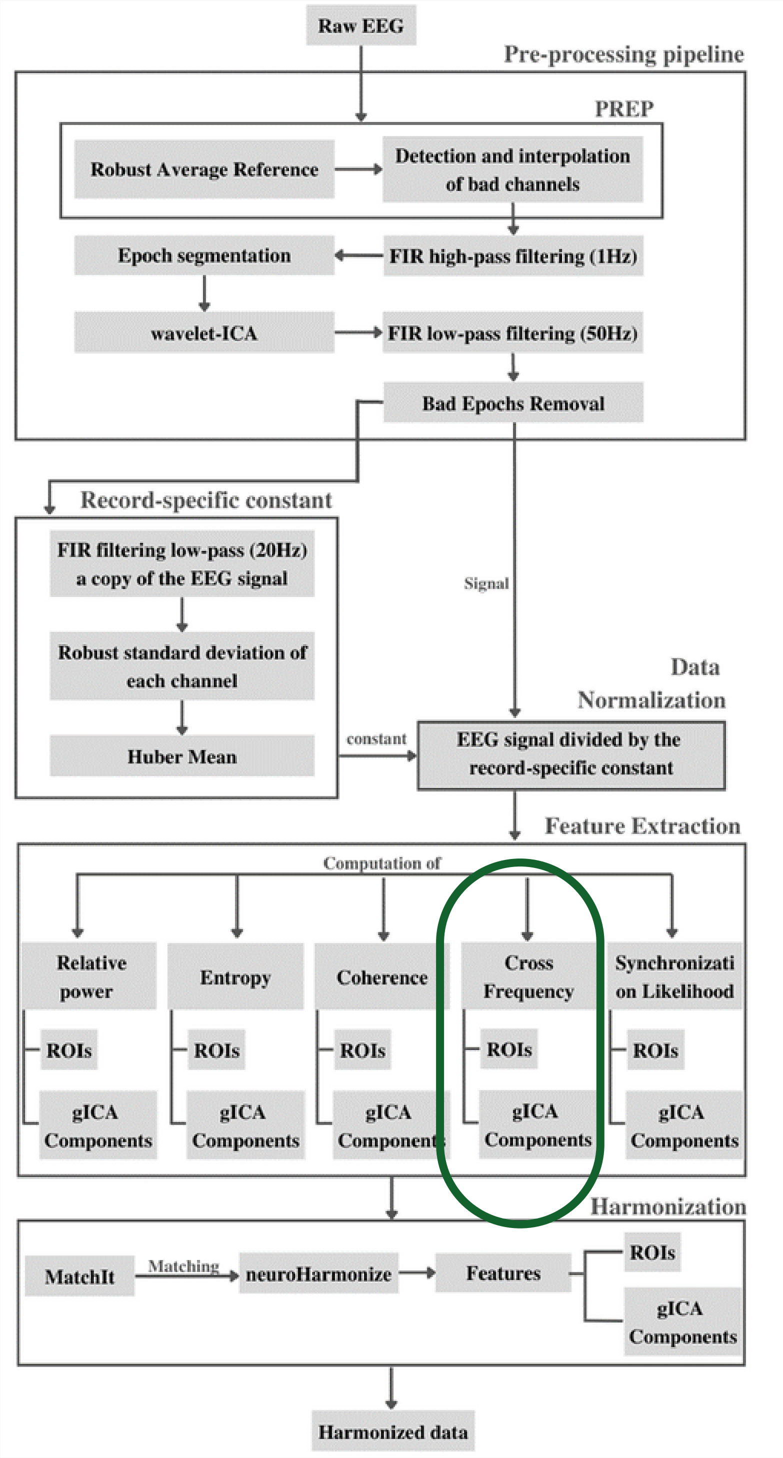




gruneco
grupo neuropsicología y conducta

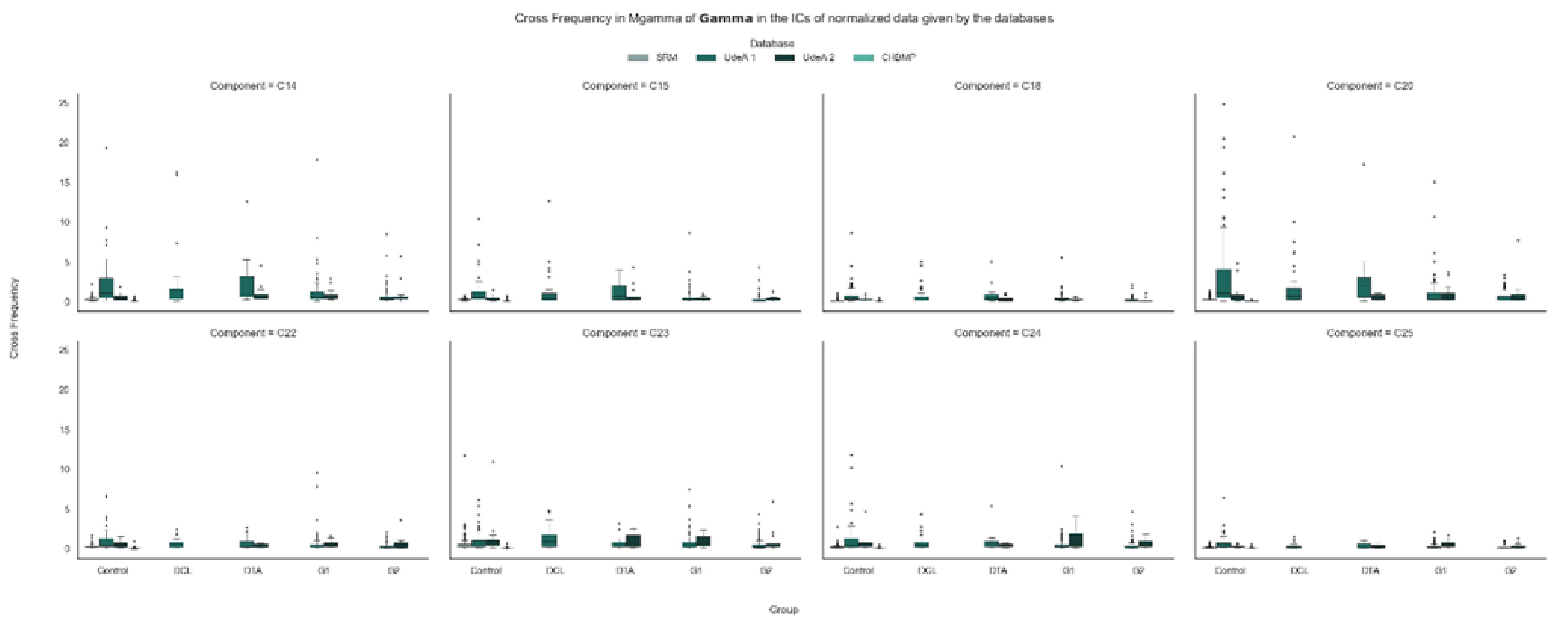
complement 9

IV Research Results
Harmonize the database

