JYU DISSERTATIONS 733

Reza Mahini Sheikhhosseini

Consensus Clustering for Group-Level Analysis of Event-Related Potential Data

UNIVERSITY OF JYVÄSKYLÄ
FACULTY OF INFORMATION TECHNOLOGY

JYU DISSERTATIONS 733

Reza Mahini Sheikhhosseini

Consensus Clustering for Group-Level Analysis of Event-Related Potential Data

Esitetään Jyväskylän yliopiston informaatioteknologian tiedekunnan suostumuksella
julkisesti tarkastettavaksi yliopiston Agora-rakennuksen auditoriossa 3
joulukuun 7. päivänä 2023 kello 12.

Academic dissertation to be publicly discussed, by permission of
the Faculty of Information Technology of the University of Jyväskylä,
in building Agora, auditorium 3, on December 7, 2023 at 12 o'clock noon.

JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 2023

Editors
Marja-Leena Rantalainen
Faculty of Information Technology, University of Jyväskylä
Ville Korkiakangas
Open Science Centre, University of Jyväskylä

ISBN 978-951-39-9863-9 (PDF)
URN:ISBN:978-951-39-9863-9
ISSN 2489-9003

Permanent link to this publication: http://urn.fi/URN:ISBN:978-951-39-9863-9

ABSTRACT

Mahini Sheikhhosseini, Reza
Consensus clustering for group-level analysis of event-related potential data
Jyväskylä: University of Jyväskylä, 2023, 71 p. (+included articles)
(JYU Dissertations
ISSN 2489-9003; 733)
ISBN 978-951-39-9863-9 (PDF)

Understanding human brain activity through spatiotemporal electroencephalogram (EEG) analysis has gained prominence, with cluster analysis emerging as a valuable tool. While traditional event-related potential (ERP) analysis techniques for identifying interesting ERPs involve subjective time window selection, conventional cluster analysis focusing on spatial dynamics amplifies the risk of component identification errors when data is imperfect. Consequently, they do not offer a unified, appropriate time window determination approach for testing experimental hypotheses.

This thesis introduces a series of consensus clustering-based approaches for examining brain responses in spatiotemporal ERP/EEG data. Specifically, the first study proposed a data-driven approach for determining the optimal number of clusters by evaluating the inner similarity of the estimated time window. A consensus clustering method from diverse clustering methods was also designed, including an M-N plot method for configuration. The second study proposed a multi-set consensus clustering approach across individual subjects to determine an appropriate (i.e., precise and stable) time window of ERP of interest. The time window determination method we developed examined two criteria for selecting a representative cluster map: inner similarity and hypothetical temporal coverage. The third study presented a multi-set consensus clustering approach for clustering analysis of single-trial EEG epochs that aimed to identify individual subjects' evoked responses (ERP components). This study also introduced a standardized approach for evaluating scores from signal processing methods. Lastly, the fourth study introduced an ensemble deep clustering pipeline for reliably determining the time window when data quality is imperfect, revealing the adeptness of deep neural networks in feature extraction and time window determination.

In conclusion, this thesis offers a promising computational framework for ERP identification in group-level analysis. The aforementioned studies enhance our understanding of human brain function, have broad implications for computational neuroscience, and suggest adaptable solutions for future neuroimaging investigations.

Keywords: Electroencephalography (EEG), Event-related potentials (ERPs), ensemble learning, consensus clustering, time window, cognitive process, deep clustering, cluster aggregation.

TIIVISTELMÄ (ABSTRACT IN FINNISH)

Mahini Sheikhhosseini, Reza
Konsensusklusterointi tapahtumakohtaisten potentiaalien ryhmätason analyysiin
Jyväskylä: Jyväskylän yliopisto, 2023, 71 s. (+ artikkelit)
(JYU Dissertations
ISSN 2489-9003; 733)
ISBN 978-951-39-9863-9 (PDF)

Tämä väitöskirja esittelee konsensusklusterointipohjaisia lähestymistapoja aivojen vastausten tutkimiseen ERP/EEG-datasta saatavan tiedon avulla. Tutkimuksessa ehdotetaan datavetoista lähestymistapaa optimaalisen klusterien lukumäärän määrittämiseksi arvioimalla estimoidun aikaluokan sisäistä samankaltaisuutta. Tutkimuksessa suunniteltiin myös monipuolisia klustereita soveltava konsensusklusterointimenetelmä, joka sisältää M-N-kaaviomenetelmän konfigurointia varten. Lisäksi väitöskirjassa ehdotetaan monijoukkoista konsensusklusterointimenetelmää yksittäisille koehenkilöille, jotta löydetään sopiva (tarkka ja vakaa) aikaluokka kiinnostaville ERP:lle. Kehitetty aikaluokan määritysmenetelmä tarkastelee kahta kriteeriä edustavan klusterikartan valintaan: sisäinen samankaltaisuus ja hypoteettinen ajallinen kattavuus. Kolmas tutkimus esittelee monijoukkoisen konsensusklusterointimenetelmän yksittäisten koetilaisuuksien klusterianalyysille. Sen tavoitteena on tunnistaa yksilöllisten koehenkilöiden aiheuttamia vastauksia (ERP-komponentteja). Tämä tutkimus esitteli myös standardoidun lähestymistavan signaalinkäsittelymenetelmien pisteiden arvioimiseksi. Viimeiseksi neljäs tutkimus esitteli ryhmän syväklusterointiputken aikaluokan määrittämiseksi luotettavasti, kun datan laatu on epätäydellistä. Tutkimus paljasti syvien neuroverkkojen soveltuvuuden ominaisuuksien erotteluun ja aikaluokan määritykseen.

Tämä väitöskirja tarjoaa lupaavan laskennallisen kehyksen ERP:n tunnistamiseen ryhmätasolla. Edellä mainitut tutkimukset lisäävät ymmärrystä ihmisaivojen toiminnasta. Ne vaikuttavat laajasti laskennalliseen neurotieteeseen ehdottaessaan mukautuvia ratkaisuja tuleviin neurokuvantamistutkimuksiin.

Avainsanat: Elektroenkefalografia (EEG), tapahtumakohtainen potentiaali (ERP), ensemble-opetus, konsensusklusterointi, aikaluokka, kognitiivinen prosessi, syvä klusterointi, klusterien yhdistäminen.

Author

Reza Mahini Sheikhhosseini
Faculty of Information Technology
University of Jyväskylä
Finland
Email: remahini@jyu.fi
https://orcid.org/0000-0001-6833-1437

Supervisors

Timo Hämäläinen
Faculty of Information Technology
University of Jyväskylä
Finland

Fengyu Cong
Faculty of Information Technology
University of Jyväskylä
Finland

Tapani Ristaniemi
Faculty of Information Technology
University of Jyväskylä
Finland

Reviewers

Fabio Babiloni
Department of Molecular Medicine
University of Rome Sapienza
Italy

Zhiguo Zhang
School of Biomedical Engineering
Shenzhen University
China

Opponent

Raju S. Bapi
International Institute of Information Technology
India

ACKNOWLEDGMENTS

When all those years come in front of my eyes, it is like saying goodbye to my PhD life—from the day I landed in China to this moment in Finland. I remember and appreciate all the people who were part of this journey. I would like to express my gratitude to the people who helped and supported me, making that possible.

First of all, a sincere thank you to Prof. Fengyu Cong, who welcomed me into his lab in China. He not only allowed me to learn all the technical works but also supported me in both China and Finland. Fengyu guided me in my academic career and provided an opportunity for me to make my research network. He taught me research at different levels and opened new insights that were a turning point in my research.

I would like to express my heartfelt thanks to my supervisor, Prof. Tapani Ristaniemi (RIP). I will always cherish and not forget Tapani's kind support and open-mindedness to allow me to join his team at the University of Jyväskylä. Sadly, I missed the chance to meet him in person due to being in my country during the pandemic time before coming to Finland.

I would like to express my sincere gratitude to my supervisor, Prof. Timo Hämäläinen. Timo is such a kind and generous person who is more than a supervisor but a great friend. He never let me worry about my steps but supported me to improve all the time. His support and encouragement by replying to my email with 'Good work' let me focus on the work progress. I feel lucky to have met Timo since I have learned not only doctoral research but also working principles and genuine meaningful communication.

I express my sincere appreciation to Prof. Asoke K. Nandi, who had a huge role in my PhD study from the day I met him in Dalian, China. He treated me as his own PhD student, supporting me all the time. He was there to help whenever I ran to him asking things. I have learned from him the research spiritual.

I am also grateful to the reviewers of my dissertation, Professor Fabio Babiloni from the University of Roma Sapienza in Italy and Professor Zhiguo Zhang from Shenzhen University in China. Their constructive comments and suggestions helped to improve the impact of this dissertation. I would also like to sincerely thank Professor Raju S. Bapi from the International Institute of Information Technology in India for accepting to be my opponent.

I also have to express my gratitude to JYU Brain&Mind, especially to Prof. Tiina Parviainen, Prof. Piia Astikainen, and Prof. Simo Monto for welcoming me to CIBR to enjoy learning more practical aspects of my work.

Furthermore, I am sincerely grateful to Behrouz Jedari, Parisa Ghorbani, Mahdi Zarei, Aziz ur-Rahman Aziz, Nauman Khalid Qureshi, Guanghui Zhang, Amir Barjasteh, Fatemeh Irani, Amir Jalili, Ziba Saleki, Sayyora Rustamovna, Baxtiyor Qosimov, Dilnoza Yuvasheva, Hamid Jafarzadeh, Ayaz Karimov, Majid Haghparast, Yu-Qin Deng, Deqing Wang, Xiaoyu Wang, Xiulin Wang, Huashuai Xu, Xiaoshuang Wang, Wenya Liu, Lili Tian, Yang Tiantian, Lina Sun, Zhang Chi, Fan Li, Wei Zhao, Yuxing Hao, Xingyu Hao, Guaqiang Hu, Dongdong Zhou,

Zhonghua Chen, Fu Rao, Ying Li, Liting Song, Jiaqi Zheng, Yalin Sun, Xiangyu Rong, Xinying Chen, Chi Jin, Toni Takala, Leevi Saukkola, and all other lovely friends in Jyväskylä and Dalian who made this journey joyful.

I would like to thank the China Scholarship Council and the Dalian University of Technology for the opportunity and funding that let me start my training and learning in China. My sincere thanks to the University of Jyväskylä and the Faculty of Information Technology for providing funding that made here an enjoyable working environment for me. I sincerely appreciate the time and effort Marja-Leena Rantalainen has spent preparing this dissertation for proceeding. I would also like to thank Nina Pekkala and Elina Salo-Pöyhönen from IT faculty for their kind support all the time.

Finally, I would like to express my deepest gratitude to my parents, my lovely grandma, my brothers, and my sister for their unconditional love and support me all the time. I dedicate this dissertation to my parents for always trusting me. Thanks again to my younger brother for bringing me peace of mind by caring for my family. My special gratitude to my lovely niece, who brought joy and happiness to the whole family, and thanks for all the nice moments you gave me.

Jyväskylä 19.11.2023
Reza Mahini

LIST OF ACRONYMS

AAHC	Atomize and agglomerate hierarchical clustering
ACC	Accuracy
AE	Autoencoder
AMI	Adjusted mutual information
ANOVA	Analysis of variance
ARI	Adjusted Rand index
CC	Consensus clustering
CNN	Convolutional neural network
CSPA	Cluster-based similarity partitioning algorithm
DEC	Deep embedded clustering
DFS	Diffusion map spectral clustering
DNN	Deep neural network
EEG	Electroencephalogram
ERP	Event-related potential
FCM	Fuzzy c-mean
FC_MLP	Fully connected multi-level perceptron
fMRI	Functional magnetic resonance imaging
KMD	k-medoids clustering
GEV	Global explained variance
GFP	Global field power
GMD	Global map dissimilarity
GMM	Gaussian mixture mode
HC	Hierarchical clustering
HE	Hyperedge
HGPA	Hypergraph-partitioning algorithm
ICA	Independent component analysis
LSTM	long short-term memory
MCLA	Meta-clustering algorithm
MEG	Magnetoencephalogram
NMI	Normalized mutual information
PCA	Principal component analysis
SD	Standard deviation
SEM	Standard error of measurement
SME	Standard measurement error
SNR	Signal-to-noise ratio
SOM	Self-organizing map
SPC	spectral clustering
TW	Time window
VAE	Variational autoencoder

FIGURES

TABLES

CONTENTS

LIST OF INCLUDED ARTICLES

I **Reza Mahini,** Peng Xu, Guoliang Chen, Yansong Li, Weiyan Ding, Lei Zhang, Nauman Khalid Qureshi, Timo Hämäläinen, Asoke K. Nandi, and Fengyu Cong (2022). Optimal number of clusters by measuring similarity among topographies for spatio-temporal ERP analysis. *Brain Topography*, 35, 537–557. https://doi.org/10.1007/s10548-022-00903-2.

II **Reza Mahini**, Yansong Li, Weiyan Ding, Rao Fu, Tapani Ristaniemi, Asoke K. Nandi, Guoliang Chen, and Fengyu Cong (2020). Determination of the time window of event-related potential using multiple-set consensus clustering. *Frontiers in Neuroscience,* 14, 521595. https://doi.org/10.3389/fnins.2020.521595.

III **Reza Mahini**, Guanghui Zhang, Tiina Parviainen, Rainer Düsing, Asoke K. Nandi, Fengyu Cong, and Timo Hämäläinen. (2023). Brain evoked response qualification using multi-set consensus clustering: toward single-trial EEG analysis. Submitted to *Brain Topography,* preprint available. https://doi.org/10.21203/rs.3.rs-3586574/v1.

IV **Reza Mahini**, Fan Li, Mahdi Zarei, Asoke K. Nandi, Timo Hämäläinen, and Fengyu Cong (2023). Ensemble deep clustering analysis for time window determination of event-related potentials. *Biomedical Signal Processing and Control,* 86, 105202. https://doi.org/10.1016/j.bspc.2023.105202.

1 INTRODUCTION

If everything around seems dark,
look again, you may be the light.
— Rumi

Electroencephalography (EEG) is a noninvasive neuroimaging technique that records neurological brain activities via electrodes on the scalp, capturing voltage potentials generated by neurons (i.e., from the flow in and around neurons). These electrophysiological cortical activities manifest as consistent brain responses spanning tens to hundreds of milliseconds (Laganaro, 2014; Lehmann et al., 1994). Event-related potentials (ERPs) are brain electrical signals recorded in a time-locked manner to experimental or cognitive events. ERP techniques have become a popular approach for investigating brain neural activities, employing signal-averaging techniques to identify and validate evoked responses resulting from external stimuli, such as sensory, cognitive, and motor stimulation (Kappenman & Luck, 2012a). Brain activity can be characterized as voltage deflections, termed ERP components or peaks that are labeled with P and N followed by a number, indicating their polarity and time of occurrence (Luck, 2014). These components are critical for understanding brain functions and information processing examination (Brandeis et al., 1995; Lehmann et al., 1994; Makeig et al., 2002). Defining an appropriate measurement interval, known as the time window, is crucial for measuring the spatiotemporal properties of neural activity in the brain.

This thesis investigates spatiotemporal ERP data modeling across different resolutions (group and individual subjects) through cluster analysis, with the aim of identifying target ERP components.

1.1 Conventional ERP Component Identification and Limitations

This section describes two types of traditional methods of measuring ERP components from spatiotemporal ERP data: conventional and objective.

1.1.1 Conventional ERP Identification

Conventional ERP component identification consists of two main approaches: 1) quantifying peak amplitude and latency within the experimentally defined time window and 2) computing the mean amplitude over a specified time window. Both approaches assume that the brain's response in the time window or peak latency is associated with the same brain activity in individuals. Nevertheless, these approaches can overlook important cognitive information if the time window is poorly chosen (Luck, 2014). Despite their limitations, these methods have been widely employed in ERP research.

Specifically, the first approach for detecting a given ERP component involves visually inspecting a prominent peak or targeting significant effect sizes within an experimentally interesting interval (Kappenman & Luck, 2012a; Kiesel et al., 2008). This method remains susceptible to noise and baseline interference, potentially affecting accuracy. However, it can report the effect size obtained from high-frequency noise, leading to biased results. Problems arise when time window selections are predicated upon invalid assumptions. Analyzing peak latency to detect larger effect sizes can be misleading or result in problematic estimations of brain responses. Kappenman et al. (Kappenman & Luck, 2012a) highlighted that "peaks are not the same as components," underscoring these limitations and stressing that the real ERP experiment often allows the elicitation of only a few precise components due to design complexities and component overlap.

The second approach offers greater robustness against noise due to averaging. However, the mean amplitude method is sensitive to component overlap, potentially yielding false findings if component latency varies across conditions. A primary solution to this is the signed area measures method (Sawaki et al., 2012), which relies on defining positive and negative areas relative to the baseline. By applying an appropriate time window, the targeted component can be isolated based on its position relative to the waveform's zero line. A narrower time window is recommended for overlapping components, while a wider window suits cases with fewer components (Luck, 2014).

An alternative, popular approach to time window selection involves the moving fixed time window technique (Mu & Han, 2010; Qi et al., 2003; Rotshtein et al., 2010; Van Overwalle et al., 2009; Wills et al., 2014). This method often spans a reasonable window (e.g., 50 ms or a fine resolution range) following stimulus onset to identify substantial effect sizes. Statistical tools, such as analysis of variance (ANOVA; Manly, 2018) and the *t*-test, are employed to identify the time window. Additionally, high-resolution time-bin analysis (e.g., 5 ms duration or point-by-point analysis) and topographic analysis of variance (TANOVA; Koenig & Melie-García, 2010) are used in this context (Wills et al., 2014). However, it is important to note that while moving time windows provide a perceptible time window determination, they can significantly elevate the risk of Type I errors (Luck & Gaspelin, 2017b). In the following section, we discuss objective techniques for identifying ERP.

1.1.2 Objective ERP Identification

In recent decades, the cluster-based permutation method has gained popularity as an advanced technique for estimating ERP time, particularly in magnetoencephalogram (MEG)/EEG data (Maris & Oostenveld, 2007). This method explores multiple dimensions (sensor, time) to detect clusters of samples with significant *p*-values from *t*-tests, defining temporal adjacency as the time window. This method has been widely used in the neuroscience community by some open-source toolboxes, such as 'FieldTrip' (Oostenveld et al., 2011). This method provides a robust platform for investigating temporal and sensory properties of brain responses in M/EEG data (i.e., both EEG and MEG data) utilizing temporal-sensor adjacency called cluster where the effect size is significant. Nevertheless, this method has limitations, such as the potential oversight of spatial distribution changes, high-frequency noise considerations, and parameter adjustments (Sassenhagen & Draschkow, 2019).

Some sophisticated statistical techniques have been used to explore ERP components, such as principal component analysis (PCA; Donchin & Heffley, 1978) and independent component analysis (ICA; Makeig et al., 1995), which are based on the blind source separation approach. The underlying assumption for qualifying an ERP is that the principal and independent components associated with the ERP of interest are fixed across all subjects (Makeig et al., 1997). In the ICA method, statistical calculations are employed to extract shared ERP components among subjects in a group. However, validating an interesting ERP component involves the subjective determination of the spatiotemporal properties of the ERP of interest (Jung et al., 2001). Some advanced methods from this category have been developed to extract the spatiotemporal features of ERPs of interest from single-trial EEG or individual subjects (Cong et al., 2010; Huster et al., 2020; Rissling et al., 2014). Conversely, PCA decomposes the ERP waveform into isolated components from the calculated latent variables (Dien et al., 2007). ICA and PCA are particularly useful for untangling overlapping ERP components (Makeig et al., 1997).

While conventional methods have been popular for decades, they lack the systematic investigation of spatial and temporal dynamics of the brain inherent in brain information processing. The following sections present advanced methods for investigating and qualifying brain neural responses using machine learning techniques, specifically focusing on spatiotemporal cluster analysis.

1.2 Conventional EEG/ERP Microstate Analysis and Its Limitations

Lehmann (1989) introduced the intricate relationship between brain function and its functional states, illuminating a fundamental parameter in cerebral information processing. Brain electric fields, characterized by transient spatial distributions, reflect neuronal population engagement at specific temporal

moments (Lehmann et al., 1998). The rationale of the microstate analyzing method is that the electric field of the brain stays stable for milliseconds of time range (e.g., 50–150 ms on average) during a neurological response, known as a microstate/state that continuously changes along with the information processing of the brain. The EEG's fine temporal resolution (in the millisecond range) enables the investigation of cognitive functions, including attention, memory, sensory-motor activities, EEG states, spontaneous behaviors, and disorders (Khanna et al., 2015). These microstates are segmentable blocks (quasi-stable elements), often described as "atoms of thoughts" (Lehmann, 1990), that can be presented as a group of momentary topographical brain states.

1.2.1 Microstate Analysis: Insights and Limitations

Microstate classification (i.e., cluster analysis of EEG/ERP microstates) was initially introduced by Lehmann et al. (1994, 1987), and microstate clustering was formalized by Pascual-Marqui et al. (1995) through the development of a modified version of the *k*-means algorithm. Conventional microstate cluster analysis uses two clustering techniques: modified *k*-means and atomize and agglomerate hierarchical clustering (AAHC; Murray et al., 2008). This method considers global field power (GFP)—i.e., the standard deviations of electrode potentials—or GFP peaks derived from time samples as a foundation for clustering spatiotemporal EEG/ERP data. For decades, EEG microstate analysis has been extensively applied and has undergone many minor refinements (Michel & Koenig, 2018). It has been a popular tool in neuropsychological studies due to its integration into various open-source software tools, such as Cartool (Brunet et al., 2011), Ragu (Koenig et al., 2011), LORETA (Pascual-Marqui, 2002), Microstate EEGlab (Poulsen et al., 2018), and the KeyPy EEG Analysis toolbox (Milz et al., 2016). This method holds promise for potential biomarker development (Antonova et al., 2022; Khanna et al., 2015), qualifying information processing in neuroscience studies such as cognitive processes (Ruggeri et al., 2019; Zappasodi et al., 2019), and brain disorder studies (Lehmann et al., 2005; Nishida et al., 2013; Ville et al., 2010; Wunderlin et al., 2022).

While microstate analysis is widely employed to isolate brain responses, it has faced criticism for lacking thorough investigation from an information theory perspective. Moreover, the concept of a "state" remains ambiguously defined. One notable limitation related to the deterministic nature of this approach is that it neglects the high variance and multidimensional characteristics of sensor signals (Mishra, 2021). Additionally, the method's disregard for the polarity of time samples, while the time samples are in the same cluster (considering their GFP values), makes the rationale for this method unclear. Furthermore, the "winner-takes-all" strategy, predicated on the assumption of a limited set of four canonical template maps in EEG and the utilization of global explained variance (GEV) to identify dominant template maps in ERP data, imposes a constraint on this analytical method (Dinov & Leech, 2017; Shaw et al., 2019).

Moreover, spatial clustering from transformed data, particularly GFP peaks, ignores many time samples, especially in the context of low signal-to-noise ratio

(SNR) data. More specifically, noisy data and possibly bad channels influence GFP peaks, resulting in spike-like artifacts within the GFP dataset (Michel & Koenig, 2018), resulting in many narrow clusters if no post hoc method is applied. Therefore, post hoc procedures, such as smoothing and backfitting, become essential for deriving meaningful outcomes (Ahmadi et al., 2020).

1.2.2 Advancements in Time Window Estimation Using Microstate Analysis

Microstate analysis offers three approaches for determining the time window. The first approach employs standard repeated measurements ANOVA or multivariate statistical approaches (MANOVA) within a fixed time window—i.e., encompassing the interesting cluster map(s) and predefined electrode sites across subjects. The second approach involves employing statistical tests, such as *t*-tests, across time points to identify temporal areas of significant effect size, utilizing representative cluster map(s) associated with the target ERP. Subsequently, the obtained time window and preselected electrode sites (i.e., relying on a significant effect size) can result in bias when low SNR data or high-frequency noise exists in the data—i.e., a high rate of false positives (Luck & Gaspelin, 2017a). Thus, a robust statistical technique known as multivariate randomization analysis was introduced in the literature on microstate analysis (Koenig & Melie-García, 2010; Michel et al., 2009). This method conducts iterative statistical tests on GFP values (in time sample resolution) and evaluates global map dissimilarity (GMD) for temporal distribution (time window) measures (Michel et al., 2009) at various resolutions (single trials or subjects). While this approach is valuable, it may underestimate the temporal distribution and stability of neural responses, potentially limiting its ability to capture a comprehensive neural profile.

They additionally established statistical tests such as point-by-point topographic consistency (TCT; Koenig & Melie-García, 2010), which assesses map differences across subjects, and TANOVA (Koenig & Melie-García, 2010) that is utilized to compare variations in map characteristics among statistical factor levels (Habermann et al., 2018; Ruggeri et al., 2019). While these randomization techniques offer robust statistical support for exploring effect size, they are computationally expensive and may lack interpretative insights. Further investigation of component latency involving electrode-by-electrode *t*-tests on mean maps within predefined time windows is needed (Michel et al., 2009). As a third approach, some ERP studies utilize microstate clustering results to guide time window selection based on cluster maps (Bailey et al., 2019; Berchio et al., 2019; Michel & Koenig, 2018). Likewise, the first approach, this method considers statistical tests, i.e., point-by-point TANOVA, besides considering cluster map(s) as the brain response (Ruggeri et al., 2019) or solo cluster map as the brain stable response representative (Koenig et al., 2014).

1.3 Cluster Analysis of Spatiotemporal EEG/ERP Data

Spatiotemporal EEG/ERP cluster analysis explores groups of time samples (although continuous) that share similar properties. In this thesis, a time sample refers to an observation with spatial attributes (topographical map) at each timestamp. Therefore, the rationale for clustering time samples is that the consecutive time samples within a cluster exhibit corresponding stable spatial patterns, revealing a very high spatial correlation (one, in theory) over tens of milliseconds. Cluster analysis of EEG may be crucial due to the complexity of EEG signals and critical neurological activities that carry linked patterns (cluster maps) that might be associated with the brain's information processing (Milz et al., 2016). Apart from learning strategies, cluster optimization, and similarity metrics, the primary goal of cluster analysis is to optimally group time samples into K clusters, maximizing inter-cluster similarity while minimizing intra-cluster variance (Abu-Jamous, Fa, & Nandi, 2015).

1.3.1 Conventional Methods for Selecting the Number of Clusters

One challenging aspect of clustering neuroimaging data is determining an optimal number of clusters to appropriately model the data and effectively characterize the underlying structure. In neuroimaging data such as EEG/ERP, the trade-off between cluster count and data compression is crucial. Fewer clusters increase data compression but reduce interpretability, while more clusters enhance interpretability at the cost of compression (Handy, 2009; Murray et al., 2008). Evaluating the quality of clustering is also essential for estimating the optimal number of clusters. Notably, existing measurement methods for this purpose (Milligan & Cooper, 1985) may not be tailored for isolating nuanced brain neurophysiological activities. Among the most popular methods, the Dunn index (Dunn, 1974), which is based on evaluating variances between cluster members and separation, as well as methods based on explained variance measurement (Goutte et al., 1999; Lleti et al., 2004), the Silhouette index (Rousseeuw, 1987), and Gap statistics (Charrad et al., 2014), have received considerable attention from the neuroscience community.

Furthermore, microstate analysis employs cross-validation analysis (Pascual-Marqui et al., 1995) and Krzanowski-Lai (Tibshirani & Walther, 2005) to address this problem. However, while these techniques help determine the number of clusters, they may not comprehensively address the quality of the identified brain neurophysiological activities. These studies commonly validate the existence of established dominant microstate classes (four canonical EEG microstate classes). Nevertheless, some ERP studies attempt to estimate the optimal number of clusters using data-driven techniques that optimize the quality of isolated ERP components (Koenig et al., 2014; Mahini et al., 2022).

1.3.2 Strategies for Clustering Spatiotemporal ERP Data

Typically, two primary clustering strategies are used for cluster analysis of spatiotemporal ERP. The first strategy, known as polarity-invariant clustering, employs GFP representation of the data for clustering, concurrently integrating microstate analysis. This method (microstate analysis) assumes that the time samples corresponding to GFP peaks represent comprehensive brain activities encapsulated within the EEG/ERP data, disregarding sample polarity. Alongside the aforementioned microstate clustering methods (modified *k*-means and AAHC), standard clustering techniques, as shown by von Wegner et al. (2018), can effectively cluster GFP peaks, similarly to conventional microstate clustering. In the context of microstate analysis, two perspectives are significant. One approach assigns four registered canonical patterns (Britz et al., 2010; Michel & Koenig, 2018) as microstate classes to EEG time samples. However, some researchers have explored an alternative number of clusters, such as seven canonical clusters (Custo et al., 2017), to analyze EEG data. The underlying assumption is that canonical maps (patterns) explain a significant portion (65–84%) of the variance; i.e., they occur in similar experimental EEG settings (Michel & Koenig, 2018). For ERP data, template maps—dominant ERP maps based on high GEV, with, e.g., 70% from grand/group averaged ERP clustering (in the form of GFP)—are assigned to time samples (Koenig & Melie-García, 2010; Koenig et al., 2014; Murray et al., 2008; Pourtois et al., 2008). At the individual-subject level, the template maps are assigned to ERP time samples based on their spatial correlation.

By contrast, the second strategy adopts a polarity-sensitive stance, emphasizing polarity's role in spatiotemporal ERP clustering. This framework encompasses a wide range of clustering techniques, such as conventional *k*-means clustering (Poulsen et al. 2018) and fuzzy clustering on PCA features (Geva & Pratt, 1994), as well as advanced clustering methods, such as a probabilistic-based method (Dinov & Leech, 2017) involving standard *k*-means (Pena et al., 1999) and fuzzy *c*-means (FCM; Bezdek, 1981). In this sense, consensus clustering has been studied by exercising standard clustering techniques to robustly identify the ERP of interest from the group average ERP (Mahini et al., 2022) and within individual subjects' ERP data (Mahini et al., 2020), and isolating spatiotemporal patterns in EEG data (Song et al., 2019).

One limitation of both cluster analysis strategies is that their performance can be compromised in situations where data quality remains uncertain post-preprocessing. Therefore, factors such as data quality, incoherency, noise ratio, and selected features can lead to noisy and fragmented clusters, highlighting challenges due to data quality intricacies. In recent years, deep learning has demonstrated substantial power in capturing intricate data features that represent the dominant ERP components. Deep neural networks (DNNs) have excelled in various EEG designs (Bashivan et al., 2015; Cecotti & Graser, 2011; Sikka et al., 2020; Zhang et al., 2019). Within this evolution, the context of deep clustering emerged, empowering DNNs to create cluster-oriented feature representations. DNNs with an embedded clustering module are materialized to

transform data points into cluster-friendly representations (Aljalbout et al., 2018; Ren et al., 2022). However, two prominent strategies, semi-supervised and unsupervised techniques, have been introduced for deep clustering (Aljalbout et al., 2018; Min et al., 2018). Notably, deep clustering has been applied in recent EEG studies, including microstate extraction with GFP peak labeling (Sikka et al., 2020), EEG signal grouping (Peterson et al., 2022), and ensemble deep clustering for time window determination (Mahini et al., 2023).

1.4 Research Motivation

This thesis explores the cluster analysis of spatiotemporal ERP to qualify the ERP component of interest. The following studies address key research questions and motivations within this context.

Article I: This study focused on developing a data-driven mechanism for determining the optimal number of clusters in spatiotemporal ERP data through consensus clustering. Appropriate data compression was recognized as crucial for identifying distinct ERP components. This study systematically tested various clustering methods from polarity-dependent approaches to configure consensus clustering. Group-averaged ERP data underwent clustering across options (e.g., 2–15 clusters) in multiple iterations. During each iteration, the ERP-related time window was estimated using topographical similarity among time samples as a criterion. The optimal number of clusters was determined where inner-time sample similarity was consistently high across clustering options. The framework was rigorously evaluated using simulated and prospective memory ERP data (Chen et al., 2015), enhancing precision in ERP component identification and advancing spatiotemporal ERP analysis techniques.

Article II: Given determining the optimal number of clusters from Article I, this study addresses the intricate task of determining the time window of ERP through cluster analysis of individual-subject ERP data. A multi-set consensus clustering framework combines polarity-invariant and polarity-dependent methods for subject-specific consensus clustering. Building upon this, a subsequent round of consensus clustering was performed to identify shared brain electrical activity patterns across subjects, leading to the determination of a realistic time window for the target ERP component. A modified time window determination method was introduced to guide this process. The proposed approach was comprehensively evaluated using simulated ERP data for accuracy and authentic data from a prior study (Chen et al., 2015). Article II enhances precision and robustness in time window determination for spatiotemporal ERP analysis.

Article III: This study aimed to estimate the time window of the brain evoked response of individual subjects through a cluster analysis of single-trial EEG epochs. It also presented an evaluation mechanism for assessing the

reproducibility of the obtained scores corresponding to the identified cognitive processes of individuals. We designed a multi-set consensus clustering framework, enabling an in-depth investigation of target ERP component at the level of individual trials. Each trial was evaluated for similar brain response compared to the identified ERPs from grand average ERP data. A multi-level consensus clustering approach was then meticulously executed within individual subjects and across trials, yielding clustering results. This approach revealed brain responses linked to intriguing potentials in individual subjects and enhanced our understanding of brain activity at the individual level. By investigating the shared information across individual trials and subjects, as opposed to relying solely on averaged ERP data, Article III presented a perspective that enhances our understanding of neural responses in the context of single-trial EEG analysis.

Article IV: This study addressed the challenge of reliable time window determination when data quality remains imperfect post-preprocessing. Residual noise after preprocessing can lead to uncertain cluster maps during conventional cluster analysis, significantly impacting the reliability of the selected time windows. To address this issue, we proposed an ensemble deep clustering pipeline that combines diverse deep clustering models from semi-supervised and unsupervised approaches. Leveraging the robust learning capabilities of DNNs, this approach resulted in potent data feature labeling and enhanced clustering within a transformed feature space. Inspired by this dual-pronged approach, the method combines clusterings from diverse perspectives to obtain the most dependable clusters capable of pinpointing the most influential components. The proposed methodology was rigorously evaluated using simulated and real ERP data (Kappenman et al., 2021), incorporating varying noise intensity levels. This study contributes a comprehensive understanding of reliable time window determination by harnessing the capabilities of deep cluster analysis and complementing current knowledge on cognitive processes.

1.5 Research Design and Structure of the Thesis

Here, we describe the research design employed in this thesis to address the diverse challenges inherent in spatiotemporal EEG/ERP data cluster analysis. The initial investigation sought to determine the optimal number of clusters to identify an ERP of interest. The significance of this inquiry is that, without a proper number of clusters (in the worst-case scenario), the ERP components can likely be divided into subclusters or be assigned to multiple components within a single cluster map. Thus, a reliable estimation of the number of clusters is essential for precise ERP identification. Once the optimal number of clusters was determined, the next challenge was to establish a robust (i.e., stable and accurate) time window determination mechanism. The goal was to achieve stability and

accuracy in delineating time windows using a reliable clustering of individuals. Following that, we extended the investigation of cognitive processes through EEG single trials to the intricate landscape of individual subjects' brain responses to enhance our understanding of cognitive dynamics, ultimately contributing to individual subjects' neural activity. A final imperative addressed the common issue of imperfect preprocessing. Therefore, powered by DNNs, a new framework was developed to enable researchers to identify the major components of ERP data. Figure 1 presents a relationship between the studies, method, data resolution, and the goal of them.

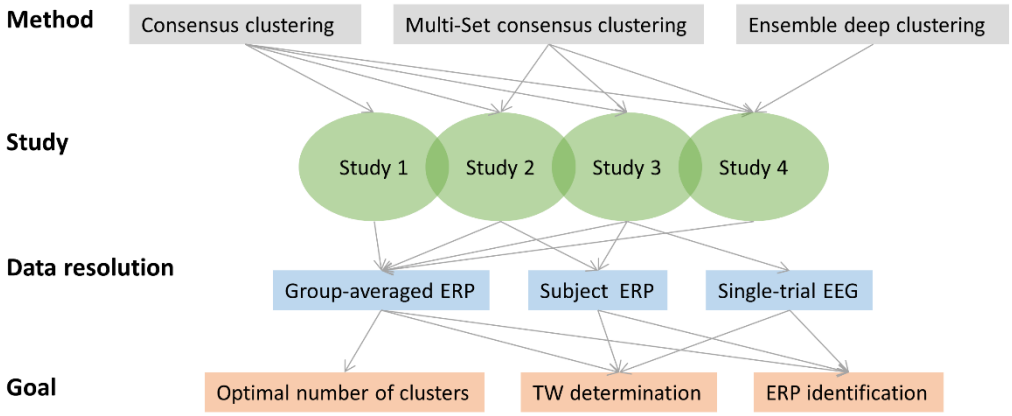


FIGURE 1 The relationships between the studies included in this thesis (methods, data resolution, and goals). TW= Time window.

The organizational structure of this thesis is as follows:

Chapter 1 introduces the fundamental concept of EEG/ERP cluster analysis, identifies limitations of existing methods, and defines the research objectives.

Chapter 2 describes ERP cluster analysis, the methods, and solutions in more detail, including the mathematical concepts of the methods used in this thesis.

Chapter 3 provides a comprehensive summary of the included studies and lists the authors' contributions to each one.

Chapter 4 concludes with a discussion that summarizes the key findings, implications, and limitations of the studies and suggests avenues for future research.

2 METHODS

This chapter presents a theoretical introduction to spatiotemporal ERP data cluster analysis methods, including spatial clustering, consensus clustering, multi-set consensus clustering, and ensemble deep clustering. Figure 2 illustrates an overview of the steps of the clustering analysis mechanism in this thesis. This chapter also provides concise explanations of the metrics used in this thesis, categorized into four groups. Furthermore, we detail the design of a specialized method for determining the time window for qualifying interesting ERPs.

2.1 Theoretical Introduction to Spatiotemporal ERP Cluster Analysis

The core objective of spatiotemporal ERP cluster analysis is to provide an abstract and informative representation of brain signals in ERP data. Despite the continuous nature of the recorded cortical signal, each time sample within the data can be conceptualized as a topographical map. The significance of cluster analysis of brain signals lies in the inherent relationship between cluster maps and their variations in brain functions (e.g., see Figure 3) and information processing mechanisms (Brandeis & Lehmann, 1989; Brandeis et al., 1995; Koenig & Lehmann, 1996). In this context, an effective clustering method is expected to reveal coherent cluster maps, aligning with the "state" concept explained in Section 1.2. The term "perfect cluster map" refers to consecutive time points within a cluster with high topographical correlation (ideally, one).

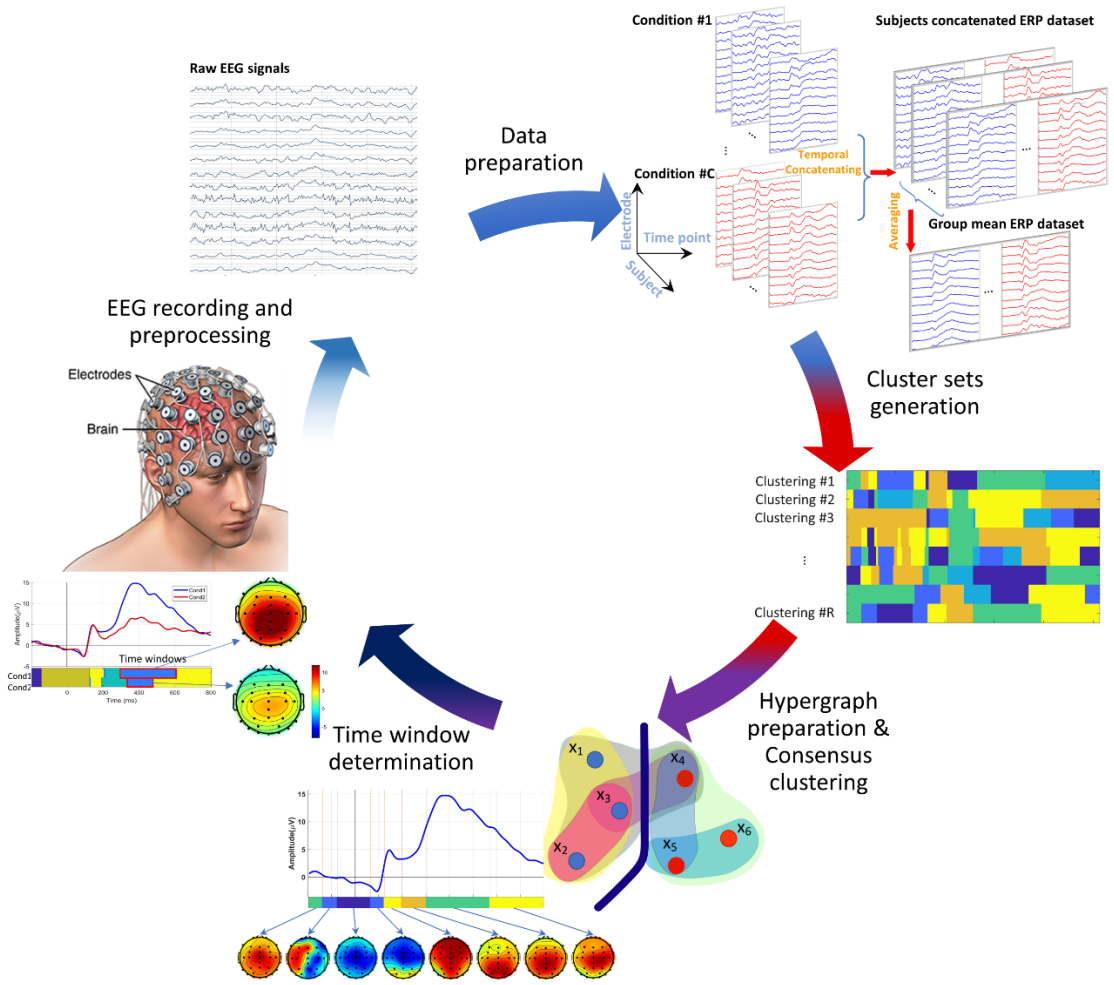


FIGURE 2 Consensus clustering steps for processing spatiotemporal ERP data.

Theoretically, cluster analysis of ERP/EEG data requires clear definitions of two key factors: the data configuration for clustering and the similarity measure used to optimize clustering assignments. In this context, time samples serve as observations, and electrode voltages act as features. For instance, microstate analysis predominantly relies on spatial similarity (topographical similarity; see Section 2.2.2). Formally, the clustering problem of N time samples, $X = \{x_1, x_2, \ldots, x_N\}$ into K groups can be formulated, where each cluster is represented by a centroid μ_k, $k = \{1,2, \ldots, K\}$ and $x_t \in \mathbb{R}^F$, $x_t = \{e_1, e_2, \ldots, e_F\}$ (i.e., a topography map), $t = \{1,2, \ldots, N\}$, and F denote the number of features (electrodes on the EEG scalp). Given that the role of a clustering method is to assign time samples to one of the K clusters $C = \{C_1, C_2, \ldots, C_K\}$, which is labeled by the cluster's number, i.e., $C_k = \{x_t | y_t = k, \forall t \in 1,2, , \ldots, N\}$.