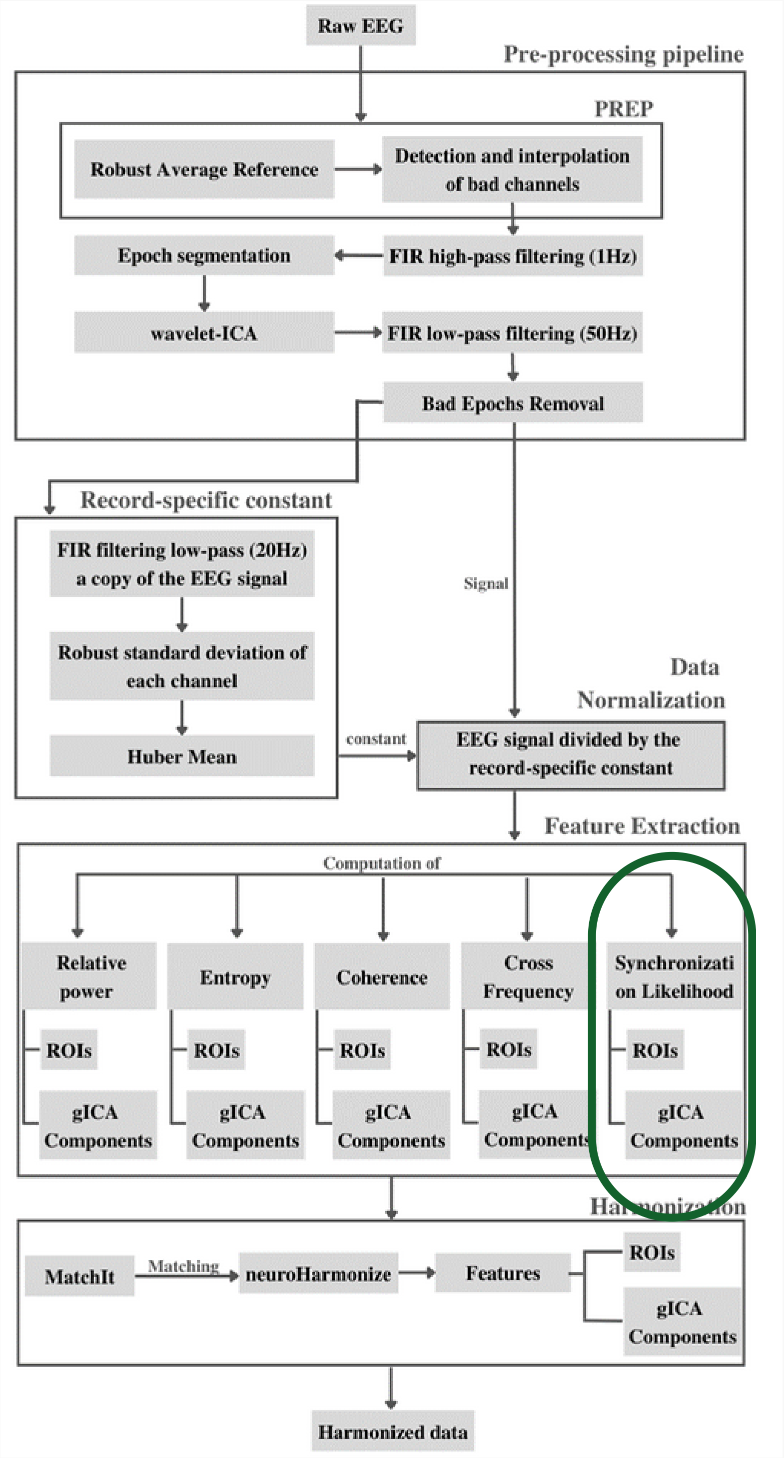


Cross Frequency

Cross-frequency interaction is believed to be important for coordinating the activity of different brain regions and for regulating neural processing



IV Research Results
Harmonize the database



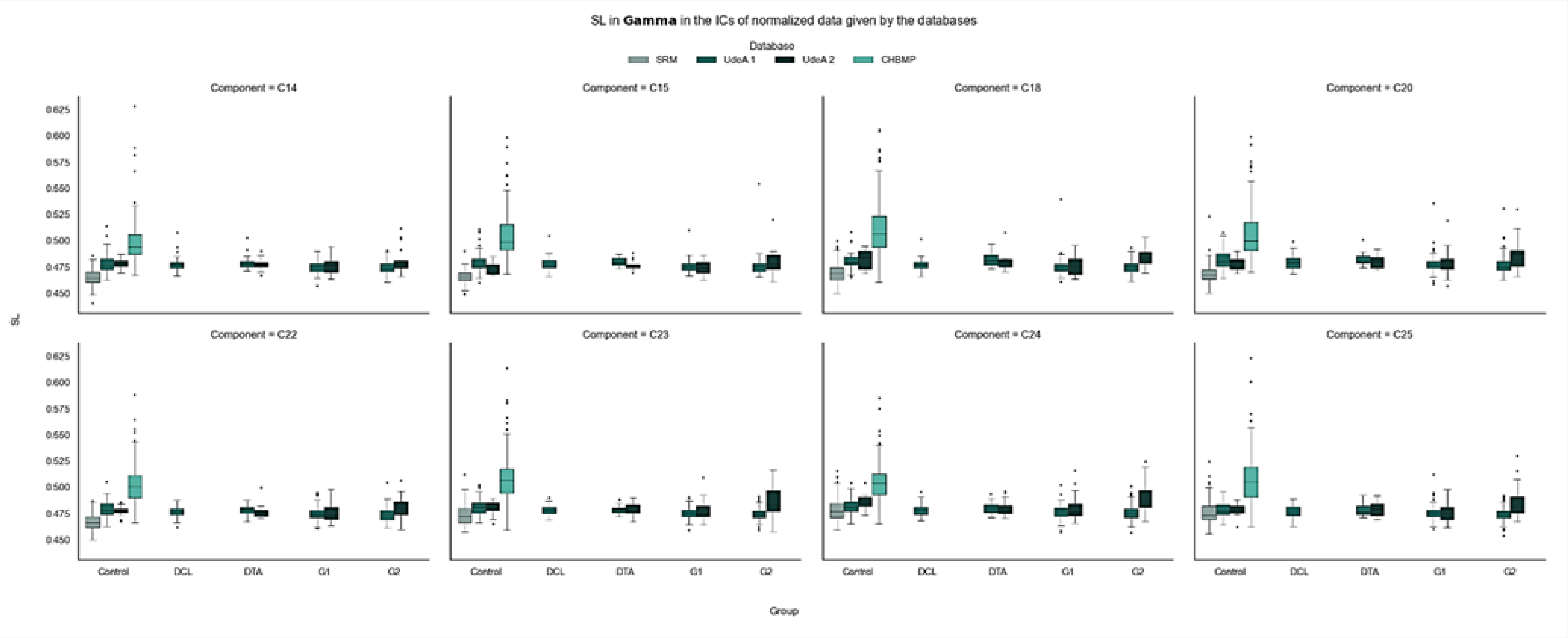
Synchronization Likelihood (SL)

The number Hi,j of channels, for which the distance of embedded vectors Xk,i and Xk,j is smaller than εk,i:

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

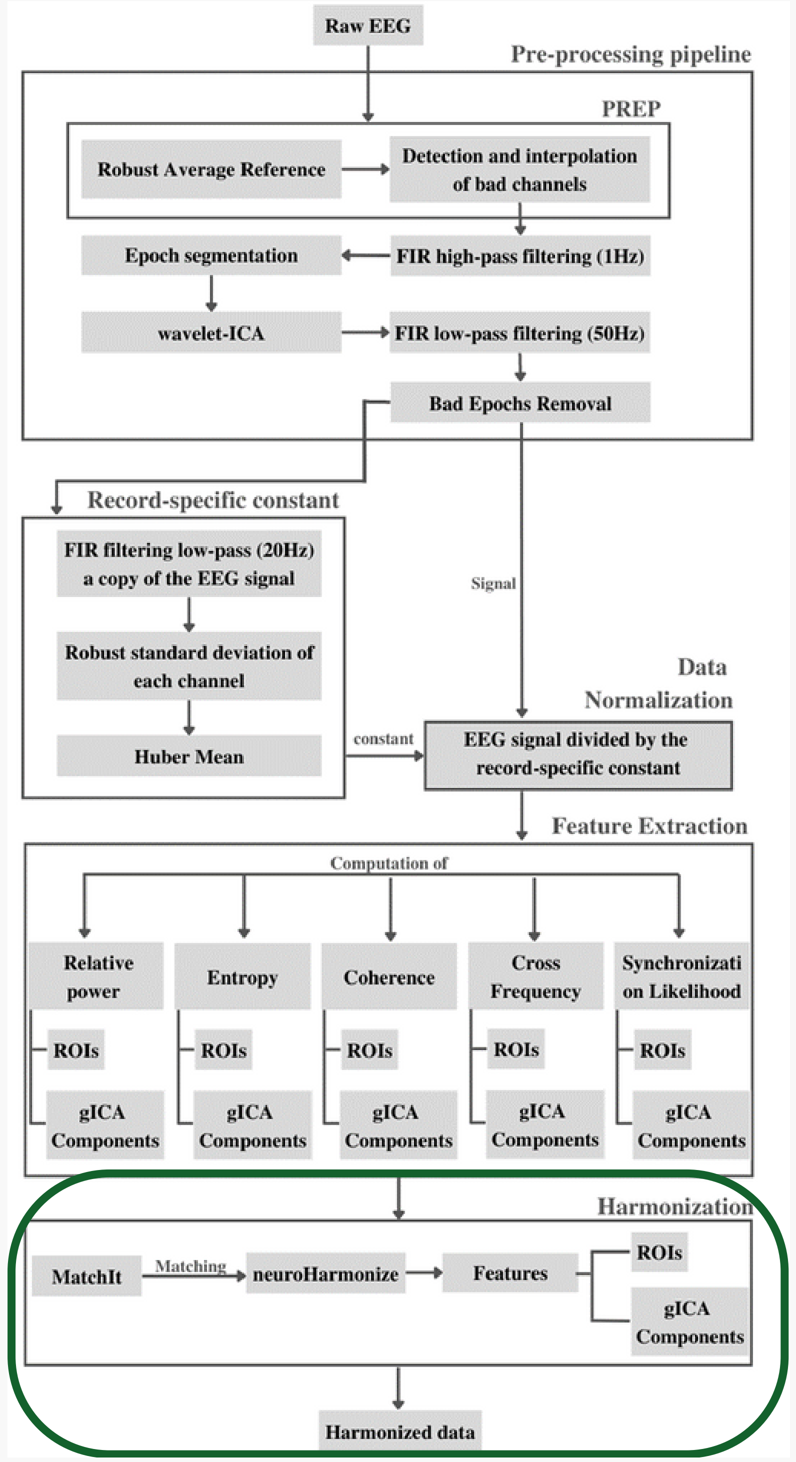
$$if\ |X_{k,i} - X_{k,j}| < \varepsilon_{k,i}: S_{k,i,j} = \frac{H_{i,j} - 1}{M - 1}$$