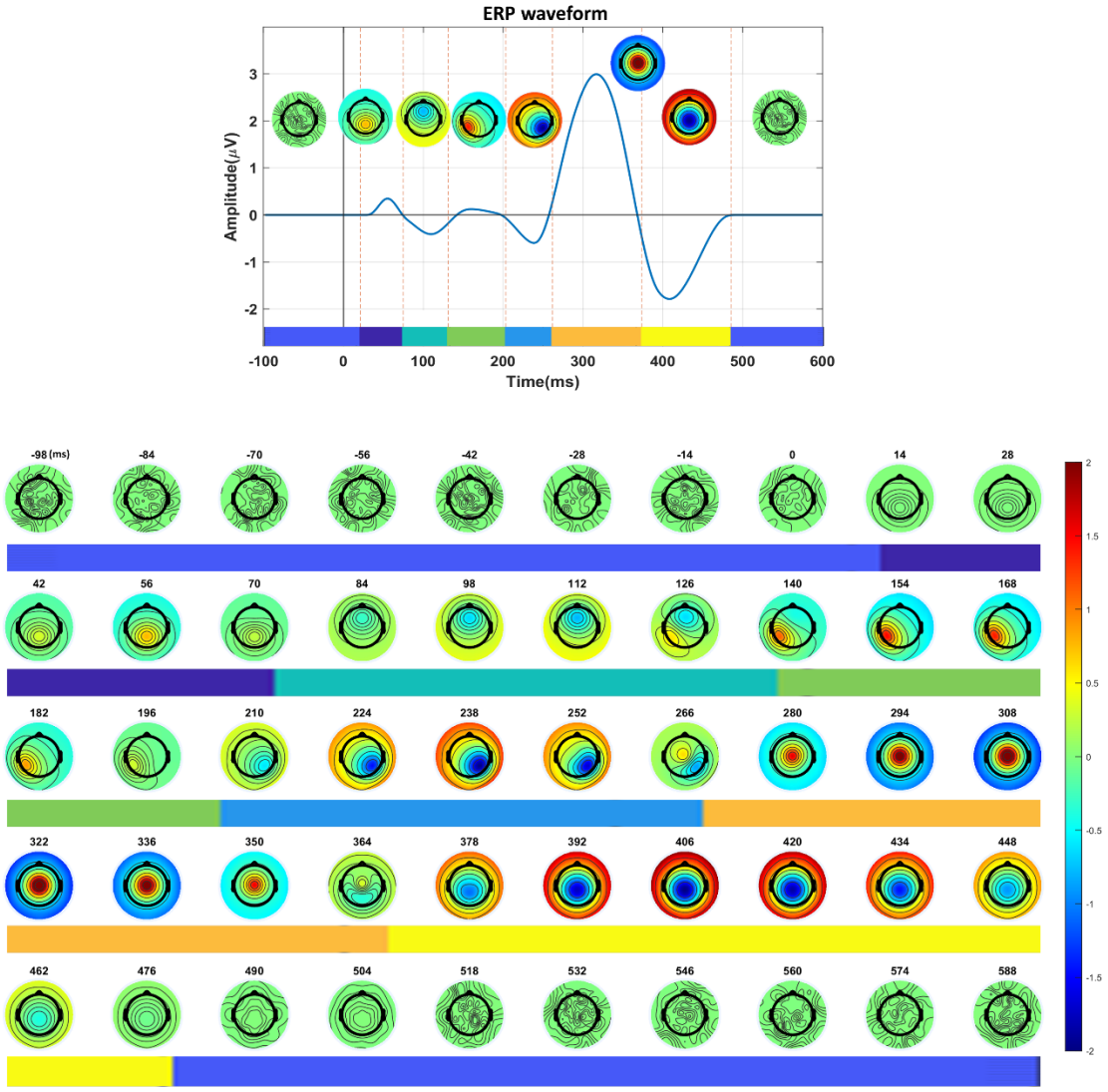


FIGURE 3 Spatiotemporal ERP clustering and topographical pattern dynamics with a 6 ms moving window.

2.2 Metrics

This section outlines the metrics utilized in cluster analysis, specifically those relevant to this thesis, which are: similarity/dissimilarity metrics, spatial analysis metrics, similarity measurement indices, and reproducibility metrics.

2.2.1 Similarity/ Dissimilarity Metrics

There are different proximity measurements, such as similarity, dissimilarity, correlation, and distance, which are conceptual measurements of the distance between two given objects (two time samples). For the cluster analysis of spatiotemporal EEG/ERP, given its non-stationary and quantitative nature,

along with multiple electrode recordings, a selection of widely used similarity measures for objects with continuous features is presented in Table 1.

TABLE 1 Similarity/dissimilarity metrics and formulas and a brief description of each metric.

No.	Measure	Formula	Comments
1	Minkowski distance	$D_{u,v} = \left(\sum_{i=1}^{F} \lvert u_i - v_i \rvert^p\right)^{1/p}$	Minkowski distance between two time points, u and v, from F electrodes is a metric of Euclidean space with a p dimension. It is also considered a generalization of Euclidean, Manhattan, and Chebyshev distances.
2	Euclidean distance	$D_{u,v} = \left(\sum_{i=1}^{F} \lvert u_i - v_i \rvert^2\right)^{1/2}$	Euclidean distance is a special form of Minkowski at p=2.
3	Manhattan distance	$D_{u,v} = \sum_{i=1}^{F} \lvert u_i - v_i \rvert$	Manhattan distance is a special form of Minkowski distance at p=1.
4	Chebyshev distance	$D_{u,v} = \max_{1 \le i \le F} \lvert u_i - v_i \rvert$	Chebyshev distance is a special case of Minkowski at $p \to \infty$.
5	Mahalanobis	$D_{u,v} = \sqrt{(u_i - v_i)\Sigma^{-1}(u_i - v_i)}$	$\sum$ is the covariance matrix.
6	Pearson correlation	$S_{u,v} = \frac{\sum_{i=1}^{F}(u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\sum_{i=1}^{F}(u_i - \bar{u})^2}\sqrt{\sum_{i=1}^{F}(v_i - \bar{v})^2}}$	This is equal to the spatial correlation between two time points. Dissimilarity (correlation distance) can be expressed by $1 - S_{u,v}$.
7	Jackknife correlation	$J_{u,v} = \min\left\{S_{u,v}^{(1)}, S_{u,v}^{(2)}, \dots, S_{u,v}^{(F)}\right\}$, $S_{u,v} = (1 + J_{u,v})/2$	Jackknife correlation is robust to the single outlier; $S_{u,v}^{(i)}$ is the correlation between u and v when ignoring the i^{th} feature.
8	Cosine similarity	$S_{u,v} = \frac{u \cdot v}{\lVert u \rVert \cdot \lVert v \rVert}$	Cosine similarity has been used in many information retrieval applications.

2.2.2 Spatial Metrics

Understanding spatial parameters is critical for EEG/ERP spatial cluster analysis, particularly in microstate analysis, as emphasized in Section 1.2. We briefly introduce the criteria frequently utilized in microstate analysis and other clustering methods presented in this thesis. Table 2 summarizes the list of spatial metrics used in this thesis.

TABLE 2 Spatial metrics for clustering EEG/ERP data.

No.	Measure	Formula	Comments
1	Average reference	$\bar{u} = \frac{1}{F}\sum_{i=1}^{F} U_i$	Average reference calculation for the recording voltages.
2	Global field power (GFP; Lehmann & Skrandies, 1980)	$GFP_u = \sqrt{\frac{1}{F}\sum_{i=1}^{F}(u_i - \bar{u})^2}$, $GFP_u = \sigma_u$	GFP at a given time point u that is the same as the standard deviation.
3	Spatial correlation	$Corr_{u,v} = \frac{\sum_{i=1}^{F} u_i \cdot v_i}{\|u\| \cdot \|v\|}$, $\|u\| = \sqrt{\sum_{i=1}^{F} {u_i}^2}$, $\|v\| = \sqrt{\sum_{i=1}^{F} {v_i}^2}$	Pearson cross-correlation coefficient between two time points.
4	Global map dissimilarity (GMD)	$GMD_{u.v} = \sqrt{\frac{1}{F}\sum_{i=1}^{F}\left(\frac{u_i - \bar{u}}{GFP_u} - \frac{v_i - \bar{v}}{GFP_v}\right)^2}$	Dissimilarity of two maps from two time points or two conditions ($u \neq v$); $\bar{u}$, $\bar{v}$ denotes the mean of electrodes for the first and second maps.
5	Global explained variance (GEV)	$GEV_c = \frac{\sum_{t=1}^{tmax}(GFP_c(t) \cdot Corr_{c,T_t})^2}{\sum_{t=1}^{tmax} GFP_c^2(t)}$, $GEV = \sum_{k=1}^{K} GEV_k$	GEV for a given cluster map. $GFP_u(t)$ is its GFP in condition u , and $Corr_{u,T_t}$ is the spatial correlation between the topography at the time point t and the assigned template map T_t . GEV_k is the GEV of cluster *k* from *K* clusters.
6	Inner similarity	$D_{cl} = d(Corr_{v,u}, Corr_{v,v})$, $u \neq v$ $InnSim = 1 - D_{avg}$ $D_{avg} = \frac{1}{n_{cl}}\sum_{i}^{n_{cl}} D_{cl}(i)$	$d(Corr_{v,u}, Corr_{v,v})$ is the distance between each element of the correlation matrix ($Corr$) with self-correlation (i.e., 1); n_{cl} is the number of elements in the cluster map.

2.2.3 Clustering Similarity Metrics

Similarity measures are defined to measure the similarity between two given clustering results (clusterings) in the same dataset. These indices quantify the information changes from clustering C to clustering C', including both lost and gained information. The comparison of clusterings is often articulated through the **confusion matrix** or the **contingency table**, also known as the **association matrix** of two clusterings C, C'. Mathematically, the confusion matrix is a $k \times k'$ matrix where kk'^{th} is the number of data points in the intersection of C_k, $C'_{k'}$ — i.e., the elements of k^{th} cluster of C_k and k'^{th} cluster of $C_{k'}$,where kk'^{th} can be calculated as:

$$nkk' = |C_k \cap C'_{k'}|. \quad (1)$$

We introduce parameters that define the criteria for agreement or disagreement between two clusterings. Given a dataset of n samples, the pairs of points are classified into four classes:

- N_{11}: the number of pairs in the same cluster under both C, C'
- N_{00}: the number of pairs in the different clusters under both C, C'
- N_{10}: the number of pairs in the same cluster under C but not in C'
- N_{01}: the number of pairs in the same cluster under C' but not in C

These four classes always satisfy the following equation:

$$N_{11} + N_{00} + N_{10} + N_{01} = n(n-1)/2, \quad (2)$$

which is derived from the confusion matrix $[nkk']$. The similarity measurements used in this thesis are listed in Table 3.

TABLE 3 Agreement/disagreement indices for two clusterings.

No.	Measure	Formula	Comments
1	Fowlkes and Mallows index (Fowlkes & Mallows, 1983)	$W_I(C,C') = \frac{N_{11}}{\sum_k n_k(n_k-1)/2}$, $W_{II}(C,C') = \frac{N_{11}}{\sum_{k'} n'_{k'}(n'_{k'}-1)/2}$, $\mathcal{F}(C,C') = \sqrt{W_I(C,C')W_{II}(C,C')}$	For given two clustering results C, C', two asymmetric indices called Wallace distance W_I *and* W_{II} for calculation of Fowlkes and Mallows index that $\mathcal{F}$ denotes.
2	Rand index (Rand, 1971)	$\mathcal{R}(C,C') = \frac{N_{11}+N_{00}}{n(n-1)/2}$	The similarity results in the range of [0, 1].
3	Adjusted Rand index (ARI) (Hubert & Arabie, 1985)	$ARI(C,C') = \frac{\mathcal{R}(C,C') - E[\mathcal{R}]}{1 - E[\mathcal{R}]}$	The reason for adjusting indices (e.g., $\mathcal{R}$ and $\mathcal{F}$) is the observation that the unadjusted $\mathcal{R}$, $\mathcal{F}$ do not range over the entire [0, 1] interval (i.e., $\min \mathcal{R}, \mathcal{F} > 0$).
4	Jaccard index (Ben-Hur et al., 2001)	$\mathcal{J}(C,C') = \frac{N_{11}}{N_{11}+N_{01}+N_{10}}$	The Jaccard index is calculated from the parameters mentioned above.
5	Accuracy (Nguyen & Caruana, 2007)	$ACC(C,C') = \max \frac{\sum_{i=1}^{N} 1\{C(i) = m(C'(i))\}}{N}$	Given the known clustering C (i.e., ground-truth) and the clustering result C'. m function provides overall possible one-to-one mappings between clusters and labels (Kuhn, 1955).
6	Adjusted mutual information (Vinh et al., 2010)	$AMI(C,C') = \frac{I(C,C') - E\{I(C,C')\}}{\max\{H(C),H(C')\} - E\{I(C,C')\}}$, $H(C) = \sum_{i=1}^{K} n_i \log \frac{n_i}{n}$	See Section 2.5.5 for more details.

2.2.4 Reproducibility Evaluation

In the context of this thesis, reproducibility refers to the predictability and consistency of estimated stimulus-locked response properties, known as scores, at the individual trial/subject level. Unlike repeatability, which assesses the consistency of repeated results, reproducibility is a measure of obtaining consistent results from different generators (e.g., trials, subjects) that are not necessarily identical. Reproducibility can be defined at both the experimental and data-processing levels. The score refers to spatial and temporal properties of the brain response in the estimated time windows, such as latency, mean amplitude, inner similarity, and time window properties of the identified ERP of interest.

Mathematically, the reproducibility of a score is a function of the standard measurement error (SME; Luck et al., 2021). Given that, we described the reproducibility of the two groups, the analytical assessment in which the error of the scores was calculated from trials of individual subjects denoted by "aSME" and error through a bootstrapping process denoted by "bSME." The standard error from N results (scores) is calculated as follows:

$$SME = \frac{SD}{\sqrt{N}}, \tag{3}$$

where SD represents the standard deviation of the scores. The idea of bootstrapping is that, given an experiment, instead of repeating the experiment many times, the experiment can be simulated by generating sufficient repeats (e.g., a minimum of 1,000) "**with replacement**" of samples and calculating the scores from them iteratively. For instance, given R repeats and the calculation of scores from single-trial epochs of a given subject $s = \{1,2, \dots, S\}$ in condition c, the bootstrap error $\widehat{bSME_s^c}$ can be calculated as follows:

$$\widehat{bSME_s^c} = \sqrt{\frac{1}{R}\sum_{r=1}^{R} \widehat{SME_r}^2}, \tag{4}$$

where the estimated standard measurement error ($\widehat{SME_r}$) for each of the repeats $r = \{1,2, \dots, R\}$ is calculated as follows:

$$\widehat{SME_r} = \frac{\widehat{SD_r}}{\sqrt{N_c^s}}, \tag{5}$$

where N_c^s denotes the number of trials for subject s in condition c in each generation of bootstrapping.

Therefore, one can calculate the scores from each generation, followed by obtaining the measurement error for all the individual subjects as:

$$MS\left(\widehat{SME}\right) = \frac{\widehat{SME_1}^2 + \widehat{SME_2}^2 + \cdots + \widehat{SME_S}^2}{S}, \tag{6}$$

where S is the number of subjects in the group. Additionally, a parameter known as total error $\widehat{Var_{all}}$ is calculated from the individual subjects $\widehat{Var}_{par}$ called true variance, and the measurement error (calculated from Eq. 6). This calculation can be expressed as follows:

$$\widehat{Var}_{all} = \widehat{Var}_{par} + MS(\widehat{SME}). \tag{7}$$

The confidence in the obtained scoring results is quantified to evaluate the reliability of the applied measurement. The reliability of the measurement can be calculated as follows:

$$\widehat{\text{Reliability}} = 1 - \frac{MS\left(\widehat{SME}\right)}{\widehat{Var_{all}}}, \tag{8}$$

Cronbach's alpha and standard error of measurement (SEM) are used to calculate the reliability, estimating the error in individual scores within the subjects. Cronbach's alpha is calculated as follows:

$$\alpha = \frac{q}{q-1}\left(1 - \frac{\sum_{i=1}^{q}\hat{V}_i}{\hat{V}_{tot}}\right), \tag{9}$$

where q is the number of items (i.e., scoring tests), $\hat{V}_i$ denotes the variance associated with each measure, and V_{tot} is the variance associated with all the scores. The $\widehat{SEM}$ is then calculated as follows:

$$\widehat{SEM} = \widehat{SD} \times \sqrt{1-\alpha}. \tag{10}$$

2.3 Data Preparation for Clustering

ERP data from a multiple sensor array represents multiple active neural sources with a time resolution of milliseconds. Generally, two prevalent data preparation methods have been used in the cluster analysis literature (Calhoun et al., 2009; Murray et al., 2008). The first involves creating a large tensor for each subject from temporal concatenating datasets across all conditions. In the second, the ERP datasets prepared for each subject using the first approach are group averaged within each group of subjects. Given a subject *s*, the temporal concatenated ERP data from C conditions is denoted by $X(s) = Conc(X_1^s, X_2^s, .., X_C^s)$, where X_c^s is the ERP dataset of condition $c = \{1,2, \dots, C\}$, $s = \{1,2, \dots, S_g\}$, and S_g denotes the number of subjects in g^{th} group. The $Conc$ function concatenates the individual datasets in the temporal domain. Averaging concatenated datasets across the subjects of each group as $X_g = mean(X(1), X(2), \dots, X(S_g))$ builds a group-averaged concatenated ERP. Figure 4 shows the temporal concatenating of the ERP datasets at the subject and group levels. Consequently, the dataset for clustering has a size of $(NC) \times F$, where N represents the number of time samples and F is the number of features (electrodes).

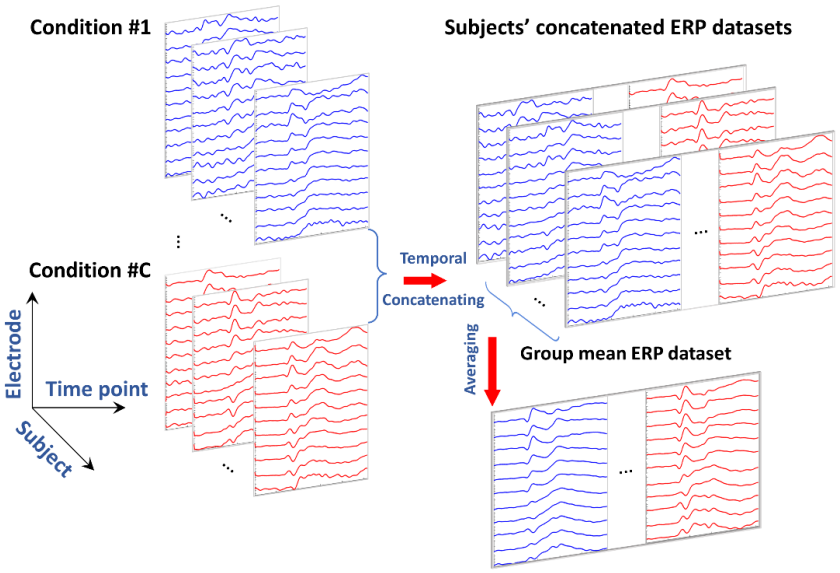


FIGURE 4 Temporal concatenating ERP datasets at the individual-subject and group levels.

2.4 Determination of the Optimal Number of Clusters

In ERP cluster analysis, one popular method for estimating an appropriate number of clusters is to calculate the explained variance by assessing eigenvalues in the dataset (Cong et al., 2014). This entails using PCA to calculate eigenvalues from the covariance matrix and selecting dominant eigenvalues that collectively explain, for example, 95% of the total variance. Mathematically, the covariance matrix, $Cov_X = XX^T$ or $Cov_X = X^TX$, is calculated in descending order, where each eigenvalue corresponds to one component. The eigenvalue order can be shown as follows:

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k = \cdots = \lambda_E = \sigma^2, \tag{11}$$

where $\{\lambda_k\}_{k=1}^{E}$ represents the eigenvalues of the covariance matrix Cov_X and E= min$\{N, F\}$. Figure 5 illustrates an example of Eq. 11, wherein the first eigenvalue signifies the highest explained variance, followed by subsequent eigenvalues in descending order.

As mentioned, conventional microstate analysis employs methods such as cumulative explained variance (Huster & Raud, 2017) and cross-validation (Pascual-Marqui et al., 1995) to optimize the ratio between the GEV and the set of cluster maps. Advanced techniques such as the Krzanowski-Lai criterion (Tibshirani & Walther, 2005), which addresses dimensionality concerns of cross-validation (Murray et al., 2008), and metacriterion (Bréchet et al., 2019; Custo et al., 2017) have been investigated in several studies. Among the conventional methods used to determine the optimal number of clusters, cross-validation (Pascual-Marqui et al., 1995) has received more attention. This method optimizes the ratio between the GEV while trying different sets of cluster maps.

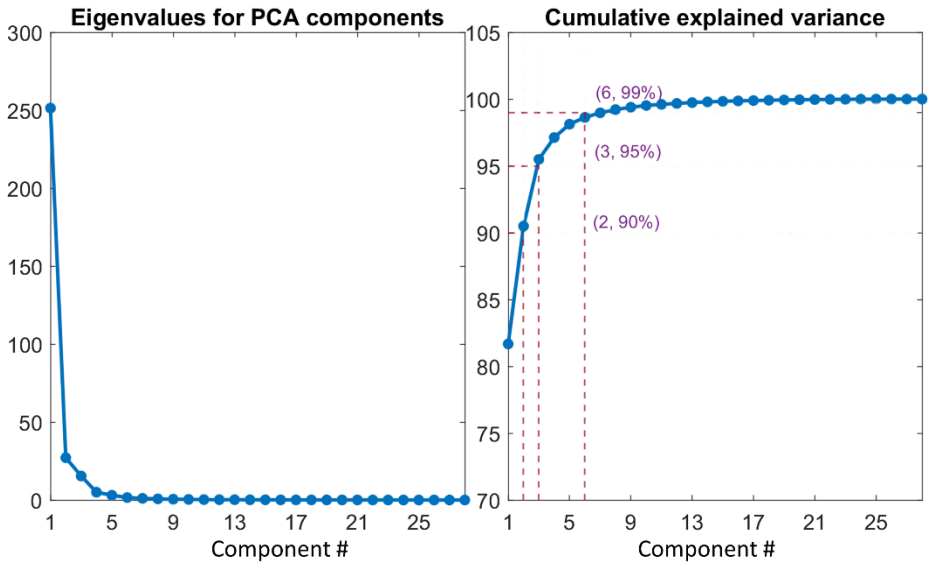


FIGURE 5 Example of estimating the number of clusters/components using the eigenvalues calculation via PCA for the first 30 components (left panel) and their explained variance illustration (right panel).

The data-driven strategies in Section 1.3.1 focus on optimally determining significant components versus modeling the entire dataset. For example, Koenig et al. (Koenig et al., 2014) introduced a method based on measuring the variance explained by n-maps (e.g., n ranging from 2 to 20 cluster maps) through randomization optimization. Another study (Mahini et al., 2022) estimated the optimal number of clusters by assessing a subset of ERP components (the ERPs of interest) using the inner similarity criterion to assess candidate cluster maps. The candidate cluster maps are the clusters with high inner similarity (e.g., > 0.95) within the experimentally relevant range. The details of this method are explained in Section 3.1.

2.5 Clustering Approaches for ERP Data

This section introduces three clustering approaches—EEG/ERP microstate clustering, consensus clustering, and ensemble deep clustering—and focuses on the consensus clustering of ERP data.

2.5.1 ERP Microstates Clustering

Originating from early studies on resting-state EEG (Brandeis & Lehmann, 1986; Brandeis & Lehmann, 1989; Lehmann & Skrandies, 1984), a dominant set of four microstate classes (traditionally labeled A, B, C, and D) reveals rapid transitions (Michel & Koenig, 2018). A primary metric, global map dissimilarity (GMD; Table 2, Eq.4), indicates shifts between successive maps (Brandeis et al., 1995).

The two-step process of EEG microstate clustering involves identifying dominant cluster maps, either predefined or map derived, accounting for a significant portion of the total variance (e.g., 65–70%). These canonical maps are then adapted to individual maps (in group-averaged or individual-subject data) based on spatial correlations via a post hoc backfitting procedure (Pascual-Marqui et al., 1995).

Likewise, for ERP microstate analysis (Koenig et al., 2014; Murray et al., 2008; Pourtois et al., 2008), the group-derived template maps (i.e., obtained from clustering of group-averaged ERP data) with an explained variance of 65–84% are assigned to individual data despite polarity differences. Figure 6 provides an overview of ERP microstate clustering, illustrating peaks of the GFP signal and neglecting the polarity of the cluster maps (e.g., assigning the same cluster map 3 to different polarities that potentially correspond to the N1 and P2 components) in the clustering results. Microstate clustering methods, including *k*-means clustering (Pascual-Marqui et al., 1995) and topographic atomize and agglomerate hierarchical clustering (TAAHC; Murray et al., 2008), leverage spatial similarity to delineate these dynamic patterns.

The methodologies are the following:

Modified *k*-means

Similar to the standard *k*-means, this method randomly initializes *k* cluster centers and iteratively refines assignments based on spatial correlations. The iterative process converges by improving the centroid by calculating new seeds by averaging the assigned map in the previous step until coverage in the optimal assignment is achieved. The clustering quality is evaluated using GEV (see Table 2, Eq.5).

Topographic Atomize and Agglomerate Hierarchical Clustering

The TAAHC method (Murray et al., 2008) begins by treating the original maps as distinct clusters. In each iteration, the worst cluster is identified and split into component cluster maps (atomized). The worst cluster map is the cluster with the lowest sum correlation (Table 2, Eq. 3) between its constituent maps and the average cluster map. Then, the "free" maps are agglomerated to any other remaining cluster maps based on the highest spatial correlation.

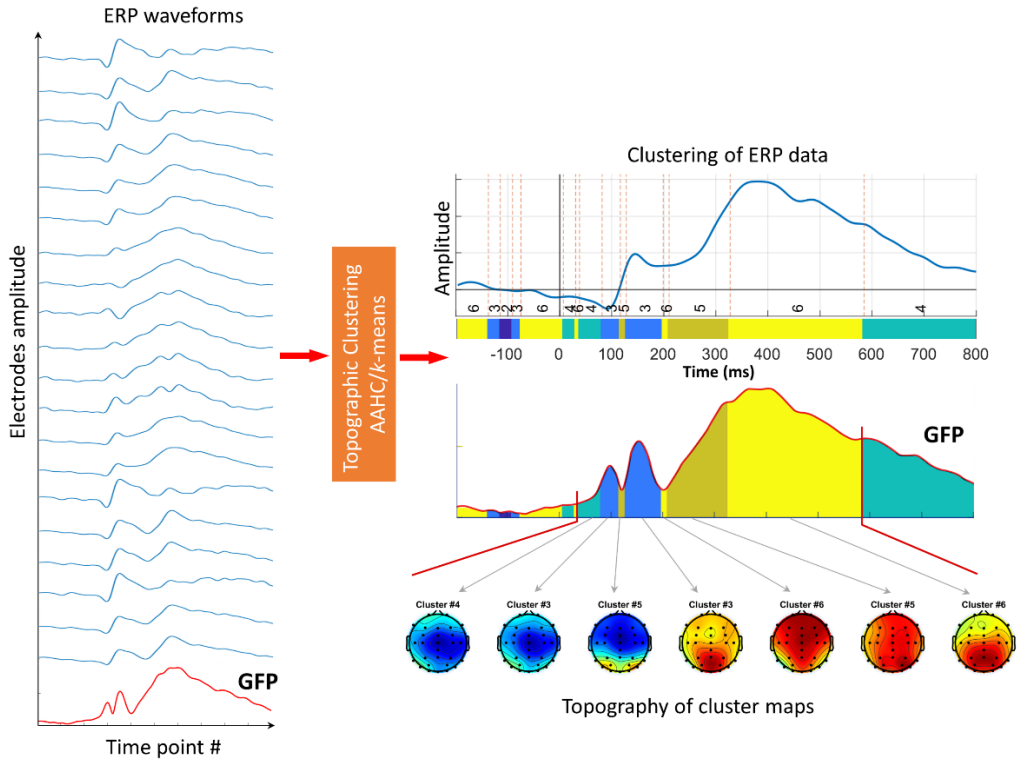


FIGURE 6 ERP microstate clustering using modified *k*-means or topographic atomize and agglomerate hierarchical clustering (TAAHC). Clustering is performed on the peak global field power (GFP) points (i.e., only the time points of the GFP maxima). For instance, some ERP components, such as large P3 components, can be elicited by cluster map 6 (indicated in yellow). Furthermore, two cluster maps from cluster 3 (indicated in blue) have been assigned to two ERP components (i.e., N1 and P2) but with opposite polarities.

2.5.2 Theory of Consensus Clustering

Consensus clustering addresses the problem of combining multiple clustering solutions into a unified, informative, and robust result, often outperforming individual methods (Strehl & Ghosh, 2003). Consensus clustering is assumed to result in robust, high-quality, multi-view, and knowledge-enriched outcomes (Acharya & Ghosh, 2011; Strehl & Ghosh, 2003). One core reason for using consensus clustering is that even popular clustering algorithms might fail spectacularly for certain datasets that do not match the corresponding modeling assumptions (Acharya & Ghosh, 2011). Hence, combining information at different levels, such as datasets, clustering methods, and multiple sources, yields novel insights into neuroimaging data processing. Another reason to use consensus clustering for ERP data is its tremendous success in processing biological data (Abu-Jamous et al., 2013; Abu-Jamous et al., 2015; Monti et al., 2003) as well as human brain functional magnetic resonance imaging (fMRI) and EEG analysis research (Liu, Abu-Jamous, et al., 2017; Liu, Brattico, et al., 2017;

Song et al., 2019). We describe the theory of consensus clustering in two primary steps: the generation and consensus phases.

2.5.2.1 Generation Mechanism

The primary role of the generation phase is to informatively supply the consensus function with clusterings from diverse strategies and methods. The main challenge is deciding what to combine and how to do so, given that even different initial conditions can result in various clustering outcomes. Specifically, the generation step involves the exploration of distinct subsets of samples, features, and parameter settings, including various subspace transformations. Different generation strategies can be designed, such as applying a single clustering method with uniform or varying parameters (e.g., number of clusters, similarity metrics) or combining multiple methods from different categories (Abu-Jamous, Fa, & Nandi, 2015; Golalipour et al., 2021; Vega-Pons & Ruiz-Shulcloper, 2011). Arguably, there is no straightforward solution for selecting appropriate clustering methods for the consensus clustering configuration (Topchy et al., 2005). This is primarily due to the absence of a ground-truth solution, which leads to uncertainty about the explanatory power of different clustering approaches (Abu-Jamous et al., 2015). Nevertheless, several general approaches based on an elimination approach to the generation problem have been employed (Onan et al., 2017; Ryali et al., 2015).

We designed two simple strategies for configuring the generation phase in ERP cluster analysis. The first method utilizes the M-N plot evaluation technique (Abu-Jamous et al., 2014) on group-averaged temporal concatenated ERP data to assess the performance of widely used clustering methods used in neuroimaging (Mahini et al., 2022). This method uses the inner similarity and duration of the estimated time windows as two criteria. ERP microstates polarity-invariant clustering methods, such as modified *k*-means and AAHC (after labeling adjustment), were also used alongside polarity-dependent clustering (Mahini et al., 2023a). Figure 7 illustrates the use of the M-N plot method for selecting the appropriate clustering configuration. The satisfied duration threshold (e.g., 50 ms) can be changed depending on the experiment and ERP of interest.

The second strategy leverages a state-of-the-art method, such as modified *k*-means, as a benchmark. It selects clustering methods with higher similarities to the benchmark across the subjects within each group or condition using a voting mechanism (Mahini et al., 2020). The adjusted Rand index (ARI; Meila, 2007; Strehl & Ghosh, 2003) is utilized to quantify the similarity between the results of individual clustering methods and benchmark clustering.

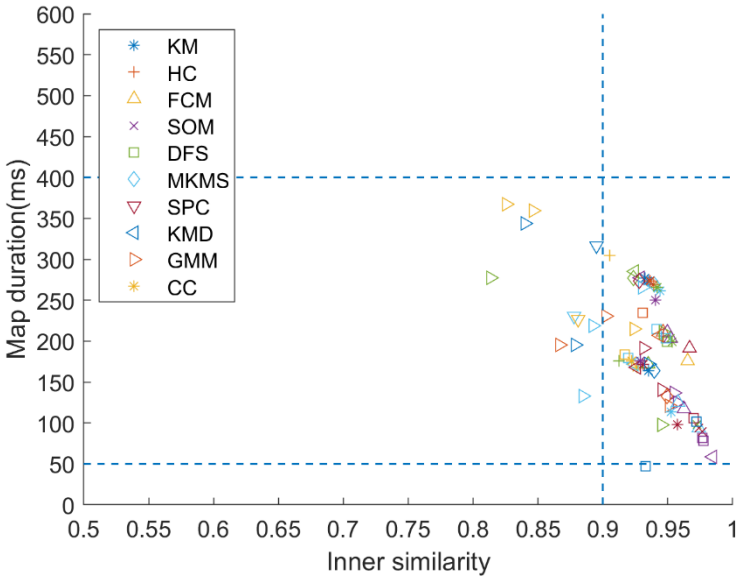


FIGURE 7 Example of M-N plot method results examining a calculated time window obtained from the presented clustering methods in, e.g., 20 repeats and 6 clusters. KM = k-means, HC = hierarchical clustering, FCM = fuzzy c-means, SOM = self-organizing map, DFS = diffusion map spectral clustering, SPC = spectral clustering, KMD = k-medoids clustering, GMM = Gaussian mixture model, and CC = consensus clustering. Notably, the GMM and SPC clustering methods provided weak cluster maps and need to be eliminated. The DFS clustering is suspicious but can be included due to overall suitable cluster map generating.

2.5.2.2 Consensus Mechanism

Once clusterings are obtained in the generation step, the consensus function explores the best-aggregated clustering outcome by assessing the similarity between the clustering sets. From a wide range of existing methodologies for ensemble clustering (Acharya & Ghosh, 2011; Vega-Pons & Ruiz-Shulcloper, 2011), four primary categories can be addressed: partition- and cluster-based approaches, voting, and matrix partitioning (Abu-Jamous, Fa, & Nandi, 2015; Boongoen & Iam-On, 2018; Golalipour et al., 2021). Alongside those methods, this thesis focuses on consensus approaches based on hypergraph partitioning, which involves assessing cluster similarity through hypergraph representations.

Within hypergraph-based strategies, three approaches can be investigated: the cluster-based similarity partitioning algorithm (CSPA), the hypergraph-partitioning algorithm (HGPA), and the meta-clustering algorithm (MCLA) (Strehl & Ghosh, 2003). The primary objective of these consensus functions is to transform the clusterings into a hypergraph representation and calculate the most aggregated clustering results upon them. A hypergraph is defined as a set of vertices and hyperedges, where a hyperedge (i.e., a generalization of an edge context in the graph theory) connects the set of vertices. For each labeling,

denoted as L_r, $r = 1,2, \dots, R$ from R clusterings, a binary membership matrix $H^{(r)}$ with a column of the cluster (named hyperedge) is defined. For example, Table 4 shows the creation of a hypergraph from five sets of clustering results with $K = 3$. An adjacency matrix of a hypergraph H is constructed from the clusterings by concatenating $H^{(r)}$, $\{r|r \in \{1,2, \dots, R\}\}$ to calculate H as follows:

$$H = H^{(1,\dots,R)} = \left(H^{(1)} \dots H^{(R)}\right), \tag{12}$$

where the number of hyperedges HE is calculated as follows:

$$m = \sum_{r=1}^{R} K^{(r)}, \tag{13}$$

where H characterizes the relationships among n objects, HE represents the hyperedges, and $K^{(r)}$ denotes the number of clusters in the r^{th} method.

So far, we have shown hypergraph calculations from the generation phase. Hypergraph clustering (dependent on the chosen policy) collapses cluster groups within the meta-cluster to explore the combined clustering outcome.

TABLE 4 Cluster ensemble problem employing five clustering methods , number of clusters (K=3), and number of time samples in the data (n=6). Original labeling (left) and the hypergraph representation with 15 hyperedges (right). Each cluster map is transformed into a hyperedge.

							$H^{(1)}$			$H^{(2)}$			$H^{(3)}$			$H^{(4)}$			$H^{(5)}$		
	l_1	l_2	l_3	l_4	l_5		h_1	h_2	h_3	h_4	h_5	h_6	h_7	h_8	h_9	h_{10}	h_{11}	h_{12}	h_{13}	h_{14}	h_{15}
x_1	1	1	1	1	2		1	0	0	1	0	0	1	0	0	1	0	0	0	1	0
x_2	1	2	1	2	1		1	0	0	0	1	0	1	0	0	0	1	0	1	0	0
x_3	2	2	1	1	2	⇔	0	1	0	0	1	0	1	0	0	1	0	0	0	1	0
x_4	1	3	3	3	2		1	0	0	0	0	1	0	0	1	0	0	1	0	1	0
x_5	3	3	3	3	3		0	0	1	0	0	1	0	0	1	0	0	1	0	0	1
x_6	3	3	2	2	3		0	0	1	0	0	1	0	1	0	0	1	0	0	0	1

We outline the principal methods employed in hypergraph clustering:

Cluster-based similarity partitioning algorithm (CSPA)

Once the hypergraph is computed using the R clustering methods, the CSPA calculates the similarity matrix from all clustering sets. This method generates a similarity graph in which clusters represent vertices, and the similarity between clusters is measured as the edge weights (Huang et al., 2017). More formally, the heuristics used in this method is that samples within the same cluster are considered fully similar, while those in different clusters are dissimilar. The $n \times n$ similarity matrix (binary matrix) is computed as $A = H^{(r)}H'^{(r)}$, for the given r^{th} clustering. Hence, entry-wise averaging of R clusterings is yielded as the overall similarity matrix S with a high granular resolution using Eq. 14. Figure 8, for example, shows the calculation of similarity matrices from six clustering methods and the combined similarity matrix for examining a new cluster set.

Each entry in S indicates the fraction of clusterings in which two objects share the same cluster, which is obtained through a sparse matrix multiplication:

$$S = \frac{1}{R}HH', \tag{14}$$

which is referred to as the cluster-based similarity matrix. The obtained large matrix $(n \times n)$ is then employed to recluster the time points into K clusters. Although CSPA offers reliability, it demands significant memory and processing resources (Karypis & Kumar, 1998).

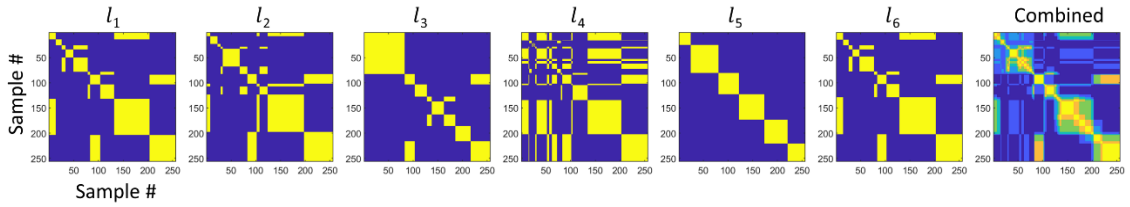


FIGURE 8 Example of a cluster-based similarity partitioning algorithm (CSPA) for hypergraph partitioning problem using six clustering methods. Each clustering method is associated with a similarity matrix. The yellow proportions indicate the similarity of the samples. The average similarity matrix is used for reclustering of the samples in the yielded consensus clusters.

Hypergraph-partitioning algorithm (HGPA)

The HGPA aims to repartition objects (time samples) into strongly connected segments by cutting the hypergraph into a minimum number of hyperedges (Strehl & Ghosh, 2003) in a process known as a minimum cut. In this method, all hyperedges have uniform weights, and all vertices are treated equally. Thus, the HGPA partitions the hypergraph into K unconnected modules in approximately balanced-size clusters. Equality is obtained by satisfying the following equation:

$$K \cdot \max_{k \in \{1,2,..,K\}} \frac{n_k}{n} \leq \beta, \tag{15}$$

where β ensures balance among clusters, given n samples and n_k denotes the number of samples at k^{th} partition. Unlike CSPA, which focuses on pairwise cluster similarity, HGPA considers the global relationships among the objects across the clusters.

Meta-clustering algorithm (MCLA)

The MCLA estimates groups of clusters (clustering clusters) based on a similarity measurement (e.g., binary Jaccard measurement) between the clusters, involving four steps: constructing a meta-graph, hyperedge clustering, meta-cluster computation, and object assignment (Strehl & Ghosh, 2003). With the prepared hypergraph, MCLA groups the related hyperedges and assigns each object/sample to the grouped hyperedges, where it participates strongly. Formally, hyperedges $h_j, j = \{1.2.3 \ldots., m\}$ serve as meta-graph vertices, and hyperedge weights express similarity. The graph partitioning (Karypis & Kumar, 1999) clusters $m = \sum_{r=1}^{R} K^{(r)}$ meta-graph hyperedges into K-balanced meta-clusters (the clusters of hyperedges). Each K meta-cluster produces a single meta-hyperedge by collapsing hyperedges and calculating object association for each meta-hyperedge. Objects are then assigned to the meta-cluster with the highest value based on the winner association confidence ratio compared to others.

However, this method does not guarantee the assignment of at least one object to each meta-cluster.

In summary, the hypergraph partitioning-based consensus functions described above exhibit complexities: CSPA $O(Kn^2R)$, HGPA $O(KnR)$, and MCLA $O(Kn^2R^2)$. CSPA is the slowest method, and HGPA is the fastest method. It has been shown that CSPA's reliability exceeds the two other methods in noisy data (Strehl & Ghosh, 2003). CSPA's memory complexity is high and may be impractical for a large n.

2.5.3 Consensus Clustering for Spatiotemporal ERP

Given preprocessed ERP data, the generation phase of consensus clustering incorporates a set of clusterings from polarity-invariant and polarity-independent clustering methods with suitable within-clustering consistency (Mahini et al., 2020; Mahini et al., 2022). Figure 9 illustrates an example of clustering results on an ERP dataset and the consensus clustering result from combining four clustering results using the CSPA consensus function.

Formally, let X represent the prepared ERP dataset (as described in Section 2.3) with N time points and F electrodes. Given that clustering outcomes yield K clusters denoted as $\{C_k|k = 1,2,\dots,K\}$ with numerical labels $L \in \mathbb{N}^N$ from R clustering methods, a set of R clusterings $\{L^r|r \in \{1,2,\dots,R\}\}$ denoted by $\wedge$ is constructed. These clusterings are used to compute the combined clustering L through an objective function defined as follows:

$$\Gamma: \{L^r|r \in \{1,2,\dots,R\}\} \rightarrow L, \tag{16}$$

where Γ represents the consensus function from $\mathbb{N}^{N\times R} \rightarrow \mathbb{N}^N$ that maps the generated cluster sets into the final clustering.

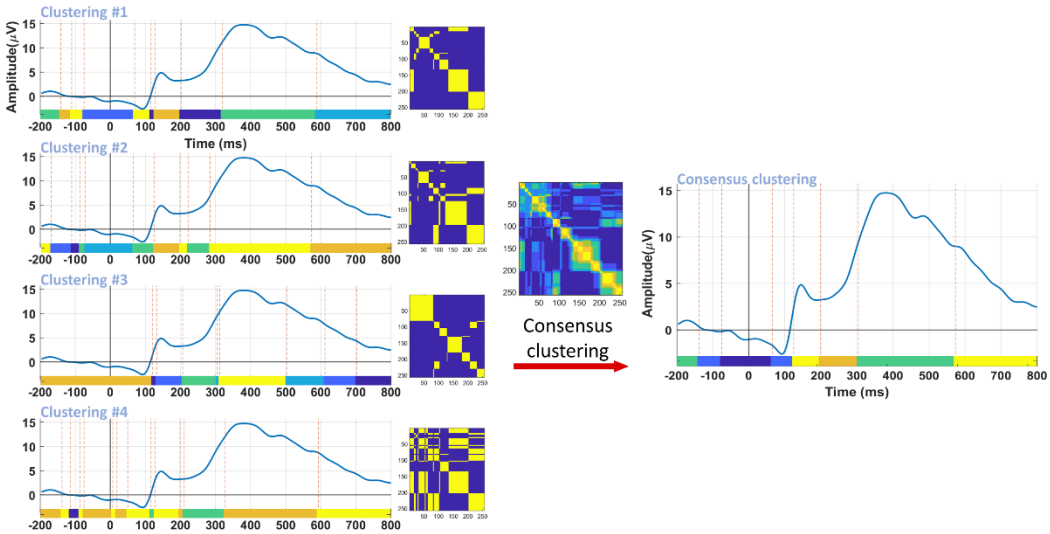


FIGURE 9 CSPA partitioning of group-averaged ERP data using four different clustering methods.

Theoretically, the consensus function is expected to identify the most robust clustering that shares maximum mutual information across all clusterings (Cover & Thomas, 1991). The mutual information between the two clustering results L^i, L^j is denoted as $I(L^i, L^j)$ (Eq. 18), with $H(L^i)$ representing the entropy of L^i. Normalized mutual information (NMI) between L^i, L^j using geometric means can be expressed as follows:

$$NMI\left(L^i, L^j\right) = \frac{I\left(L^i, L^j\right)}{\sqrt{H(L^i)H(L^j)}}, \tag{17}$$

in which

$$I\left(L^i, L^j\right) \leq \min\left(H\left(L^i\right), H\left(L^j\right)\right), \tag{18}$$

$$H\left(\hat{L}\right) = \sum_{a=1}^{K} N_a \log\frac{N_a}{N}, \tag{19}$$

where N_a denotes the number of samples in cluster C_a according to a single clustering like $\hat{L}$. Therefore, for two clustering outcomes L^i, L^j, the mutual information (ranging between 0 and 1) is computed as follows:

$$\Gamma^{(NMI)}\left(L^i, L^j\right) = \frac{\sum_{a=1}^{K}\sum_{b=1}^{K} N_{a,b} \log\left(\frac{N.N_{a,b}}{N_a N_b}\right)}{\sqrt{\left(\sum_{a=1}^{K} N_a \log\frac{N_a}{N}\right)\left(\sum_{b=1}^{K} N_b \log\frac{N_b}{N}\right)}}, \tag{20}$$

where N_a , N_b indicate the number of samples in clusters C_a, C_b based on L^i, L^j, respectively. $N_{a,b}$ represents the number of samples in cluster a according to C_a and cluster b according to C_b. Thus, the mutual information among R clusterings (Λ) can be defined as the average NMI (ANMI):

$$\Gamma^{(ANMI)}\left(\Lambda, \hat{L}\right) = \frac{1}{R}\sum_{r=1}^{R} \Gamma^{(NMI)}\left(\hat{L}, L^r\right). \tag{21}$$

As a result, the optimal labeling from R clusterings can be expressed as follows:

$$L^* = \text{argmax}_{L\in\mathbb{L}} \sum_{r=1}^{R} \Gamma^{(NMI)}\left(L^r\right), \tag{22}$$

where Γ denotes a similarity measurement (e.g., NMI), evaluating mutual information between a set of R clusterings, and L^* represents the optimal combined clustering outcome that exhibits maximum average similarity to all other clusterings L^r. Importantly, L^* shares the same size with individual labeling L^r. In this thesis, Γ represents an unsupervised method known as the "supra" consensus function (Ghosh et al., 2002), which can embody the selected consensus function with the highest ANMI from the hypergraph-based consensus functions.

2.5.4 Multi-Set Consensus Clustering

Multi-set consensus clustering (Abu-Jamous et al., 2015; Filkov & Skiena, 2004; Hoshida et al., 2007; Liu et al., 2015) involves combining cluster sets obtained from a prior ensemble clustering procedure across multiple datasets. This process is aimed at extracting a new set of collective insights (clusters) from all individuals/datasets. In the context of EEG/ERP data from multiple subjects

across different conditions and numerous trials, multi-set consensus clustering is used to capture the most informative patterns shared among similar sources (multiple recordings from individuals/trials; Mahini et al., 2020). In contrast, a conventional cluster-based approach (microstates analysis) to address this challenge involves assigning template cluster maps derived from group-averaged data or canonical microstate classes to individuals (Berchio et al., 2019; Koenig et al., 2014; Michel & Koenig, 2018; Murray et al., 2008; Ruggeri et al., 2019).

Mathematically, let $L_s^r = \{C_{1,s}^r, C_{2,s}^r, \dots, C_{K,s}^r\}$ represent the clustering results for the r^{th} clustering method $r = 1,2, \dots, R$, for the i^{th} subject, $s = 1,2, \dots, S$ with K clusters. Here, $C_{k,s}^r$ indicates the k^{th} cluster, $k = 1,2, \dots, K$ from the r^{th} method for the s^{th}subject. The first level of consensus clustering for each individual dataset is performed as described in Section 2.5.5:

$$L_s^{*-opt} = \text{argmax}_{L \in \mathbb{L}_X} \sum_{r=1}^{R} \Gamma(L_s^r), \tag{23}$$

where L_s^{*-opt} represents the optimal consensus clustering results of the s^{th} subject from all possible k-partitions on dataset X. Second-level consensus clustering involves clustering the results obtained at the first level across all datasets:

$$L^{**-opt} = \text{argmax}_{L \in \mathbb{L}_{Subj}} \sum_{s=1}^{S} \Gamma\left(L_s^{*-opt}\right), \tag{24}$$

where L^{**-opt} denotes the result of consensus clustering across the subjects/datasets. Together, the optimal ensemble clustering across the subjects can be denoted as follows:

$$L^{**-opt} = \text{argmax}_{L \in \mathbb{L}_{X,Subj}} \sum_{s=1}^{S} \sum_{r=1}^{R} \Gamma(L_s^r). \tag{25}$$

This approach facilitates the integration of information from multiple homogeneous sources (e.g., single trials, subjects, clusterings), aiding in identifying interesting ERP components from a set of those sources (Mahini et al., 2020; Mahini et al., 2023b).

2.5.5 Deep Clustering Analysis for ERP Data

Deep clustering involves encouraging DNNs to learn a cluster-oriented feature representation that is suitable for cluster analysis. DNNs with clustering modules group similar features, enhancing cluster assignments (Aljalbout et al., 2018; Min et al., 2018). Two primary strategies exist. The first is a two-step process in which DNNs are trained in the initialized labels or input data, with a focus on minimizing non-clustering losses (exclusive to DNN's loss). Afterward, a clustering method such as *k*-means is used for clustering transformed data in the latent space. The second approach involves joint training of the DNN and clustering to minimize both DNN and clustering losses, often using measures such as Kullback-Leibler divergence (Kullback & Leibler, 1951) to optimize clustering.

Specifically, for spatiotemporal ERP/EEG data X from N time points and F electrodes, along with initial clustering labels $Y = \{y_1, y_2, \ldots, y_N\}$ corresponding to time points, the transformation function is denoted as $S_\emptyset: X \rightarrow Y$, in which $\emptyset$ representing learnable parameters by the network. In the first step, the DNN learns to map each time point $x_t = \{e_1, e_2, \ldots, e_F\}$ (i.e., a topography map) to a cluster label y_t, $t \in 1,2, \ldots, N$, ensuring high similarity between shared features. Hence, the input space X is assigned to K clusters $L = \{C_1, C_2, \ldots, C_K\}$, where $C_k = \{x_t | y_t = k, \forall t \in 1,2, , \ldots, N\}$ in the labeling space Y. The transformation from the input space to the latent space is facilitated by a nonlinear mapping function $f_\vartheta : X \rightarrow Z$, where ϑ represents learnable parameters and Z represents an embedded feature space in $Z \in \mathbb{R}^K$. Together, the role of DNN is to predict clustering labels using a classifier such as g_ω on top of the features $f_\theta(x_t)$ to minimize network loss:

$$\min_{\theta,\omega} \frac{1}{N} \sum_{t=1}^{N} Loss_{net}\big(g_\omega(f_\theta(x_t)), y_t\big), \tag{26}$$

where $Loss_{net}$ denotes the multinomial logistic loss (e.g., negative log-softmax function) and ω and θ are classifiers and mapping parameters, respectively.

In an unsupervised context, a DNN model, often of the end-to-end autoencoder (AE) structure, learns to represent the input space. In unsupervised models, the network is optimized using the $Loss_{net}$ known as reconstruction loss:

$$\min_{\theta_1,\omega_1} L_{rec} = \min \frac{1}{N} \sum_{t=1}^{N} \left\| x_t - g_{\omega_1}\left(f_{\theta_1}(x_t)\right) \right\|^2, \tag{27}$$

where the network consists of two groups of layers corresponding to the encoder $f_{\theta_1}(.)$ and decoder $g_{\omega_1}(.)$, often with bottleneck layer(s) in between. During cluster analysis and fine-tuning, the encoder is isolated to transfer weighted data to the DNN's latent space, which assumes a size of $N \times K$ for the spatiotemporal ERP dataset.

Given the two general strategies of DNN designs, in semi-supervised methods, an initialized clustering such as k-means configures the DNN. Afterward, transformed weighted data from the latent layer are used for fine clustering (e.g., using k-means for final clustering). In unsupervised deep clustering, an embedded clustering layer optimizes clusters within the DNN. Apart from the design of the clustering module, deep clustering methods minimize both network and clustering losses. The combined loss is given by

$$Loss = Loss_{net} + \gamma Loss_{cl}, \tag{28}$$

where $Loss_{cl}$ denotes the clustering loss and γ is a hyper-parameter used to balance the two learning components.

2.6 Time Window Determination

Conventional methods for time window selection have often relied on identifying pronounced peaks and effect sizes, a practice that hinges on statistical significance. However, a paradigm shift toward objectivity has given rise to more robust techniques. Although objective solutions such as cluster permutation (Maris & Oostenveld, 2007) and multivariate randomization analyses (Michel et al., 2009) have led to the exploration of a robust time window relying on quantifying significant effect size, these methods result in a fixed time window and are computationally expensive.

Clustering analysis, however, selects the time window of ERP using two strategies. The first group, particularly in microstate analysis studies (Bailey et al., 2019; Berchio et al., 2019; Murray et al., 2008), identifies an ERP of interest in a fixed time window for all conditions that might involve multiple cluster maps. The second group identifies the time window of each condition from the group-averaged ERP data. Different time windows relying on cluster maps can be obtained for conditions/groups in this category (Mahini et al., 2020; Mahini et al., 2022; Pascual-Marqui et al., 1995; Ruggeri et al., 2019).

From the second category, we proposed a pipeline to examine clustering results to determine the time window for each condition in two steps (Mahini et al., 2022): 1) Candidate cluster maps in the experimental interval (i.e., a roughly expected interval for the target ERP component) are sought. Candidate cluster maps are clusters with very high inner similarity (e.g., > 0.95, depending on data quality). 2) The cluster map with sufficient duration (e.g., > 50 ms, depending on the component of interest) and high overlap with the experimental interval is selected as the representer map of the time window. Nevertheless, this method determines a fixed time window for all the subjects from the group-averaged concatenated ERP data.

A consensus clustering method and a time window determination method have been introduced to combine clustering results across subjects, resulting in a more precise time window using a similar algorithm (Mahini et al., 2020). However, the methods described above underestimate the imperfection of the data, and the remaining noise, even after preprocessing and averaging, inherently helps improve the signal's SNR. A modified time window determination method has been reported based on adjusting the sensitivity parameters if needed (Mahini et al., 2023a). Moreover, time window determination for individual subjects has also been considered in the clustering of single-trial EEG epochs of individual subjects. This approach uses the spatial correlation of candidate cluster maps with the identified ERP from the group-averaged ERP data as additional criteria to deduce the risk of selecting inappropriate components (Mahini et al., 2023b).

3 OVERVIEW OF INCLUDED ARTICLES

This chapter presents an overview of the involved articles, including the study objective, developed methods, and the results and conclusions. The authors' contributions in each article have been described as well.

3.1 Article I: Optimal number of clusters by measuring similarity among topographies for spatio-temporal ERP analysis

Reza Mahini, Peng Xu, Guoliang Chen, Yansong Li, Weiyan Ding, Lei Zhang, Nauman Khalid Qureshi, Timo Hämäläinen, Asoke K. Nandi, and Fengyu Cong (2022). Optimal number of clusters by measuring similarity among topographies for spatio-temporal ERP analysis. *Brain Topography*, 35, 537–557. https://doi.org/10.1007/s10548-022-00903-2.

Objective

The conventional approaches for selecting the optimal number of clusters for spatiotemporal ERP data rely on intra-cluster tightness and inter-cluster separation to assess the whole dataset (Goutte et al., 1999; Lleti et al., 2004; Mur et al., 2016). Given the nature of ERP data, it has been shown that only a few components can be distinctly elicited (Kappenman & Luck, 2012b) due to the overlapping of components. Article I explores the optimal number of clusters through a data-driven technique to identify an ERP of interest, examining the impact of selecting an appropriate number of clusters on the quality of estimated ERPs. Moreover, we introduce a novel consensus clustering mechanism for group-averaged spatiotemporal ERP data (see Figure 10). To this end, topographical similarity was used as a criterion to qualify the estimated time windows when varying cluster numbers were examined.

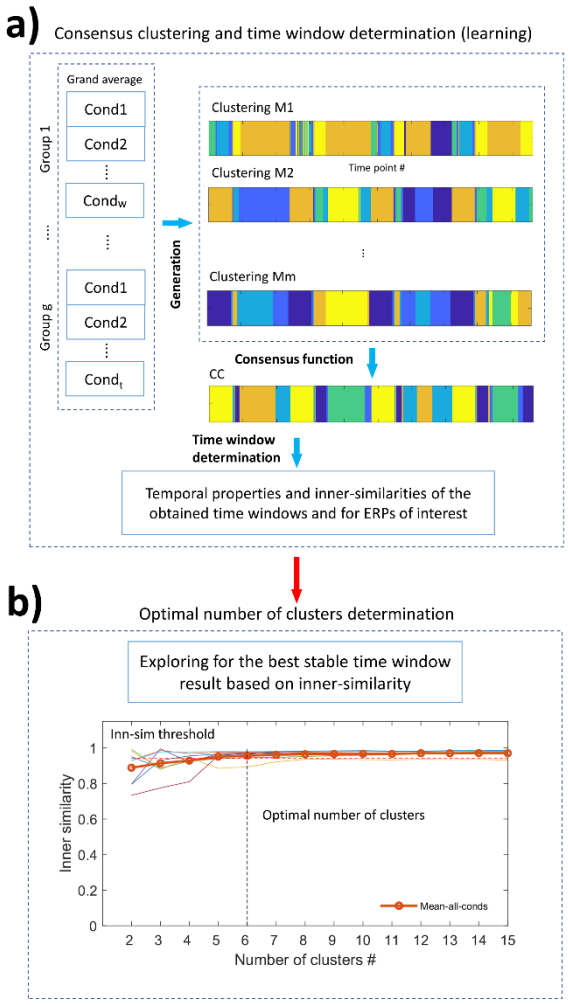


FIGURE 10 Optimal number of clusters determination using consensus clustering of spatiotemporal group-averaged ERP data. a) Generation step of consensus clustering and combining the results. b) Example of seeking an optimal number of clusters where the inner similarity of estimated time windows is high and stable.

Methods

Two simulated and real ERP datasets were used to assess the proposed pipeline. We used simulated data involving six predefined components (P1, N1, P2, N2, P3, and N4) within "Cond1" and "Cond2" conditions from a group of 20 subjects and a simulated scalp with 65 electrodes. Two state-of-the-art ERP components, N2 and P3, were tested for real ERP data that illuminated the prospective memory domain (Chen et al., 2015) within EEG records adorned with 10/20 configuration and 32 electrodes. This exploration involved 20 symptomatically remitted patients with schizophrenia (RS) and 20 healthy controls participants engaged in prospective memory and ongoing tasks, revealing insights within our proposed pipeline.

First, an M-N plot (Abu-Jamous et al., 2014) was developed to configure consensus clustering from the polarity-independent clustering methods by considering two criteria: the inner similarity (e.g., > 0.95) and the duration

(e.g., > 50 ms) of the examined time windows. A set of clustering algorithms was selected for the real ERP data, including *k*-means (Pena et al., 1999) with correlation similarity, FCM (Bezdek, 1981), self-organizing maps (SOMs; Kohonen, 1990), diffusion map spectral (DFS) clustering (Sipola et al., 2013)—consisting of *k*-means with Euclidean similarity—and *k*-medoids (KMD) clustering (Park & Jun, 2009) with correlation similarity. For the simulated data, the set of clustering methods involved *k*-means, hierarchical clustering (HC; Tibshirani & Walther, 2005) with correlation similarity, SOMs, spectral clustering (SPC), and KMD clustering.

The hypergraph partitioning CSPA was employed to explore the best cluster representation, which was carefully selected using the "supra" test on hypergraph-based consensus functions (Strehl & Ghosh, 2003). When a range of clusters (e.g., 2–15) is applied through numerous iterations (e.g., up to 100), the optimal number of clusters is selected where the average inner similarity of estimated time windows across conditions, groups, and repetitions converges on satisfactory thresholds. This convergence adheres to predefined thresholds of inner similarity (e.g., ≥ 0.95) and stability, ensuring minimal change (e.g., ≤ 0.03) in the amplitude of the inner similarities.

Results and Conclusions

The optimal number of clusters was estimated in six clusters for both simulated and real ERP data. Notably, the effect of the condition was significant for both the identified N2 and P3 components in the simulated data, which, as expected, measured a larger response in the second condition than in the first condition. For the real ERP data, our method confirms previous experimental findings (Chen et al., 2015) identifying interesting ERPs (N300 and prospective positivity) at predefined electrode sites. Furthermore, our method revealed the relationship between the number of cluster determinations and the quality of the identified ERP components. The proposed consensus clustering design from polarity-dependent clustering methods provided a novel complementary understanding regarding modeling spatiotemporal ERP data using a synergetic combination of diverse clustering strategies.

Contributions

Reza Mahini conceived and developed the algorithm, processed the data, coded the software, and wrote and revised the paper. Peng Xu expanded the main idea, contributed experimental considerations, and collected data. Guoliang Chen contributed to data collection, revising the manuscript, and providing experimental details. Yansong Li contributed to writing and reviewing the manuscript, provided experimental support, and engaged in discussions. Weiyan Ding and Lei Zhang participated in the data collection. Nauman Khalid Qureshi provided technical support and revised the manuscript. Timo Hämäläinen conceptualized and supervised the project and also wrote and revised the manuscript. Asoke K. Nandi conceptualized and supervised the whole project, provided technical support, and wrote and revised the manuscript.

Fengyu Cong conceptualized and supervised the whole project, provided technical support, and wrote and revised the manuscript.

3.2 Article II: Determination of the time window of event-related potential using multiple-set consensus clustering

Reza Mahini, Yansong Li, Weiyan Ding, Rao Fu, Tapani Ristaniemi, Asoke K. Nandi, Guoliang Chen, and Fengyu Cong (2020). Determination of the time window of event-related potential using multiple-set consensus clustering. *Frontiers in Neuroscience,* 14, 521595. https://doi.org/10.3389/fnins.2020.521595.

Objective

Cluster analysis methods are pivotal in estimating the time window of ERP components from spatiotemporal ERP data (Bailey et al., 2019; Berchio et al., 2019; Koenig et al., 2014; Ruggeri et al., 2019). However, these methods potentially overlook individual subjects' brain responses due to reliance on the averaging process in ERP and the mechanism of determining template maps (dominant clusters). This study introduces a novel approach based on multi-set consensus clustering, which involves brain response identification from a group of subjects and aims to derive a robust and dependable time window.

Methods

A multiple-set consensus clustering pipeline was designed using polarity-invariant and polarity-independent clustering methods to identify ERP(s) of interest in spatiotemporal ERP data. Figure 11 shows the design of the proposed pipeline in this study. Using multi-set consensus clustering consists of combining information from cluster sets from diverse clustering methods at the individual level and across the individual subjects at the group level. Like in Article I, two datasets, the simulated and the real ERP data, were investigated with two target ERPs, N2 and P3, in the simulated data, and N300 and prospective positivity components in the real data.

The initial generation phase involved evaluating prominent neuroimaging clustering techniques compared to the modified *k*-means (used as a benchmark), using the ARI similarity measurement (see Table 3, Eq. 3) at the individual-subject level. Then, a voting design was applied to selecting clustering methods that achieve sufficient similarity (e.g., > 0.70) from the majority of subjects (not necessarily all). We chose *k*-means, HC, AAHC, and modified *k*-means for the real data, and *k*-means, FCM, SOMs, DFS clustering, AAHC, and modified *k*-means methods to configure the consensus clustering. It is noteworthy that, to enhance stability, a consensus clustering-based stabilization method was applied (whenever required) to guarantee stable clustering results from the individual clustering methods. Subsequently, the second-level consensus clustering was executed on the individuals' clustering results across the subjects using the CSPA consensus function.

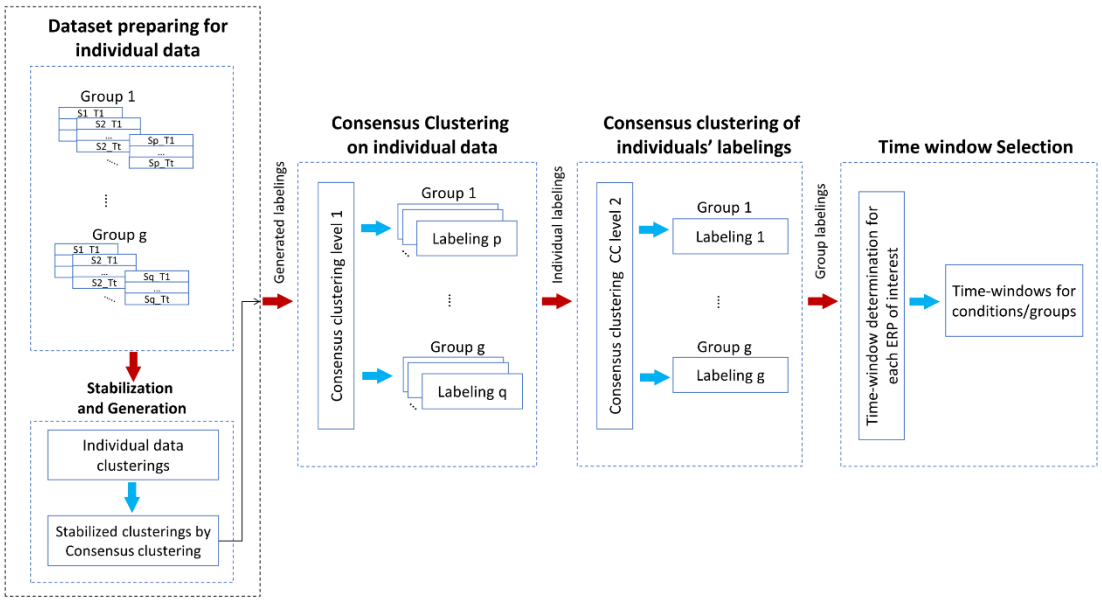


FIGURE 11 Proposed multi-set consensus clustering (i.e., consensus clustering in and across the subject levels) to determine the time window of ERPs from each group/condition.

Finally, the time window determination method examined the clustering results to identify the temporal and spatial properties of ERPs. This method tested inner similarity (e.g., ≥ 0.95) and coverage with a sufficient response duration (e.g., ≥ 50 ms, depending on the experiment) of candidate cluster maps that most likely represent the ERP component.

Results and Discussion

After obtaining the optimal number of clusters at six clusters for both ERP datasets (Mahini et al., 2022), our findings revealed the remarkable stability and efficacy of multi-set consensus clustering in identifying the ERP components. Compared to the state-of-the-art clustering methods, enhanced precision and stability were achieved in identifying the N2 and P3 components in the simulated data as well as more stable identification of the N300 and prospective positivity components in iterative tests. These findings highlight the reliability of multi-set consensus clustering in analyzing spatiotemporal ERP data. This study introduces a promising approach that emphasizes the value of combining information from individual subjects, departing from the conventional practice of processing group-averaged ERP data. It promotes the integration of a powerful ensemble of clustering methods to model complex ERP data using a range of clustering optimization techniques.

Contributions

Reza Mahini conceptualized and conducted the study, including developing the algorithm, data processing, coding, and software development, as well as writing and revising the paper. Yansong Li contributed to writing and reviewing the manuscript, offered experimental support, and engaged in discussions. Weiyan

Ding participated in data collection. Rao Fu provided the simulated data and engaged in discussions. Tapani Ristaniemi conceptualized the study, provided technical support, and wrote the manuscript. Asoke K. Nandi conceptualized the study, provided technical support, and wrote and revised the manuscript. Guoliang Chen collected data, revised it, and provided experimental details. Fengyu Cong conceptualized and supervised the whole project, provided technical support, and wrote and revised the manuscript.

3.3 Article III: Brain evoked response qualification using multi-set consensus clustering: toward single-trial EEG analysis

Reza Mahini, Guanghui Zhang, Tiina Parviainen, Rainer Düsing, Asoke K. Nandi, Fengyu Cong, and Timo Hämäläinen. (2023). Brain evoked response qualification using multi-set consensus clustering: toward single-trial EEG analysis. Submitted to *Brain Topography*, preprint available.
https://doi.org/10.21203/rs.3.rs-3586574/v1.

Objective

Single-trial EEG data includes repetitive neural information about the brain's neurological activities. Due to its crucial role, it becomes particularly significant when developing a computational data-driven model to capture brain dynamics from many similar datasets. Conventional EEG cluster analysis, notably EEG microstate analysis, has been extensively investigated in clinical (Khanna et al., 2015; Lehmann et al., 2005; Nishida et al., 2013) and cognitive (Britz & Michel, 2011; Caldara et al., 2004; Ruggeri et al., 2019) neuroscience studies. The prevailing approach for obtaining the dominant microstate classes' EEG data and the strategy of "winner-takes-all," while fitting back the canonical classes to data points, can ignore considerable information and obscure essential nuances of individual neural processes.

This study seeks to establish an effective clustering model based on multi-set consensus clustering of individual subjects' single-trial EEG epochs to identify brain-evoked responses. Additionally, it presents a novel evaluation method for evaluating the scores from single trials and individual subjects.

Methods

In this study, initiating a trial selection step, we excluded trials that exhibited no or low correlation with the identified component from grand average ERP data. We retained a minimum of 50% of the trials, adjusting the correlation threshold (e.g., >0.40) as necessary. First, trial-level consensus clustering was utilized to select trials of each subject, facilitating the exploration of correlated brain responses within individual trials. Then, consensus clustering was applied across the selected trial clustering (see Figure 12). After obtaining subject-specific clustering results, a modified time window was developed, with an additional criterion that investigated the spatial correlation between the candidate cluster

maps and the identified ERP from group-averaged data, to explore the interesting ERPs of individual subjects.

To design a standardized evaluation (Franceschiello et al., 2022; Luck et al., 2021), we established a bootstrapping of the obtained scores. We tested the scores, including mean amplitude magnitude, the properties of the time window (start, end, and duration), the inner similarity of the time window, and spatial correlation between the template map and the mean topographical map within the time window. Analytical scores from single trials and bootstrapping scores from the generated trial clusterings (from each repeat) were evaluated to calculate the standard error of the measurements ("aSME" and "bSME"). We evaluated our pipeline with the visual oddball paradigm experiment (Kappenman et al., 2021) to obtain scores of the identified P3 component.

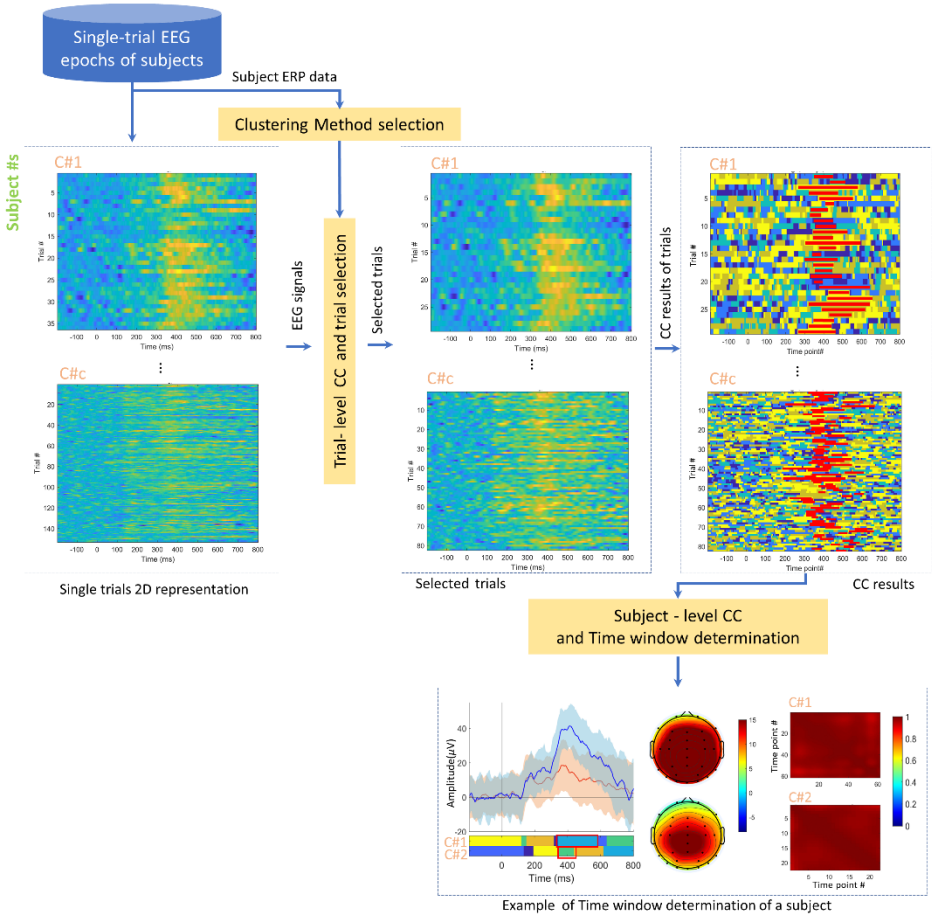


FIGURE 12 Proposed single-trial EEG epochs multi-set consensus clustering pipeline to determine the time window of ERPs at the individual-subject level.

Results and Conclusions

This study revealed the feasibility of exploring an ERP of interest from EEG single trials and elucidating individual cognitive responses. We observed an average spatial correlation of 0.65 between the P3 components identified from the individual subjects and the group P3 component. Both bootstrapping and analytical results consistently supported this finding. Furthermore, the reliability tests yielded a Cronbach's alpha of 0.70, indicating strong test reliability. These results enhance our understanding of individual-level cognitive contributions to group-level cognitive responses, thus validating the effectiveness of our approach.

Contributions

Reza Mahini conceptualized and conducted the study, including developing the algorithm, data processing, coding, and software development, as well as writing and revising the paper. Guanghui Zhang contributed to writing and reviewing the manuscript, offered experimental support, and engaged in discussions. Tiina Parviainen conceptualized the study, provided technical support, and revised the manuscript. Rainer Düsing provided technical support and revised the manuscript. Asoke K. Nandi conceptualized the study, provided technical support, and wrote and revised the manuscript. Fengyu Cong conceptualized and supervised the whole project, provided technical support, and wrote and revised the manuscript. Timo Hämäläinen supervised the whole project, provided technical support, and wrote and revised the manuscript.

3.4 Article IV: Ensemble deep clustering analysis for time window determination of event-related potentials

Reza Mahini, Fan Li, Mahdi Zarei, Asoke K. Nandi, Timo Hämäläinen, and Fengyu Cong (2023). Ensemble deep clustering analysis for time window determination of event-related potentials. *Biomedical Signal Processing and Control, 86*, 105202. https://doi.org/10.1016/j.bspc.2023.105202.

Objective

One important challenge when processing neuroimaging data (particularly EEG) is the uncertainty of data quality (after preprocessing), which causes poor results. This uncertainty can lead to unreliable identification and interpretation of brain neural activities. Cluster analysis of neuroimaging data, such as ERP data, often assumes that the data is preprocessed. However, various sources of uncertainty, especially noise in the signal, can cause cluster analysis failure, particularly due to low spatial correlation (for microstate analysis) or unreliable similarity definitions in popular clustering methods.

Methods

We developed an ensemble deep clustering pipeline from semi-supervised and unsupervised deep clustering methods (as shown in Figure 13) to identify an ERP of interest from group-averaged ERP data with additional noise (e.g., adding 20 dB to -5 dB white Gaussian noise) applied. We used the M-N plot (Mahini et al., 2022) to configure the consensus function in both initializing consensus clustering and deep clustering ensemble. Semi-supervised deep clustering with fully connected multi-level perceptron (FC_MLP), long short-term memory (LSTM), and one-dimensional convolutional neural network (1DCNN) DNN models were designed and initialized through consensus clustering. The unsupervised deep clustering methods were designed with autoencoder (AE), variational AE (VAE), and deep embedded clustering (DEC) models. The latent layer of the semi-supervised models was connected to the clustering module for clustering of the transformed weighted data. The transformed data from the encoder part of the unsupervised models (except DEC, which had an integrated clustering layer) was used for clustering.

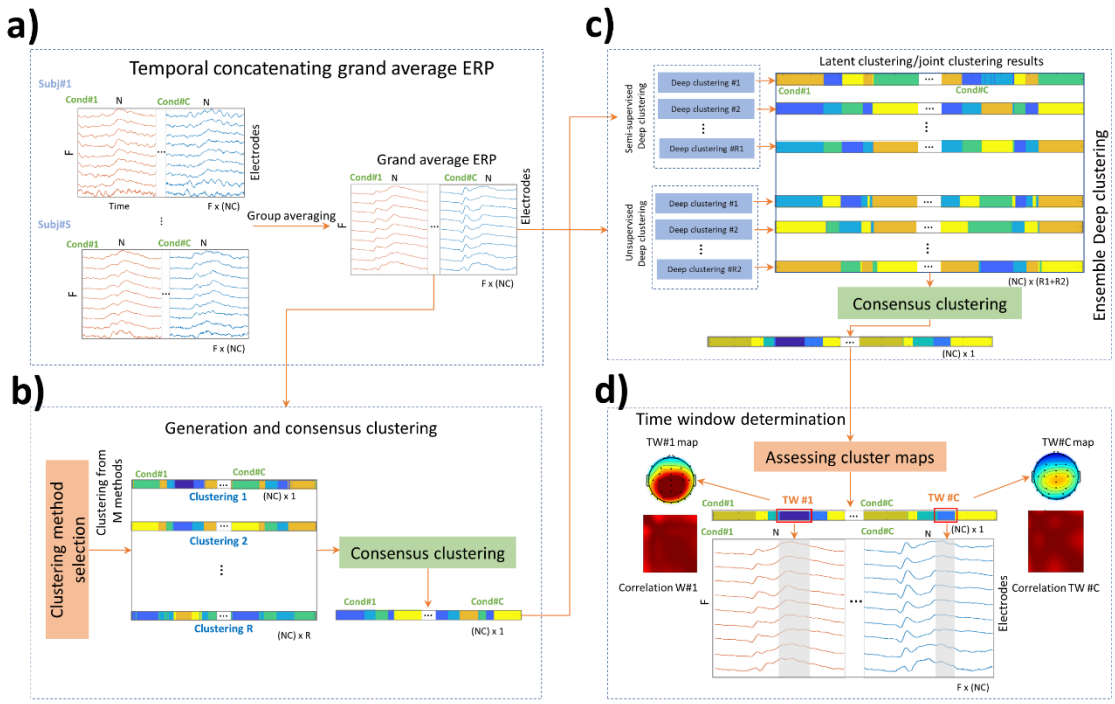


FIGURE 13 Ensemble deep clustering pipeline for determining the time window of an ERP in group mean spatiotemporal ERP data. TW=time window.

Among different strategies for ensemble clustering (Cao et al., 2020; Sagi & Rokach, 2018), we combined the clustering results of individual methods to calculate the ensemble result from non-heterogeneous elements (i.e., various DNNs strategies). Once the results from the deep clustering methods were obtained, the hypergraph-based consensus function CSPA was used to explore the most aggregated clustering results. We assessed our approach by applying the simulated data (from Article I) and the real ERP data (used in Article III). The

modified time window determination method was adapted by automatically adjusting the inner similarity threshold (e.g., 0.7 ≤ minimum inner similarity ≤ 0.95) and the duration of candidate cluster maps (e.g., 30 ms ≤ minimum number of time points ≤ 50 ms) as required.

Results and Conclusions

Our findings revealed the power of deep clustering methods in isolating the ERP components, including both semi-supervised and unsupervised designs. Compared to conventional clustering methods, the proposed ensemble method yielded more robust clustering results in terms of the spatiotemporal properties of estimated time windows considering the added noise in data. The new design also disclosed the untapped potential of deep learning methods as tools for exploring intricate, interesting patterns within neuroimaging data and offering new insights into the complexities of the human brain.

Contributions

Reza Mahini conceptualized and conducted the study, including developing the algorithm, data processing, coding, and software development, as well as writing and revising the manuscript. Fan Li provided technical support and tested the method and software. Mahdi Zarei contributed to discussions as well as writing and reviewing the manuscript. Asoke K. Nandi conceptualized the study, provided technical support, and wrote the manuscript. Timo Hämäläinen conceptualized the study, provided technical support, and wrote and revised the manuscript. Fengyu Cong conceptualized and supervised the whole project, provided technical support, and wrote and revised the manuscript.

4 CONCLUSION AND DISCUSSION

This thesis discussed cluster analysis in the context of a group-level analysis of ERP data. In this chapter, we provide an overview of our findings for each research study, address methodological limitations, and consider avenues for future research.

4.1 Summary of New Findings in Group-level ERP Analysis

Article I introduced a novel data-driven approach for determining the optimal number of clusters and proposed a consensus clustering pipeline to process spatiotemporal ERP data and identify the ERP of interest. One key challenge in determining the appropriate number of clusters is the overlapping nature of ERP components, as improper clustering can result in the combination or division of distinct components. Article I addressed estimating the optimal number of clusters by proposing a data-driven approach, investigating two ERP components. Our results highlight the relationship between the determining number of clusters and the identifying time window of ERP. This is achieved by evaluating the inner similarity of candidate cluster maps under different clustering options (e.g., 2–15 clusters).

Article II proposed developing a time window determination method and introduced a multi-set consensus clustering approach in individual subjects and group levels. This method captures mutual information by combining clustering results from individual subjects, yielding a robust estimation of ERP time windows. Article III took a step further by establishing a multi-set consensus clustering mechanism for single-trial EEG epochs to isolate the ERP response of individual subjects. A standardization mechanism was developed to evaluate signal processing on single-trial EEG data through a bootstrapping test to test the reproducibility of the scoring, i.e., obtained from designed multi-set consensus clustering for processing single-trial EEG data. This mechanism aims to explore realistic evoked responses from individual subjects. While the methodologies

presented in Articles I and II underscore the reliability of consensus clustering, it is imperative to note that even the most robust clustering method can encounter challenges in noisy data, which leads to unstable clusters.

Article IV addressed this challenge by applying an ensemble deep clustering approach based on various deep clustering strategies, focusing on clustering group-averaged data, especially in scenarios where significant noise persists after preprocessing and averaging. Our findings in Article IV highlight the effectiveness of deep clustering methods in identifying the major ERP components under varying levels of additive noise.

4.2 Limitations of Methodological Designs

The primary contribution of this thesis is the design of a reliable clustering approach for qualifying an ERP of interest, assuming a foundational understanding of existing ERPs and their spatial properties. While our studies and methods contribute significantly to addressing spatiotemporal ERP cluster analysis, certain limitations warrant consideration.

Specifically, in Article I, a consensus clustering framework was applied to group-averaged ERP data to determine the optimal number of clusters. The challenge of selecting appropriate configurations for generating consensus clusters persists. Although this study adapted configurations using the M-N plot technique to assess inner similarity and estimated time window duration for different ERP data, a straightforward solution to the consensus clustering configuration problem remains elusive.

Another limitation is that although group-averaged ERPs encapsulate dominant information from individual subjects' brain responses, this overlooks potential variations stemming from physiological differences and dataset-specific characteristics. These aspects, crucial for accurate representation, can be neglected when calculating group-averaged ERPs. Notably, group-averaged ERP signals represented the average of trials within each condition across the participant group and thus did not reflect their actual brain responses. However, Article II addressed this issue by proposing a multi-set consensus clustering (Mahini et al., 2020) approach as a more cautious approach, assuming that combining brain responses from similar sources (individual subjects) would be feasible via this method. However, both consensus clustering-based solutions (Mahini et al., 2020; Mahini et al., 2022) assume a fixed configuration for all individual subjects, omitting systematic investigation to determine the optimal number of clusters and cluster configurations for each subject. While modified multi-set consensus clustering for single-trial EEG analysis in Article III delved deeper into investigating individual subjects' brain responses (Mahini et al., 2023b), this should be further explored in future research.

In Article IV, we applied various standard deep clustering methods to determine a reliable time window for group-averaged ERP data with additional noise. However, this approach may have underestimated complex data

uncertainties, such as the data recorded from the disorder groups. While demonstrating reliability in detecting larger effects, a more comprehensive exploration is required to identify the limits and accuracy of deep learning in identifying neurophysiological activities. Lastly, across Article I–IV, a consistent ensemble learning mechanism utilizing the hypergraph-based consensus function CSPA was applied at multiple levels. Despite CSPA's reliability in terms of tolerance for missing label problem, as well as its consistency across varying cluster counts, its high computational and memory demands necessitate advanced hardware, particularly for local machines.

An additional drawback of the consensus clustering mechanism employed in these studies is its shortcomings in guaranteeing optimal clustering results when employing clustering approaches, including the standard and microstate clustering methods. This issue is known as the exploitation problem in machine learning, where consensus clustering aims to identify the optimal clustering solution from various clustering solutions by exploring the most robust aggregation from diverse clustering approaches. However, consensus functions such as CSPA inherently consider temporal dynamics more than spatial dynamics of cluster maps.

4.3 Future Directions

There are several promising avenues to explore based on the frameworks presented for group-level EEG/ERP analysis. First, the techniques developed in this thesis can be extended to process MEG data (Hansen et al., 2010; Lopes da Silva, 2013).

Specifically, while Article I and II investigated consensus clustering using standard and microstate clustering methods, one direction to enhance the proposed consensus clustering is to optimize the configuration mechanism during the generation phase. This optimization is expected to overcome individual clustering limitations and provide an adaptive configuration mechanism, such as testing the consistency of employed clustering methods (Mahini et al., 2022) and involving the advanced clustering method (i.e., incorporating variations of microstate clustering) as suggested by von Wegner et al. (2018). Furthermore, the determination of the optimal number of clusters can also be extended to individual-subject data. In addition, identifying the ERP of interest for each subject could constitute a more realistic investigation.

Cluster analysis of EEG signals (evoked and resting states) is crucial due to the quality of the data, resulting in lower sample similarities and unstable clusters. This arises from the inherent limitation of the proposed multi-set consensus function to directly account for changes in microstate dynamics and spatial configuration within the obtained cluster maps. To enhance the efficacy of the proposed multi-set consensus function in Articles I, II, and III, a post hoc procedure could be implemented to enhance the quality of cluster maps derived

from clustering outcomes. Therefore, additional refinement based on spatial correlation analysis would optimize the proposed consensus clustering.