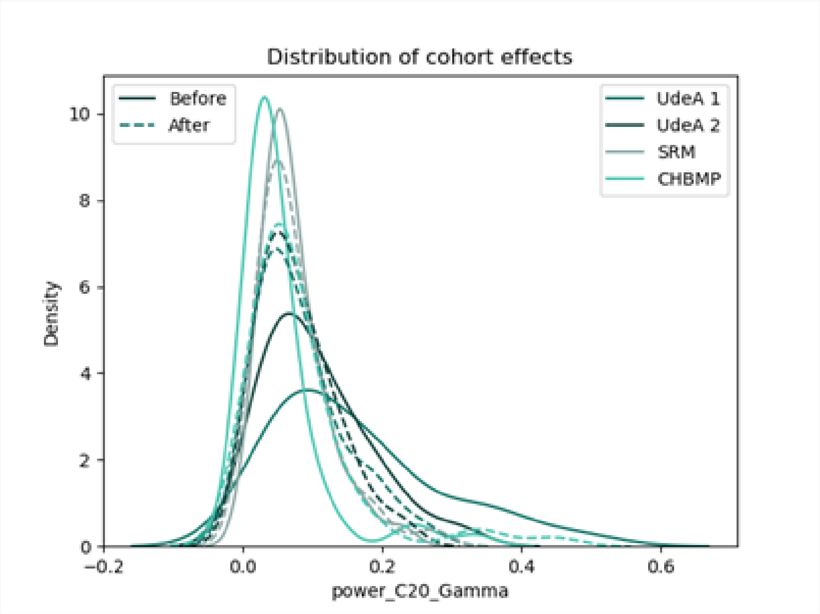
$$if\ |X_{k,i} - X_{k,j}| \geq \varepsilon_{k,i}: S_{k,i,j} = 0$$

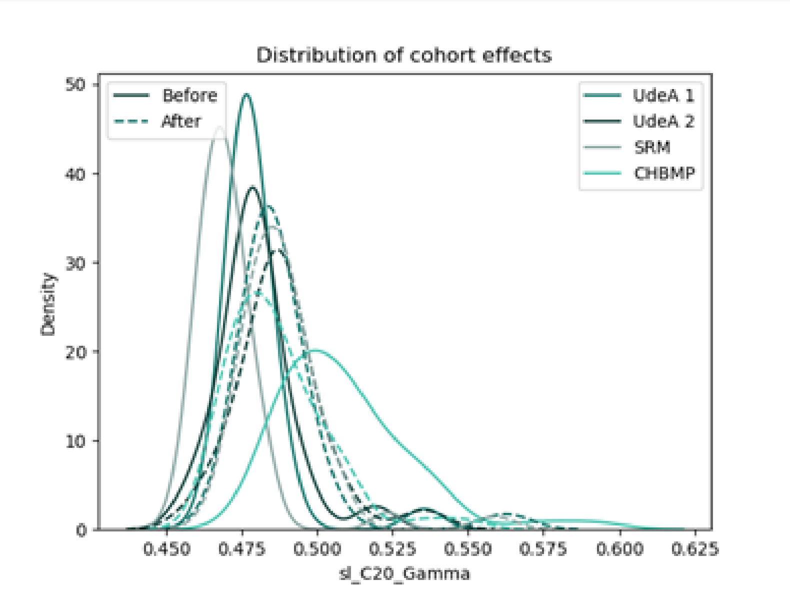


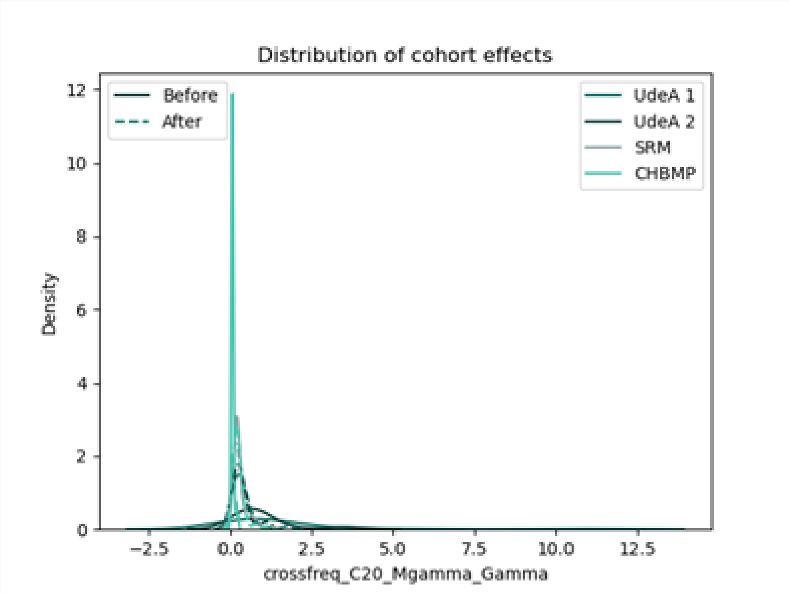
gruneco
grupo neuropsicología y conducta