An important way to build on Article IV would involve investigating the synergy between deep learning and clustering methods. Further inquiry is needed to analyze advanced learning strategies for DNN models and the design of ensemble deep clustering to optimize clustering quality. More specifically, divisions and ensembles can be performed on datasets, on training models, or at the clustering level. While the quality of extracted features from model training can be evaluated using methods such as PCA or t-distributed stochastic neighbor embedding (Van der Maaten & Hinton, 2008) representation, clustering quality evaluation lacks a straightforward method.

Expanding these methodologies to other domains (e.g., time-frequency, spectral analysis) and types of neuroimaging data, such as MEG and fMRI, would open exciting avenues for future research that could continue to advance the field of neuroimaging cluster analysis.

YHTEENVETO (SUMMARY IN FINNISH)

Tässä väitöskirjassa käsitellään klusterianalyysiä konsensusklusterointipohjaisessa (ERP) ryhmätason analyysissä. Väitöskirjan päätulos on luotettavan klusterointimenetelmän suunnittelu ERP:n laadulliseen arviointiin olettaen, että olemassa olevat ERP:t ja niiden spatiaaliset ominaisuudet tunnetaan perusteellisesti.

Artikkelissa I esitellään uudentyyppinen datavetoinen lähestymistapa optimaalisen klusterien lukumäärän määrittämiseen ja ehdotetaan yhteisymmärrysklusterointiputkea, jolla käsitellään spatiaalis-temporaalisia ERP-tietoja ja tunnistetaan kiinnostava ERP-komponentti. Yksi keskeinen haaste optimaalisen klusterien lukumäärän määrittämisessä on ERP-komponenttien päällekkäisyys, sillä virheellinen klusterointi voi johtaa erillisten komponenttien sulautumiseen tai jakautumiseen. Artikkelissa käsitellään optimaalisen klusterien lukumäärän arviointia ehdottamalla datavetoista lähestymistapaa ja tutkimalla kahta ERP-komponenttia. Tulokset korostavat yhteyttä klusterien lukumäärän ja ERP:n aikaikkunan tunnistamisen välillä. Tämä saavutetaan arvioimalla ehdokasklusterikarttojen sisäistä samankaltaisuutta eri klusterointivaihtoehdoilla (esim. 2–15 klusteria).

Artikkelissa II ehdotetaan aikajakson määritysmenetelmän kehittämistä ja esitellään monijoukkoiseen konsensusklusterointiin perustuva lähestymistapa yksittäisille koehenkilöille ja ryhmätasoille. Tämä menetelmä sieppaa yhteistä tietoa yhdistämällä klusterointitulokset yksittäisiltä koehenkilöiltä, tuottaen vakaan arvion ERP:n aikaluokista. Artikkeli III esittää monijoukkoista konsensusklusterointimekanismia yksittäisten koetilaisuuksien EEG-epookkeihin eristääkseen yksittäisten koehenkilöiden ERP-vasteen. Standardointimekanismi kehitettiin arvioimaan signaalinkäsittelyä yksittäisillä EEG-tietojaksoilla bootstrapping-testin kautta testaamaan niiden pisteiden toistettavuutta, jotka saadaan suunnitellusta monijoukkoisesta konsensusklusteroinnista yksittäisten EEG-tietojen käsittelyssä. Tämä mekanismi pyrkii tutkimaan realistisia aiheutettuja vastauksia yksittäisiltä koehenkilöiltä.

Artikkeleissa I ja II esitellyt menetelmät korostavat konsensusklusteroinnin luotettavuutta. On tärkeää huomata, että jopa tehokkain klusterointimenetelmä voi kohdata haasteita kohinaisessa datassa, mikä johtaa epävakaaseen klusterointiin.

Artikkeli IV käsittelee tätä haastetta soveltamalla ryhmäkeskitetyn datan syvään klusterointiin perustuvaa monijoukkoista lähestymistapaa erityisesti tilanteissa, joissa merkittävää kohinaa esiintyy esikäsittelyn ja keskiarvoistamisen jälkeen. Artikkelin tulokset korostavat syvien klusterointimenetelmien tehokkuutta merkittävien ERP-komponenttien tunnistamisessa vaihtelevilla kohinatasoilla.

Tutkimuksen myötä on löytynyt useita tutkimussuuntia tässä väitöskirjassa esitetyn EEG/ERP-analyysin viitekehyksien perusteella. Kehitetyt tekniikat voidaan laajentaa prosessoimaan MEG-tietoja (Hansen et al., 2010; Lopes da Silva, 2013). Lisäksi näiden metodologioiden laajentaminen muihin alueisiin (esim. aika-taajuus, spektrianalyysi) ja neurokuvantamistietoihin (kuten MEG ja fMRI) avaa mielenkiintoisia mahdollisuuksia tulevalle tutkimukselle neurokuvantamisen klusterianalyysin kehittämiseksi.

REFERENCES

Abu-Jamous, B., Fa, R., & Nandi, A. K. (2015). *Integrative cluster analysis in bioinformatics*. Copyright © 2015 John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118906545

Abu-Jamous, B., Fa, R., Roberts, D. J., & Nandi, A. K. (2013). Paradigm of Tunable Clustering Using Binarization of Consensus Partition Matrices (Bi-CoPaM) for Gene Discovery. *PloS One, 8*(2), Article e56432. https://doi.org/10.1371/journal.pone.0056432

Abu-Jamous, B., Fa, R., Roberts, D. J., & Nandi, A. K. (2014). M-N scatter plots technique for evaluating varying-size clusters and setting the parameters of Bi-CoPaM and Uncles methods. 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), https://doi.org/10.1109/ICASSP.2014.6854902

Abu-Jamous, B., Fa, R., Roberts, D. J., & Nandi, A. K. (2015). UNCLES: method for the identification of genes differentially consistently co-expressed in a specific subset of datasets. *BMC Bioinformatics, 16*, Article 184. https://doi.org/10.1186/s12859-015-0614-0

Acharya, A., & Ghosh, J. (2011). Cluster Ensembles. In *Wiley StatsRef: Statistics Reference Online* (pp. 1-20). https://doi.org/10.1002/9781118445112.stat08170

Aljalbout, E., Golkov, V., Siddiqui, Y., & Cremers, D. J. a. p. a. (2018). Clustering with Deep Learning: Taxonomy and New Methods. https://doi.org/arXiv:1801.07648v2

Antonova, E., Holding, M., Suen, H. C., Sumich, A., Maex, R., & Nehaniv, C. (2022). EEG microstates: Functional significance and short-term test-retest reliability. *Neuroimage: Reports, 2*(2), 100089. https://doi.org/10.1016/j.ynirp.2022.100089

Bailey, N. W., Freedman, G., Raj, K., Sullivan, C. M., Rogasch, N. C., Chung, S. W., Hoy, K. E., Chambers, R., Hassed, C., & Van Dam, N. T. J. P. o. (2019). Mindfulness meditators show altered distributions of early and late neural activity markers of attention in a response inhibition task. *14*(8), e0203096. https://doi.org/10.1371/journal.pone.0203096

Bashivan, P., Rish, I., Yeasin, M., & Codella, N. (2015). Learning representations from EEG with deep recurrent-convolutional neural networks. *arXiv preprint arXiv:1511.06448*. https://doi.org/10.48550/arXiv.1511.06448

Ben-Hur, A., Elisseeff, A., & Guyon, I. (2001). A stability based method for discovering structure in clustered data. In *Biocomputing 2002* (pp. 6-17). WORLD SCIENTIFIC. https://doi.org/10.1142/9789812799623_0002

Berchio, C., Küng, A.-L., Kumar, S., Cordera, P., Dayer, A. G., Aubry, J.-M., Michel, C. M., & Piguet, C. (2019). Eye-gaze processing in the broader bipolar phenotype revealed by electrical neuroimaging. *Psychiatry Research: Neuroimaging, 291*, 42-51. https://doi.org/10.1016/j.pscychresns.2019.07.007

Bezdek, J. C. (1981). *Pattern recognition with fuzzy objective function algorithms*. https://doi.org/10.1007/978-1-4757-0450-1

Boongoen, T., & Iam-On, N. (2018). Cluster ensembles: A survey of approaches with recent extensions and applications. *Computer Science Review, 28*, 1-25. https://doi.org/10.1016/j.cosrev.2018.01.003

Brandeis, D., & Lehmann, D. (1986). Event-Related Potentials of the Brain and Cognitive Processes: Approaches And Applications. In M. A. Jeeves & G. Baumgartner (Eds.), *Methods in Neuropsychology* (pp. 151-168). Pergamon. https://doi.org/10.1016/B978-0-08-032026-7.50014-9

Brandeis, D., & Lehmann, D. (1989). Segments of event-related potential map series reveal landscape changes with visual attention and subjective contours. *Electroencephalography and Clinical Neurophysiology, 73*(6), 507-519. https://doi.org/10.1016/0013-4694(89)90260-5

Brandeis, D., Lehmann, D., Michel, C. M., & Mingrone, W. (1995). Mapping event-related brain potential microstates to sentence endings. *Brain Topography, 8*(2), 145-159. https://doi.org/10.1007/BF01199778

Britz, J., & Michel, C. (2011). State-Dependent Visual Processing [Perspective]. *Frontiers in Psychology, 2*. https://doi.org/10.3389/fpsyg.2011.00370

Britz, J., Van De Ville, D., & Michel, C. M. (2010). BOLD correlates of EEG topography reveal rapid resting-state network dynamics. *Neuroimage, 52*(4), 1162-1170. https://doi.org/10.1016/j.neuroimage.2010.02.052

Brunet, D., Murray, M. M., & Michel, C. M. (2011). Spatiotemporal Analysis of Multichannel EEG: CARTOOL. *Computational Intelligence and Neuroscience*, Article 813870. https://doi.org/10.1155/2011/813870

Caldara, R., Deiber, M.-P., Andrey, C., Michel, C. M., Thut, G., & Hauert, C.-A. J. E. B. R. (2004). Actual and mental motor preparation and execution: a spatiotemporal ERP study [journal article]. *159*(3), 389-399. https://doi.org/10.1007/s00221-004-2101-0

Cao, Y., Geddes, T. A., Yang, J. Y. H., & Yang, P. (2020). Ensemble deep learning in bioinformatics. *Nature Machine Intelligence, 2*(9), 500-508. https://doi.org/10.1038/s42256-020-0217-y

Cecotti, H., & Graser, A. (2011). Convolutional Neural Networks for P300 Detection with Application to Brain-Computer Interfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 33*(3), 433-445. https://doi.org/10.1109/TPAMI.2010.125

Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2014). Nbclust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *Journal of Statistical Software, 61*(6), 1-36. https://doi.org/10.18637/jss.v061.i06

Chen, G., Zhang, L., Ding, W., Zhou, R., Xu, P., Lu, S., Sun, L., Jiang, Z., Li, H., Li, Y., & Cui, H. (2015). Event-related brain potential correlates of prospective memory in symptomatically remitted male patients with schizophrenia. *Frontiers in Behavioral Neuroscience, 9*, Article 262. https://doi.org/10.3389/fnbeh.2015.00262

Cong, F., Kalyakin, I., Huttunen-Scott, T., Li, H., Lyytinen, H., & Ristaniemi, T. (2010). Single-Trial Based Independent Component Analysis on Mismatch Negativity in Children. *International Journal of Neural Systems, 20*(04), 279-292. https://doi.org/10.1142/s0129065710002413

Cong, F., Ristaniemi, T., & Lyytinen, H. (2014). *Advanced Signal Processing on Brain Event-Related Potentials* [doi:10.1142/9306]. WORLD SCIENTIFIC. https://doi.org/doi:10.1142/9306

Cover, T. M., & Thomas, J. A. (1991). Entropy, relative entropy and mutual information. *Elements of information theory, 2*(1), 12-13.

Custo, A., Van De Ville, D., Wells, W. M., Tomescu, M. I., Brunet, D., & Michel, C. M. (2017, 2017-12). Electroencephalographic Resting-State Networks: Source Localization of Microstates. *Brain connectivity, 7*(10), 671-682. https://doi.org/10.1089/brain.2016.0476

Dien, J., Khoe, W., & Mangun, G. R. (2007). Evaluation of PCA and ICA of simulated ERPs: Promax vs. infomax rotations. *Human Brain Mapping, 28*(8), 742-763. https://doi.org/10.1002/hbm.20304

Dinov, M., & Leech, R. (2017). Modeling Uncertainties in EEG Microstates: Analysis of Real and Imagined Motor Movements Using Probabilistic Clustering-Driven Training of Probabilistic Neural Networks [Methods]. *11*(534). https://doi.org/10.3389/fnhum.2017.00534

Donchin, E., & Heffley, E. (1978). Multivariate analysis of event-related potential data: A tutorial review.

Dunn, J. C. (1974). Well-separated clusters and optimal fuzzy partitions. *Journal of Cybernetics, 4*(1), 95-104. <Go to ISI>://INSPEC:808480

Filkov, V., & Skiena, S. (2004). Integrating microarray data by consensus clustering. *International Journal on Artificial Intelligence Tools (Architectures, Languages, Algorithms), 13*(4), 863-880. https://doi.org/10.1142/s0218213004001867

Fowlkes, E. B., & Mallows, C. L. (1983). A Method for Comparing Two Hierarchical Clusterings. *Journal of the American Statistical Association, 78*(383), 553-569. https://doi.org/10.1080/01621459.1983.10478008

Franceschiello, B., Lefebvre, J., Murray, M. M., & Glomb, K. (2022). A Roadmap for Computational Modelling of M/EEG. *Brain Topography*. https://doi.org/10.1007/s10548-022-00889-x

Geva, A. B., & Pratt, H. (1994). Unsupervised clustering of evoked potentials by waveform. *Medical and Biological Engineering and Computing, 32*(5), 543-550. https://doi.org/10.1007/BF02515313

Ghosh, J., Strehl, A., & Merugu, S. (2002). A consensus framework for integrating distributed clusterings under limited knowledge sharing. Proc. NSF Workshop on Next Generation Data Mining,

Golalipour, K., Akbari, E., Hamidi, S. S., Lee, M., & Enayatifar, R. (2021). From clustering to clustering ensemble selection: A review. *Engineering Applications of Artificial Intelligence, 104*, 104388. https://doi.org/10.1016/j.engappai.2021.104388

Goutte, C., Toft, P., Rostrup, E., Nielsen, F. Å., & Hansen, L. K. (1999). On Clustering fMRI Time Series. *Neuroimage, 9*(3), 298-310. https://doi.org/10.1006/nimg.1998.0391

Habermann, M., Weusmann, D., Stein, M., & Koenig, T. (2018). A Student's Guide to Randomization Statistics for Multichannel Event-Related Potentials Using Ragu. *12*(355). https://doi.org/10.3389/fnins.2018.00355

Handy TC, e. (2009). Brain Signal Analysis: Advances in Neuroelectric and Neuromagnetic Methods. *MIT Press Scholarship Online*, 21-53. https://doi.org/10.7551/mitpress/9780262013086.001.0001

Hansen, P., Kringelbach, M., & Salmelin, R. (2010). *MEG: An Introduction to Methods*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195307238.001.0001

Hoshida, Y., Brunet, J.-P., Tamayo, P., Golub, T. R., & Mesirov, J. P. (2007). Subclass Mapping: Identifying Common Subtypes in Independent Disease Data Sets. *PloS One, 2*(11), Article e1195. https://doi.org/10.1371/journal.pone.0001195

Huang, D., Wang, C.-D., & Lai, J.-H. (2017). LWMC: A Locally Weighted Meta-Clustering Algorithm for Ensemble Clustering. Neural Information Processing, Cham.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification, 2*(1), 193-218. https://doi.org/10.1007/BF01908075

Huster, R. J., Messel, M. S., Thunberg, C., & Raud, L. (2020). The P300 as marker of inhibitory control – Fact or fiction? *Cortex, 132*, 334-348. https://doi.org/10.1016/j.cortex.2020.05.021

Huster, R. J., & Raud, L. (2017). A Tutorial Review on Multi-subject Decomposition of EEG. *Brain Topography*. https://doi.org/10.1007/s10548-017-0603-x

Jung, T.-P., Makeig, S., Westerfield, M., Townsend, J., Courchesne, E., & Sejnowski, T. J. (2001). Analysis and visualization of single-trial event-related potentials. *Human Brain Mapping, 14*(3), 166-185. https://doi.org/10.1002/hbm.1050

Kappenman, E. S., Farrens, J. L., Zhang, W., Stewart, A. X., & Luck, S. J. (2021). ERP CORE: An open resource for human event-related potential research. *Neuroimage, 225*, 117465. https://doi.org/10.1016/j.neuroimage.2020.117465

Kappenman, E. S., & Luck, S. J. (2012a). ERP components: The ups and downs of brainwave recordings. 3-30. https://doi.org/10.1093/oxfordhb/9780195374148.013.0014

Kappenman, E. S., & Luck, S. J. (2012b). Manipulation of orthogonal neural systems together in electrophysiological recordings: the MONSTER approach to simultaneous assessment of multiple neurocognitive dimensions. *Schizophrenia Bulletin, 38*(1), 92-102. https://doi.org/10.1093/schbul/sbr147

Karypis, G., & Kumar, V. (1998). Multilevelk-way Partitioning Scheme for Irregular Graphs. *Journal of Parallel and Distributed Computing, 48*(1), 96-129. https://doi.org/10.1006/jpdc.1997.1404

Karypis, G., & Kumar, V. (1999). Multilevel k-way hypergraph partitioning. Proceedings of the 36th annual ACM/IEEE design automation conference,

Khanna, A., Pascual-Leone, A., Michel, C. M., & Farzan, F. (2015, 2015/02/01/). Microstates in resting-state EEG: Current status and future directions. *Neuroscience and Biobehavioral Reviews, 49*, 105-113. https://doi.org/10.1016/j.neubiorev.2014.12.010

Kiesel, A., Miller, J., Jolicœur, P., & Brisson, B. (2008). Measurement of ERP latency differences: A comparison of single-participant and jackknife-based scoring methods. *45*(2), 250-274. https://doi.org/10.1111/j.1469-8986.2007.00618.x

Koenig, T., Kottlow, M., Stein, M., Melie- Garc´ıa, L. (2011). Ragu: a free tool for the analysis of EEG and MEG event-related scalp field data using global randomization statistics. Neuroscience. *2011*, 1-14. https://doi.org/10.1155/2011/938925

Koenig, T., & Lehmann, D. (1996). Microstates in Language-Related Brain Potential Maps Show Noun–Verb Differences. *Brain and Language, 53*(2), 169-182. https://doi.org/10.1006/brln.1996.0043

Koenig, T., & Melie-García, L. (2010). A Method to Determine the Presence of Averaged Event-Related Fields Using Randomization Tests. *Brain Topography, 23*(3), 233-242. https://doi.org/10.1007/s10548-010-0142-1

Koenig, T., & Melie-García, L. J. B. T. (2010). A Method to Determine the Presence of Averaged Event-Related Fields Using Randomization Tests [journal article]. *23*(3), 233-242. https://doi.org/10.1007/s10548-010-0142-1

Koenig, T., Stein, M., Grieder, M., & Kottlow, M. (2014). A Tutorial on Data-Driven Methods for Statistically Assessing ERP Topographies. *Brain Topography, 27*(1), 72-83. https://doi.org/10.1007/s10548-013-0310-1

Kohonen, T. (1990). THE SELF-ORGANIZING MAP. *Proceedings of the Ieee, 78*(9), 1464-1480. https://doi.org/10.1109/5.58325

Kuhn, H. W. J. N. r. l. q. (1955). The Hungarian method for the assignment problem. *2*(1-2), 83-97.

Kullback, S., & Leibler, R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics, 22*(1), 79-86. http://www.jstor.org/stable/2236703

Laganaro, M. (2014). ERP topographic analyses from concept to articulation in word production studies. *Frontiers in Psychology, 5*, Article 493. https://doi.org/10.3389/fpsyg.2014.00493

Lehmann, D. (1989). Microstates of the brain in EEG and ERP mapping studies. In *Brain Dynamics* (pp. 72-83). Springer. https://doi.org/10.1007/978-3-642-74557-7_6

Lehmann, D. (1990). Brain Electric Microstates and Cognition: The Atoms of Thought. In E. R. John, T. Harmony, L. S. Prichep, M. Valdés-Sosa, & P. A. Valdés-Sosa (Eds.), *Machinery of the Mind: Data, Theory, and Speculations About Higher Brain Function* (pp. 209-224). Birkhäuser Boston. https://doi.org/10.1007/978-1-4757-1083-0_10

Lehmann, D., Faber, P. L., Galderisi, S., Herrmann, W. M., Kinoshita, T., Koukkou, M., Mucci, A., Pascual-Marqui, R. D., Saito, N., Wackermann, J., Winterer, G., & Koenig, T. (2005). EEG microstate duration and syntax in acute, medication-naïve, first-episode schizophrenia: a multi-center study. *Psychiatry Research: Neuroimaging, 138*(2), 141-156. https://doi.org/10.1016/j.pscychresns.2004.05.007

Lehmann, D., Michel, C. M., Pal, I., & Pascual-marqui, R. D. (1994). Event-Related Potential Maps Depend on Prestimulus Brain Electric Microstate Map. *International Journal of Neuroscience, 74*(1-4), 239-248. https://doi.org/10.3109/00207459408987242

Lehmann, D., Ozaki, H., & Pal, I. (1987). EEG alpha map series: brain micro-states by space-oriented adaptive segmentation. *67*(3), 271-288. https://doi.org/10.1016/0013-4694(87)90025-3

Lehmann, D., & Skrandies, W. (1980). Reference-free identification of components of checkerboard-evoked multichannel potential fields. *Electroencephalography and Clinical Neurophysiology, 48*(6), 609-621. https://doi.org/10.1016/0013-4694(80)90419-8

Lehmann, D., & Skrandies, W. (1984). Spatial analysis of evoked potentials in man—a review. *Progress in Neurobiology, 23*(3), 227-250. https://doi.org/10.1016/0301-0082(84)90003-0

Lehmann, D., Strik, W. K., Henggeler, B., Koenig, T., & Koukkou, M. (1998). Brain electric microstates and momentary conscious mind states as building blocks of spontaneous thinking: I. Visual imagery and abstract thoughts. *International Journal of Psychophysiology, 29*(1), 1-11. https://doi.org/10.1016/S0167-8760(97)00098-6

Liu, C., Abu-Jamous, B., Brattico, E., & Nandi, A. (2015). Clustering Consistency in Neuroimaging Data Analysis [Proceedings Paper]. *2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, 1118-1122. https://doi.org/10.1109/FSKD.2015.7382099

Liu, C., Abu-Jamous, B., Brattico, E., & Nandi, A. K. (2017). Towards Tunable Consensus Clustering for Studying Functional Brain Connectivity During Affective Processing. *International Journal of Neural Systems, 27*(02), 1650042. https://doi.org/10.1142/S0129065716500428

Liu, C., Brattico, E., Abu-jamous, B., Pereira, C. S., Jacobsen, T., & Nandi, A. K. (2017). Effect of Explicit Evaluation on Neural Connectivity Related to Listening to Unfamiliar Music [Original Research]. *Frontiers in Human Neuroscience, 11*(611). https://doi.org/10.3389/fnhum.2017.00611

Lleti, R., Ortiz, M. C., Sarabia, L. A., & Sanchez, M. S. (2004). Selecting variables for k-means cluster analysis by using a genetic algorithm that optimises

the silhouettes. *Analytica Chimica Acta, 515*(1), 87-100. https://doi.org/10.1016/j.aca.2003.12.020
Lopes da Silva, F. (2013). EEG and MEG: Relevance to Neuroscience. *Neuron, 80*(5), 1112-1128. https://doi.org/10.1016/j.neuron.2013.10.017
Luck, S. J. (2014). *An introduction to the event-related potential technique* (Second edition ed.). MIT press. (MIT press)
Luck, S. J., & Gaspelin, N. (2017a). How to get statistically significant effects in any ERP experiment (and why you shouldn't). *54*(1), 146-157. https://doi.org/10.1111/psyp.12639
Luck, S. J., & Gaspelin, N. (2017b). How to get statistically significant effects in any ERP experiment (and why you shouldn't). *Psychophysiology, 54*(1), 146-157. https://doi.org/10.1111/psyp.12639
Luck, S. J., Stewart, A. X., Simmons, A. M., & Rhemtulla, M. (2021). Standardized measurement error: A universal metric of data quality for averaged event-related potentials. *Psychophysiology, 58*(6), e13793. https://doi.org/10.1111/psyp.13793
Mahini, R., Li, F., Zarei, M., Nandi, A. K., Hämäläinen, T., & Cong, F. (2023a). Ensemble deep clustering analysis for time window determination of event-related potentials. *Biomedical Signal Processing and Control, 86*, 105202. https://doi.org/10.1016/j.bspc.2023.105202
Mahini, R., Li, Y., Ding, W., Fu, R., Ristaniemi, T., Nandi, A. K., Chen, G., & Cong, F. (2020). Determination of the Time Window of Event-Related Potential Using Multiple-Set Consensus Clustering [Methods]. *Frontiers in Neuroscience, 14*(1047). https://doi.org/10.3389/fnins.2020.521595
Mahini, R., Xu, P., Chen, G., Li, Y., Ding, W., Zhang, L., Qureshi, N. K., Hämäläinen, T., Nandi, A. K., & Cong, F. (2022). Optimal Number of Clusters by Measuring Similarity Among Topographies for Spatio-Temporal ERP Analysis. *Brain Topography*. https://doi.org/10.1007/s10548-022-00903-2
Mahini, R., Zhang, G., Parviainen, T., Düsing, R., Nandi, A. K., Cong, F., & Hämäläinen, T. (2023b). Brain Evoked Response Qualification Using Multi-set Consensus Clustering: Toward Single-trial EEG Analysis. https://doi.org/10.21203/rs.3.rs-3586574/v1
Mahini, R., Zhou, T., Li, P., Nandi, A. K., Li, H., Li, H., & Cong, F. (2017). Cluster Aggregation for Analyzing Event-Related Potentials. In F. Cong, A. Leung, & Q. Wei (Eds.), *Advances in Neural Networks - ISNN 2017: 14th International Symposium, ISNN 2017, Sapporo, Hakodate, and Muroran, Hokkaido, Japan, June 21–26, 2017, Proceedings, Part II* (pp. 507-515). Springer International Publishing. https://doi.org/10.1007/978-3-319-59081-3_59
Makeig, S., Bell, A., Jung, T.-P., & Sejnowski, T. J. (1995). Independent component analysis of electroencephalographic data. *Advances in Neural Information Processing Systems, 8*.
Makeig, S., Jung, T.-P., Bell, A. J., Ghahremani, D., & Sejnowski, T. J. (1997). Blind separation of auditory event-related brain responses into

independent components. *Proceedings of the National Academy of Sciences, 94*(20), 10979-10984. https://doi.org/doi:10.1073/pnas.94.20.10979
Makeig, S., Westerfield, M., Jung, T. P., Enghoff, S., Townsend, J., Courchesne, E., & Sejnowski, T. J. (2002). Dynamic brain sources of visual evoked responses [Article]. *Science, 295*(5555), 690-694. https://doi.org/10.1126/science.1066168
Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods, 164*(1), 177-190. https://doi.org/10.1016/j.jneumeth.2007.03.024
Meila, M. (2007). Comparing clusterings - an information based distance. *Journal of Multivariate Analysis, 98*(5), 873-895. https://doi.org/10.1016/j.jmva.2006.11.013
Michel, C. M., & Koenig, T. (2018). EEG microstates as a tool for studying the temporal dynamics of whole-brain neuronal networks: A review. *Neuroimage, 180*, 577-593. https://doi.org/10.1016/j.neuroimage.2017.11.062
Michel, C. M., Koenig, T., & Brandeis, D. (2009). Electrical neuroimaging in the time domain. In C. M. Michel, D. Brandeis, J. Wackermann, L. R. R. Gianotti, & T. Koenig (Eds.), *Electrical Neuroimaging* (pp. 111-144). Cambridge University Press. https://doi.org/DOI: 10.1017/CBO9780511596889.007
Michel, C. M., Koenig, T., Brandeis, D., Gianotti, L. R., & Wackermann, J. (2009). *Electrical neuroimaging*. Cambridge University Press.
Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set [journal article]. *Psychometrika, 50*(2), 159-179. https://doi.org/10.1007/bf02294245
Milz, P., Faber, P. L., Lehmann, D., Koenig, T., Kochi, K., & Pascual-Marqui, R. D. (2016). The functional significance of EEG microstates—Associations with modalities of thinking. *Neuroimage, 125*, 643-656. https://doi.org/10.1016/j.neuroimage.2015.08.023
Min, E., Guo, X., Liu, Q., Zhang, G., Cui, J., & Long, J. J. I. A. (2018). A survey of clustering with deep learning: From the perspective of network architecture. *6*, 39501-39514. https://doi.org/10.1109/ACCESS.2018.2855437
Mishra, A. (2021). *Multiscale Microstates Uniform spatiotemporal analysis across spatial scales* sn: Sl].
Monti, S., Tamayo, P., Mesirov, J., & Golub, T. (2003). Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning, 52*(1-2), 91-118. https://doi.org/10.1023/a:1023949509487
Mu, Y., & Han, S. (2010). Neural oscillations involved in self-referential processing. *Neuroimage, 53*(2), 757-768. https://doi.org/10.1016/j.neuroimage.2010.07.008
Mur, A., Dormido, R., Duro, N., Dormido-Canto, S., & Vega, J. (2016). Determination of the optimal number of clusters using a spectral

clustering optimization [Article]. *Expert Systems with Applications, 65*, 304-314. https://doi.org/10.1016/j.eswa.2016.08.059
Murray, M. M., Brunet, D., & Michel, C. M. (2008). Topographic ERP analyses: A step-by-step tutorial review. *Brain Topography, 20*(4), 249-264. https://doi.org/10.1007/s10548-008-0054-5
Nguyen, N., & Caruana, R. (2007). Consensus Clusterings. Seventh IEEE International Conference on Data Mining (ICDM 2007), https://doi.org/10.1109/ICDM.2007.73
Nishida, K., Morishima, Y., Yoshimura, M., Isotani, T., Irisawa, S., Jann, K., Dierks, T., Strik, W., Kinoshita, T., & Koenig, T. (2013). EEG microstates associated with salience and frontoparietal networks in frontotemporal dementia, schizophrenia and Alzheimer's disease. *Clinical Neurophysiology, 124*(6), 1106-1114. https://doi.org/10.1016/j.clinph.2013.01.005
Onan, A., Korukoğlu, S., & Bulut, H. (2017). A hybrid ensemble pruning approach based on consensus clustering and multi-objective evolutionary algorithm for sentiment classification. *Information Processing & Management, 53*(4), 814-833. https://doi.org/10.1016/j.ipm.2017.02.008
Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2011). FieldTrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data. *Computational Intelligence and Neuroscience, 2011*, 9, Article 156869. https://doi.org/10.1155/2011/156869
Park, H.-S., & Jun, C.-H. (2009). A simple and fast algorithm for K-medoids clustering. *Expert Systems with Applications, 36*(2, Part 2), 3336-3341. https://doi.org/10.1016/j.eswa.2008.01.039
Pascual-Marqui, R. D. (2002). Standardized low-resolution brain electromagnetic tomography (sLORETA): Technical details [Article; Proceedings Paper]. *Methods and Findings in Experimental and Clinical Pharmacology, 24*, 5-12. <Go to ISI>://WOS:000180426300003
Pascual-Marqui, R. D., Michel, C. M., & Lehmann, D. J. I. T. o. B. E. (1995). Segmentation of brain electrical activity into microstates: model estimation and validation. *42*(7), 658-665. https://doi.org/10.1109/10.391164
Pena, J. M., Lozano, J. A., & Larranaga, P. J. P. r. l. (1999). An empirical comparison of four initialization methods for the k-means algorithm. *20*(10), 1027-1040. https://doi.org/10.1016/S0167-8655(99)00069-0
Peterson, S. M., Rao, R. P., & Brunton, B. W. (2022). Learning neural decoders without labels using multiple data streams. *Journal of Neural Engineering, 19*(4), 046032. https://doi.org/10.1088/1741-2552/ac857c
Poulsen, A. T., Pedroni, A., Langer, N., & Hansen, L. K. J. b. (2018). Microstate EEGlab toolbox: An introductory guide. 289850. https://doi.org/https://doi.org/10.1101/289850
Pourtois, G., Delplanque, S., Michel, C., & Vuilleumier, P. (2008). Beyond Conventional Event-related Brain Potential (ERP): Exploring the Time-course of Visual Emotion Processing Using Topographic and Principal

Component Analyses [journal article]. *Brain Topography, 20*(4), 265-277. https://doi.org/10.1007/s10548-008-0053-6

Qi, Y., Luo, F., Zhang, W., Wang, Y., Chang, J., Woodward, D. J., Chen, A. C., & Han, J. (2003). Sliding-window technique for the analysis of cerebral evoked potentials. *Beijing Da Xue Xue Bao Yi Xue Ban, 35*(3), 231-235.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association, 66*(336), 846-850. https://doi.org/10.1080/01621459.1971.10482356

Ren, Y., Pu, J., Yang, Z., Xu, J., Li, G., Pu, X., Yu, P. S., & He, L. (2022). Deep clustering: A comprehensive survey. *arXiv preprint arXiv:2210.04142*. https://doi.org/10.48550/arXiv.2210.04142

Rissling, A. J., Miyakoshi, M., Sugar, C. A., Braff, D. L., Makeig, S., & Light, G. A. (2014). Cortical substrates and functional correlates of auditory deviance processing deficits in schizophrenia. *NeuroImage: Clinical, 6*, 424-437. https://doi.org/10.1016/j.nicl.2014.09.006

Rotshtein, P., Richardson, M. P., Winston, J. S., Kiebel, S. J., Vuilleumier, P., Eimer, M., Driver, J., & Dolan, R. J. (2010). Amygdala damage affects event-related potentials for fearful faces at specific time windows. *Human Brain Mapping, 31*(7), 1089-1105. https://doi.org/10.1002/hbm.20921

Rousseeuw, P. J. (1987). SILHOUETTES - A GRAPHICAL AID TO THE INTERPRETATION AND VALIDATION OF CLUSTER-ANALYSIS. *Journal of Computational and Applied Mathematics, 20*, 53-65. https://doi.org/10.1016/0377-0427(87)90125-7

Ruggeri, P., Meziane, H. B., Koenig, T., & Brandner, C. (2019). A fine-grained time course investigation of brain dynamics during conflict monitoring. *Scientific Reports, 9*, Article 3667. https://doi.org/10.1038/s41598-019-40277-3

Ryali, S., Chen, T., Padmanabhan, A., Cai, W., & Menon, V. (2015). Development and validation of consensus clustering-based framework for brain segmentation using resting fMRI. *Journal of Neuroscience Methods, 240*, 128-140. https://doi.org/10.1016/j.jneumeth.2014.11.014

Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *WIREs Data Mining and Knowledge Discovery, 8*(4), e1249. https://doi.org/10.1002/widm.1249

Sassenhagen, J., & Draschkow, D. (2019). Cluster-based permutation tests of MEG/EEG data do not establish significance of effect latency or location. *Psychophysiology, 56*(6), e13335. https://doi.org/10.1111/psyp.13335

Sawaki, R., Geng, J. J., & Luck, S. J. J. J. o. N. (2012). A common neural mechanism for preventing and terminating the allocation of attention. *32*(31), 10725-10736. https://doi.org/10.1523/JNEUROSCI.1864-12.2012

Shaw, S. B., Dhindsa, K., Reilly, J. P., & Becker, S. (2019). Capturing the Forest but Missing the Trees: Microstates Inadequate for Characterizing Shorter-Scale EEG Dynamics. *Neural Computation, 31*(11), 2177-2211. https://doi.org/10.1162/neco_a_01229

Sikka, A., Jamalabadi, H., Krylova, M., Alizadeh, S., van der Meer, J. N., Danyeli, L., Deliano, M., Vicheva, P., Hahn, T., Koenig, T., Bathula, D. R., & Walter, M. (2020). Investigating the temporal dynamics of electroencephalogram (EEG) microstates using recurrent neural networks. *Human Brain Mapping, n/a*(n/a). https://doi.org/10.1002/hbm.24949

Sipola, T., Cong, F., Ristaniemi, T., Alluri, V., Toiviainen, P., Brattico, E., & Nandi, A. K. (2013). Diffusion map for clustering fMRI spatial maps extracted by independent component analysis. 2013 IEEE International Workshop on Machine Learning for Signal Processing (MLSP), https://doi.org/10.1109/MLSP.2013.6661923

Song, Y., Zhang, Z., Hu, T., Gong, X., & Nandi, A. K. (2019). Identify of Spatial Similarity of Electroencephalography (EEG) during Working-Memory Maintenance. 2019 27th European Signal Processing Conference (EUSIPCO), https://doi.org/10.23919/EUSIPCO.2019.8902595

Strehl, A., & Ghosh, J. (2003). Cluster ensembles- a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research, 3*(3), 583-617. https://doi.org/10.1162/153244303321897735

Tibshirani, R., & Walther, G. (2005). Cluster Validation by Prediction Strength. *Journal of Computational and Graphical Statistics, 14*(3), 511-528. https://doi.org/10.1198/106186005X59243

Topchy, A., Jain, A. K., & Punch, W. (2005). Clustering ensembles: models of consensus and weak partitions. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27*(12), 1866-1881. https://doi.org/10.1109/TPAMI.2005.237

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research, 9*(11).

Van Overwalle, F., Van den Eede, S., Baetens, K., & Vandekerckhove, M. (2009). Trait inferences in goal-directed behavior: ERP timing and localization under spontaneous and intentional processing. *Social Cognitive and Affective Neuroscience, 4*(2), 177-190. https://doi.org/10.1093/scan/nsp003

Vega-Pons, S., & Ruiz-Shulcloper, J. (2011). A SURVEY OF CLUSTERING ENSEMBLE ALGORITHMS. *International Journal of Pattern Recognition and Artificial Intelligence, 25*(3), 337-372. https://doi.org/10.1142/s0218001411008683

Ville, D. V. D., Britz, J., & Michel, C. M. (2010). EEG microstate sequences in healthy humans at rest reveal scale-free dynamics. *Proceedings of the National Academy of Sciences, 107*(42), 18179-18184. https://doi.org/doi:10.1073/pnas.1007841107

Vinh, N. X., Epps, J., & Bailey, J. J. T. J. o. M. L. R. (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *11*, 2837-2854.

von Wegner, F., Knaut, P., & Laufs, H. (2018). EEG Microstate Sequences From Different Clustering Algorithms Are Information-Theoretically Invariant [Methods]. *12*(70). https://doi.org/10.3389/fncom.2018.00070

Wills, A. J., Lavric, A., Hemmings, Y., & Surrey, E. (2014). Attention, predictive learning, and the inverse base-rate effect: Evidence from event-related potentials. *Neuroimage, 87*, 61-71. https://doi.org/10.1016/j.neuroimage.2013.10.060

Wunderlin, M., Koenig, T., Zeller, C., Nissen, C., & Züst, M. A. (2022). Automatized online prediction of slow-wave peaks during non-rapid eye movement sleep in young and old individuals: Why we should not always rely on amplitude thresholds. *Journal of Sleep Research*. https://doi.org/10.1111/jsr.13584

Zappasodi, F., Perrucci, M. G., Saggino, A., Croce, P., Mercuri, P., Romanelli, R., Colom, R., & Ebisch, S. J. (2019). EEG microstates distinguish between cognitive components of fluid reasoning. *Neuroimage, 189*, 560-573. https://doi.org/10.1016/j.neuroimage.2019.01.067

Zhang, G. (2021). Methods to extract multi-dimensional features of event-related brain activities from EEG data. *JYU dissertations*.

Zhang, P., Wang, X., Zhang, W., & Chen, J. (2019). Learning Spatial–Spectral–Temporal EEG Features With Recurrent 3D Convolutional Neural Networks for Cross-Task Mental Workload Assessment. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 27*(1), 31-42. https://doi.org/10.1109/TNSRE.2018.2884641

ORIGINAL PAPERS

I

OPTIMAL NUMBER OF CLUSTERS BY MEASURING SIMILARITY AMONG TOPOGRAPHIES FOR SPATIO-TEMPORAL ERP ANALYSIS

by

Reza Mahini, Peng Xu, Guoliang Chen, Yansong Li, Weiyan Ding, Lei Zhang, Nauman Khalid Qureshi, Timo Hämäläinen, Asoke K. Nandi, and Fengyu Cong
2022

Brain Topography, 35, 537–557

https://doi.org/10.1007/s10548-022-00903-2

Brain Topography
https://doi.org/10.1007/s10548-022-00903-2

ORIGINAL PAPER

Optimal Number of Clusters by Measuring Similarity Among Topographies for Spatio-Temporal ERP Analysis

Reza Mahini[1,2] · Peng Xu[3] · Guoliang Chen[3] · Yansong Li[4,5] · Weiyan Ding[3] · Lei Zhang[3] · Nauman Khalid Qureshi[9] · Timo Hämäläinen[2] · Asoke K. Nandi[6] · Fengyu Cong[1,2,7,8]

Received: 23 July 2019 / Accepted: 11 May 2022

Abstract

Averaging amplitudes over consecutive time samples (i.e., time window) is widely used to calculate the peak amplitude of event-related potentials (ERPs). Cluster analysis of the spatio-temporal ERP data is a promising tool to determine the time window of an ERP of interest. However, determining an appropriate number of clusters to optimally represent ERPs is still challenging. Here, we develop a new method to estimate the optimal number of clusters utilizing consensus clustering. Various polarity dependent clustering methods, namely, *k*-means, hierarchical clustering, fuzzy *c*-means, self-organizing map, spectral clustering, and Gaussian mixture model, are used to configure consensus clustering after assessing them individually. When a range of clusters is applied many times, the optimal number of clusters should correspond to the expectation, which is the average of the obtained mean inner-similarities of estimated time windows across all conditions and groups converge in the satisfactory thresholds. In order to assess our method, the proposed method has been applied to simulated data and prospective memory experiment ERP data aimed to qualify N2 and P3, and N300 and prospective positivity components, respectively. The results of determining the optimal number of clusters meet at six cluster maps for both ERP data. In addition, our results revealed that the proposed method could be reliably applied to ERP data to determine the appropriate time window for the ERP of interest when the measurement interval is not accurately defined.

Keywords Event-related potentials · Optimal number of clusters · Topographical analysis · Time window · Microstates · Consensus clustering

Handling Editor: Thomas Koenig.

Reza Mahini and Peng Xu have equal contribution.

✉ Asoke K. Nandi
asoke.nandi@brunel.ac.uk

✉ Fengyu Cong
cong@dlut.edu.cn

[1] School of Biomedical Engineering, Faculty of Electronic and Electrical Engineering, Dalian University of Technology, Dalian, China

[2] Faculty of Information Technology, University of Jyvaskyla, Jyvaskyla, Finland

[3] Department of Psychiatry, Chinese PLA 967Th Hospital, LieNing Street Courtyard, Dalian, China

[4] Reward, Competition and Social Neuroscience Lab, Department of Psychology, School of Social and Behavioral Sciences, Nanjing University, Nanjing, China

[5] Institute for Brain Sciences, Nanjing University, Nanjing, China

[6] Department of Electronic and Electrical Engineering, Brunel University London, Uxbridge, UK

[7] School of Artificial Intelligence, Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian, China

[8] Key Laboratory of Integrated Circuit and Biomedical Electronic System, Liaoning Province, Dalian University of Technology, Dalian, China

[9] Department of Architecture, ETH Zurich, Zurich, Switzerland

Introduction

Event-related potentials (ERPs) have been considered as a fundamental neuroimaging technique for cognitive neuroscience. Measuring the mean peak amplitude of an ERP in a specific temporal interval so-called time window, undertakes a significant role in the statistical power analysis (Luck 2014). The underlying assumption of this measurement is that the brain response in the time window is associated with the same brain activity as the individuals. This response can be considered as a quasi-stable synchronized network activation (topographical maps) at this certain time (Lehmann 1990). Inspiring by exploring stable brain response, clustering of spatio-temporal electroencephalogram (EEG)/ERP data is employed as a promising tool to identify the time window of the ERP of interest (Brunet et al. 2011; Koenig et al. 2014; Murray et al. 2008). Therefore, the clustering quality for qualifying ERP components has become more and more critical. The vital issue in clustering is the trade-off between the number of clusters and segmentation quality. Indeed, the number of clusters impacts data compression level following the qualifying ERP components. If the number of clusters is low, the dataset will be highly compressed (based on the explained variance), whereas increasing the number of clusters decreases data compression (Murray et al. 2008).

Numerous traditional methods were used to determine the optimal number of clusters in the data analysis literature. Various policies such as similarity within the time points (Kaufman and Rousseeuw 2009; Rousseeuw 1987), distance within each cluster and between the clusters (Dunn 1974), explained variance measurement (Goutte et al. 1999; Lleti et al. 2004) were widely used as the classic methods for finding the optimal number of clusters. Furthermore, hybrid methods using thirty different indices (Milligan and Cooper 1985), combining Silhouette validity index and local scaling (Mur et al. 2016) were considered. Numerous of those methods were addressed in the *R* software package *NbClust* (Charrad et al. 2014; Kassambara 2017), which are available for the researchers. More approaches based on information theory (Jonnalagadda and Srinivasan 2009; Pelleg and Moore 2000; Sugar and James 2003) were proposed to assess clustering quality. For example, *x*-means (Pelleg and Moore 2000) and Gap statistics (Charrad et al. 2014) found more attention in the recent decade by employing evaluation within-group dispersion. Together, the studies mentioned above focused on analyzing the quality of clustering by evaluating the tightness and distance between all the clusters.

For EEG brain imaging clustering, cross-validation (Pascual-Marqui et al. 1995) and Krzanowski–Lai Index (Murray et al. 2008) were used as the popular methods for determining the optimal number of clusters. Some other studies (Bréchet et al. 2019; Custo et al. 2017) considered a meta-criterion from multiple indices for a better determination of the appropriate number of clusters. Yet, the predefined four canonical template maps have been widely observed in the clustering of resting-state EEG data (Michel and Koenig 2018). In this approach, the template maps are assigned to the microstate maps called "Back-fitting" (Bréchet et al. 2019) based on spatial correlation. Then, the temporal smoothing needs to be performed due to neglecting the temporal order.

To the best of our knowledge, two groups of partitioning methods exist for EEG/ERP data depending on either considering or ignoring polarity in data. In the first and prevalent group, the calculated global field power (GFP) from the data (i.e., GFP peaks or the entire GFPs) is used for clustering EEG (especially resting-state EEG) or ERP microstates. Two of the most popular clustering methods, modified *k*-means (Pascual-Marqui et al. 1995) and hierarchical clustering (Murray et al. 2008), are used to cluster EEG/ERP data by ignoring the polarity of time points. Likewise, clustering spatio-temporal ERP data (i.e., from the GFP peaks) is investigated in numerous studies using polarity-invariant clustering methods (Koenig et al. 2014; Murray et al. 2008; Pourtois et al. 2008; Ruggeri et al. 2019) on mostly grand average ERP (i.e., to determine template maps). In this method, a post-hoc procedure is required to reassign the time samples (i.e., based on spatial correlation) to the dominant template maps. Those template maps are commonly selected by considering the maxima global explained variance (GEV) and dissimilarity between them in ERP and using four predefined classic microstate template maps in EEG analysis (Michel and Koenig 2018).

In the second group, recently, a number of researchers argued the limitations of the microstate analysis on GFPs, which potentially increases the uncertainty of the results (Dinov and Leech 2017; Mishra et al. 2019; Shaw et al. 2019). Therefore, some researchers applied polarity dependent clustering methods such as probabilistic-based clustering (Dinov and Leech 2017). The standard *k*-means clustering for ERP (Poulsen et al. 2018) was also used in the Microstate EEGlab toolbox in addition to popular microstate analyzing methods. Furthermore, a clustering method on multiple domains, i.e., (sensor, time)-samples, called cluster-based permutation (Maris and Oostenveld 2007), was proposed to determine the time window in the popular toolbox called FieldTrip (Oostenveld et al. 2011).

Together, both mentioned clustering strategies assumed that a typical clustering method could be sufficiently suitable to be applied on EEG/ERP data without considering the inconsistency and quality of the datasets. However, even the popular clustering algorithms could

Journal : Large 10548 | Article No : 903 | Pages : 19 | MS Code : 903 | Dispatch : 26-5-2022

fail spectacularly for certain datasets that do not match the corresponding modeling assumptions (Acharya and Ghosh 2011). Therefore, consensus clustering has received remarkable attention in biological data processing (Abu-Jamous et al. 2014, 2015a, b), particularly in brain imaging, e.g., functional magnetic resonance imaging (fMRI) and EEG/ERP data processing (Liu et al. 2015, 2017a, b; Mahini et al. 2020; Song et al. 2019). We used consensus clustering from the polarity independent clustering methods to investigate ERPs in the entire dataset. The idea is to provide the most reliable and firmest mutual clustering among the suitable polarity dependent clustering methods without being influenced by their differences in optimization for ERP data.

The rationale of the proposed method is that in an ERP experiment, several ERP components are inevitably generated; however, a few of them are targeted ERP components, which are more probably elicited if the ERP experiment is rerun. Those targeted ERP components are more likely elicited among multiple subjects. Besides, for an interesting ERP component, clustering the ERP dataset into different numbers of clusters may affect its analysis. This is because of two reasons: the first reason is that the ERP component can be associated with a certain brain activity that has its own topography in the spatial (i.e., topographic) domain and its own starting time point and the ending time-point in the time domain. The other reason is that the inappropriate number of clusters used for the clustering may result in separating one true cluster into two or more clusters in practice. Finally, the ideal number of clusters used for the clustering can result in the perfect cluster of interest in theory. The topographies of different time points in a time window will be identical for the perfect cluster since the cluster represents an ERP component of certain brain activity. Therefore, the correlation coefficient of the topographies between any two time points in the time window (called inner-similarity), which is found by the clustering, should be 1 in theory. Our proposed method investigates the mean inner-similarity of the selected time windows (i.e., using a newly developed time window determination method) from many times running consensus clustering for different options (e.g., from 2 to 15 clusters based on the past experiences, which can be changed if needed) to find an optimal number of clusters. Further, we assess our proposed method on two ERP datasets: simulated ERP data and prospective memory (PM) ERP data. We illustrate that the proposed consensus clustering provides a robust clustering result for identifying the most suitable time window for ERPs of interest. On this basis, we also demonstrate the determination of the optimal number of clusters to be carried out via assessing the quality of obtaining time windows.

Materials and Methods

ERP Data

Simulated ERP Data

In order to validate the proposed methodology, simulated data was conducted using Berg's Dipo (2006) simulator (http://www.besa.de/updates/tools/). We defined six components, namely, P1, N1, P2, N2, P3, and N4, and two conditions named 'Cond1' and 'Cond2' from a group of 20 subjects. A simulated scalp with 65 electrodes was defined for representing the spatial dimension. Each epoch started from 100 ms pre-stimulus to 600 ms post-stimulus with a 429 Hz sampling rate. The averaged reference method was used for the referencing. The waveforms of the defined components and corresponding topographical maps have been illustrated in Fig. 1. Two state-of-the-art ERP components, the N2 component that refers to the maximum negative peak in 183–278 ms and the P3 component that refers to the positive response in 231–350 ms, were selected for a more detailed investigation. Meanwhile, the signal was manipulated using the MATLAB function *awgn* (i.e., adding white Gaussian noise) to add some noise, i.e., signal-to-noise ratio (SNR) = 20 dB, to the signal power measured for each individual dataset. Further, a random silent movement of two ERP components (e.g., by randomly increasing/decreasing a maximum of 5 time points) was applied to the signal. Thereby, the conducted ERP data was supposed to be preprocessed, time-locked, and phase-locked. The electrode sites for measuring statistical amplitude power differences were defined in P6/PO4 for N2 and CPz/Cz for P3, which was associated with the definition of each component.

Real ERP Data

The prospective memory experiment (Chen et al. 2015) data was employed from the publication of our group as real ERP data. The real data contained 20 symptomatically remitted patients, i.e., with schizophrenia (RS) and 20 healthy control (HC) participants. Two tasks were investigated, namely, prospective memory (PM) and ongoing task. The EEG data was recorded with 32 electrodes (SynAmps amplifier, NeuroScan) and epoched from 200 ms pre-stimulus to 1000 ms post-stimulus. Furthermore, a 30 Hz (24 dB/octave) digital low-pass filter was applied. Two target ERP components were investigated. Following the previous study, the N300 component refers to the maximum negative voltage over the occipital region (hypothetically between 190 and 400 ms), and

Journal : Large 10548 | Article No : 903 | Pages : 19 | MS Code : 903 | Dispatch : 26-5-2022

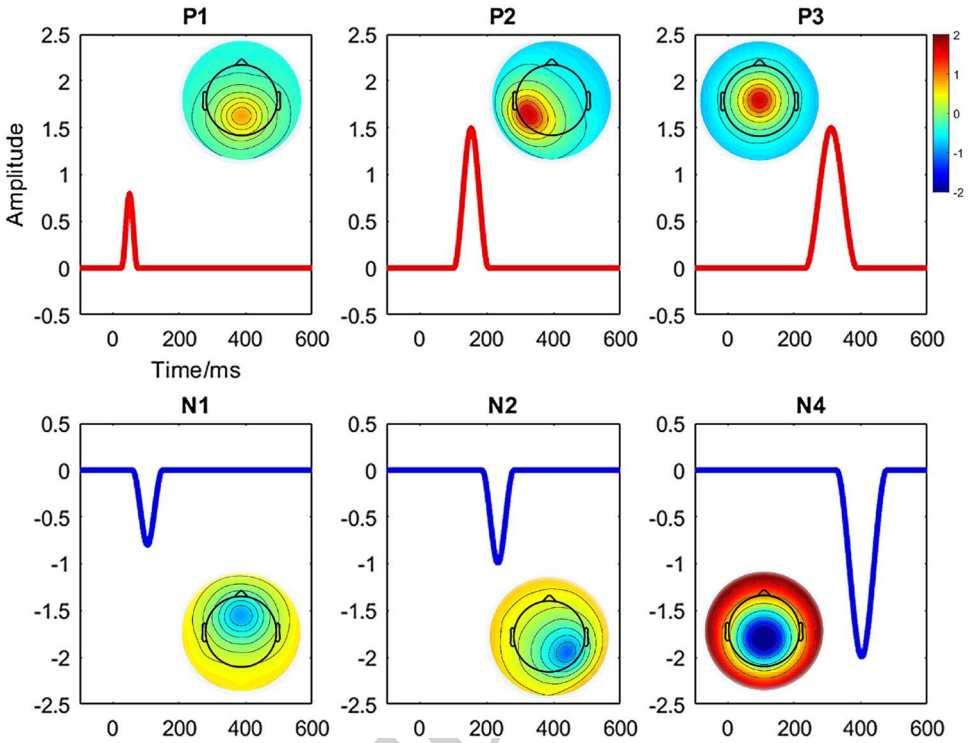


Fig.1 Visualizing the topographical maps and corresponding waveforms of six predefined ERP components, namely, P1, N1, P2, N2, P3, and N4, in the simulated ERP data

the prospective positivity represents the maximum positive voltage over the parietal region (between 400 and 1000 ms).

Proposed Method

The proposed method contains the main procedure (Procedure 1) and a subprocedure for the time window determination (Procedure 2). The graphical representation of the steps of the proposed method is illustrated in Fig. 2, and the pseudo-code is shown in Procedure 1. A more detailed explanation is as follows:

Procedure 1: Optimal number of clusters
Input: ERP data, ERPs of interest info. (Experimental intervals)
Output: Optimal number of clusters, time window
Procedure
  Step1. Temporal concatenating datasets;
  FOR 100 independent runs
    FOR each ERP component
      FOR Number of clusters 2 to 15
        Step2. Consensus clustering;
        Step3. Time window determination (Procedure 2);
      End of FOR (Number of clusters)
    End of FOR (ERP)
  End of FOR (100 runs)
  Step4. Optimal number of clusters determination;
End of Procedure

Data for Clustering

First, a temporal concatenated dataset is provided using the group-averaged ERP data across all conditions/groups of the experiment (Calhoun et al. 2009; Murray et al. 2008). Thereby, samples for clustering are the time points, and the features are represented by the topographical map (i.e., the electrodes' value). The goal of clustering is to find the consecutive time points sharing similar topographies in which the neural responses remain stable for a period of time (i.e., time window).

Consensus Clustering

Consensus clustering refers to synergistically combining multiple clusterings of a dataset(s) into a consolidated clustering result (Acharya and Ghosh 2011). There is, however, no predefined or straightforward solution for selecting clustering methods in consensus clustering literature. It is due to the fact that there is no ground-truth solution (in the clustering) when the data generation process (the recorded data) is complicated (i.e., there is no confirmed information about how well clustering explains the data).

We configured our proposed consensus clustering from the widely used clustering algorithms in neuroimaging, particularly polarity independent clustering methods. This was

Journal : Large 10548 | Article No : 903 | Pages : 19 | MS Code : 903 | Dispatch : 26-5-2022

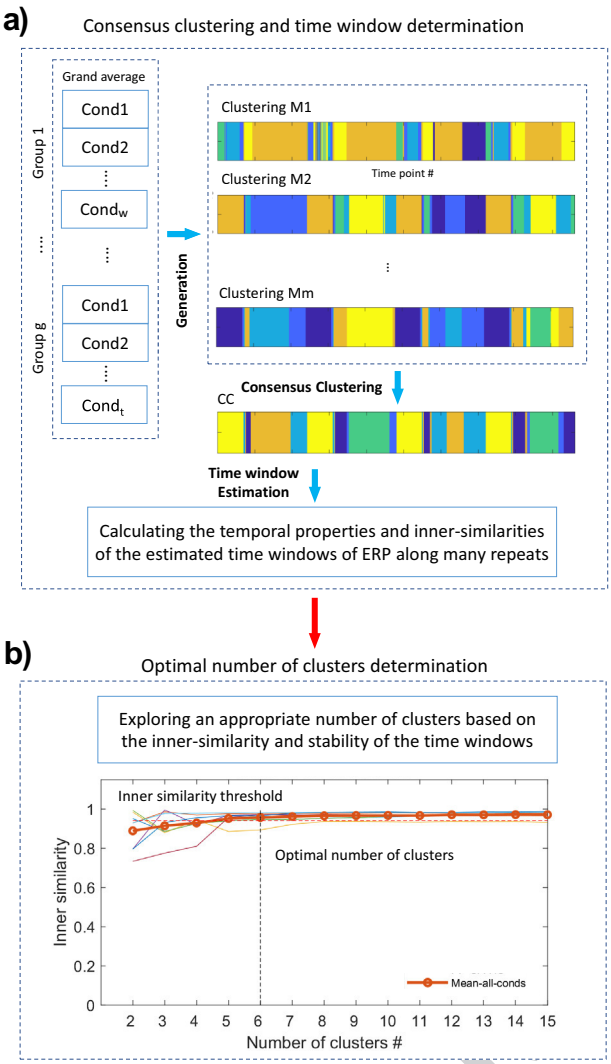


Fig. 2 Illustration of the proposed method for determining the optimal number of clusters. **a** The learning step includes data preparation (temporal concatenating), generation phase of consensus function, and the time window estimation. **b** Processing of the estimated time windows properties along with the previous step's results and determining the optimal number of clusters. The optimal number of clusters is obtained from analyzing the average of the means of inner-similarities among the conditions (i.e., obtained from many times running the method). *CC* consensus clustering, *Cond* condition

because of two reasons; first, our method for processing ERP data uses the entire time points, taking their polarity into account; second, to ensure consistency between the methods in the generation step. We used the M–N plot evaluation technique (Abu-Jamous et al. 2014) from the individual clustering methods on the grand average ERP data to qualify clustering methods. Two criteria were considered, the inner-similarity and number of samples in the estimated time windows. The procedure was repeated several times, testing each method (e.g., 20 to keep the plot readable) with the number of clusters obtained from the explained variance analysis, i.e., four and six cluster maps for the simulated and the real data. A number of polarity dependent clustering methods were used for assessing and configuration of consensus clustering, namely, *k*-means (Pena et al. 1999) and hierarchical clustering (Tibshirani and Walther 2005) with correlation similarity function, fuzzy *c*-means (FCM) (Bezdek 1981), self-organizing maps (SOM) (Kohonen 1990), diffusion map spectral clustering (Sipola et al. 2013) consisting of *k*-means with Euclidean similarity, standard spectral clustering (SPC) (Ng et al. 2002), *k*-medoids clustering (KMD) (Park and Jun 2009) with correlation similarity function, and Gaussian mixture model clustering (GMM) (De Lucia et al. 2007; Dempster et al. 1977) with repetitive *k*-means structure.

Afterward, the proposed method generates many partitions in over 100 independent runs (via the selected clustering methods) for each option (e.g., 2–15 clusters) and applies the consensus clustering for each independent run. In the ensembling phase, cluster-based similarity partitioning algorithm (CSPA) (Karypis and Kumar 1998; Nguyen and Caruana 2007), based on pairwise similarity, is utilized as a measurement between partitions. The goal of the consensus clustering method is to find aggregate labeling such as L^* which could better represent the properties of each labeling in L in terms of specificity and coverage of the information in the dataset. Mathematically, L^* can be defined as:

$$L^* = argmax_{L\in\mathbb{L}} \sum_{r=1}^{R} \Gamma(L_r), \tag{1}$$

where Γ denotes a similarity measurement (i.e., cluster-based similarity), which can measure mutual information between a set of R clusterings. From Eq. 1, the cluster ensemble L^* is an optimal clustering with maximum similarity to other clusterings. Hence, the consensus function puts the samples in clusters where they have clustered in the same group from most clusterings. Once the clustering labels are assigned via clustering methods (generation phase), the consensus function (i.e., CSPA) explores the maximum aggregation between the clusterings. Therefore, the final clustering carries the mutual information (Eq. 1) from different methods (i.e., topographical configuration, polarity, and similarity in temporal domain with configuration changes). In other words, the consensus function combines mutual information about the cognitive processes in the data from different results.

To assess the role of the selected clustering methods, ARI was used to measure the mutual similarity (Meila 2007; Strehl and Ghosh 2003) between each clustering method and consensus clustering. Rand index (1971) can be calculated using the following equation:

Journal : Large 10548 | Article No : 903 | Pages : 19 | MS Code : 903 | Dispatch : 26-5-2022

$$\mathcal{R}(L,L') = \frac{N_{11}+N_{00}}{n(n-1)/2}, \quad (2)$$

and by calculating the expectation of $\mathcal{R}$ ($E[\mathcal{R}]$), ARI is calculated as the following:

$$ARI(L,L') = \frac{\mathcal{R}(L,L') - E[\mathcal{R}]}{1 - E[\mathcal{R}]}, \quad (3)$$

where n denotes the number of observations and N_{00} denotes the number of object pairs in different clusters from both L and L' clusterings. While N_{11} denotes the number of object pairs in the same clusters in L and L'.

Time Window Determination

The time window determination method analyzes the temporal and spatial characteristics of the result cluster maps for estimating an appropriate time window (see Procedure 2). The candidate cluster maps' inner-similarity and overlapping (with the experimental defined ERP component), i.e., the maps in the experimental measurement area, were investigated for qualifying the time windows. It is noteworthy that, following the literature (Kappenman and Luck 2012) and our experience, we rely on the experiment mechanism that undertakes an important role in defining the experimental interval of the components. The inner-similarity of consecutive time points is calculated in the candidate cluster maps, which is aimed at selecting the cluster maps with a higher spatial correlation. The inner-similarity of a cluster map refers to the mean of correlation coefficients between topographical maps of each two different time points.

Procedure 2: Time window determination
Input: Clustering result, ERPs of interest info. (Experimental intervals)
Output: Time windows, Inner-similarities
Procedure
  Step1. Detecting the candidate cluster maps;
  FOR Each candidate map
    Step2. Calculating inner-similarity and overlapping;
    Step3. Detecting cluster maps with high inner-similarity;
    Step4. Selecting higher overlapping within maps;
  End of FOR
End of Procedure

In order to calculate the inner-similarity of a cluster map, first, the spatial correlation coefficient (Micah et al. 2009; Murray et al. 2008) of the time points is calculated. Then, for each row (in the correlation matrix), the distance matrix is calculated as:

$$D_v = d(Cor_{v,u}, Cor_{v,v}), u \neq v, \quad (4)$$

where D denotes the distance matrix in which each row is the distance between each element in the row and $Cor_{v,v}$ (i.e., self-correlation) from the correlation matrix (Cor).$Cor_{v,u}$ denotes the correlation coefficient between the topographical maps of u and v as two time points in the cluster map. For the variance-stabilizing transformation of the calculated correlation, Fisher z-transform (1921) was used for each vector D_v (i.e., every row of distance matrix) before calculating the mean of the distance matrix D_{avg}. Finally, an inverse z-transform of D_{avg} was used for calculating the inner-similarity as shown below:

$$InnSim = 1 - D_{avg}. \quad (5)$$

Therefore, among the candidate cluster maps, the cluster maps with a higher inner-similarity than the threshold (e.g., ≥ 0.95) were selected for overlap testing. Hypothetically, in the ERP component, the spatial correlation between the time points is close to 1, indicating consecutive time points representing a cognitive process. Specifically, the minimum acceptable time interval, which is a sufficient number of time points for selecting the candidate cluster maps, was determined, e.g., ≥ 50 ms (Luck 2014) depending on the experimental goal, to avoid noise effect in measuring peak latency of components. Next, among the candidate cluster maps, the cluster map with a higher inner-similarity and overlap (i.e., with the experimental interval) was selected as the suitable cluster map for representing the time window. The temporal properties of the selected cluster map (start, end, and duration) are used for further steps. The implementable procedure of determining the time window was presented in Procedure 2.

Optimal Number of Clusters

The inner-similarities of the estimated time windows are calculated from many times independent running of consensus clustering. Afterward, the optimal number of clusters for the prepared ERP dataset is calculated from the average inner-similarities' mean across the conditions and groups. In addition to the parameter with the inner-similarity threshold (e.g., ≥ 0.95), the stability threshold, which is a very small change (e.g., ≤ 0.03) in the amplitude of inner-similarities, has been considered to determine the appropriate number of clusters. In other words, the optimal number of clusters can be expected when inner-similarity is satisfactorily high and stable (i.e., the minimum change concerning other nearby options) from the clustering options. Noteworthy to mention that, in our updated toolbox, we have provided a dynamic mechanism for adjusting the sensitivity parameters (i.e., the inner-similarity and stability thresholds) whenever needed to provide a better adaptation when the data is not well preprocessed.

Journal : Large 10548 | Article No : 903 | Pages : 19 | MS Code : 903 | Dispatch : 26-5-2022

Statistical Analysis

In order to assess the statistical performance of the new methodology, a repeated-measures ANOVA was performed with the within-subject factor, *Task* ('Cond1' and 'Cond2') in the electrode sites (P6/PO4 for N2 and CPz/Cz for P3) in the simulated data. This was set up to the mean amplitude of N2 and P3 in the estimated time windows. The standard analysis was carried out to determine whether these effects of the noted factors for each study were statistically significant. Likewise, another statistical analysis (for N300 and prospective positivity) was performed by a repeated-measures ANOVA (i.e., mixed 2×2) with the addition of a between-subject factor: *Group* (RS and HC) and within-subject factors, *Task* (PM and Ongoing) in the electrode sites following the previous study (Chen et al. 2015) (i.e., electrode sites are O1/Oz/O2 for N300 and P3/Pz/P4 for prospective positivity). This was accomplished for the mean amplitude of N300 and prospective positivity measured in the selected time windows. The statistical comparisons were made at p-values of $p < 0.05$.

Results

As explained earlier, the M–N plot method has been used for selecting the clustering methods that configure consensus clustering. Figure 3 illustrates the test results for each dataset. Therefore, in order to provide an appropriate configuration of consensus clustering, we eliminated GMM, HC, and SPC methods from clustering of the real data (Fig. 3a), and GMM, DFS, and FCM from simulated data (Fig. 3b).

Results for Simulated ERP Data

We provided a feasibility test for the proposed clustering performance on the simulated data. Figure 4 illustrates the clustering result and the topography maps of identified ERP components by the cluster maps. Observably, the proposed method has successfully isolated all the predefined ERP components assigning seven cluster maps (i.e., all the components with the corresponding topography maps). Therefore, the P1, N1, P2, N2, P3, and N4 components are qualified with the cluster maps 1, 4, 3, 5, 6, and 7, respectively. Therefore, the predefined ERP components are qualified based on the spatial correlation between the predefined spatial configuration (topographic maps) and the obtained cluster maps from the proposed clustering method. It should be stated that cluster map 2 refers to the brain state before stimulus onset and does not present any predefined ERP component. However, in practice, it is not always possible to distinctly qualify all the ERP components in the real experimental data due to the overlapping between the components and experimental conditions.

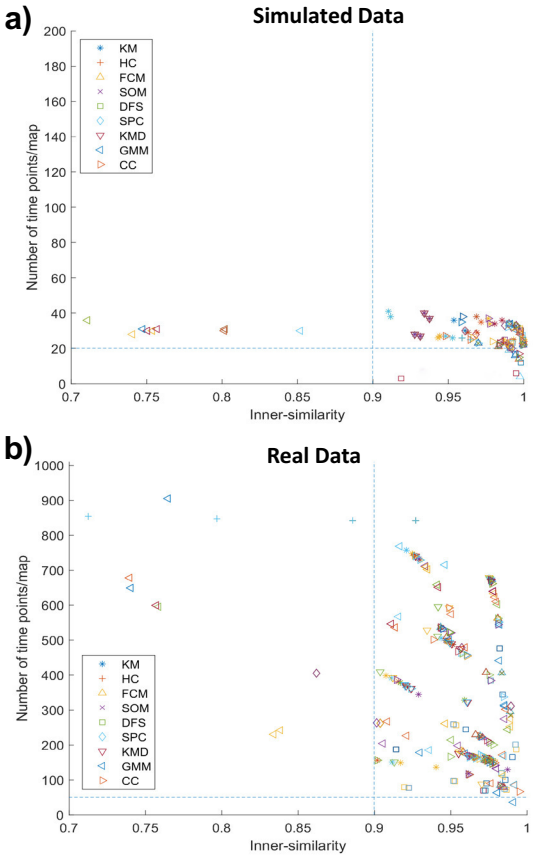


Fig. 3 The results of assessing clustering methods for consensus clustering configuration. The M–N plot method from the obtained time window results by each clustering method for the ERPs of interest from 20 repeats and the number of clusters calculated from the expressed maximum 90% of explaining variance, i.e., four and six cluster maps for the simulated and the real data. **a** GMM, DFS, and FCM for simulated data, and **b** GMM, HC, and SPC for the real data, are excluded from the consensus clustering configuration. *KM* *k*-means, *HC* hierarchical clustering, *FCM* fuzzy c-means, *SOM* self-organizing map, *DFS* diffusion map spectral clustering, *SPC* spectral clustering, *KMD* *k*-medoids clustering, *GMM* Gaussian mixture model, *CC* consensus clustering

Inner-similarity and the Optimal Number of Clusters

Figure 5 shows the changes of the inner-similarity of the estimated time windows in over 100 independent runs from the clustering options (i.e., from 2 to 15 clusters). As outlined before, our strategy is to find the optimal number of clusters wherein the mean of the inner-similarities (of the obtained time windows) tends to be reasonably high and stable. Therefore, the proposed method explores the optimal

Journal : Large 10548 | Article No : 903 | Pages : 19 | MS Code : 903 | Dispatch : 26-5-2022

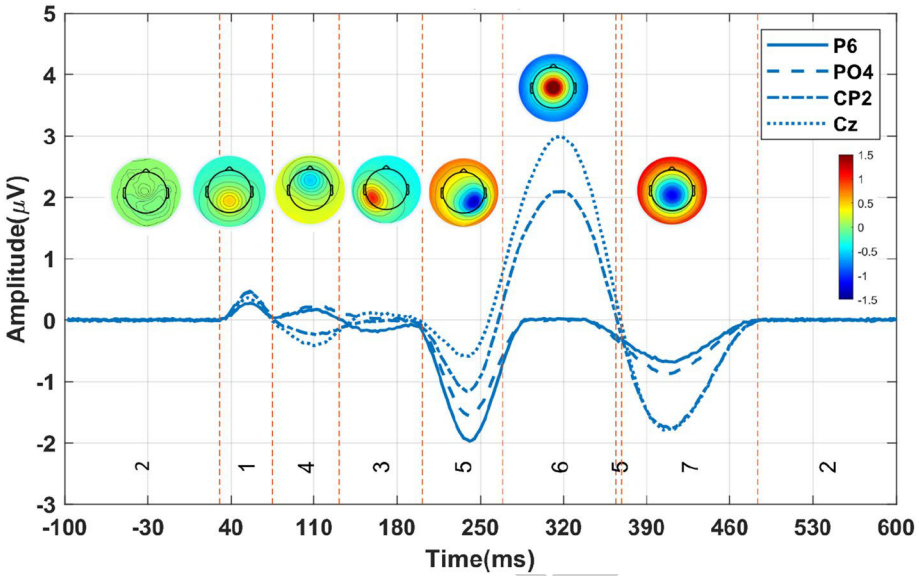


Fig. 4 The proposed consensus clustering results on grand-averaged ERP (i.e., condition 2) in the selected electrode sites (Pz, Fz, CP5, P6, Cz, and CPz). The qualified P1, N1, P2, N2, P3, and N4 components correspond to the cluster maps 1, 4, 3, 5, 6, and 7, respectively (cluster map 2 refers to the brain state before stimulus onset)

number of clusters from the average inner-similarities mean of all conditions (see Fig. 5c). Accordingly, the optimal number of clusters for N2 and P3 components was obtained in six maps. Table 1 shows the temporal properties (in average) of the estimated time windows applying the obtained optimal number of clusters. As can be seen from the table, there is no reasonable aggregation between the results of various methods, which can lead to unreliable interpretation while investigating those results. This phenomenon indicates the effect of the existing overlapping between the components and the compression level of clusterings. Particularly, Gap statistic and Dunn methods carried out a trivial performance (i.e., the optimal number of clusters cannot be determined in the defined range); thus, we eliminated their results from the table.

Cluster Analysis and Time Window Determination

Figure 6 illustrates the clustering result (one randomly selected result) when applying the optimal number of clusters to isolate the N2 and P3 components. It is observable that the N2 component (marked by blue color) is distinctly isolated by cluster map 4 for both conditions from 203 to 262 ms and 201 to 262 ms, respectively. Likewise, P3 (marked by pink color) is isolated by cluster map 1 from 268 to 353 ms and 268 to 360 ms post-stimulus for Cond1 and Cond2, respectively. Comparison between the temporal properties of the estimated time window results in Table 1 and the ground-truth time window reveals that the N2 and P3 components were successfully estimated using the proposed method. Notably, we have reported the mean of the obtained time windows over 100 independent runs in Table 1.

Statistical Results of Peak Amplitudes

We measured the mean amplitude of grand-means data in the selected time windows from the previous step for each condition/group in the defined electrode site. Regardless of the methods (for determining the optimal number of clusters), the statistical power analysis results revealed that the main effect of *Task* was significant ($p < 0.0001$) in the identified N2 component. Similarly, a significant ($p < 0.0001$) main effect of *Task* was detected in the P3 component. For both components, the measured amplitudes were larger in Cond2 as our expectation (i.e., from the simulation mechanism). Notably, these results disclose that the proposed time window determination successfully isolates the N2 and P3 components.

Static Properties Analysis

The mean of inner-similarities from the obtained time windows over many runs of the studied clustering methods is illustrated in Fig. 7. It illustrates the role of studied clustering methods in the proposed clustering, disclosing a similar trend in studied clustering methods and consensus clustering for both components of interest. This also indicates the stable performance of consensus clustering in terms of the inner-similarity of the estimated time windows. Besides, Table 2 provides more apparent evidence for the existing similarity between the studied clustering methods (i.e., the aggregation between the clusterings). This provides information about the contribution of each clustering method to the proposed clustering. The ARI measurement (see section "Consensus Clustering") was used to calculate the similarity between the clustering results.

Journal : Large 10548 | Article No : 903 | Pages : 19 | MS Code : 903 | Dispatch : 26-5-2022

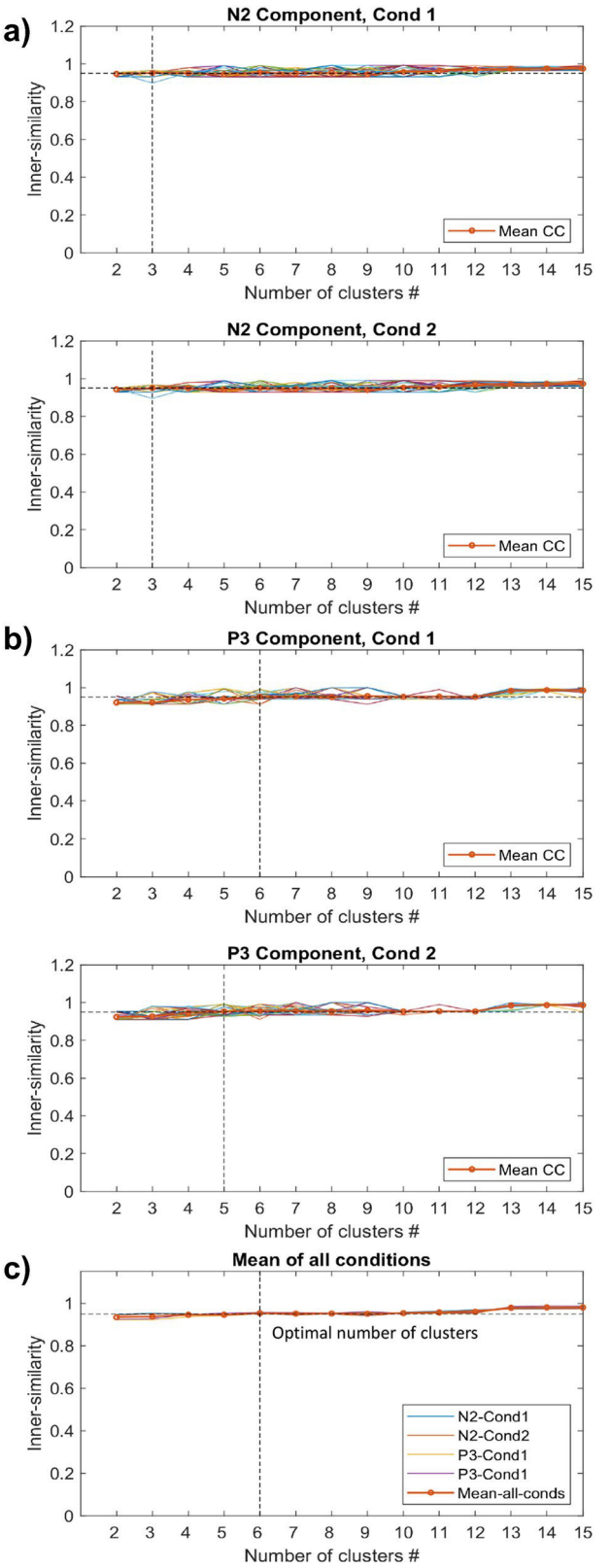


Fig. 5 Illustration of the optimal number of clusters estimation from the mean of inner-similarities of the time windows over 100 independent runs for N2 and P3 in the simulated ERP data. **a** The inner-similarities of the obtained time windows for N2 in two conditions ('Cond 1' and 'Cond 2'). Likewise, **b** the inner-similarities for P3 in two conditions. **c** The yielded the optimal number of clusters (indicated with vertical black hidden line), using the threshold of 0.95 (highlighted with black horizontal hidden line) and the stability threshold of 0.03 (difference between the previous and next values), from the average of inner-similarities from conditions (illustrated by the dark orange line). *N2-Cond1* mean inner-similarity in condition 1 and N2, *N2-Cond2* mean inner-similarity in condition 2 and N2, *P3-cond1* mean inner-similarity in condition 1 and P3, *P3-Cond 2* mean inner-similar in condition 2 and P3, *Mean-all-Conds* average inner-similarities mean across all the conditions

Results for Real Data

Inner Similarity Analysis and the Optimal Number of Clusters

Figure 8 shows the inner-similarity variations (i.e., in the estimated time windows) with the number of cluster options for each condition and group. As mentioned before, the optimal number of clusters can be estimated from the average of mean inner-similarities across the conditions and groups. Hence, the optimal number of clusters for real data met in six clusters (Fig. 8e) by satisfying the inner-similarity and stability thresholds. On the other hand, the optimal number of clusters from the studied conventional methods is reported in Table 3. Observably, missing a suitable aggregation discloses challenges about the reliability and probable misleading interpretation issues for the clustering results upon the selected number of clusters via the conventional methods.

Cluster Analysis and Time Window Determination

The clustering results, including the identified time windows (indicated by the colored areas), the corresponding topographical maps, and the spatial correlation of time samples (from one randomly selected clustering result), have been illustrated in Fig. 9. As shown in the figure, N300 is qualified with cluster map 5 for RS and HC and both tasks. Accordingly, N300 (the marked area with blue color) is identified in 164–244 ms and 165–319 ms post-stimulus for PM and ongoing tasks in the RS group, where it is identified in 156–312 ms and 162–333 ms post-stimulus from the HC group, respectively. Similarly, the prospective positivity (marked by pink color) is identified in 256–1000 ms and

Journal : Large 10548 | Article No : 903 | Pages : 19 | MS Code : 903 | Dispatch : 26-5-2022

Table 1 The obtained number of clusters from the studied methods and the average time window properties results, i.e., from 100 runs, to qualify the N2 and P3 components in the simulated ERP data

Method	OptNC	Properties	N2		P3	
			Cond1	Cond2	Cond1	Cond2
Ground truth	–	Start	201.0	201.0	266.0	266.0
		End	261.0	261.0	357.0	364.0
		Duration	60.0	60.0	91.0	98.0
Silhouette NbClust optimal	3	Start	203.7/0.8	201.5/1.0	268.6/2.3	268.0/2.2
		End	264.3/1.0	264.3/1.0	354.2/1.4	359.1/2.9
		Duration	**60.6/1.4**	62.8/1.5	85.5/3.4	91.1/5.0
x-means elbow	4	Start	203.7/0.8	201.7/1.1	268.6/1.9	268.5/1.9
		End	264.4/1.6	264.4/1.5	354.6/1.2	359.4/2.2
		Duration	60.7/2.0	62.8/2.0	86.0/2.6	91.0/3.5
Proposed method	6	Start	**203.3/0.0**	201.4/0.9	**268.5/0.5**	**267.9/0.9**
		End	**262.3/1.6**	262.4/1.6	**354.9/0.5**	**360.7/1.3**
		Duration	59.0/1.6	60.9/1.9	**86.3/0.7**	**92.8/1.7**
Cross-validation	11	Start	203.5/0.6	201.2/0.6	268.8/0.7	268.8/0.7
		End	262.6/2.2	**261.6/2.7**	353.0/1.2	359.6/0.7
		Duration	59.1/2.3	**60.4/3.1**	84.3/1.7	90.8/1.3
Modified *k*-means	11	Start	**203.3/0.0**	**201.0/0.0**	271.2/1.5	271.2/1.5
		End	265.59/2.3	265.6/2.3	352.6/1.2	359.2/2.0
		Duration	62.25/2.3	64.6/2.3	81.4/2.7	88.0/3.4

The bold font marks the results representing the significant outcomes considering the ground-truth time windows. The reported format is, averaged time windows (ms)/standard deviation error, SD (ms)

Cond1 condition 1, *Cond2* condition 2, *OptNC* optimal number of clusters

360–922 ms post-stimulus for PM and ongoing tasks in the RS group (i.e., isolated by map 2 and map 1, respectively), and 313–695 ms and 376–783 ms for PM and ongoing tasks in HC group by map 4 and map 1, respectively. Table 3 illustrates the mean of time windows temporal properties applying the obtained number of clusters from different methods.

Statistical Results of Peak Amplitudes

The mean of *p*-values in over 100 independent tests was reported in Table 4. The results showed that the main effect of *Group* ($p < 0.0001$) was significant. Likewise, the significant main effect of *Task* type ($p < 0.002$) was observed for qualifying the N300 component. Importantly, the interaction between *Task* and *Group* was also significant ($p < 0.048$). However, the interaction between *Task* and *Group* was not sufficiently stable (SD = 0.020) along with the results. Furthermore, the statistical analysis revealed that N300 in HC was qualified by a significantly more negative potential over the occipital-central electrodes ($p < 0.001$). Besides, a silently more negative potential was observed over occipital-central electrodes ($p < 0.001$) in the ongoing task from both RS and HC groups in the N300 component. Likewise, investigating the prospective positivity component showed that the main effect of *Task* was significant ($p < 0.0001$). However, the main effect of *Group* ($p < 0.393$) and the interaction effect between *Task* and *Group* ($p < 0.085$) were not significant. Furthermore, a larger positive potential was localized over central electrodes ($p < 0.0001$) in the ongoing task compared to the PM task for both groups.

Static Properties Analysis

Table 4 shows the mean *p*-value regarding the studied factors (*Group*, *Task*, and interaction between *Task* and *Group*) for the N300 and prospective positivity components. As observable in Table 4, the proposed method seems to afford relatively more stable analyzing results in the estimated optimal number of clusters than the result yielded from other methods. Note that the clustering results from modified *k*-means (in many times) were applied to the time window determination method to calculate the statistical analysis results.