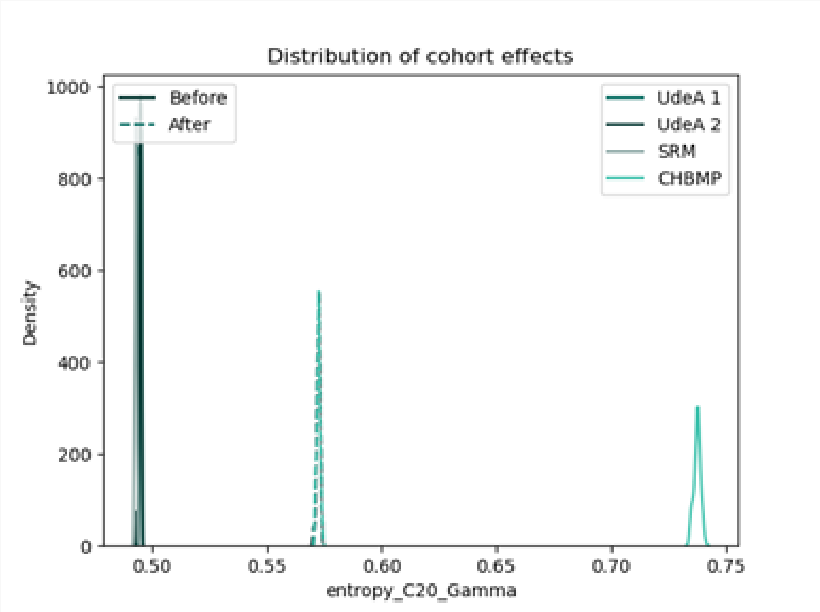
IV Research Results
Harmonize the database

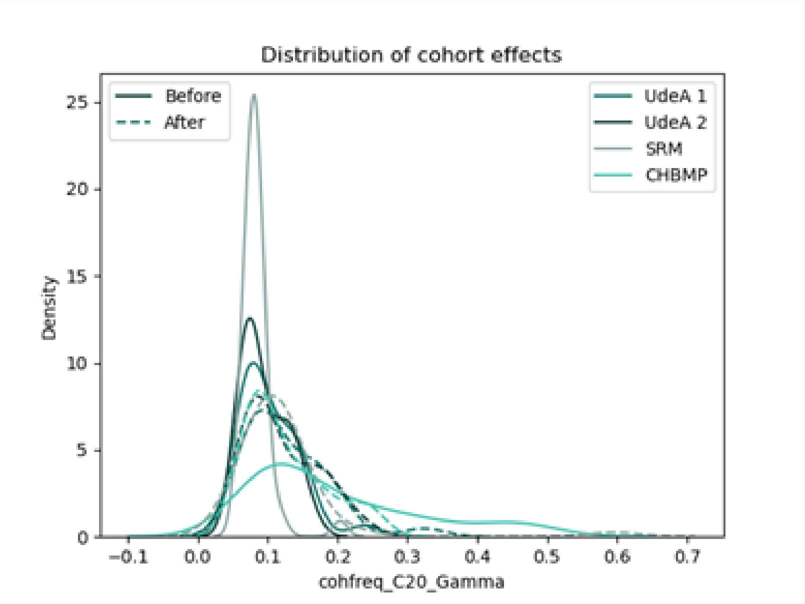












gruneco
grupo neuropsicología y conducta

IV Research Results

Harmonize the database

Adjusting batch effects in microarray expression data using empirical Bayes methods

W. EVAN JOHNSON, CHENG LI*

Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA, USA and Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA
cli@hsph.harvard.edu

ARIEL RABINOVIC

Department of Genetics and Complex Diseases, Harvard School of Public Health, Boston, MA, USA

(Johnson W, 2007)

$$\underset{1}{y_{ijv}} = \overset{2}{\alpha_v} + \underset{3}{\mathbf{X}_{ij}^T\beta_v} + \overset{4}{\gamma_{iv}} + \underset{5}{\delta_{iv}\epsilon_{ijv}}$$

1. Harmonized feature, 2. reference feature, 3. feature-associated vector, 4,5. parameters estimated with empirical Bayesian.

Harmonization of cortical thickness measurements across scanners and sites

Jean-Philippe Fortin[1,*], Nicholas Cullen[2,3,*], Yvette I. Sheline[3,4,5], Warren D. Taylor[6], Irem Aselcioglu[3], Philip A. Cook[3,4], Phil Adams[7], Crystal Cooper[8], Maurizio Fava[9], Patrick J. McGrath[7], Melvin McInnis[10], Mary L. Phillips[11], Madhukar H. Trivedi[8], Myrna M. Weissman[7,12,13], and Russell T. Shinohara[1,3,†]

Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan

Raymond Pomponio[a,*], Guray Erus[a], Mohamad Habes[a,b], Jimit Doshi[a], Dhivya Srinivasan[a], Elizabeth Mamourian[a], Vishnu Bashyam[a], Ilya M. Nasrallah[a,g], Theodore D. Satterthwaite[l], Yong Fan[a], Lenore J. Launer[c], Colin L. Masters[d], Paul Maruff[d], Chuanjun Zhuo[e,f], Henry Völzke[h], Sterling C. Johnson[i], Jurgen Fripp[j], Nikolaos Koutsouleris[k], Daniel H. Wolf[l], Raquel Gur[g,l], Ruben Gur[g,l], John Morris[m], Marilyn S. Albert[n], Hans J. Grabe[o], Susan M. Resnick[p], R. Nick Bryan[q], David A. Wolk[b], Russell T. Shinohara[a,r,s], Haochang Shou[a,r,2], Christos Davatzikos[a,**,1,2]