From the information theory perspective, Fig. 10 demonstrates similar behavior of the studied clustering methods, with observing their mean inner-similarity in both N300

Journal : Large 10548 | Article No : 903 | Pages : 19 | MS Code : 903 | Dispatch : 26-5-2022

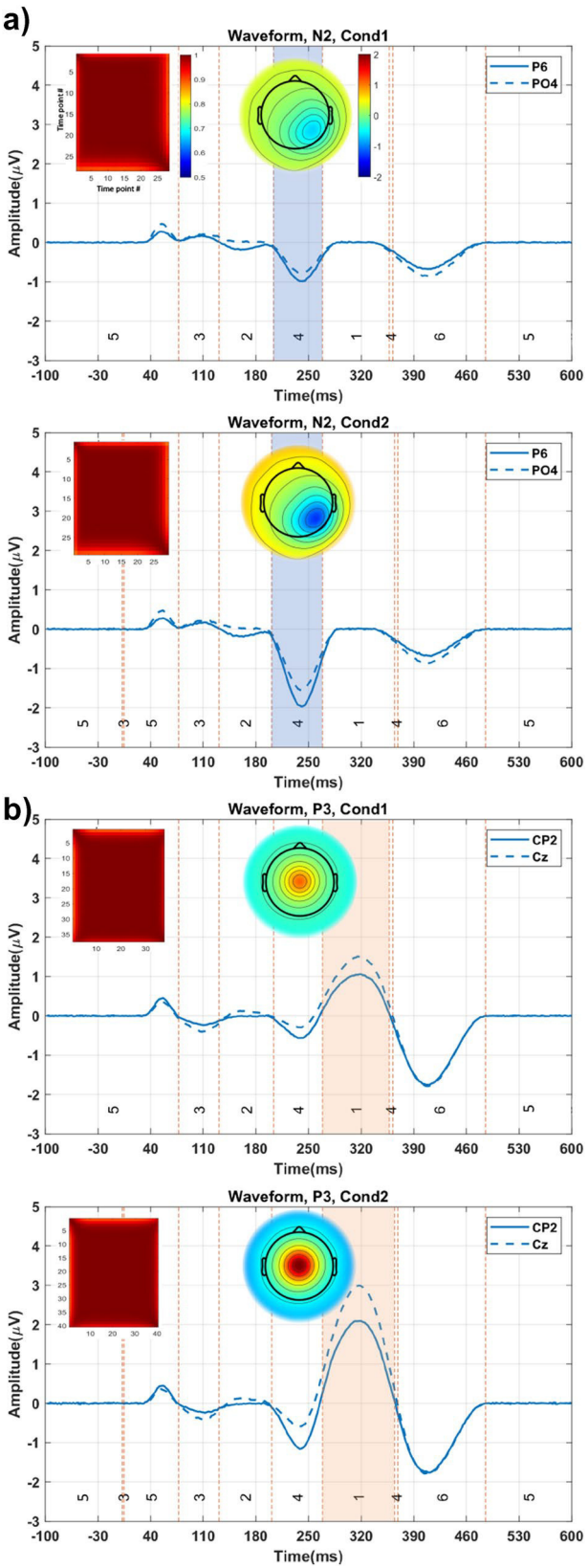


Fig. 6 Clustering and time window determination results for the N2 and P3 components. **a** The qualified N2 component by map 4 (colored blue) for both conditions, the corresponding topography map, and the correlation between time points for each condition. **b** The qualified P3 component by map 1 (colored pink) for both conditions

and prospective positivity components. This can be attributed to the role of each individual clustering method in the proposed consensus clustering approach for real ERP data. The smoothness and overall high performance of consensus clustering (i.e., indicated with bold light blue line) can be observed in terms of robustness and the obtained high inner-similarities. The calculated averaged similarity between the clusterings, i.e., in the obtained optimal number of clusters (see Table 5), provides a suitable criterion for evaluating the contribution of individual clustering to consensus clustering.

Discussion

The present study shows that the optimal number of clusters for spatio-temporal ERP can be determined based on observing the phenomena that correspond with the quality of identifying the ERPs of interest. A consensus clustering was designed with the aim of obtaining a reliable clustering (series of stable intervals in time). The time window estimation method was applied to identify the ERPs of interest from the clustering result. The main objective of this study was to accomplish a robust clustering mechanism for determining the optimal identification of interesting ERP by studying the temporal dynamic and sensory information about group brain response.

The simulated and real ERP data were used to assess the performance of the proposed method. Our results declared that the proposed method successfully determines the optimal number of clusters by analyzing the quality of the identified ERPs of interest (i.e., a few target components) in both ERP data. Our findings also revealed that analyzing the quality of time window (i.e., high inner similarity in the experimental interest area), which indicates the stable brain state (highly probable to be a brain response), can be a useful tool for selecting an appropriate number of clusters. Furthermore, we found that the isolated ERP components differ on the properties of the time window (i.e., in the start, end, and duration) in different conditions/groups, which reveals the variety in the spatial and temporal dynamics of the brain response in different conditions and groups.

Journal : Large 10548 | Article No : 903 | Pages : 19 | MS Code : 903 | Dispatch : 26-5-2022

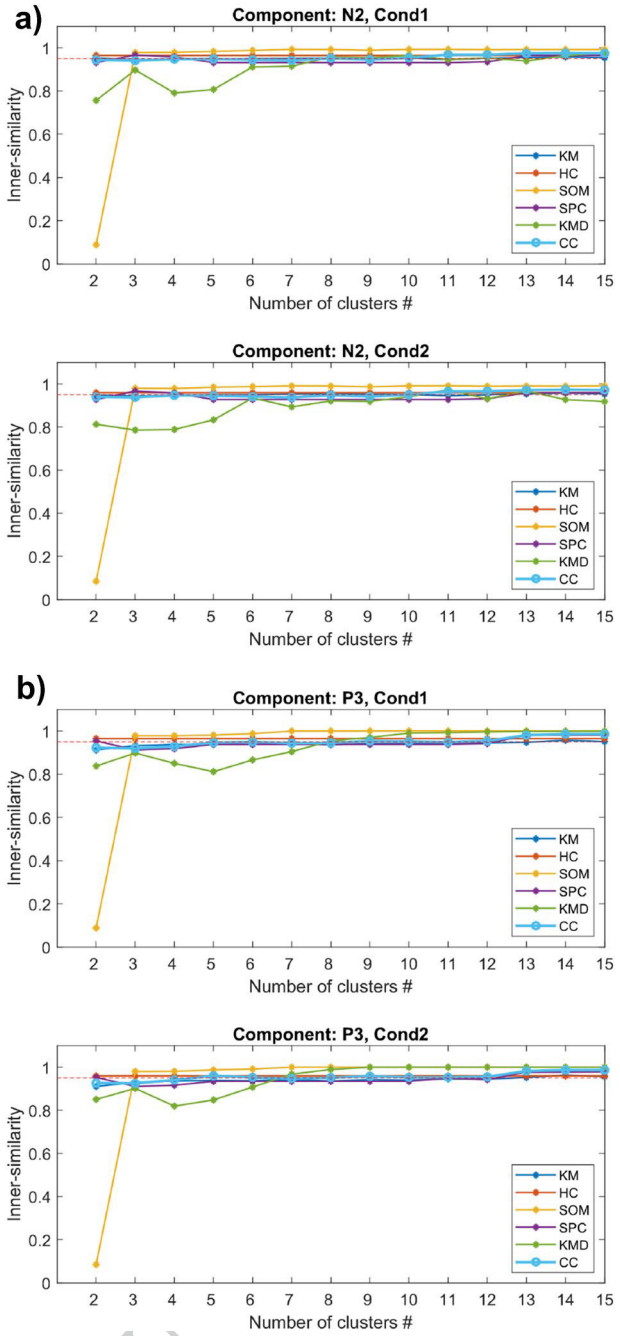


Fig. 7 Mean inner-similarities of the estimated time windows using the studied clustering methods from clustering options (2–15 clusters) in the simulated data. **a** The mean inner-similarity of each method for N2. **b** The mean inner-similarity of each method for P3

Compared with the conventional methods to find the optimal number of clusters and the combination of various indexes in literature (Murray et al. 2008; Pourtois et al. 2008; Ruggeri et al. 2019; von Wegner et al. 2018; Zappasodi et al. 2019) for brain imaging clustering, we showed that the compression level (provided by the number of clusters determination) affects the quality of the estimated time window and the statistical power analyzing results. As such, some recent works focused on data-driven methods based on using cross-validation methods for optimizing the quality of the results, e.g., based on spatial correlation (Koenig et al. 2014). In the discussed method by Koenig et al. (2014), the maximum mean correlation obtained by calculating the mean correlation between template maps was used for estimating the optimal number of template maps from many times clustering for each clusters option (e.g., 2–20). This study inspired us to design a data-driven analysis by investigating different quantifies. Although this method is practically useful, the authors investigated the whole cluster maps on grand average spatio-temporal ERP, applying, e.g., the cross-validation method and assessing the clustering quality. Besides, following the literature (Kappenman and Luck 2012; Michel and Pascual-Leone 2020), only a few ERPs can be distinctly elicited in the real experiment. This is because of the existing overlap between the components, affecting the accuracy of isolating the ERP of interest. Thus, the proposed method investigates the candidate cluster maps for ERP of interest, despite the literature that analyze the whole cluster maps in spatio-temporal ERP. Nevertheless, the obtained time windows quality needs to be carefully studied in terms of reliability and interpretability of the results.

As mentioned in the introduction, both polarity-invariant and polarity dependent clustering methods were used to cluster spatio-temporal ERP with the aim of exploring the time window of interest. One should note that merely ensembling clustering methods without investigating the polarity of samples, or composing a consensus of polarity-invariant with polarity dependent clustering methods, will increase the risk of assigning the samples with different polarities into identical maps. (i.e., a post-hoc procedure might be required). Although clustering of GFP values in the polarity-invariant methods provides a suitable tool for spatial dynamic analysis of the brain, the winner-takes-all strategy for determining the dominant template maps and GFP peak analysis increase uncertainty in the data and overlapping within the ERP components (Dinov and Leech 2017; Shaw et al. 2019). However, limited clustering methods are available from both strategies. Therefore, considering the fact that no clustering method matches the corresponding modeling assumptions and to minimize cluster ensemble inconsistency, we propose the firmest mutual clustering from polarity dependent clustering methods without being influenced by their differences in optimization for ERP data.

Journal : Large 10548 | Article No : 903 | Pages : 19 | MS Code : 903 | Dispatch : 26-5-2022

Table 2 The mean measured similarity and SD value between the clusterings in many times (up to 100 times) running the studied clustering methods for the simulated ERP data

Clustering method	KM	HC	SOM	SPC	KMD	CC
KM	–	0.81//0.02	0.74/0.02	0.84/0.03	0.74/0.03	0.82/0.03
HC	–	–	0.72/0.01	0.80/0.00	0.72/0.01	0.79/0.03
SOM	–	–	–	0.76/0.01	0.86/0.04	0.89/0.04
SPC	–	–	–	–	0.74/0.02	0.80/0.02
KMD	–	–	–	–	–	0.89//0.04
CC	–	–	–	–	–	–

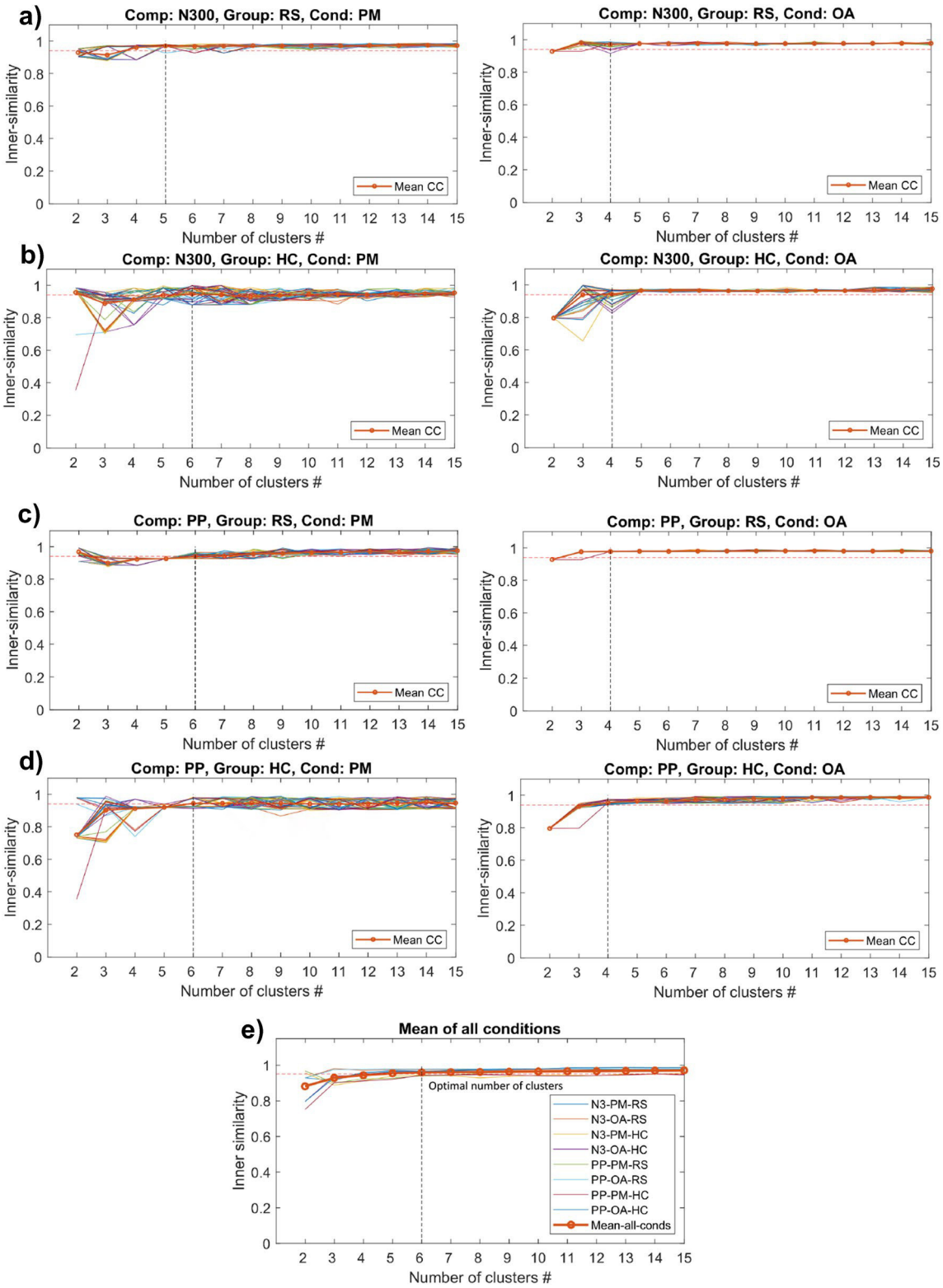
Another parallel issue is the investigation of the configuration of consensus clustering considering how well it represents the ERPs of interest, besides the lack of ground-truth segmentation and a straightforward method for selecting clustering methods. Our method selects the clustering methods satisfying two criteria: inner-similarity and sufficient time window duration. Those selected clustering methods are most likely to be able to isolate ERPs. Furthermore, we provided the evaluation by testing the performance of single clustering methods based on the inner-similarity criteria (Figs. 7, 10). Our results revealed that almost all the selected single clustering methods followed a similar behavior, which is important to obtain more accurate results from the consensus clustering. Meanwhile, it was also observed that consensus clustering showed more stability and a higher inner-similarity (i.e., in most conditions). Furthermore, the statistical power analysis in over 100 runs of the proposed method entailed suitable robust results that made the proposed method be able to be applied to different ERP data. However, in some of the statistical results, the proposed method is not the best (e.g., compared to the previous study), resulting from subjectively selecting a specific time range to get a larger difference of the peaks.

According to the obtained results, two major differences were noticed between the proposed method and conventional methods: (i) the comparison results revealed that qualifying the ERP components is sensitive to the determination of the appropriate number of clusters. Thus, the inappropriate number of clusters can influence isolating the interesting components. We designed a data-driven approach investigating the experiment goal for identifying the ERP components. Applying the proposed method to the real data disclosed that the obtained cluster maps are interpretable as the real cognitive processes. (ii) Considering two important criteria for identifying the ERP of interest, the time window parameters (the start, end, and duration) and the stability of topographical configuration (i.e., high inner-similarity) disclosed the fact that the new methodology is complementary to the conventional microstates analyzing in terms of isolating the ERP of interest. The drawback of the proposed method is that if the preprocessing of the data was not perfectly performed or the data has large combined noises, the clustering of such data can lead to uninterpretable results. Additionally, the high overlapping of the ERP component can challenge similar newly developed methodologies.

From the clinical perspective, identifying two ERP components in real data can be interpreted as the fact of the variety of brain responses from the subjects for different conditions/groups. In N300 component isolation, for example, the difference was shown in cluster maps 1 (i.e., between RS and HC groups) in PM tasks. Likewise, the duration differed in cluster maps 1 and 2 in the ongoing task between the groups. Again, a silently larger negative response was observed at the source level in the ongoing than PM task in both RS and HC groups. This was reflected by the significant main effect of task type among these two groups. As a result, this finding showed a complementary viewpoint to the prior studies (Chen et al. 2015; Fukumoto et al. 2014). Therefore, our results can be employed for interpreting the advantage of the treatment in RS patients in terms of measuring/identifying the difference in ERPs of interest in the observations. This can provide further evidence for recent research demonstrating that symptomatic remission in schizophrenia is associated with a degree of functional recovery of attentional processes.

In summary, our findings are anticipated to be a welcomed addition to EEG/ERP studies in terms of applying the consensus clustering technique and the optimal number of clusters determination. This is attained due to two main reasons: firstly, analyzing spatio-temporal ERP by focusing on a few components of interest instead of fitting microstates to template microstate classes. Secondly, despite conventional methods, which have used a single clustering method, the proposed method utilizes a robust clustering strategy representing neurophysiologically interpretable ERP identification. In order to provide access to the new methodology, a toolbox has been developed under the MATLAB platform named OptNC_ERP (https://github.com/remahini/OptNC_ERP) available with the simulated ERP data (i.e., can be used with another ERP data), which can be used beside EEGLAB (Delorme and Makeig 2004) for testing the researchers' hypotheses.

Journal : Large 10548 | Article No : 903 | Pages : 19 | MS Code : 903 | Dispatch : 26-5-2022



Journal : Large 10548 | Article No : 903 | Pages : 19 | MS Code : 903 | Dispatch : 26-5-2022

◂**Fig. 8** Estimating the optimal number of clusters in the real ERP data. **a** The inner-similarities of estimated time windows in over 100 runs in the RS group and conditions (PM and OA) for the N300 component. **b** Results for the HC group and both conditions for the N300 component. **c** The inner-similarities of estimated time windows in the RS group and both conditions (PM and OA) for prospective positivity. **d** Results in the HC group and both conditions for prospective positivity. **e** The estimated optimal number of clusters, i.e., six cluster maps (showed with vertical black hidden line). The inner-similarity and stability thresholds on the average mean inner-similarities across the conditions and groups are 0.95 (indicated with red horizontal hidden line) and 0.03, respectively. *PP* positivity component, *PM* prospective memory, *OA* ongoing task, *RS* remitted schizophrenia, *HC* healthy control

Conclusions and Outlook

This work illustrated the determination of the optimal number of clusters using consensus clustering. The proposed method investigated both dynamics of cluster maps, i.e., spatial correlation of microstates and temporal properties (i.e., start, end, and duration), to determine the proper time window for an interesting ERP component.

Table 3 The obtained number of clusters from the studied methods and the average time window properties results in over 100 runs

Method	OptNC	Properties	N300				Prospective positivity			
			RS		HC		RS		HC	
			PM	OA	PM	OA	PM	OA	PM	OA
Silhouette	3	Start	160.6/5.6	187.3/19.9	223.4/81.03	178.6/35.8	192.7/46.9	329.0/24.1	331.0/38.0	348.9/16.6
Dunn		End	943.3/99.5	300.0/20.9	346.2/37.0	294.6/48.3	991.0/27.2	1000.0/00.0	715.1/29.3	869.4/24.9
NbClust optimal		Duration	782.6/100.7	112.7/31.4	122.8/69.3	116.0/63.6	798.3/53.1	670.0/24.1	384.1/41.4	520.5/24.7
Gap statistic	4	Start	164.3/5.7	166.7/7.0	156.1/8.2	160.3/8.0	259.3/4.7	337.3/11.9	318.4/17.6	355.1/19.2
x-means		End	255.7/3.8	329.3/14.4	291.4/26.7	339.3/16.7	1000.0/00.0	995.9/15.1	699.1/5.6	820.8/17.3
		Duration	91.3/6.9	162.6/18.9	135.3/29.3	179.0/20.9	739.7/4.7	658.6/22.4	380.7/13.2	465.7/31.5
Previous_study	–	Start	190.0	190.0	190.0	190.0	400.0	400.0	400.0	400.0
		End	400.0	400.0	400.0	400.0	1000.0	1000.0	1000.0	1000.0
		Duration	210.0	210.0	210.0	210.0	600.0	600.0	600.0	600.0
Proposed method	6	Start	169.6/7.2	171.1/6.4	160.0/6.1	165.3/4.2	260.2/3.7	334.7/8.7	321.4/2.5	345.4/13.7
Elbow		End	255.2/3.4	327.1/4.7	317.3/11.2	331.8/3.6	1000.0/00.0	979.9/32.8	685.2/3.9	811.3/8.4
Cross-validation		Duration	85.6/9.0	156.1/7.7	157.3/11.6	166.5/6.2	738.2/3.7	645.2/36.1	363.8/5.3	465.9/15.0
Modified *k*-means	6	Start	163.6/2.1	166.3/2.6	157.8/2.9	164.4/2.7	250.2/19.7	341.9/43.8	315.9/4.0	341.5/11.0
		End	254.2/19.1	339.7/43.9	301.2/21.1	326.8/6.4	1000.0/0.0	1000.0/0.0	673.0/55.1	814.3/42.6
		Duration	90.6/21.1	173.5/46.4	143.4/23.9	162.4/9.0	748.8/19.7	657.1/43.8	357.0/57.6	472.8/52.6

We cluster the dataset using the proposed consensus clustering except for modified *k*-means. The proposed time window determination was used to estimate each method's time windows. The reported format is averaged time windows (ms)/SD (ms)

PM prospective memory, *OA* ongoing task, *RS* remitted schizophrenia, *HC* healthy control

Table 4 The statistical power analysis results (mean *p*-value/SD) from 100 runs for the real data

Method	OptNC	N300			Prospective positivity		
		Group	Task	Group × Task	Group	Task	Group × Task
Silhouette	3	0.002/0.002	0.003/0.005	0.063/0.060	0.771/0.121	**0.000/0.000**	0.104/0.077
Dunn							
NbClust optimal							
Gap statistic	4	0.003/0.003	0.003/0.010	**0.019/0.033**	0.445/0.073	**0.000/0.000**	0.083/0.048
x-means							
Previous study	–	0.001/NA	0.050/NA	NS	**0.005**/NA	0.001/NA	**0.010**/NA
Proposed method	6	**0.000/0.001**	**0.002/0.003**	0.048/0.020	0.393/0.046	**0.000/0.000**	0.085/0.025
Elbow							
Cross-validation							
Modified *k*-means	6	0.001/0.001	0.006/0.009	0.280/0.109	0.528/0.075	**0.000/0.000**	0.115/0.058

The boldly marked results indicate superior performance

NA not available, *NS* nonsignificant

Journal : Large 10548 | Article No : 903 | Pages : 19 | MS Code : 903 | Dispatch : 26-5-2022

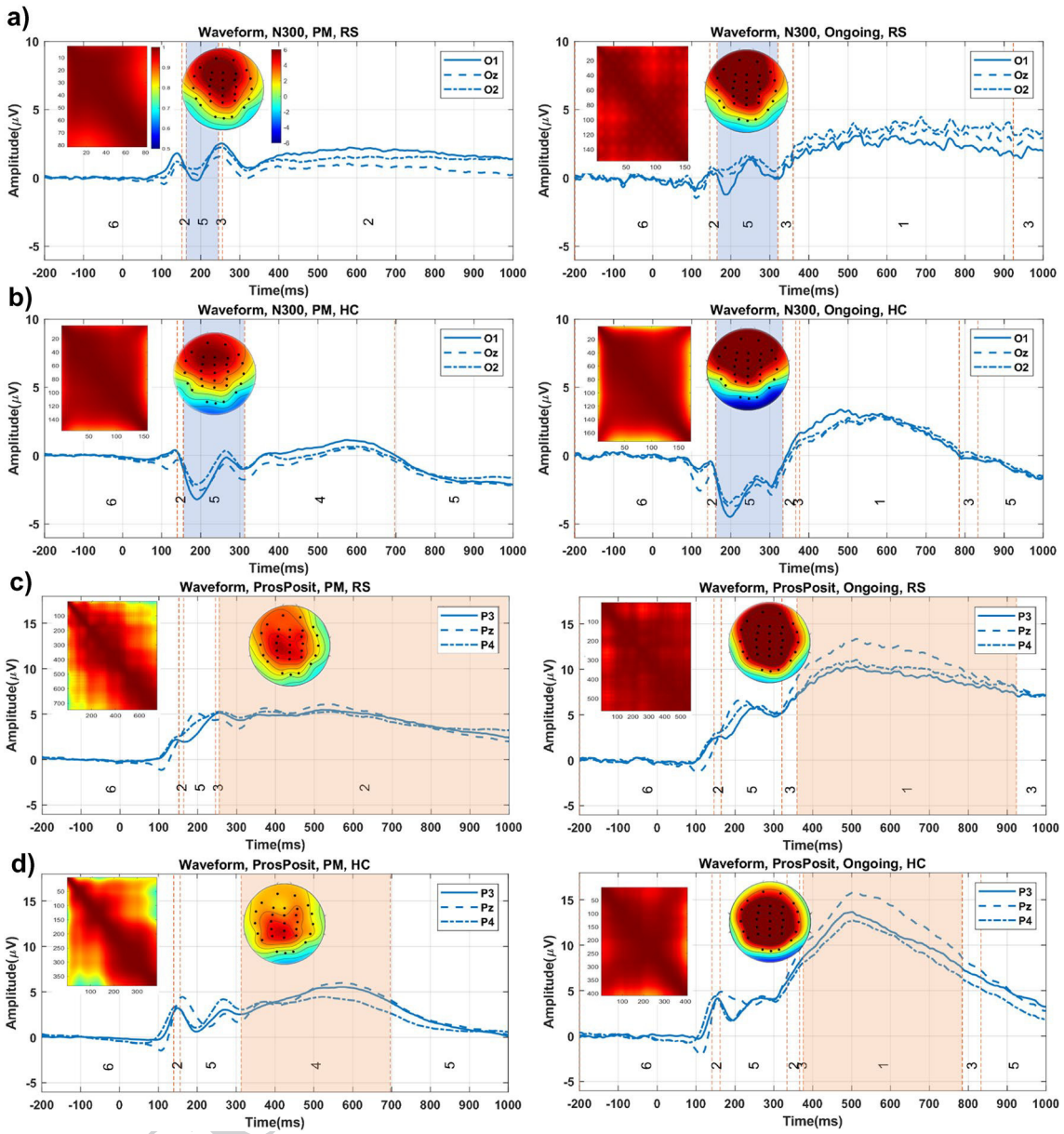


Fig. 9 Consensus clustering and time window determination result in the real data. **a** The identified N300 component by map 5 (colored blue), the corresponding topography maps, and the correlation between time points for both conditions (PM and OA) in the RS group. **b** The identified N300 for the HC group by cluster map 5 (colored pink) for both conditions, the corresponding topography maps, and correlation between the time points. **c** Identified PP component by cluster maps 2 and 1 (colored orange) in the RS group for PM and OA conditions, respectively. **d** Identified prospective positivity by cluster maps 4 and 1 in the HC group for PM and OA conditions, respectively

The proposed method has successfully extended the previous research findings on determining the optimal number of clusters and the ERPs of interest qualification. It is worth mentioning that applying the proposed method to the simulated and real ERP data revealed that the studied standard clustering methods (i.e., polarity dependent) could be combined in a synergistic clustering (consensus clustering). Furthermore, the proposed method is appropriate for either single-trial EEG by considering clustering in a higher resolution (single-trials) or investigating different domains (i.e., frequency, time–frequency) in the future.

Journal : Large 10548 | Article No : 903 | Pages : 19 | MS Code : 903 | Dispatch : 26-5-2022

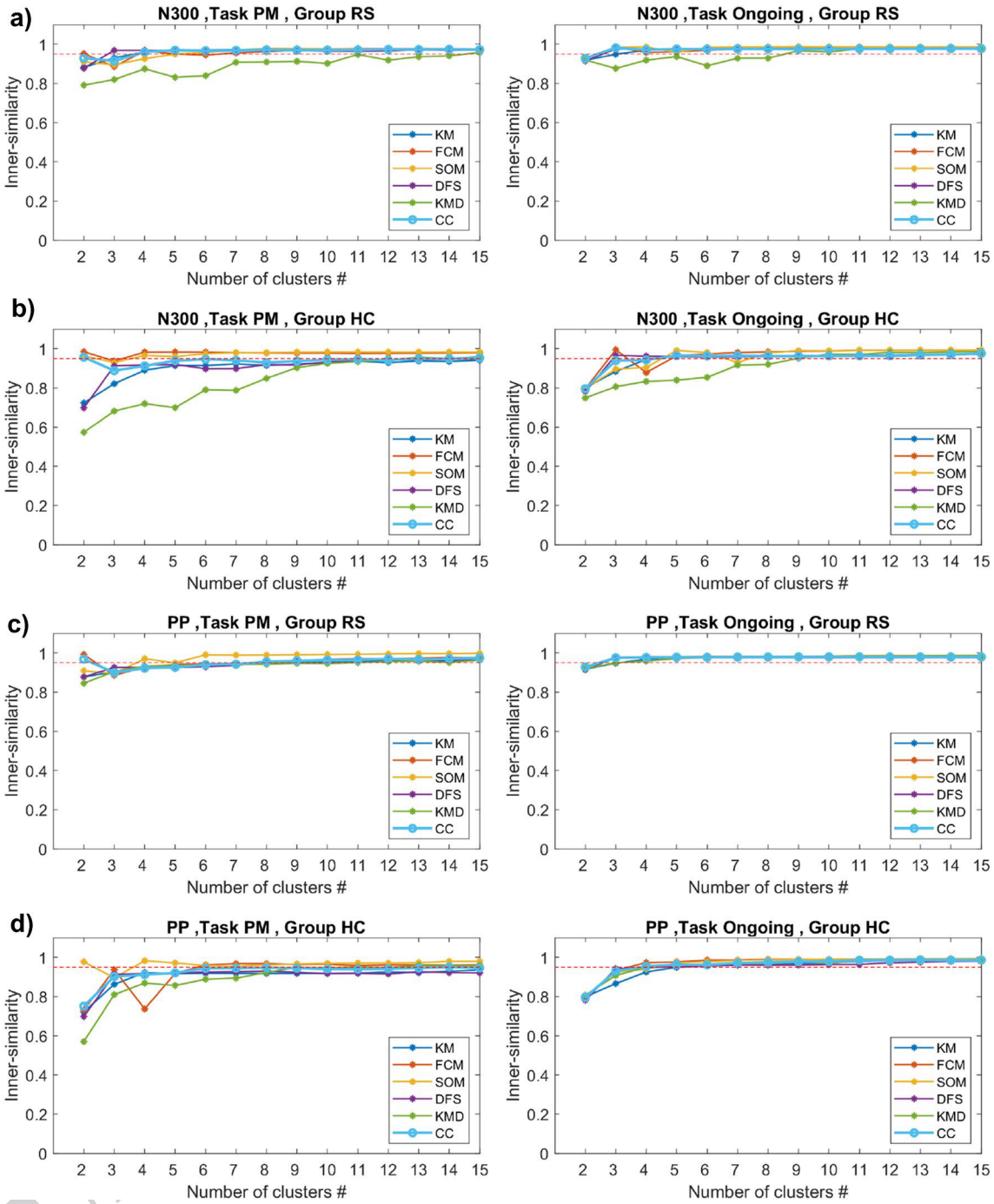


Fig. 10 Mean inner-similarities from the estimated time windows using the studied clustering methods in clustering options (2–15 clusters) and over 100 independent runs on the real ERP data. **a** Results in the RS group and two conditions (PM and Ongoing) for the N300 component. **b** Results for the HC group and both conditions for the N300 component. **c** Results in the RS group and two conditions for the prospective positivity. **d** Results in the HC group for two conditions for the prospective positivity

Journal : Large 10548 | Article No : 903 | Pages : 19 | MS Code : 903 | Dispatch : 26-5-2022

Table 5 The mean measured similarity and SD value between the clusterings in many times (up to 100 times) running of the studied clustering methods for the real ERP data

Clustering method	KM	FCM	SOM	DFS	KMD	CC
KM	–	0.74/0.03	0.72/0.03	0.90/0.05	0.74/0.03	0.82/0.03
FCM	–	–	0.94/0.01	0.71/0.02	0.88/0.0.02	0.89/0.04
SOM	–	–	–	0.70/0.03	0.88/0.02	0.88/0.03
DFS	–	–	–	–	0.72/0.03	0.81/0.04
KMD	–	–	–	–	–	0.86/0.04
CC	–	–	–	–	–	–

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10548-022-00903-2.

Acknowledgements This study would like to remember Prof. Tapani Ristaniemi, who was involved in this study and passed away in 2020, for his great help to all the authors, especially Fengyu Cong, Asoke K. Nandi, Timo Hämäläinen, and Reza Mahini.

Funding This work was supported by National Natural Science Foundation of China (Grant No. 91748105), National Foundation in China (Nos. JCKY2019110B009 & 2020-JCJQ-JJ-252) and the Fundamental Research Funds for the Central Universities [DUT2019] in Dalian University of Technology in China.

Data Availability The datasets analyzed in this article are not publicly available. Requests to access the datasets should be directed to Guoliang Chen.

References

Abu-Jamous B, Fa R, Roberts DJ, Nandi AK (2014) Comprehensive analysis of forty yeast microarray datasets reveals a novel subset of genes (APha-RiB) consistently negatively associated with ribosome biogenesis. BMC Bioinform. https://doi.org/10.1186/1471-2105-15-322

Abu-Jamous B, Fa R, Nandi AK (2015a) Integrative cluster analysis in bioinformatics. Copyright © 2015a Wiley, Chichester. https://doi.org/10.1002/9781118906545

Abu-Jamous B, Fa R, Roberts DJ, Nandi AK (2015b) UNCLES: method for the identification of genes differentially consistently co-expressed in a specific subset of datasets. BMC Bioinform. https://doi.org/10.1186/s12859-015-0614-0

Acharya A, Ghosh J (2011) Cluster ensembles. In: Wiley StatsRef: statistics reference online. pp 1–20. https://doi.org/10.1002/9781118445112.stat08170

Berg P (2006) Dipole simulator (version 3.3. 0.4).

Bezdek JC (1981) Pattern recognition with fuzzy objective function algorithms. https://doi.org/10.1007/978-1-4757-0450-1

Bréchet L, Brunet D, Birot G, Gruetter R, Michel CM, Jorge J (2019) Capturing the spatiotemporal dynamics of self-generated, task-initiated thoughts with EEG and fMRI. NeuroImage 194:82–92. https://doi.org/10.1016/j.neuroimage.2019.03.029

Brunet D, Murray MM, Michel CM (2011) Spatiotemporal analysis of multichannel EEG: CARTOOL. Comput Intell Neurosci. https://doi.org/10.1155/2011/813870

Calhoun VD, Liu J, Adalı T (2009) A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data. NeuroImage 45:S163–S172. https://doi.org/10.1016/j.neuroimage.2008.10.057

Charrad M, Ghazzali N, Boiteau V, Niknafs A (2014) Nbclust: an R package for determining the relevant number of clusters in a data set. J Stat Softw 61:1–36. https://doi.org/10.18637/jss.v061.i06

Chen G et al (2015) Event-related brain potential correlates of prospective memory in symptomatically remitted male patients with schizophrenia. Front Behav Neurosci. https://doi.org/10.3389/fnbeh.2015.00262

Custo A, Van De Ville D, Wells WM, Tomescu MI, Brunet D, Michel CM (2017) Electroencephalographic resting-state networks: source localization of microstates. Brain Connect 7:671–682. https://doi.org/10.1089/brain.2016.0476

De Lucia M, Michel CM, Clarke S, Murray MM (2007) Single-trial topographic analysis of human EEG: a new 'image' of event-related potentials. IEEE. https://doi.org/10.1109/itab.2007.4407353

Delorme A, Makeig S (2004) EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J Neurosci Methods 134:9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B (Methodol) 39:1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

Dinov M, Leech R (2017) Modeling uncertainties in EEG microstates: analysis of real and imagined motor movements using probabilistic clustering-driven training of probabilistic neural networks. Front Hum Neurosci. https://doi.org/10.3389/fnhum.2017.00534

Dunn JC (1974) Well-separated clusters and optimal fuzzy partitions. J Cybern 4:95–104

Fisher RA (1921) On the probable error of a coefficient of correlation deduced from a small sample. Metron 1:3–32

Fukumoto M et al (2014) Relation between remission status and attention in patients with schizophrenia. Psychiatry Clin Neurosci 68:234–241. https://doi.org/10.1111/pcn.12119

Goutte C, Toft P, Rostrup E, Nielsen FÅ, Hansen LK (1999) On clustering fMRI time series. NeuroImage 9:298–310. https://doi.org/10.1006/nimg.1998.0391

Jonnalagadda S, Srinivasan R (2009) NIFTI: an evolutionary approach for finding number of clusters in microarray data. BMC Bioinform. https://doi.org/10.1186/1471-2105-10-40

Kappenman ES, Luck SJ (2012) Manipulation of orthogonal neural systems together in electrophysiological recordings: the MONSTER approach to simultaneous assessment of multiple neurocognitive dimensions. Schizophr Bull 38:92–102. https://doi.org/10.1093/schbul/sbr147

Karypis G, Kumar V (1998) Multilevelk-way partitioning scheme for irregular graphs. J Parallel Distrib Comput 48:96–129. https://doi.org/10.1006/jpdc.1997.1404

Kassambara A (2017) Practical guide to cluster analysis in R: unsupervised machine learning, vol 1. STHDA, (ISBN-10: 1542462703)

Kaufman L, Rousseeuw PJ (2009) Finding groups in data: an introduction to cluster analysis, vol 344. Copyright © 2005 Wiley. https://doi.org/10.1002/9780470316801

Koenig T, Stein M, Grieder M, Kottlow M (2014) A tutorial on data-driven methods for statistically assessing ERP topographies. Brain Topogr 27:72–83. https://doi.org/10.1007/s10548-013-0310-1

Kohonen T (1990) The self-organizing map. Proc IEEE 78:1464–1480. https://doi.org/10.1109/5.58325

Lehmann D (1990) Brain electric microstates and cognition—the atoms of thought. In: John ER, Harmony T, Prichep LS, Valdés-Sosa M, Valdés-Sosa PA (eds) Machinery of the mind: data, theory, and speculations about higher brain function. Birkhauser, Boston

Liu C, Abu-Jamous B, Brattico E, Nandi A (2015) Clustering consistency in neuroimaging data analysis. In: 2015 12th international conference on fuzzy systems and knowledge discovery (FSKD). pp 1118–1122. https://doi.org/10.1109/FSKD.2015.7382099

Liu C, Abu-Jamous B, Brattico E, Nandi AK (2017a) Towards tunable consensus clustering for studying functional brain connectivity during affective processing. Int J Neural Syst 27:1650042. https://doi.org/10.1142/s0129065716500428

Liu C, Brattico E, Abu-jamous B, Pereira CS, Jacobsen T, Nandi AK (2017b) Effect of explicit evaluation on neural connectivity related to listening to unfamiliar music. Front Hum Neurosci. https://doi.org/10.3389/fnhum.2017.00611

Lleti R, Ortiz MC, Sarabia LA, Sanchez MS (2004) Selecting variables for k-means cluster analysis by using a genetic algorithm that optimises the silhouettes. Anal Chim Acta 515:87–100. https://doi.org/10.1016/j.aca.2003.12.020

Luck SJ (2014) An introduction to the event-related potential technique, 2nd edn. MIT Press, Cambridge

Mahini R et al (2020) Determination of the time window of event-related potential using multiple-set consensus clustering. Front Neurosci. https://doi.org/10.3389/fnins.2020.521595

Maris E, Oostenveld R (2007) Nonparametric statistical testing of EEG- and MEG-data. J Neurosci Methods 164:177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024

Meila M (2007) Comparing clusterings—an information based distance. J Multivar Anal 98:873–895. https://doi.org/10.1016/j.jmva.2006.11.013

Micah MM, Lucia MD, Brunet D, Michel CM (2009) Principles of topographic analyses for electrical neuroimaging. MIT Press Scholarship Online. https://doi.org/10.7551/mitpress/9780262013086.003.0002

Michel CM, Koenig T (2018) EEG microstates as a tool for studying the temporal dynamics of whole-brain neuronal networks: a review. NeuroImage 180:577–593. https://doi.org/10.1016/j.neuroimage.2017.11.062

Milligan GW, Cooper MC (1985) An examination of procedures for determining the number of clusters in a data set. Psychometrika 50:159–179. https://doi.org/10.1007/bf02294245

Mishra A, Englitz B, Cohen MX (2019) EEG microstates as a continuous phenomenon. NeuroImage. https://doi.org/10.1016/j.neuroimage.2019.116454

Mur A, Dormido R, Duro N, Dormido-Canto S, Vega J (2016) Determination of the optimal number of clusters using a spectral clustering optimization. Expert Syst Appl 65:304–314. https://doi.org/10.1016/j.eswa.2016.08.059

Murray MM, Brunet D, Michel CM (2008) Topographic ERP analyses: a step-by-step tutorial review. Brain Topogr 20:249–264. https://doi.org/10.1007/s10548-008-0054-5

Ng AY, Jordan MI, Weiss Y (2002) On spectral clustering: analysis and an algorithm. In: Advances in neural information processing systems. pp 849–856

Nguyen N, Caruana R (2007) Consensus clusterings. In: Seventh IEEE international conference on data mining (ICDM 2007), 28–31 October 2007. pp 607–612. https://doi.org/10.1109/ICDM.2007.73

Oostenveld R, Fries P, Maris E, Schoffelen J-M (2011) FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. Comput Intell Neurosci. https://doi.org/10.1155/2011/156869

Park H-S, Jun C-H (2009) A simple and fast algorithm for K-medoids clustering. Expert Syst Appl 36:3336–3341. https://doi.org/10.1016/j.eswa.2008.01.039

Pascual-Marqui RD, Michel CM, Lehmann D (1995) Segmentation of brain electrical activity into microstates: model estimation and validation. IEEE Trans Biomed Eng 42:658–665. https://doi.org/10.1109/10.391164

Pelleg D, Moore AW (2000) X-means: extending k-means with efficient estimation of the number of clusters. In: Icml. pp 727–734

Pena JM, Lozano JA, Larranaga P (1999) An empirical comparison of four initialization methods for the k-means algorithm. Pattern Recogn Lett 20:1027–1040. https://doi.org/10.1016/S0167-8655(99)00069-0

Pourtois G, Delplanque S, Michel C, Vuilleumier P (2008) Beyond conventional event-related brain potential (ERP): exploring the time-course of visual emotion processing using topographic and principal component analyses. Brain Topogr 20:265–277. https://doi.org/10.1007/s10548-008-0053-6

Rand WM (1971) Objective criteria for the evaluation of clustering methods. J Am Stat Assoc 66:846–850

Rousseeuw PJ (1987) Silhouettes—a graphical aid to the interpretation and validation of cluster-analysis. J Comput Appl Math 20:53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Ruggeri P, Meziane HB, Koenig T, Brandner C (2019) A fine-grained time course investigation of brain dynamics during conflict monitoring. Sci Rep. https://doi.org/10.1038/s41598-019-40277-3

Shaw SB, Dhindsa K, Reilly JP, Becker S (2019) Capturing the forest but missing the trees: microstates inadequate for characterizing shorter-scale EEG dynamics. Neural Comput 31:2177–2211. https://doi.org/10.1162/neco_a_01229

Sipola T, Cong F, Ristaniemi T, Alluri V, Toiviainen P, Brattico E, Nandi AK (2013) Diffusion map for clustering fMRI spatial maps extracted by independent component analysis. In: 2013 IEEE international workshop on machine learning for signal processing (MLSP), 22–25 September 2013. pp 1–6. https://doi.org/10.1109/MLSP.2013.6661923

Song Y, Zhang Z, Hu T, Gong X, Nandi AK (2019) Identify of spatial similarity of electroencephalography (EEG) during working-memory maintenance. In: 2019 27th European signal processing conference (EUSIPCO), 2–6 September 2019. pp 1–5. https://doi.org/10.23919/EUSIPCO.2019.8902595

Strehl A, Ghosh J (2003) Cluster ensembles—a knowledge reuse framework for combining multiple partitions. J Mach Learn Res 3:583–617. https://doi.org/10.1162/153244303321897735

Sugar CA, James GM (2003) Finding the number of clusters in a dataset: an information-theoretic approach. J Am Stat Assoc 98:750–763. https://doi.org/10.1198/016214503000000666

Tibshirani R, Walther G (2005) Cluster validation by prediction strength. J Comput Graph Stat 14:511–528. https://doi.org/10.1198/106186005X59243

von Wegner F, Knaut P, Laufs H (2018) EEG microstate sequences from different clustering algorithms are information-theoretically invariant. Front Comput Neurosci. https://doi.org/10.3389/fncom.2018.00070

Zappasodi F et al (2019) EEG microstates distinguish between cognitive components of fluid reasoning. NeuroImage 189:560–573. https://doi.org/10.1016/j.neuroimage.2019.01.067

II

DETERMINATION OF THE TIME WINDOW OF EVENT-RELATED POTENTIAL USING MULTIPLE-SET CONSENSUS CLUSTERING

Reza Mahini, Yansong Li, Weiyan Ding, Asoke K. Nandi, Guoliang Chen, Rao Fu, Tapani Ristaniemi, and Fengyu Cong, 2020

Frontiers in Neuroscience, 14, 521595

https://doi.org/10.3389/fnins.2020.521595

frontiers in Neuroscience

METHODS
published: 21 October 2020
doi: 10.3389/fnins.2020.521595

Determination of the Time Window of Event-Related Potential Using Multiple-Set Consensus Clustering

Reza Mahini[1,2], Yansong Li[3,4], Weiyan Ding[5], Rao Fu[1], Tapani Ristaniemi[2†], Asoke K. Nandi[6], Guoliang Chen[5] and Fengyu Cong[1,2,7,8*]*

[1] School of Biomedical Engineering, Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian, China, [2] Faculty of Information Technology, University of Jyvaskyla, Jyvaskyla, Finland, [3] Reward, Competition and Social Neuroscience Lab, Department of Psychology, School of Social and Behavioral Sciences, Nanjing University, Nanjing, China, [4] Institute for Brain Sciences, Nanjing University, Nanjing, China, [5] Department of Psychiatry, Chinese PLA 967th Hospital, Dalian, China, [6] Department of Electronic and Computer Engineering, Brunel University London, Uxbridge, United Kingdom, [7] School of Artificial Intelligence, Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian, China, [8] Key Laboratory of Integrated Circuit and Biomedical Electronic System, Liaoning Province, Dalian University of Technology, Dalian, China

OPEN ACCESS

Edited by:
Kamran Avanaki, University of Illinois at Chicago, United States

Reviewed by:
Ashkan Faghiri, Georgia State University, United States
Armin Iraji, Georgia State University, United States

***Correspondence:**
Fengyu Cong
cong@dlut.edu.cn
Guoliang Chen
chenguoliangsmmu@126.com

†Deceased

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *19 December 2019*
Accepted: *09 September 2020*
Published: *21 October 2020*

Citation:
Mahini R, Li Y, Ding W, Fu R, Ristaniemi T, Nandi AK, Chen G and Cong F (2020) Determination of the Time Window of Event-Related Potential Using Multiple-Set Consensus Clustering. Front. Neurosci. 14:521595. doi: 10.3389/fnins.2020.521595

Clustering is a promising tool for grouping the sequence of similar time-points aimed to identify the attention blocks in spatiotemporal event-related potentials (ERPs) analysis. It is most likely to elicit the appropriate time window for ERP of interest if a suitable clustering method is applied to spatiotemporal ERP. However, how to reliably estimate a proper time window from entire individual subjects' data is still challenging. In this study, we developed a novel multiset consensus clustering method in which several clustering results of multiple subjects were combined to retrieve the best fitted clustering for all the subjects within a group. Then, the obtained clustering was processed by a newly proposed time-window detection method to determine the most suitable time window for identifying the ERP of interest in each condition/group. Applying the proposed method to the simulated ERP data and real data indicated that the brain responses from the individual subjects can be collected to determine a reliable time window for different conditions/groups. Our results revealed more precise time windows to identify N2 and P3 components in the simulated data compared to the state-of-the-art methods. Additionally, our proposed method achieved more robust performance and outperformed statistical analysis results in the real data for N300 and prospective positivity components. To conclude, the proposed method successfully estimates the time window for ERP of interest by processing the individual data, offering new venues for spatiotemporal ERP processing.