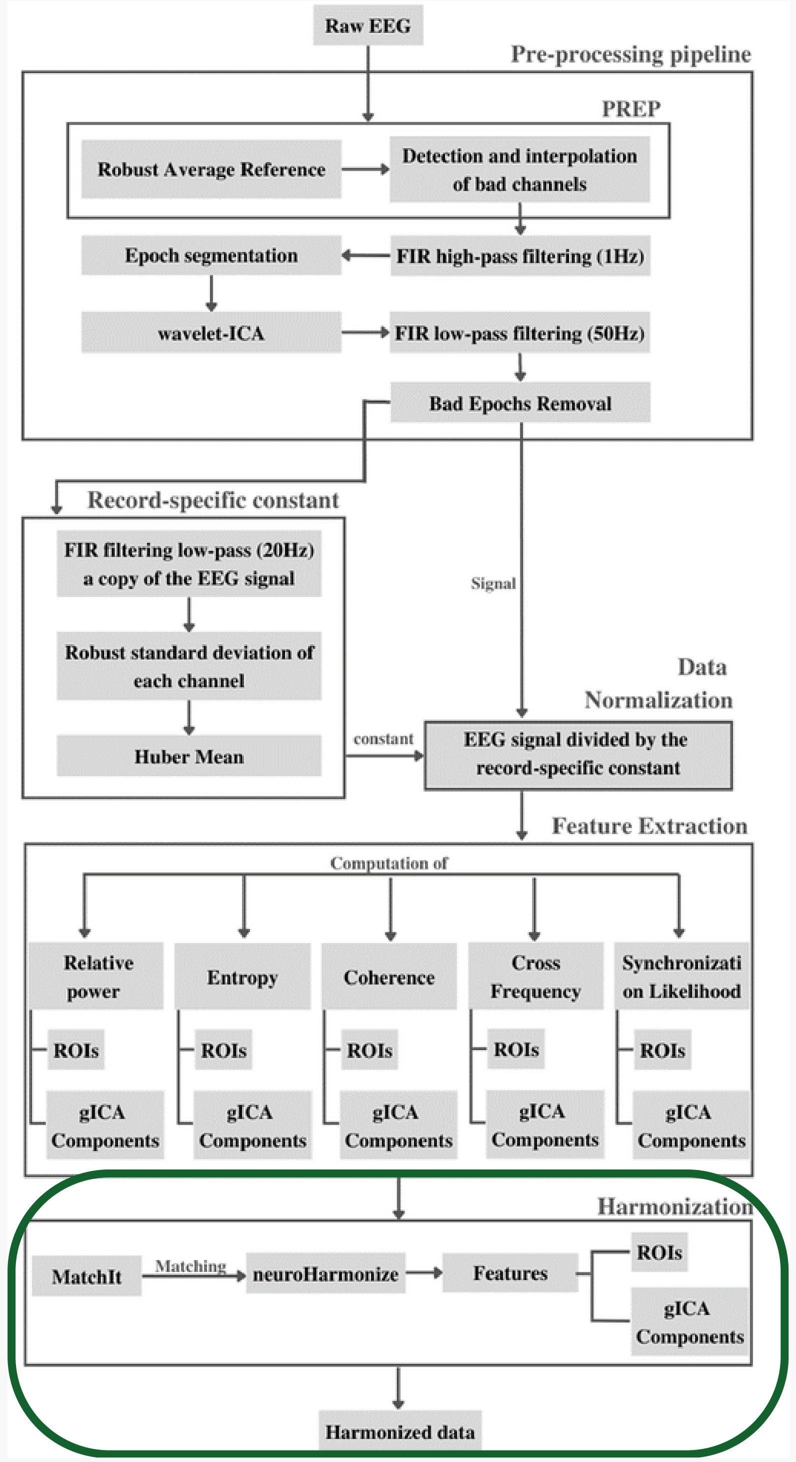
(Pomponio R, 2019)

$$\overset{5}{Y_{ijk}^*} = \left(\frac{\overset{1}{Y_{ijk}} - f_k(X_{ij}, Z_{ij}, W_{ij}) - \overset{4}{g_{ik}^*}}{\underset{3}{d_{ik}} + \underset{2}{f_k(X_{ij}, Z_{ij}, W_{ij})}}\right)$$

1. Reference feature, 2. generalized additive model and covariates, 3. conditional posterior estimation, 4. posterior estimation of the effect, 5. adjusted feature.