Keywords: multi-set consensus clustering, time window, event-related potentials, microstates analysis, cognitive neuroscience

INTRODUCTION

The event-related potentials (ERPs) carry important information about the cognitive process evoked by the brain response in milliseconds of the temporal domain. Almost all the ERP components are influenced by the attention corresponding to the latencies from the individual and a group of subjects (Luck and Kappenman, 2012). The latencies of ERP components can be

considered as a stable brain electric field configuration (topography map) in milliseconds associated with the specific psychological process (i.e., attention module) (Lehmann, 1990). Moreover, measuring the ERP of interest undertakes a fundamental role in identifying and interpreting the cognitive process in the experiment. The most common approach to measure the magnitude and timing of the ERP of interest is to investigate the amplitude and the latency of peak voltage in the experimentally defined time window. Thereby, an important issue in the analysis of ERPs is how to define or select time windows. This influences both identifying components and performing statistical analyses. Hence, if the time window is not appropriately defined, the comparison between different conditions/groups can lead to unreliable and wrong psychological interpretations (Luck and Gaspelin, 2017).

The traditional ERP approach is to obtain the mean of measured potentials over a fixed and/or experimenter defined time window. The assumption is that the brain electric field configuration is stable for different conditions/groups, although this assumption is not empirically verified. Apart from widely used conventional ERP techniques such as latency peak and mean amplitude, numerous studies have used moving time-window technique and high-resolution time-bin analysis (e.g., each 5 ms) for measuring the peak (Van Overwalle et al., 2009; Mu and Han, 2010; Wills et al., 2014). Although moving time-window or point-by-point analysis in spatiotemporal ERP can provide more fine-grained temporal characterization and significant statistical results (Rotshtein et al., 2010), they can dramatically increase the probability of reporting errors (Luck and Gaspelin, 2017). In the above reviewed methods, the variety of responses, which dynamically influence the duration of time windows in different conditions/groups, are neglected.

Another group of researchers investigated the brain response states by analyzing the topographical changes (Lehmann, 1989, 1990; Lehmann et al., 1994; Micah et al., 2009) to determine the components of interest. The underlying assumption is that the electric field configuration does not change randomly as a function of time, despite exhibiting stability for tens to hundreds of milliseconds involving intervals of topographic instability (Lehmann et al., 1987; Murray et al., 2008). The clustering of spatiotemporal electroencephalogram (EEG)/ERP was used to capture template maps (i.e., topographies found by the clustering) which identifies the recorded signal (Lehmann, 1989, 1990). Hypothetically, the brain state (i.e., the brain electric field configuration) does not change during a specific response time (Lehmann, 1990; Pascual-Marqui et al., 1995; Lehmann et al., 2009). Consequently, the spatial correlation of corresponding topographies of the time-points in the cluster map is close to 1 (Pourtois et al., 2008). Two clustering algorithms in EEG/ERP research, namely, modified *k*-means (Pascual-Marqui et al., 1995) and agglomerate hierarchical clustering (AAHC; Tibshirani and Walther, 2005; Murray et al., 2008) were predominantly used in EEG/ERP researches. Two global measurements together, namely, global field power (GFP) and the global map dissimilarity (GMD), and the global explained variance (GEV) of the template maps (the most important cluster maps), for quantifying the template maps, were applied. Furthermore, the topographical analysis for spatiotemporal ERPs using clustering methods has been explored in several studies (Murray et al., 2008; Micah et al., 2009; Koenig et al., 2014). So far in the aforementioned microstates analysis studies (Michel and Koenig, 2018), determination of template cluster maps with higher explained variance and *post hoc* determination of microstates by fitting those maps to the data (topography maps) were used. As a result, the time-points are clustered based on their similarity in the electrode field configuration. Alternative methods, for cluster or factor analysis, such as optimized *k*-means with genetic algorithm and principal component analysis (PCA) (Williams et al., 2015), topographic pattern analysis, and PCA in high-density ERP (Pourtois et al., 2008) were utilized to determine the most dominant spatial components from the map series. Although independent component analysis and PCA are standard methods and are used for decomposition of the EEG/ERP with cluster analysis, the determination of the event of the interest is subjective instead of being the objective exploration of ERP.

Importantly, finding the suitable time window for measuring the ERP of interest using microstates analysis has also been studied in the numerous literature (Tzovara et al., 2012; Cacioppo et al., 2014; Koenig et al., 2014; Khanna et al., 2015; Mahe et al., 2015). The time window has been determined by testing time-point by time-point, the topographical ANOVA analysis, and microstate classes on momentary grand-mean maps (Koenig et al., 2011). Some recent studies, for example, have explored the most suitable time window from the most fitted microstate maps via the clustering of spatiotemporal ERP by comparing the ERPs of individual subjects with the obtained ERPs of clustering from grand average data (Bailey et al., 2019; Berchio et al., 2019; Ruggeri et al., 2019). Although obtaining global optimal cluster maps by clustering both group (grand average) and individual datasets assigning time-points to template maps is a straightforward solution (Michel and Koenig, 2018), it is challenging to set of template maps from grand average ERP, which reliably represent individual subjects brain responses.

Consensus clustering, as a reliable and stable clustering method, has been successfully used for processing biological data (Monti et al., 2003; Abu-Jamous et al., 2013, 2015a; Liu et al., 2015; Mahini et al., 2017), human brain functional magnetic resonance imaging and EEG data processing (Liu et al., 2017a,b; Song et al., 2019), and multidataset consensus clustering (Filkov and Skiena, 2004; Hoshida et al., 2007; Abu-Jamous et al., 2015b; Liu et al., 2015). However, there has been little discussion about the role of multidataset consensus clustering on individual data from spatiotemporal ERP aimed to identify the ERP components. This is critical because of the difference between the subjects regarding the response time and delay and difference in the quality of recorded data. Therefore, a robust method is required for processing information about the subjects.

The rationale of the current study is to investigate three major points; first, in the ERP experiment, several ERP components are inevitably generated; however, a few of them are targeted, which are more probably elicited if the ERP experiment is run again (Kappenman and Luck, 2012b). Those targeted ERP components are more probably elicited among multiple subjects.

The proposed method isolates reliable time windows for ERP of interest for each condition/group. Second, essentially, even after the well-done preprocessing of the collected data, there are still some remaining interferences and some overlapped brain activity with the ERP of interest in the time domain. Therefore, it is practically expected the time window for measuring the amplitude of the ERP of interest includes information of the ERP. One strategy is to check whether the consecutive multiple topographies of time-points are similar or not. If they are similar enough, they come from the same brain activity of the ERP in terms of the linear transformation model of EEG. Thus, such a time window should be determined. Since the time window contains mostly the ERP of interest, the analysis of the brain response can be more accurate. This can result in a better understanding of cognitive processes. Finally, the ERP signal is elicited from numerous similar responses from the subjects. Defining the ERP of interest from the clustering of grand average data neglects the information about individual subjects. Thereby, the new methodology explores the ERP of interest from individual subjects using a multisubject consensus clustering.

In this article, we develop a stabilized multiple-subject consensus clustering (from the multiset consensus clustering family) approach for reliably clustering spatiotemporal ERP data in both individual subjects and group levels. This can provide a novel mechanism to explore the cognitive functions in ERP/EEG data. Furthermore, we use a newly proposed time-window determination method to obtain the most suitable time window for a given ERP of interest. We do expect the new methodology can retrieve the consistent response among the subjects in a group to discover a reliable time window for the ERP of interest. To assess the efficiency and reliability of our method, the proposed method is applied to simulated and the prospective memory experiment data (Chen et al., 2015). The proposed method has been tested to identify two state-of-the-art ERPs, namely, N2 and P3 components in simulated data, and isolating N300 and prospective positivity components in the real data.

MATERIALS AND METHODS

This section describes first two ERP datasets including conducted simulated data and real data. Then, our proposed method is described in detail. Finally, two classes of statistical analysis for assessing the studied methods are explained.

ERP Studies

Simulated ERP Data

We conducted a simulated ERP data using the BESA dipole simulator[1] for assessing the performance of the studied clustering methods aimed to identify the predefined ERP components. Entirely, six components (i.e., P1, N1, P2, N2, P3, and N4) and two conditions (i.e., "Cond1" and "Cond2") from a group of 20 subjects were defined. A simulated scalp with 65 electrodes was used for representing the spatial (i.e., topographic) information. Each trial was epoched from 100-ms prestimulus to 600-ms poststimulus at a sampling rate of 429 Hz. The averaged reference method was used for referencing. The topography maps of the components and corresponding waveforms are

[1]https://www.besa.de/products/besa-simulator/besa-simulator-overview/

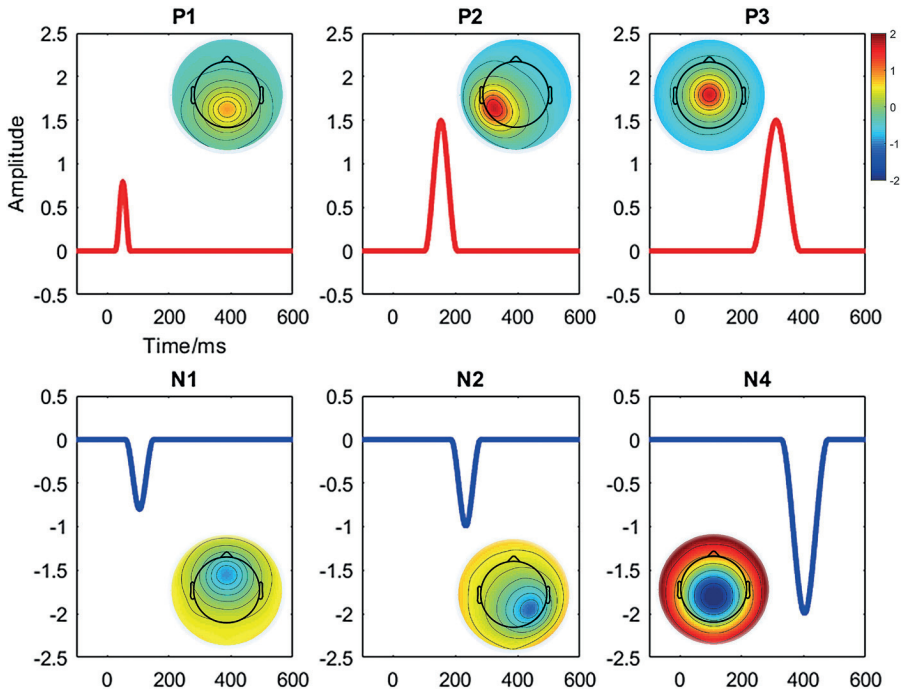


FIGURE 1 | Illustration of topography maps and waveforms for the defined six components (i.e., P1, N1, P2, N2, P3, and N4) in simulated ERP data.

shown in **Figure 1**. Among the defined components, we studied N2 referring to the maximum negative voltage in 201- to 265-ms poststimulus [i.e., it was defined in 175–292 ms via simulator (**Figure 1**)]. The time window was calculated using the signed area measurement method (Sawaki et al., 2012). Similarly, P3 component refers to the positive response (266–357 ms) poststimulus (i.e., defined in 240–385 ms according to **Figure 1**). Meanwhile, the signal was manipulated using the MATLAB function *awgn* (i.e., adding white Gaussian noise) to add a reasonable noise (i.e., signal-to-noise ratio = 20 dB) on signal power measured for each simulated dataset as a whole. Furthermore, random movement of two ERPs (e.g., changing the original signal by randomly increasing/decreasing maximum five time-points) was applied to the original signal from the 20 individuals' data. The electrode sites for measuring statistical amplitude power differences were defined as P6/PO4 and CPz/Cz for N2 and P3, respectively.

Real ERP Data

The prospective memory experiment (Chen et al., 2015) data were used as real ERP data to assess the performance of the proposed method. Following the prior study, the experiment data included 20 symptomatically remitted patients, i.e., with schizophrenia (RS) and 20 healthy control (HC) participants. Two tasks, namely, prospective memory (PM) and ongoing task, were investigated. The EEG data were recorded with 32 electrodes (SynAmps amplifier, NeuroScan) and epoched from 200-ms prestimulus to 1,000-ms poststimulus. Furthermore, a 30 Hz (24 dB/octave) digital low-pass filter was applied. Two target ERP components, N300 and prospective positivity components, were studied. The N300 referred to the maximum negative voltage, over the occipital region, hypothetically between 190 and 400 ms, and the prospective positivity represented the maximum positive voltage, over the parietal region, and between 400 and 1,000 ms.

Proposed Method

The graphical explanation of the proposed method is illustrated in **Figure 2**. Besides, Procedure 1 and Procedure 2 are presented for a better representation of the new methodology. Noteworthy to mention that we have employed a mechanism to obtain the optimal number of clusters by, first, running the consensus clustering many times followed by determining the optimal number of clusters based on the quality of obtaining time windows (Mahini et al., 2019). The details of the proposed method are given as follows:

Procedure 1: Proposed Method
Inputs: ERP data, ERPs of interest (experimental intervals)
Outputs: Time windows
Procedure {
Step 1: Temporal concatenating datasets for each individual subject;
Step 2: Stabilization and generation;
Step 3: Multilevel consensus clustering
Individual level consensus clustering;
Group level consensus clustering;
FOR each ERP of interest
Step 4. Time-window determination;
End of FOR
} End of Procedure

Dataset for Clustering

The collected multiple data points by a high-dense EEG sensor array consist of the spatial topographies of brain activities (i.e., each time-point corresponds to a topography). We have investigated the spatiotemporal ERP data where the time-points are clustered based on their topographical similarity. For each subject, a larger dataset was yielded from temporal concatenating (Murray et al., 2008; Calhoun et al., 2009) the associated datasets from all conditions together. For example, given a subject's ERP data from 300 time-points, 2 conditions, and 65 electrodes, the temporal concatenated dataset with a dimension of 600 × 65 is used for clustering. Therefore, the samples for clustering individual data are the time-points, and the features are represented by the topography (i.e., the electrode field configuration). The goal of clustering is to find the consecutive time-points sharing similar topographies in which the neural responses remain stable for periods of time called time window.

Stabilization and Generation

We utilized the cluster-based similarity partitioning algorithm method (Karypis and Kumar, 1998; Nguyen and Caruana, 2007) as the consensus function based on pairwise similarity measurement between partitions. This function was used for each level of consensus clustering and the stabilization step of the proposed method. Before the generation step, two important issues, consensus clustering configuration and stabilized generation, are necessary to be investigated. Several clustering methods were considered for selecting the appropriate configuration of consensus clustering. Hence, *k*-means (Pascual-Marqui et al., 1995; Pena et al., 1999) and hierarchical clustering (Tibshirani and Walther, 2005) with correlation similarity function, fuzzy *c*-means (FCM; Bezdek, 1981), self-organizing maps (SOMs; Kohonen, 1990), diffusion map spectral clustering (Sipola et al., 2013) consisting *k*-means with Euclidean similarity, and modified *k*-means (Pascual-Marqui et al., 1995), and AAHC (Murray et al., 2008) using spatial correlation, were used for the generation purpose. Thereby, for appropriate consensus clustering configuration, modified *k*-means was used as a benchmark [i.e., the accepted clustering method in many studies (Michel and Koenig, 2018)] to be compared with other studied clustering methods. The clustering methods with higher mutual similarities with modified *k*-means in the majority of clustering results of individuals data (e.g., ≥50% of the subjects), were selected using in the generation phase. Rand index (Strehl and Ghosh, 2003; Meila, 2007) was used to measure the mutual similarity between the results of each clustering method on individual data and modified *k*-means. Rand index can be calculated using the following equation:

$$\mathcal{R}\left(L, L^{'}\right) = \frac{N_{11} + N_{00}}{n\,(n-1)/2} \quad (1)$$

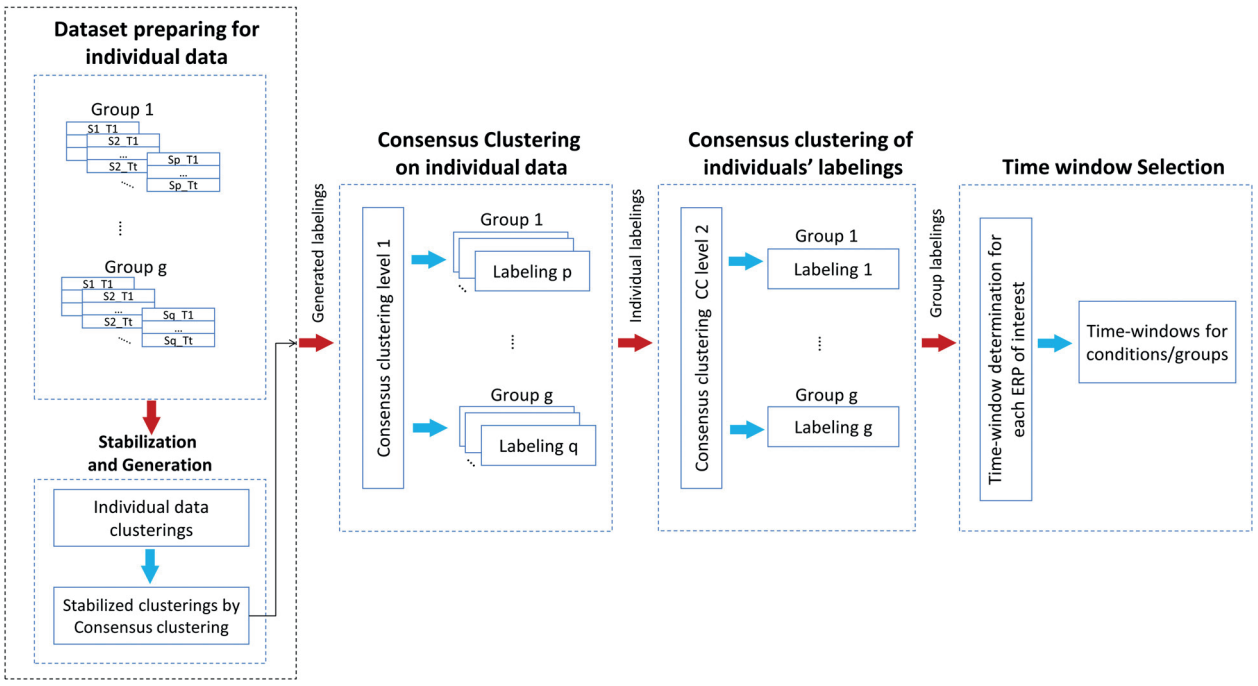


FIGURE 2 | Illustration of the steps of the proposed multisubject consensus clustering for processing the ERPs of interest. The scheme of the proposed method on *g* groups of subjects is demonstrated. S = subject, T = task.

where n denotes the number of observations and N_{00} denotes the number of object pairs in different clusters from both L and L' clusterings. While N_{11} denotes the number of object pairs in the same clusters in L and L'.

Additionally, a stabilization procedure based on consensus clustering was designed for the clustering generation of consensus clustering (at the subject level). The stable clustering refers to the clustering results in which the mutual similarity between two or more clustering results is closed to 1 in theory. To measure stability, a mechanism based on the testing similarity of two clustering results was utilized. If they are highly similar, the clustering method is robust. The consensus clustering of grand average ERP data from multiple runs of each stochastic clustering method (e.g., from 2 to 20 repeats that can be changed if necessary) was employed to find the appropriate number of repetitions to get stable clustering. The optimal number of repetitions should satisfy the following two conditions:

$$\max(|R_r - R_{r-1}|, |R_r - R_{r+1}|) \leq \varepsilon \tag{2}$$

where

$$R_r = \mathcal{R}\left(L^{*-r}, L^{*-(r-1)}\right)$$

$$L^{*-r} = \arg\max_{L \in \mathbb{L}_{\bar{X}}} \sum_{r=2}^{Mr} \Gamma\left(L_r\right)$$

and

$$\min(R_{r-1}, R_r, R_{r+1}) \geq \tau \tag{3}$$

where Γ denotes the consensus function, L^{*-r} denotes the consensus clustering results from r repetitive results (i.e., maximum repeats denotes by Mr) of stochastic clustering method, which is indicated by L_r, and $\bar{X}$ denotes the grand average from the individual datasets. Furthermore, R_r denotes the mutual similarity between the consensus clustering results from r and $r-1$ repetitions. Thus, a proper number of repetitions is determined by measuring the mutual similarity among the results of consensus clustering. In other words, the optimal repetition option is selected when the mutual similarity between $r-1$ and r, and between r and $r+1$ reaches a suitable similarity threshold (e.g., $\tau \geq 90$), and the change among mutual similarities tends to very small values (e.g., $\varepsilon \leq 0.03$).

Multilevel Consensus Clustering

A two-level consensus clustering was utilized for finding the best fitted clustering from individual subjects. The proposed two-level multisubject consensus clustering is explained by the following notations:

Let $S = \{S_1, S_2, \ldots, S_p\}$ denotes a set of subjects from a group, and $X = \{x_1, x_2, \ldots, x_n\}$ denotes a set of time-points for individual data, in which each time-point $x_s = \{e_1, e_2, \ldots, e_f\}$, $s = 1, 2, \ldots, n$ (f denotes the number of electrodes) is a vector of features/channels (i.e., it can be represented in the spatial dimension as a topography map). Besides, $L_i^j = \{C_{1,i}^j, C_{2,i}^j, \ldots, C_{k,i}^j\}$ represents the clustering results for jth clustering method $j = 1, 2, \ldots, m$, for ith subject, $i = 1, 2, \ldots, p$ with k number of clusters. Thus, $C_{w,i}^j$ is defined as wth cluster, $w = 1, 2, \ldots, k$ from jth method for ith subject. The

result of the first-level clustering for each of individual datasets is denoted as:

$$L_i^{*-opt} = \arg\max_{L\in\mathbb{L}_X} \sum_{j=1}^{m} \Gamma\left(L_j^i\right) \quad (4)$$

where, L_i^{*-opt} denotes the consensus clustering results of ith subject from all possible k-partitions on X. At the second level, another consensus clustering is used on the first level clustering results across the subjects (i.e., in the group level), which is defined as:

$$L^{**-opt} = \arg\max_{L\in\mathbb{L}_S} \sum_{i=1}^{p} \Gamma\left(L_i^{*-opt}\right) \quad (5)$$

where, L^{**-opt} denotes the result of consensus clustering across the subjects.

Taken as a whole, the optimal ensemble clustering across the subjects can be noted by:

$$L^{**-opt} = \arg\max_{L\in\mathbb{L}_{X,S}} \sum_{i=1}^{p}\sum_{j=1}^{m} \Gamma\left(L_j^i\right) \quad (6)$$

To provide a better sense of implementation of the proposed method, the multisubject consensus clustering was implemented in MATLAB platform, as demonstrated in **Figure 2** and Procedure 1.

Time Window Determination

The time window determination procedure explores the measurement time window by analyzing the temporal and spatial characteristics of the result cluster maps. The inner-similarity of the candidate cluster map (the maps in the experimental measurement area) and their overlapping with the defined experimental time interval, were considered to estimate the proper time windows. First, the inner-similarity of candidate maps is calculated aimed to detect those with the consecutive time-points with a high spatial correlation. The inner-similarity of a cluster map is the mean of correlation coefficients between topography maps of each two different time-points. More in detail, to calculate the inner-similarity of a cluster map, first, the spatial correlation coefficient (Murray et al., 2008; Micah et al., 2009) of time-points was calculated. Therefore $Cor_{v,u}$ denotes the correlation coefficient between the topographical maps of u and v as two time-points in the cluster map. Then, for each row, the distance matrix can be calculated as:

$$D_v = d\left(Cor_{v,u}, Cor_{v,v}\right), \quad u \neq v \quad (7)$$

where, D denotes the distance matrix in which each row is calculated by the distance between each element in the row and $Cor_{v,v}$ (i.e., self-correlation) in correlation matrix (Cor). To variance-stabilizing transformation of the calculated correlation, fisher z-transform (Fisher, 1921) was used for each vector D_v (i.e., every row of distance matrix) before calculating the mean of the distance matrix D_{avg}. Finally, an inverse z-transform of D_{avg} was used for calculating inner-similarity as shown below:

$$InnSim = 1 - D_{\text{avg}} \quad (8)$$

Hypothetically, in the ERP component, the spatial correlation between the time-points is close to 1 indicating consecutive time-points that represent a cognitive process. Therefore, among the candidate cluster maps, the cluster maps with higher inner similarity than the threshold (e.g., ≥0.90) were selected for overlap testing. We have selected a realistic choice of 0.9 as a satisfactory threshold for time-window qualification. Next, among those cluster maps, the cluster map with the greatest inner-similarity and overlapping was selected as the best suitable cluster map for representing the time window [i.e., via the properties (start, end, and duration)]. More details for implementing the time-window selection method are presented in Procedure 2.

Procedure 2: Time-Window Determination
Input: Clustering result, ERPs of interest (experimental intervals)
Output: Time windows
Procedure
Step 1. Detecting the candidate cluster maps;
FOR each candidate map
Step 2. Calculating inner-similarity and overlapping;
Step 3. Detecting cluster maps with high inner-similarity;
Step 4. Selecting higher overlapping within maps;
End of FOR
} End of Procedure

Statistical Analysis

Two classes of p values based statistical measurements were used to evaluate the performance of the proposed method. First, two one-sided tests (TOST; Rogers et al., 1993; Harms and Lakens, 2018) was performed on simulated data to test the similarity between ground truth and estimated time windows by measuring the obtained time-window properties (start, end, and duration). Second, a statistical power analysis was used by employing repeated measures ANOVA for both simulated and real data. Further, for testing the robustness of those methods, the statistical analysis results were calculated on over 50 independent runs of the studied methods. Overall, we tried to assess the meaningfulness, accuracy, and robustness of the proposed methodology.

The TOST test was accomplished by setting equivalence margin $[-\delta\ \delta]$ in $[-5\ 5]$ ms (can vary depending on the dataset and quality of discriminability). Two composite null hypotheses tested the assumption of the differences: $H0_1: (\mu_1 - \mu_2) \leq -\delta$ and $H0_2: (\mu_1 - \mu_2) \geq \delta$, where μ_1, μ_2 are the mean of each series in the comparison (e.g., the estimated start points from all the individual subjects in a group and corresponding ground truth start points). When both null hypotheses can be statistically rejected, it can be concluded that the observed effect falls within the equivalence margins and practically equivalent (Seaman and Serlin, 1998). In other words, the difference between the mean of the estimated values and the corresponding ground truth values should not exceed the equivalence margins. Furthermore, a repeated-measures ANOVA for the simulated data with the within-subject factor: task ("Cond1" and "Cond2") was considered for statistically analyzing N2 component in the

electrode sites: P6/PO4 for N2 and CPz/Cz for P3. The test was applied to the mean amplitude of N2 and P3 in the estimated time windows separately. Similarly, the statistical power analysis for real data was carried out via repeated measures ANOVA (i.e., mixed 2 × 2) with the addition of a between-subject factor: group (RS and HC) and the within-subject factor: task (PM and ongoing). The test was applied to the mean amplitude of N300 and prospective positivity. The selection of electrodes was based on prior ERP findings (Chen et al., 2015). Specifically, the amplitude of N300 over the occipital region (electrodes: O1/Oz/O2) and prospective positivity over the parietal region (electrodes: P3/Pz/P4) were measured. Statistical comparisons were made at p values of $p < 0.05$ for both data.

RESULTS

To achieve the appropriate clustering result, several important parameters were adjusted, (i) determination of the optimal number of clusters: following our previous study (Mahini et al., 2019), the appropriate number of clusters for simulated and real data was determined in five and six cluster maps, respectively. (ii) The configuration of the proposed consensus clustering: among the studied clustering methods (addressed in "Stabilization and Generation"), k-means, hierarchical clustering, AAHC, and modified k-means methods were applied to the simulated data. Similarly, k-means, FCM, SOMs, diffusion map spectral clustering, AAHC, and modified k-means methods were selected for the clustering of real data (**Table 1**). (iii) Generating stabilized clustering from stochastic clustering methods: following (section "Stabilization and Generation") the optimal repeat for modified k-means and standard k-means was obtained in five and seven repeats for the simulated data (**Figure 3**). Likewise, those clustering methods met stability in seven repetitions in real data. Furthermore, a realistic inner-similarity threshold (e.g., ≥0.90) and a sufficient number of time-points for selecting the candidate cluster maps, e.g., a minimum of 60 to 100 ms (Grieder et al., 2016; Koenig and Brandeis, 2016) were determined.

Results of Simulated ERP Data

We applied the proposed consensus clustering in the simulated data aimed to illustrate all the predefined ERP components. The clustering in seven cluster maps successfully isolated all predefined six components (**Figure 4**) P1, N1, P2, N2, P3, and N4 correspond with the cluster maps 3, 5, 6, 1, 7, and 2, respectively. Note that cluster map 4 refers to the brain state before stimulus onset and does not present any predefined ERP component.

TABLE 1 | The illustration of the clustering method selection by calculating the similarity of the results with the modified k-means method for individual data.

Data	Group	KMS	HC	FCM	SOM	DSC	AAHC
Simulated data	**G1**	**19**	**14**	0	0	0	**20**
Real data	**RS**	**19**	9	**17**	**17**	**15**	**20**
	HC	**19**	11	**19**	**19**	**15**	**18**

The marked methods with bold font are selected where they achieved higher similarity (rand index) for the majority of individual data (e.g., ≥50% of subjects and similarity ≥ 0.7). RS, remitted schizophrenia; HC, healthy control; G1, simulated group; KMS, k-means; HC, hierarchical clustering; FCM, fuzzy c-means; SOM, self-organizing map; DSC, diffusion maps spectral clustering; and AAHC, atomize and agglomerate hierarchical clustering.

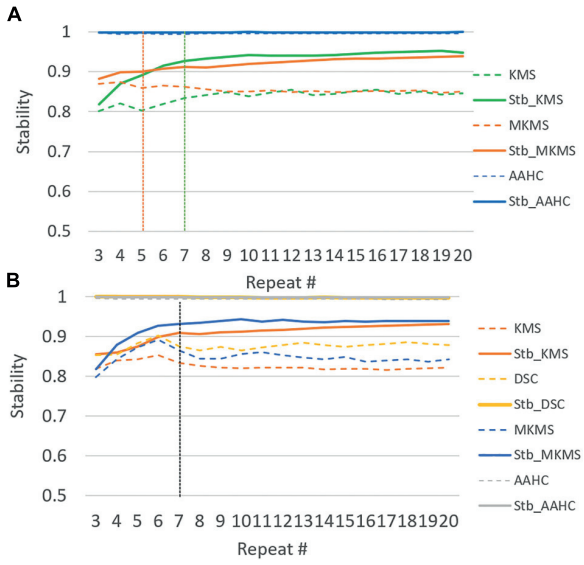


FIGURE 3 | The illustration of the stability test in 20 runs of studied clustering methods to the grand average ERP data. Dash lines demonstrate the original clustering method stability and the continuous lines illustrate the corresponding stabilized version for each studied clustering method behavior for the range of repetition (e.g., from 2 to 20). **(A)** Stabilizing in the simulated data and **(B)** stabilizing in the real data.

Time Windows and Topographies for ERPs of Interest

Figure 5 illustrates the clustering results and the elicited N2 and P3 components (from one random execution), including the corresponded topography maps and the spatial correlation of time-points obtained by the proposed method on the simulated data. **Figure 5A** indicates that the N2 component in Cond1 and Cond2 are elicited by cluster maps 5 (marked blue). Likewise, **Figure 5B** illustrates that the P3 component is identified by the microstate map 1 (marked orange) in both conditions. These results reveal that a significant main effect of task ($p < 0.0001$) was identified in N2 in the duration of microstate maps. Similarly, a significant main effect of task ($p < 0.0001$) was detected in the P3 component. For both components, the measured amplitudes were greater in Cond2. This reveals that the N2 and P3 components seem to be distinctly elicited by the proposed method in the simulated data.

Comparison Between Estimated and Ground Truth Time Windows

The proposed method was compared with the state-of-the-art clustering methods, namely, modified k-means and AAHC, in spatiotemporal ERP clustering. Our time-window selection method was applied to the clustering results (i.e., proposed consensus clustering, modified k-means, and AAHC results) to identify each ERP of interest. The Start, End, and Duration

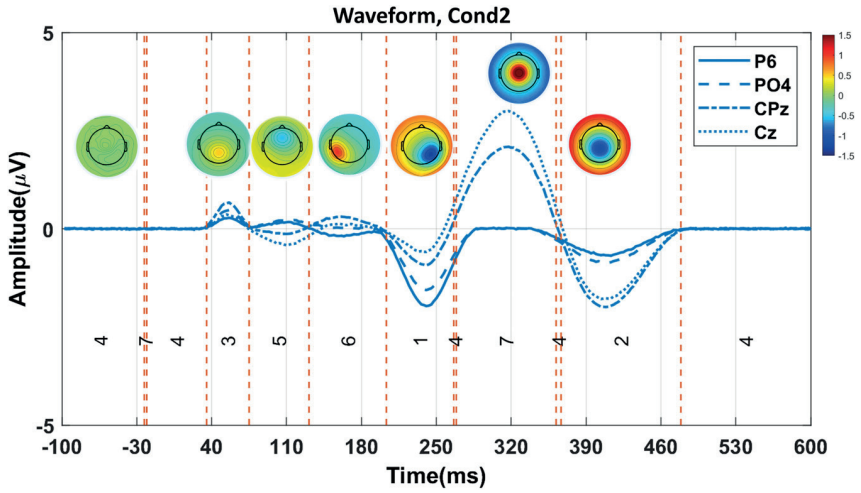


FIGURE 4 | The proposed consensus clustering results in the simulated Cond2 data. Red and blue colors indicate positive and negative potential values, respectively. Each topography map represents a cluster map. P1, N1, P2, N2, P3, and N4 components are presented with cluster maps 3, 5, 6, 1, 7, and 2, respectively. Note that cluster map 4 does not show any component in the simulated data.

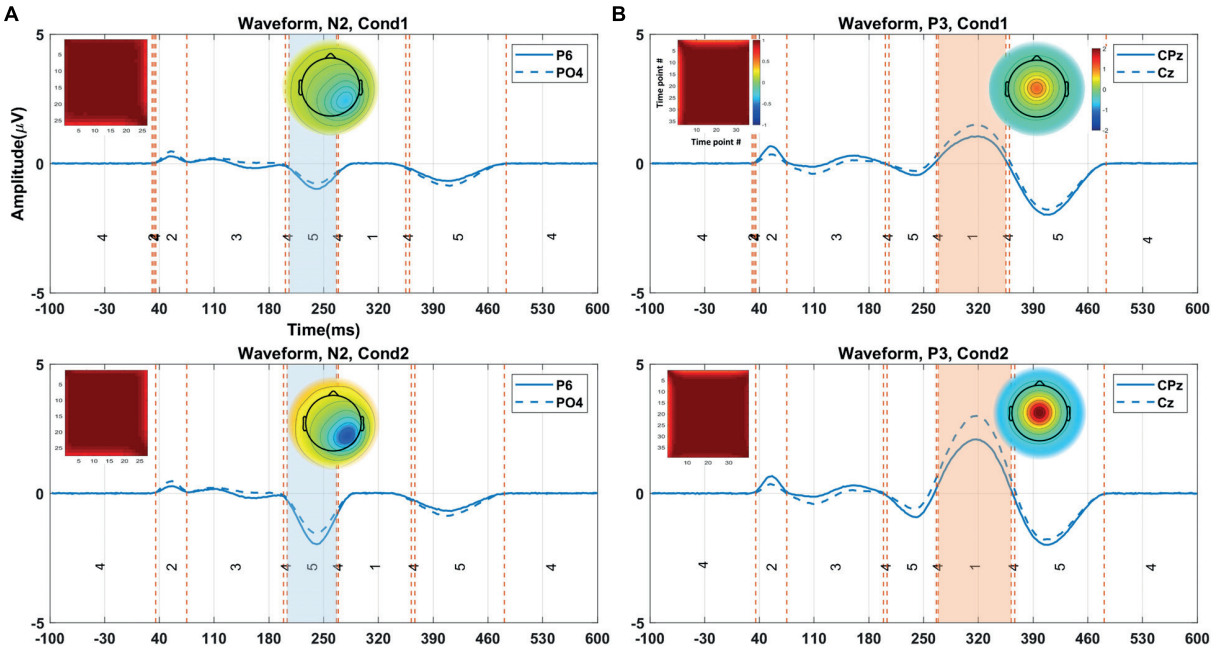


FIGURE 5 | The illustration of the clustering results (shown in the grand average data) for eliciting N2 and P3, including the corresponded topography maps and the spatial correlation of time-points, obtained by the proposed method on the simulated data. **(A)** Selected time window identified by cluster map 5 (i.e., blue area from 205 to 264 ms) for N2 in Cond1 **(upper panel)**. The selected time window by cluster map 5 (i.e., orange area from 203 to 264 ms) for N2 in Cond2 **(lower panel)**. **(B)** Selected time window with cluster map 1 (i.e., colored area from 264 to 350 ms) for P3 in Cond 1 **(upper panel)**. Selected time window identified by cluster map 1 (i.e., colored area from 268 to 357 ms), the topographic map for P3 in Cond 2 **(lower panel)**. The range of the color bars is equally associated with the plot sections. Cond1 = condition 1, Cond2 = condition 2.

parameters of estimated time windows were compared with that of the ground truth time windows (obtained from the simulation) on the clustering results of individual data for testing the accuracy. The TOST result (**Table 2**) for N2 component from clustering methods illustrates that the null hypothesis was rejected for the proposed method, modified *k*-means, and AAHC for all parameters in both conditions except End in Cond1 for AAHC. Similarly, the null hypothesis was rejected in all parameters except Duration in both conditions for the proposed method. It was, however, not rejected in either of the criteria in P3 for modified *k*-means and AAHC. Taken as a whole, the proposed method achieved a more precise estimation of time

TABLE 2 | Descriptive two one-sided tests (TOST) equivalence tests between ground truth TWs (time windows) and estimated TWs by the proposed consensus clustering (CC), modified *k*-means (MKMS), and atomize and agglomerate hierarchical clustering (AAHC) in individual subjects' data from simulated ERP data.