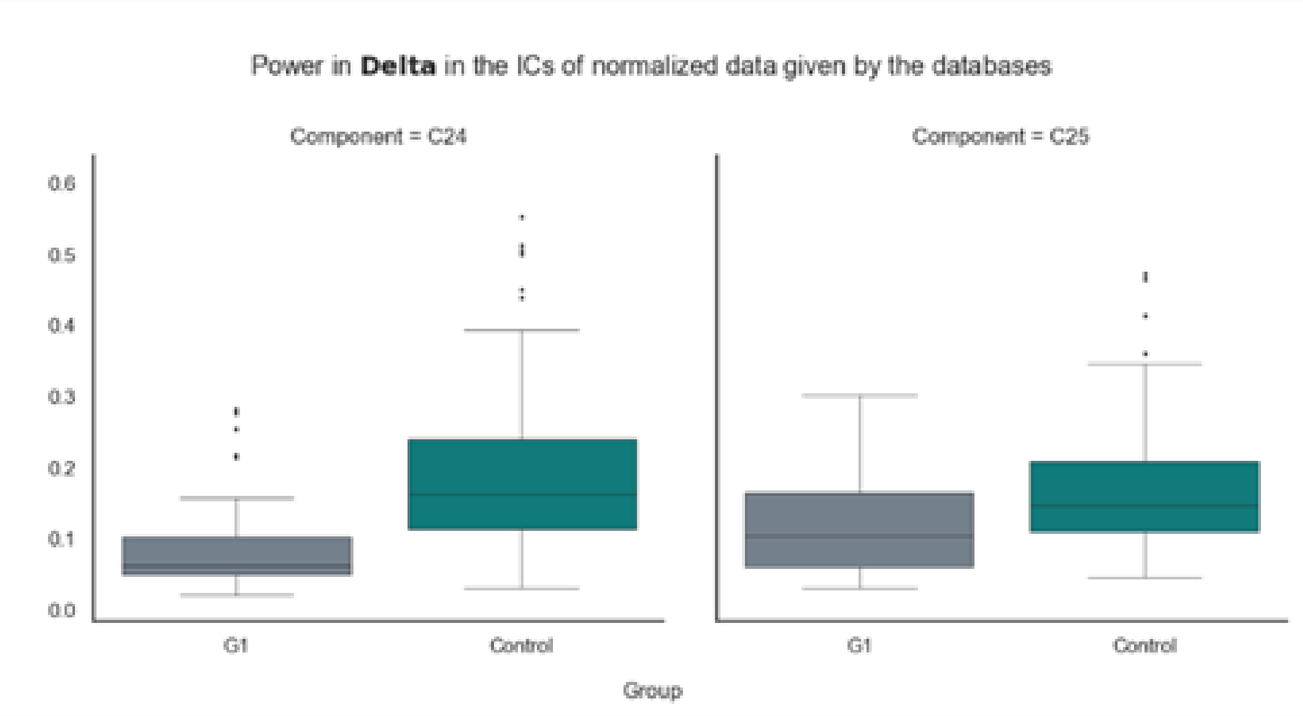
IV Research Results

Harmonize the database

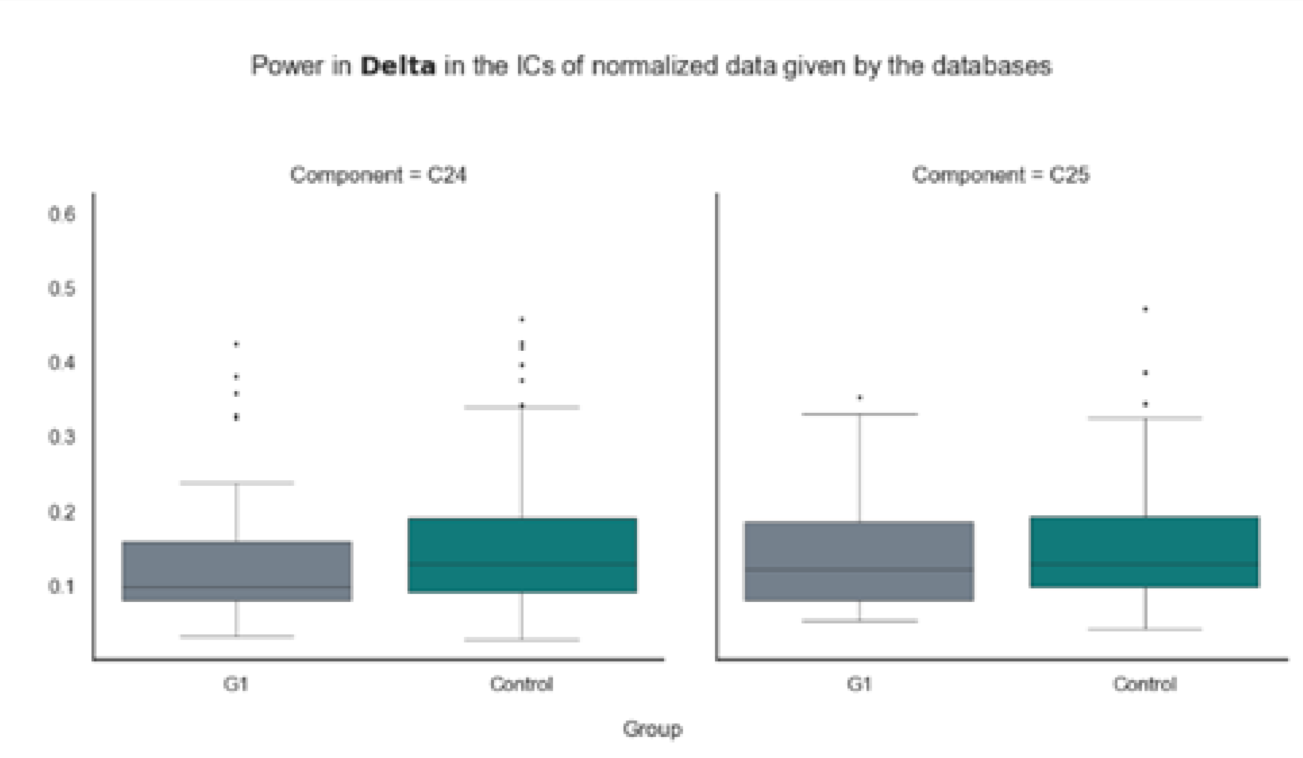


G1 with Controls

After the matching stage



After using Neuroharmonization



grupo neuropsicología y conducta

IV Research Results

Design a machine learning model

with neuroHarmonize

Grid Search (Random Forest Classifier)
bootstrap=False
criterion='log_loss'
max_depth=100
min_samples_split=10
n_estimators=558

DecisionTree

Support Vector Machine (SVM)
C = 0.1, gamma = 0.001

Boruta

estimator= Random Forest Classifier
bootstrap=False
criterion='log_loss'
max_depth=20
min_samples_split=5
n_estimators=1000

TPOT

cv=10, generations=5
n_jobs=-1
population_size=58
Random_state=10
verbosity=3

DecisionTree
DecisionTree
DecisionTree
XGB & SGD
GradientBoosting

XGB & SGD
XGB & SGD
GradientBoosting
GradientBoosting
DecisionTree

RandomForest

without neuroHarmonize

Grid Search (Random Forest Classifier)
criterion='log_loss'
max_depth=90
min_samples_leaf=2
min_samples_split=5
n_estimators=296

DecisionTree

Support Vector Machine (SVM)
C = 0.1, gamma = 0.001

Boruta

estimator= Random Forest Classifier
criterion='log_loss'
max_depth=90
min_samples_leaf=2
min_samples_split=5
n_estimators=1000

TPOT

cv=10, generations=5
n_jobs=-1
population_size=58
Random_state=10
verbosity=3

ExtraTrees

bootstrap=True
criterion='entropy'
max_depth=0.75
min_samples_leaf=14
min_samples_split=3
n_estimators=100

gruneco
grupo neuropsicología y conducta

V Research Results

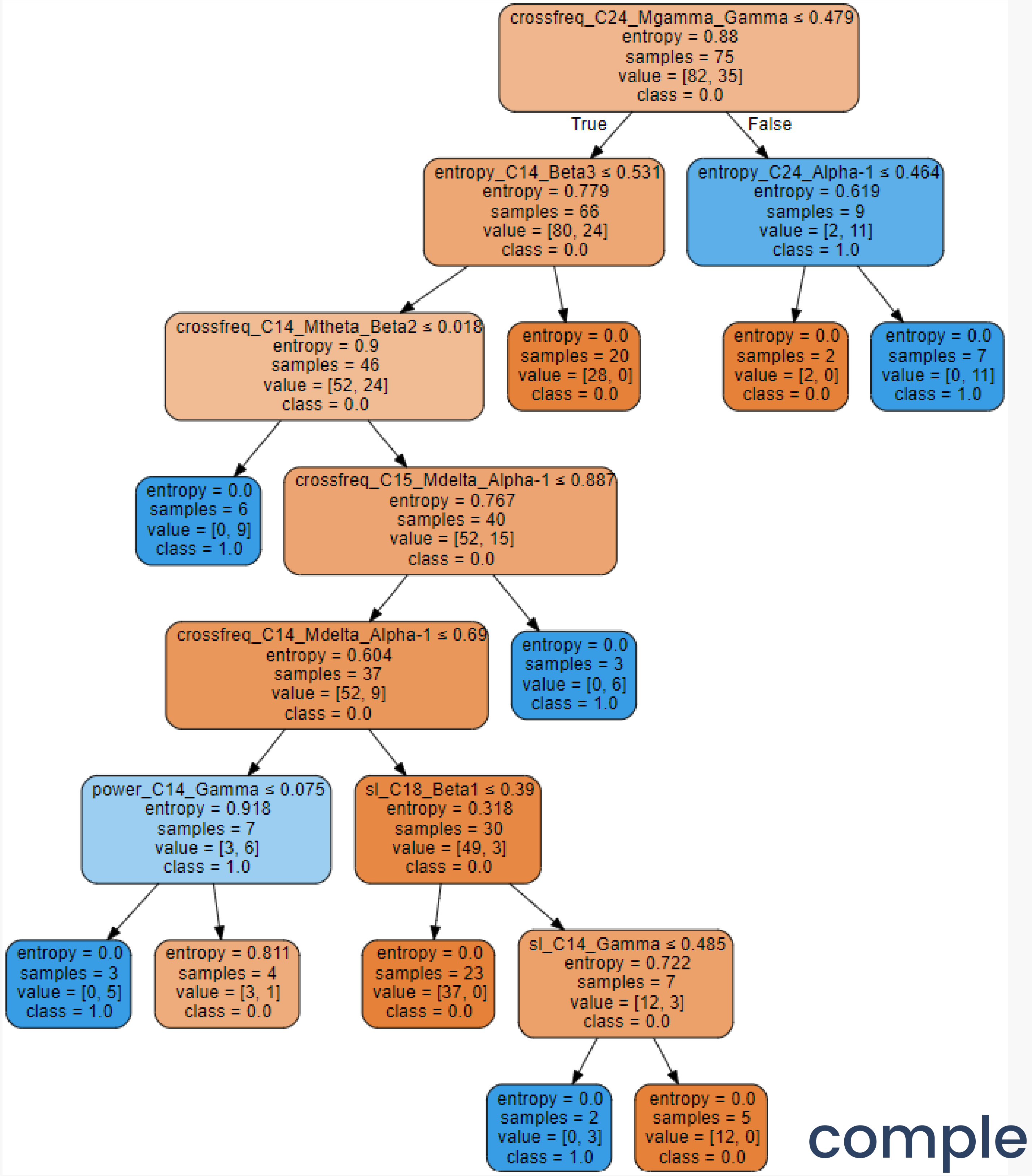
Design a machine learning model

gini: The Gini index is a measure of leaf impurity. There is no mixture of classes in this leaf; all samples in the leaf belong to the same class.

samples: Indicates how many samples are in this leaf.

value: It is a list showing the class distribution in the leaf.

class: Indicates the majority class in the leaf.



gruneco
grupo neuropsicología y conducta

complement 16