Comp-Meth	Cond	Criteria	*p1*	*p2*	DiffMu (ms)	EQ_interval (ms)	
N2_CC	C1	Start	0.000	0.003	2.2	0.4	4.1
		End	0.000	0.003	2.6	1.0	4.1
		Duration	0.000	0.000	0.4	−1.6	2.3
	C2	Start	0.000	0.001	1.6	−0.4	3.6
		End	0.000	0.002	2.3	0.7	4.0
		Duration	0.000	0.000	0.7	−1.2	2.6
P3_CC	C1	Start	0.000	0.000	1.9	0.5	3.2
		End	0.018	0.000	2.8	−4.7	−0.9
		Duration	**0.380**	0.000	4.7	−6.8	−2.6
	C2	Start	0.000	0.000	1.8	0.3	3.2
		End	0.008	0.000	2.6	−4.4	−0.7
		Duration	**0.266**	0.000	4.3	−6.4	−2.2
N2_MKMS	C1	Start	0.000	0.001	2.1	0.3	3.9
		End	0.000	0.041	3.5	1.9	5.1
		Duration	0.000	0.000	1.4	−0.3	3.1
	C2	Start	0.000	0.001	1.4	−0.6	3.4
		End	0.000	0.030	3.5	2.0	5.0
		Duration	0.000	0.001	2.1	0.5	3.7
P3_MKMS	C1	Start	0.000	**0.543**	5.1	2.8	7.5
		End	**0.548**	0.000	5.1	−7.3	−3.0
		Duration	**0.997**	0.000	10.3	−13.8	−6.7
	C2	Start	0.000	**0.554**	5.1	3.2	7.0
		End	**0.654**	0.000	5.5	−7.8	−3.1
		Duration	**0.999**	0.000	10.6	−14.0	−7.3
N2_AAHC	C1	Start	0.000	0.001	1.8	0.0	3.5
		End	0.000	**0.104**	4.1	2.7	5.5
		Duration	0.000	0.000	2.3	0.9	3.7
	C2	Start	0.000	0.000	1.3	−0.6	3.2
		End	0.000	0.039	3.6	2.1	5.1
		Duration	0.000	0.001	2.3	0.8	3.9
P3_AAHC	C1	Start	0.000	**0.162**	4.2	2.7	5.7
		End	**0.244**	0.000	4.3	−6.2	−2.4
		Duration	**0.999**	0.000	8.5	−10.6	−6.4
	C2	Start	0.000	**0.167**	4.2	2.6	5.8
		End	**0.276**	0.000	4.4	−6.3	−2.6
		Duration	**0.999**	0.000	8.6	−10.7	−6.6

Bold marked represent nonsignificant results. Comp-Meth, component of interest and the method; Cond, condition; C1, condition 1; C2, condition 2; p1, p value of lower bound; p2, p value of upper bound; DiffMu, difference of mean of two sets; and EQ_interval, confident equivalence interval.

windows in individual data. Moreover, for a better sense of comparison between studied clustering methods, the accuracy of estimated (i.e., based on Start and End parameters) time windows for the subjects is exhibited in **Figures 6**, **7**. It is observable that the consensus clustering method outperforms modified *k*-means and AAHC in terms of accuracy of estimation, especially in P3 component.

Results of Real ERP Data

Time Windows and Topographies for ERPs of Interest

The clustering results (randomly selected) from running the proposed method on real data for N300 and prospective positivity components, the corresponding topography maps, and the spatial correlation of time-points are illustrated in **Figure 8**. N300 identified by the cluster maps 1 and 2 in the RS group, is illustrated by the colored area in **Figure 8A** for both PM and ongoing tasks. Furthermore, N300 identified by cluster map 1 in the HC group and two tasks (PM and ongoing), is illustrated in **Figure 8B**. Similarly, the prospective positivity component is isolated by the cluster maps 6 and 5 in the RS group for PM and ongoing tasks, respectively (**Figure 8C**). The identified prospective positivity by cluster maps 4 and 5 in the HC group for PM and ongoing tasks are illustrated, respectively (**Figure 8D**). The average topographies shown in **Figure 8** are obtained from the selected time windows identified by the cluster maps. Hence, the statistical power analysis revealed that HC was

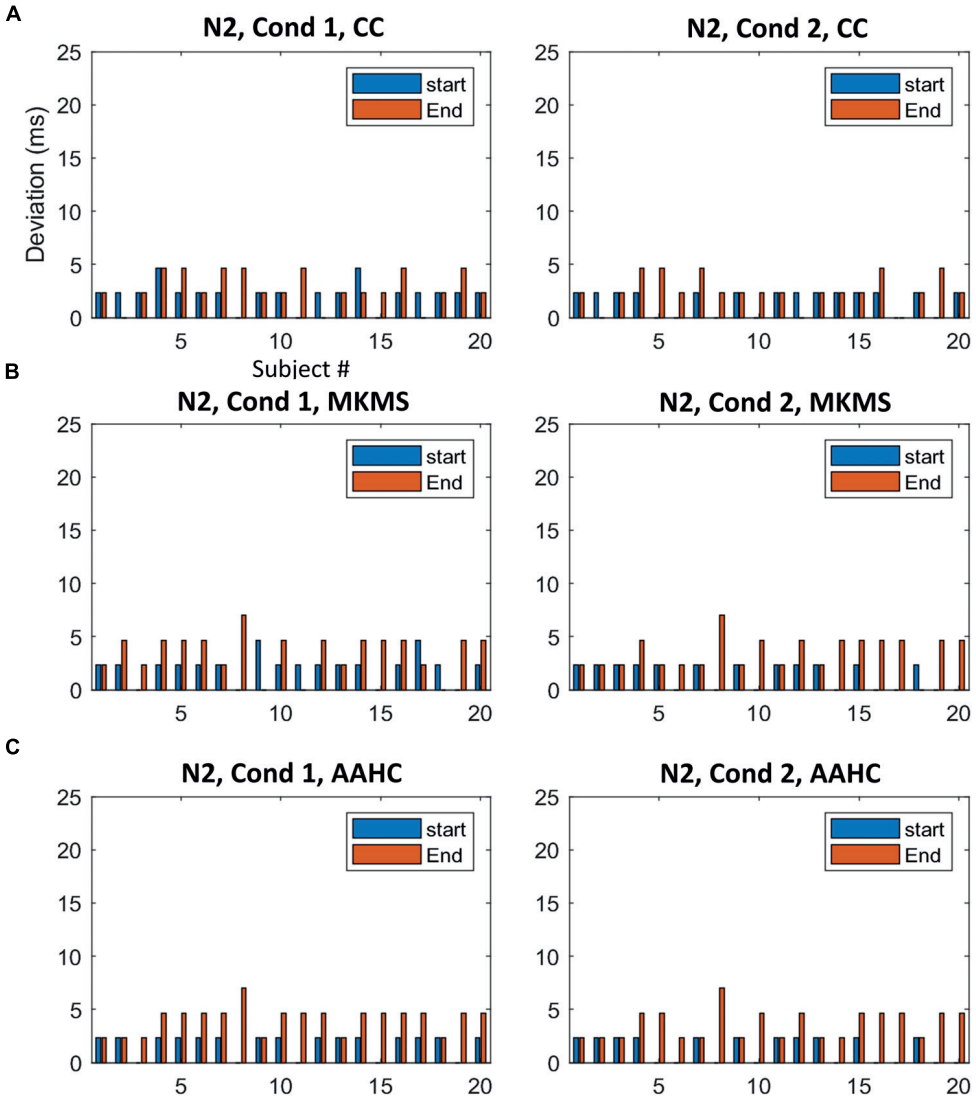


FIGURE 6 | The difference between the estimated time windows from corresponding ground truth time windows (i.e., based on start and end criteria) for individual data in N2. **(A)** The estimation results of the proposed method. **(B)** The estimation results by MKMS method. **(C)** The estimation results by AAHC method. MKMS, modified *k*-means; AAHC, atomize, and agglomerate hierarchical clustering.

characterized by a more negative potential over the occipital-central electrodes ($p < 0.001$). Additionally, a silently larger positive potential was localized over frontal-central electrodes compared to the RS group in N300. Moreover, a slightly more negative potential was observed over occipital-central electrodes ($p < 0.001$) in the ongoing task from both RS and HC groups in the N300 component. Our results revealed no significant difference for prospective positivity regarding group factor; however, a larger positive potential was localized over central electrodes ($p < 0.0001$) in the ongoing task comparing to the PM task.

Statistical Analysis and Stability Test Results

The mean p value and standard deviation (SD) were obtained from over 50 independent runs of the studied clustering methods and statistical analysis on the individual data (**Table 3**). Rendering to stability analysis, the proposed method (SD = 0.003) was more stable compared to modified *k*-means (SD = 0.006) for the main effect of group and less stable than AAHC (SD = 0.002) for N300 component. Interestingly, it was the most stable method compared to other studied clustering methods for both the main effect of task (SD = 0.002) and interaction between group and task (SD = 0.043). Besides, the statistical power analysis results showed that the main effects of group and task by the proposed method were significant ($p < 0.002$ for both factors). Likewise, the main effect of group was significant by the modified *k*-means ($p < 0.017$) and AAHC ($p < 0.004$). The main effect of task, however, was significant only via AAHC ($p < 0.013$). Meanwhile, the interaction between group and task was not significant in

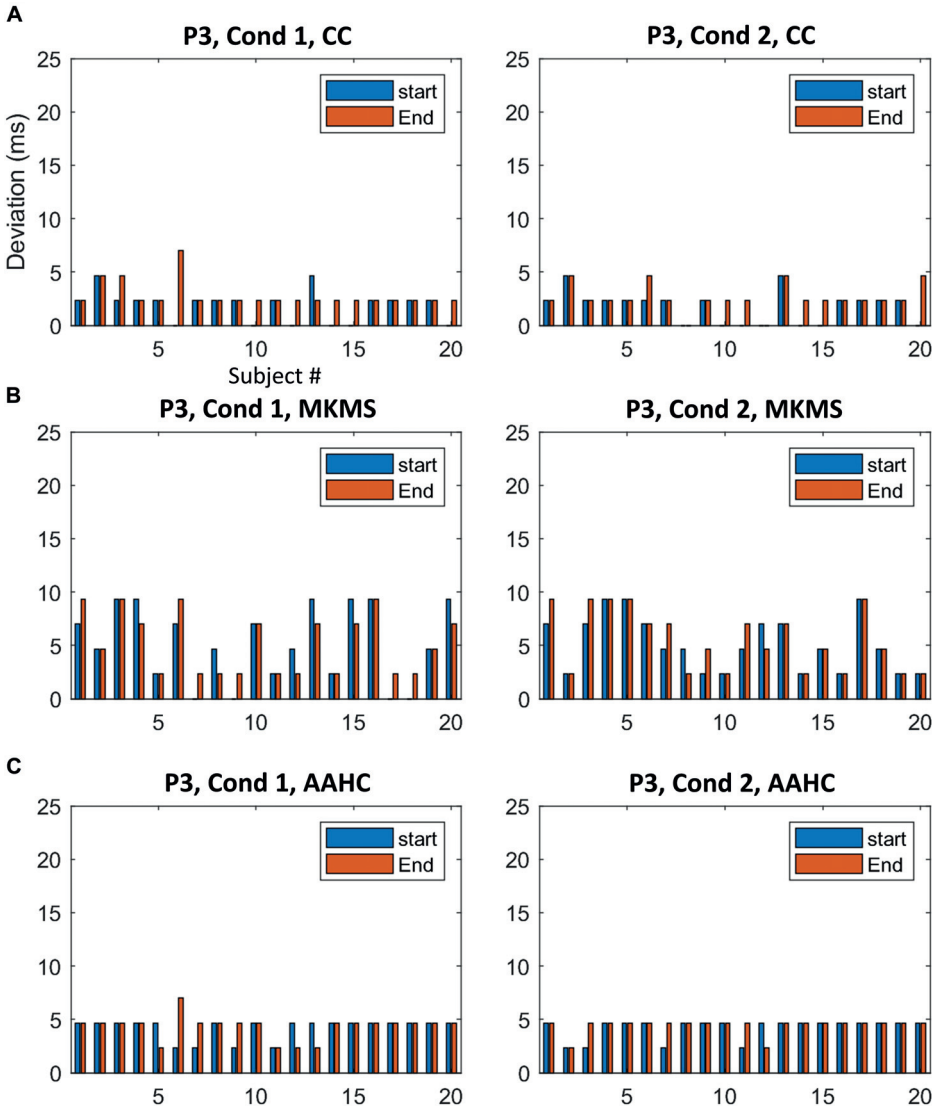


FIGURE 7 | The difference between the estimated time windows from corresponding ground truth time windows for individual data in P3. **(A)** The results from the proposed method. **(B)** The estimation results by MKMS method. **(C)** The estimation results by AAHC method.

both modified *k*-means and AAHC. Similarly, the proposed method was statistically the most stable for the interaction between group and task (SD = 0.011) comparing to other studied clustering methods in prospective positivity (**Table 3**). Additionally, the main effect of task was significant ($p < 0.0001$), and, more importantly, the interaction between group and task was also significant ($p < 0.007$) by the proposed method. However, the main effect of group was not significant by the proposed method. The main effect of task was also significant by modified *k*-means ($p < 0.0001$) and AAHC ($p < 0.0001$), whereas, the main effect in group, and the interactions between group and task were not significant by both modified *k*-means and AAHC methods.

DISCUSSION

This study proposed a new methodology based on multisubject consensus clustering on spatiotemporal ERP data for the suitable time-window determination. To this end, we designed the stabilized multisubject consensus clustering in two levels described as follows: (i) subject resolution in which the stabilized consensus clustering was used to combine the results of various clusterings on each subject's data in the group; (ii) group resolution in which the most suitable clustering for each group was obtained by consensus clustering of the clustering results of individual data. From the ERP technique point of view, the researchers using the ERP technique for the cognitive

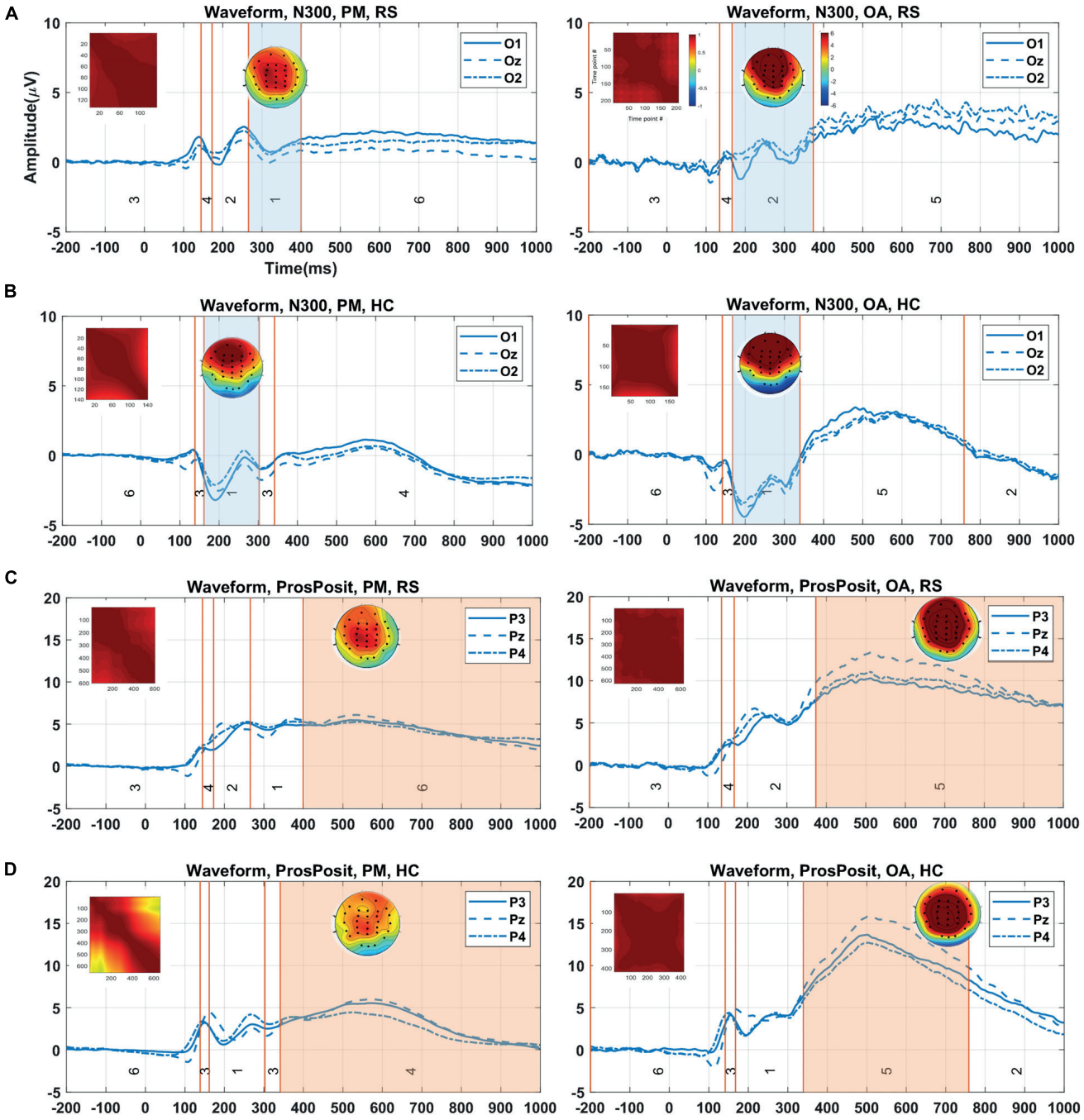


FIGURE 8 | Demonstration of the clustering result (showed in the grand average data), identified time window for ERP of interest, and corresponding topography map and spatial correlation of the time-points in each group/condition via the proposed method. **(A)** Identified time window by cluster maps 1 and 2 (i.e., colored areas) for two tasks (PM and ongoing) for N300 in RS group. **(B)** Selected time windows identified with map 1 for both conditions in the HC group. **(C)** The isolated time windows by cluster maps 6 and 5 for the tasks (PM and ongoing) in RS group. (**D**) Equally, the time windows identified by cluster maps 4 and 5 for the tasks (PM and ongoing) in the HC group. The visual comparison between two groups in panels **(A,B)** for N300 and in panels **(C**,**D)** for prospective positivity shows the difference in the waveforms in the selected time windows. The color bars are equally associated with the plot sections. PM, prospective memory; OA, ongoing task; Pros.Pos, prospective positivity; RS, remitted schizophrenia; HC, healthy control.

neuroscience research often face up the challenge to determine a time window for an ERP, since the most popular textbook of ERP recommends the readers averaging the amplitudes in the time window as the measurement of the ERP peak amplitude (Luck, 2014). In terms of previous publications, we found that the determination of such a time window has mostly

TABLE 3 | Mean *p* value and standard deviation (SD) calculations of statistical power analysis results in over 50 runs of study clustering methods on the individual data for the real data.

	N300			Pros.Pos.		
Method	**Group**	**Task**	**intGrTsk**	**Group**	**Task**	**intGrTsk**
Proposed (*p* value)	**0.002**	**0.002**	0.058	0.590	**0.000**	**0.007**
SD	0.003	**0.002**	**0.043**	0.227	0.000	**0.011**
MKMS(*p* value)	**0.017**	0.101	0.303	0.614	**0.000**	0.150
SD	0.006	0.075	0.225	0.199	0.000	0.156
AAHC (*p* value)	**0.004**	**0.013**	0.145	0.662	**0.000**	0.246
SD	**0.002**	0.009	0.133	0.201	0.000	0.131

IntGrTsk, interaction between group and task; Pros.Pos., prospective positivity. Bold marked represent significant results.

relied on the visual inspection, which can be subjective and bring bias to conclusions and difficulty for the readers to repeat the experiment. Therefore, the main objective of this work was to provide a reliable clustering-based mechanism (objective approach) for studying the temporal dynamic and sensory information about the subjects (i.e., brain responses). This was accomplished with the multilevel clustering mechanism and the time-window determination method. The clustering result from entire subjects entails important information about group response which is critical for studying the cognitive processes in ERP.

One issue in processing individual data is, apart from the need for sufficient trials for obtaining reliable ERP (Boudewyn et al., 2018) and the variety of brain responses in the trials, the variability associated with individual subjects' brain responses, which is observable when ERPs are used to assess cognitive functions. The underlying assumption is that the variety in the trials and subject responses are involved in ERP, although in the ERP techniques, the assumption is that the ERP is phase-locked and time-locked. Therefore, each subject grants value to the statistical test in terms of differences between conditions or groups, which is through the variance across subjects assisting in the ability to detect a significant experimental effect (Kappenman and Luck, 2012a). Yet, in the literature, the individual responses were mostly addressed by fitting the cluster maps of individual data to the cluster maps of group average data (Murray et al., 2008; Koenig et al., 2014; Michel and Koenig, 2018; Berchio et al., 2019; Ruggeri et al., 2019). To cover this gap, we strived to cluster individual subject data in the first level and map the entire individual clusterings into a group as the ultimate clustering.

From the cluster analysis view of point, the various clustering strategies such as using the single clustering method on the different types of datasets; repeated clustering with a single clustering method and combining the results; and the multiple-clustering methods applied to the individual dataset potentially affect the clustering quality (Abu-Jamous et al., 2013, 2015b; Liu et al., 2015; von Wegner et al., 2018). To investigate this issue and reliably feeding consensus clustering, two data-driven based mechanisms were appropriated before multilevel cluster analysis. First, consensus clustering configuration was performed aim to find the appropriate clustering methods. This was recognized by calculating the similarity between candidate clustering methods and modified *k*-means (benchmark) from individual data. Second, the stabilized clusterings were carried out by stabilizing the stochastic clusterings. Taken as a whole, these two procedures can make an additional sense of obtaining reliable and stable results instead of using a single clustering method or the conventional consensus clustering platform. Noteworthy to mention that clustering selection and stabilization can result in different configurations for various ERP data.

In accordance with the obtained results, two major differences were noticed between the proposed method and conventional clustering methods:

(i) The statistical test in this study revealed that the proposed method estimates a more precise time windows for individual subjects in comparison with the other conventional clustering methods in simulated data for both ERPs of interest (N2 and P3). The foremost reason is that our method uses the strength of multiple clustering methods and data-driven processing individual subject data to fit the suitable time windows for each condition/group, despite with using spatial consistency comparison between ERPs of individual and grand average data (Habermann et al., 2018; Michel and Koenig, 2018; Berchio et al., 2019).

(ii) According to the statistical analysis results (**Table 3**), the proposed method outperformed other benchmark methods regarding achieving more stability in the real data. Over 50 independent runs of the clustering on the same datasets, the estimation of the proposed method was with a much smaller variance. This indicates that the estimation of the ERP time window was much closer to the ground truth time window of the ERP, in contrast to the other methods. Such results from the real data also correspond to the ones from the simulation data, i.e., the estimation of the time window of an ERP was more accurate by the proposed method. Therefore, the results of the current study, based on analyzing the brain dynamics from the stimuli onset to the brain response, successfully explored the attention effect on the neural responses from the subjects in real ERP data.

The drawback of the proposed method, however, is that if the real ERP component is still embedded in ERP waveforms the determination of the time window of an ERP component cannot be precise. Indeed, this also happens in the visual inspection method to determine the time window of an ERP. Therefore, in order to determine the time window of an ERP component more precisely, the EEG preprocessing is very critical. The better the preprocessing is, the more precise and objective determination of the time window of an EPR is carried out in terms of the proposed method.

The results of analyzing the brain dynamic responses revealed that the brain electrical dynamics in obtained time windows were comparatively different in time-window properties (start, end, and duration) for different conditions/groups. Therefore,

from the clinical point of view, the brain responses from two groups (RS and HC) to the stimuli onset were investigated to identify N300 and prospective positivity components. This can be interpreted as the fact of the variety of brain response for the subjects in different condition/group. In N300 component isolation, for example, the difference was shown in cluster maps 1 (i.e., between RS and HC groups) in PM tasks. Likewise, the duration differed in cluster maps 1 and 2 in the ongoing task between the groups. Again, at the source level, a silently larger negative response was observed in ongoing than PM task in both RS and HC groups. These results demonstrate that RS patients with schizophrenia showed a functional recovery of PM cue detection during the event-based PM task. Consequently, the electrophysiological data revealed the ability of symptomatically remitted patients with schizophrenia to distinguish the PM task from the ongoing task. This was reflected by the significant main effect of task type among these two groups. As a result, this finding showed a complementary viewpoint to the prior studies (Fukumoto et al., 2014; Chen et al., 2015). Our results can be employed for interpreting the advantage of the treatment in RS patients in terms of measuring/identifying the difference in ERPs of interest in the observations. Therefore, this may indicate a degree of functional recovery of preparatory attentional processes that helps the processing of PM task in these subjects (RS patients) during clinical remission. Thus, providing further evidence for the recent researches demonstrating symptomatic remission in schizophrenia is associated with a degree of functional recovery of attentional processes.

CONCLUSION AND FUTURE WORKS

This work presents a multisubject consensus clustering technique to explore spatiotemporal ERP by extracting group-level information from individual responses. Our proposed methodology has successfully extended the previous research findings (Murray et al., 2008; Koenig et al., 2014; Michel and Koenig, 2018) of cluster analysis of EEG/ERP. Noteworthy to mention that we have proposed the multiset consensus clustering method in the present study which can work better for the group-level analysis. Since the proposed method is not limited to just ERP data, it is very interesting to apply the proposed method on other brain imaging modalities for investigating the various types of brain dynamics. Furthermore, the proposed method can also be used as an appropriate tool to analyze the single-trial EEG by considering suitable roles for the trials in higher resolution (single-trials) in the future. Taken together, this work emphasizes that, in the time-window determination from spatiotemporal ERP, the temporal dynamics can be extremely influenced through the measurement interval. It is noteworthy that this methodology can be investigated on different levels (i.e., groups, subjects, trials). The current study also highlights that the obtained time windows are sensitive to the responses from the subjects, which can provide a better sense of understanding in information processing of the neural responses. In order to show the effectiveness of the proposed method, we have used the simulated ERP dataset and the real ERP dataset. Indeed, the selection of the real ERP dataset does not mean that the proposed method only works for such an attention-related ERP experiment. The proposed method has no limitation on the experiment types of ERPs. Thereby, a toolbox has been developed under the MATLAB platform, named ERP_CC.[2] Taken as a whole, we can rely on the information retrieved by the new method, which reflects the attention mechanism regarding the response to the stimuli in the real data. We therefore believe that the EEG neuroimaging method can be studied by the proposed methodology in various dimensions to accomplish useful results in cognitive neuroscience studies.

DATA AVAILABILITY STATEMENT

The datasets analyzed in this article are not publicly available. Requests to access the datasets should be directed to GC.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the ethics committee of PLA general hospital. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

RM designed the methodology, implemented the algorithms, performed data analysis, produced tables and figures, and wrote the manuscript. YL performed statistical analysis, collected real data, and analyzed the results. WD collected the real data. RF prepared the simulated data. AN designed the methodology and contributed to the final manuscript. GC designed the study (real data), collected real data, and analyzed the results. FY designed the study, analyzed the results, and wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by National Natural Science Foundation of China (Grant No. 91748105), National Foundation in China (No. JCKY2019110B009) and the Fundamental Research Funds for the Central Universities (DUT2019) in Dalian University of Technology

[2]https://doi.org/10.5281/zenodo.3541576

in China. YL was supported by the National Natural Science Foundation of China (Grant No. 31600929) and the Fundamental Research Funds for the Central Universities (010914380002).

ACKNOWLEDGMENTS

This study is to memorize TR for his great help to all the authors, especially FC, AN, and RM.

REFERENCES

Abu-Jamous, B., Fa, R., and Nandi, A. K. (2015a). *Integrative Cluster Analysis in Bioinformatics. Copyright © 2015*. Hoboken, NJ: John Wiley & Sons, Ltd, doi: 10.1002/9781118906545

Abu-Jamous, B., Fa, R., Roberts, D. J., and Nandi, A. K. (2015b). UNCLES: method for the identification of genes differentially consistently co-expressed in a specific subset of datasets. *BMC Bioinformatics* 16:184. doi: 10.1186/s12859-015-0614-0

Abu-Jamous, B., Fa, R., Roberts, D. J., and Nandi, A. K. (2013). Paradigm of tunable clustering using binarization of consensus partition matrices (Bi-CoPaM) for gene discovery. *PLoS One* 8:e0056432. doi: 10.1371/journal.pone.0056432

Bailey, N. W., Freedman, G., Raj, K., Sullivan, C. M., Rogasch, N. C., Chung, S. W., et al. (2019). Mindfulness meditators show altered distributions of early and late neural activity markers of attention in a response inhibition task. *PLoS One* 14:e0203096. doi: 10.1371/journal.pone.0203096

Berchio, C., Küng, A.-L., Kumar, S., Cordera, P., Dayer, A. G., Aubry, J.-M., et al. (2019). Eye-gaze processing in the broader bipolar phenotype revealed by electrical neuroimaging. *Psychiatry Res.* 291, 42–51. doi: 10.1016/j.pscychresns.2019.07.007

Bezdek, J. C. (1981). *Pattern Recognition With Fuzzy Objective Function Algorithms*. Berlin: Springer, doi: 10.1007/978-1-4757-0450-1

Boudewyn, M. A., Luck, S. J., Farrens, J. L., and Kappenman, E. S. (2018). How many trials does it take to get a significant ERP effect? It depends. *Psychophysiology* 55:e13049. doi: 10.1111/psyp.13049

Cacioppo, S., Weiss, R. M., Runesha, H. B., and Cacioppo, J. T. (2014). Dynamic spatiotemporal brain analyses using high performance electrical neuroimaging: theoretical framework and validation. *J. Neurosci. Methods* 238, 11–34. doi: 10.1016/j.jneumeth.2014.09.009

Calhoun, V. D., Liu, J., and Adalı, T. (2009). A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data. *NeuroImage* 45(Suppl. 1), S163–S172. doi: 10.1016/j.neuroimage.2008.10.057

Chen, G., Zhang, L., Ding, W., Zhou, R., Xu, P., Lu, S., et al. (2015). Event-related brain potential correlates of prospective memory in symptomatically remitted male patients with schizophrenia. *Front. Behav. Neurosci.* 9:262. doi: 10.3389/fnbeh.2015.00262

Filkov, V., and Skiena, S. (2004). Integrating microarray data by consensus clustering. *Int. J. Artif. Intell. Tools* 13, 863–880. doi: 10.1142/s0218213004001867

Fisher, R. A. (1921). On the probable error of a coefficient of correlation deduced from a small sample. *Metron* 1, 3–32.

Fukumoto, M., Hashimoto, R., Ohi, K., Yasuda, Y., Yamamori, H., Umeda-Yano, S., et al. (2014). Relation between remission status and attention in patients with schizophrenia. *Psychiatry Clin. Neurosci.* 68, 234–241. doi: 10.1111/pcn.12119

Grieder, M., Koenig, T., Kinoshita, T., Utsunomiya, K., Wahlund, L.-O., Dierks, T., et al. (2016). Discovering EEG resting state alterations of semantic dementia. *Clin. Neurophysiol.* 127, 2175–2181. doi: 10.1016/j.clinph.2016.01.025

Habermann, M., Weusmann, D., Stein, M., and Koenig, T. (2018). A Student's guide to randomization statistics for multichannel event-related potentials using ragu. *Front. Neurosci.* 12:355. doi: 10.3389/fnins.2018.00355

Harms, C., and Lakens, D. (2018). Making 'null effects' informative: statistical techniques and inferential frameworks. *J. Clin. Transl. Res.* 3(Suppl. 2), 382–393. doi: 10.17605/OSF.IO/WPTJU

Hoshida, Y., Brunet, J.-P., Tamayo, P., Golub, T. R., and Mesirov, J. P. (2007). Subclass mapping: identifying common subtypes in independent disease data sets. *PLoS One* 2:1195. doi: 10.1371/journal.pone.0001195

Kappenman, E. S., and Luck, S. J. (2012a). "ERP components: the ups and downs of brainwave recordings," in *The Oxford Handbook of Event-Related Potential Components*, eds S. J. Luck and E. S. Kappenman (Oxford: Oxford University Press), 3–30. doi: 10.1093/oxfordhb/9780195374148.013.0014

Kappenman, E. S., and Luck, S. J. (2012b). Manipulation of orthogonal neural systems together in electrophysiological recordings: the MONSTER approach to simultaneous assessment of multiple neurocognitive dimensions. *Schizophr. Bull.* 38, 92–102. doi: 10.1093/schbul/sbr147

Karypis, G., and Kumar, V. (1998). Multilevelk-way partitioning scheme for irregular graphs. *J. Parallel Distrib. Comput.* 48, 96–129. doi: 10.1006/jpdc.1997.1404

Khanna, A., Pascual-Leone, A., Michel, C. M., and Farzan, F. (2015). Microstates in resting-state EEG: current status and future directions. *Neurosci. Biobehav. Rev.* 49, 105–113. doi: 10.1016/j.neubiorev.2014.12.010

Koenig, T., and Brandeis, D. (2016). Inappropriate assumptions about EEG state changes and their impact on the quantification of EEG state dynamics. *Neuroimage* 125, 1104–1106. doi: 10.1016/j.neuroimage.2015.06.035

Koenig, T., Kottlow, M., Stein, M., and Melie-Garc, L. (2011). Ragu: a free tool for the analysis of EEG and MEG event-related scalp field data using global randomization statistics %J Intell. *Neuroscience* 2011, 1–14. doi: 10.1155/2011/938925

Koenig, T., Stein, M., Grieder, M., and Kottlow, M. (2014). A tutorial on data-driven methods for statistically assessing ERP topographies. *Brain Topogr.* 27, 72–83. doi: 10.1007/s10548-013-0310-1

Kohonen, T. (1990). The self-organizing map. *Proc. IEEE* 78, 1464–1480. doi: 10.1109/5.58325

Lehmann, D. (1989). "Microstates of the brain in EEG and ERP mapping studies," in *Brain Dynamics*, eds E. Başar and T. H. Bullock (Berlin: Springer), 72–83. doi: 10.1007/978-3-642-74557-7_6

Lehmann, D. (1990). "Brain electric microstates and cognition: the atoms of thought," in *Machinery of the Mind: Data, Theory, and Speculations About Higher Brain Function*, eds E. R. John, T. Harmony, L. S. Prichep, M. Valdés-Sosa, and P. A. Valdés-Sosa (Boston, MA: Birkhäuser Boston), 209–224. doi: 10.1007/978-1-4757-1083-0_10

Lehmann, D., Michel, C. M., Pal, I., and Pascual-marqui, R. D. (1994). Event-related potential maps depend on prestimulus brain electric microstate map. *Int. J. Neurosci.* 74, 239–248. doi: 10.3109/00207459408987242

Lehmann, D., Ozaki, H., and Pal, I. (1987). EEG alpha map series: brain micro-states by space-oriented adaptive segmentation. *Electroencephalogr. Clin. Neurophysiol.* 67, 271–288. doi: 10.1016/0013-4694(87)90025-3

Lehmann, D., Pascual-Marqui, R. D., and Michel, C. (2009). EEG microstates. *Scholarpedia* 4:7632. doi: 10.4249/scholarpedia.7632

Liu, C., Abu-Jamous, B., Brattico, E., and Nandi, A. (2015). "Clustering consistency in neuroimaging data analysis," in *2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, New York, NY, 1118–1122. doi: 10.1109/FSKD.2015.7382099

Liu, C., Abu-Jamous, B., Brattico, E., and Nandi, A. K. (2017a). Towards tunable consensus clustering for studying functional brain connectivity during affective processing. *In. J. Neural Syst.* 27:1650042. doi: 10.1142/s0129065716500428

Liu, C., Brattico, E., Abu-jamous, B., Pereira, C. S., Jacobsen, T., and Nandi, A. K. (2017b). Effect of explicit evaluation on neural connectivity related to listening to unfamiliar music. *Front. Hum. Neurosci.* 11:611. doi: 10.3389/fnhum.2017.00611

Luck, S. J. (2014). *An Introduction to the Event-Related Potential Technique*. Cambridge, MA: MIT press.

Luck, S. J., and Gaspelin, N. (2017). How to get statistically significant effects in any ERP experiment (and why you shouldn't). *Psychophysiology* 54, 146–157. doi: 10.1111/psyp.12639

Luck, S. J., and Kappenman, E. S. (2012). "ERP components and selective attention," in *The Oxford Handbook of Event-Related Potential Components*, eds S. J. Luck and E. S. Kappenman (New York, NY: Oxford University Press), 295–327.

Mahe, G., Zesiger, P., and Laganaro, M. (2015). Beyond the initial 140 ms, lexical decision and reading aloud are different tasks: an ERP study with

topographic analysis. *Neuroimage* 122, 65–72. doi: 10.1016/j.neuroimage.2015.07.080

Mahini, R., Xu, P., Chen, G., Li, Y., Ding, W., Zhang, L., et al. (2019). Optimal number of clusters by measuring similarity among topographies for spatio-temporal ERP analysis. *arXiv* [Preprint], Available online at: https://zenodo.org/badge/latestdoi/197860407

Mahini, R., Zhou, T., Li, P., Nandi, A. K., Li, H., Li, H., et al. (2017). "Cluster Aggregation for analyzing event-related potentials," in *Advances in Neural Networks - ISNN 2017. ISNN 2017. Lecture Notes in Computer Science*, eds F. Cong, A. Leung, and Q. Wei (Cham: Springer International Publishing), 507–515. doi: 10.1007/978-3-319-59081-3_59

Meila, M. (2007). Comparing clusterings - an information based distance. *J. Mult. Anal.* 98, 873–895. doi: 10.1016/j.jmva.2006.11.013

Micah, M. M., Lucia, M. D., Brunet, D., and Michel, C. M. (2009). *Principles of Topographic Analyses for Electrical Neuroimaging*. Cambridge, MA: MIT Press, doi: 10.7551/mitpress/9780262013086.003.0002

Michel, C. M., and Koenig, T. (2018). EEG microstates as a tool for studying the temporal dynamics of whole-brain neuronal networks: a review. *NeuroImage* 180, 577–593. doi: 10.1016/j.neuroimage.2017.11.062

Monti, S., Tamayo, P., Mesirov, J., and Golub, T. (2003). Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Mach. Learn.* 52, 91–118. doi: 10.1023/a:1023949509487

Mu, Y., and Han, S. (2010). Neural oscillations involved in self-referential processing. *Neuroimage* 53, 757–768. doi: 10.1016/j.neuroimage.2010.07.008

Murray, M. M., Brunet, D., and Michel, C. M. (2008). Topographic ERP analyses: a step-by-step tutorial review. *Brain Topogr.* 20, 249–264. doi: 10.1007/s10548-008-0054-5

Nguyen, N., and Caruana, R. (2007). "Consensus clusterings," in *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, Omaha, NE, 607–612. doi: 10.1109/ICDM.2007.73

Pascual-Marqui, R. D., Michel, C. M., and Lehmann, D. (1995). Segmentation of brain electrical activity into microstates: model estimation and validation. *IEEE Trans. Biomed. Eng.* 42, 658–665. doi: 10.1109/10.391164

Pena, J. M., Lozano, J. A., and Larranaga, P. (1999). An empirical comparison of four initialization methods for the k-means algorithm. *Pattern Recogn Lett.* 20, 1027–1040. doi: 10.1016/S0167-8655(99)00069-0

Pourtois, G., Delplanque, S., Michel, C., and Vuilleumier, P. (2008). Beyond conventional event-related brain potential (ERP): exploring the time-course of visual emotion processing using topographic and principal component analyses. *Brain Topogr.* 20, 265–277. doi: 10.1007/s10548-008-0053-6

Rogers, J. L., Howard, K. I., and Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychol. Bull.* 113, 553–565. doi: 10.1037/0033-2909.113.3.553

Rotshtein, P., Richardson, M. P., Winston, J. S., Kiebel, S. J., Vuilleumier, P., Eimer, M., et al. (2010). Amygdala damage affects event-related potentials for fearful faces at specific time windows. *Hum. Brain Mapp.* 31, 1089–1105. doi: 10.1002/hbm.20921

Ruggeri, P., Meziane, H. B., Koenig, T., and Brandner, C. (2019). A fine-grained time course investigation of brain dynamics during conflict monitoring. *Sci. Rep.* 9:3667. doi: 10.1038/s41598-019-40277-3

Sawaki, R., Geng, J. J., and Luck, S. J. (2012). A common neural mechanism for preventing and terminating the allocation of attention. *J. Neurosci.* 32, 10725–10736. doi: 10.1523/JNEUROSCI.1864-12.2012

Seaman, M. A., and Serlin, R. C. (1998). Equivalence confidence intervals for two-group comparisons of means. *Psychol. Methods* 3, 403–411. doi: 10.1037/1082-989X.3.4.403

Sipola, T., Cong, F., Ristaniemi, T., Alluri, V., Toiviainen, P., Brattico, E., et al. (2013). "Diffusion map for clustering fMRI spatial maps extracted by independent component analysis," in *2013 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, Southampton, 1–6. doi: 10.1109/MLSP.2013.6661923

Song, Y., Zhang, Z., Hu, T., Gong, X., and Nandi, A. K. (2019). "Identify of spatial similarity of electroencephalography (EEG) during working-memory maintenance," in *2019 27th European Signal Processing Conference (EUSIPCO)*, A Coruna, 1–5. doi: 10.23919/EUSIPCO.2019.8902595

Strehl, A., and Ghosh, J. (2003). Cluster ensembles- a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.* 3, 583–617. doi: 10.1162/153244303321897735

Tibshirani, R., and Walther, G. (2005). Cluster validation by prediction strength. *J. Comput. Graph. Stat.* 14, 511–528. doi: 10.1198/106186005X59243

Tzovara, A., Murray, M. M., Plomp, G., Herzog, M. H., Michel, C. M., and De Lucia, M. (2012). Decoding stimulus-related information from single-trial EEG responses based on voltage topographies. *Pattern Recogn.* 45, 2109–2122. doi: 10.1016/j.patcog.2011.04.007

Van Overwalle, F., Van den Eede, S., Baetens, K., and Vandekerckhove, M. (2009). Trait inferences in goal-directed behavior: ERP timing and localization under spontaneous and intentional processing. *Soc. Cogn. Affect. Neurosci.* 4, 177–190. doi: 10.1093/scan/nsp003

von Wegner, F., Knaut, P., and Laufs, H. (2018). EEG microstate sequences from different clustering algorithms are information-theoretically invariant. *Front. Comput. Neurosci.* 12:70. doi: 10.3389/fncom.2018.00070

Williams, N., Nasuto, S. J., and Saddy, J. D. (2015). Method for exploratory cluster analysis and visualisation of single-trial ERP ensembles. *J. Neurosci. Method* 250, 22–33. doi: 10.1016/j.jneumeth.2015.02.007

Wills, A. J., Lavric, A., Hemmings, Y., and Surrey, E. (2014). Attention, predictive learning, and the inverse base-rate effect: evidence from event-related potentials. *Neuroimage* 87, 61–71. doi: 10.1016/j.neuroimage.2013.10.060

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

III

BRAIN EVOKED RESPONSE QUALIFICATION USING MULTI-SET CONSENSUS CLUSTERING: TOWARD SINGLE-TRIAL EEG ANALYSIS

by

Reza Mahini, Guanghui Zhang, Tiina Parviainen, Rainer Düsing, Asoke K. Nandi, Fengyu Cong, and Timo Hämäläinen 2023

To be submitted

https://doi.org/10.21203/rs.3.rs-3586574/v1

Brain Evoked Response Qualification Using Multi-Set Consensus Clustering: Toward Single-Trial EEG Analysis

Reza Mahini[1], Guanghui Zhang[2], Tiina Parviainen[4], Rainer Düsing[6], Asoke K. Nandi[5], Fengyu Cong[1,3,7,8], Timo Hämäläinen[1*]

[1]Faculty of Information Technology, University of Jyväskylä, Jyväskylä, Finland.

[2]Center for Mind and Brain, University of California -Davis, Davis, 95618, USA.

[3]School of Biomedical Engineering, Faculty of Electronic and Electrical Engineering, Dalian University of Technology, China.

[4]Department of Psychology, Centre for Interdisciplinary Brain Research, University of Jyväskylä, Jyväskylä, Finland.

[5]Department of Electronic and Electrical Engineering, Brunel University London, Uxbridge UB8 3PH, UK.

[6]Institute of Psychology, Department of Research Methods, Diagnostics and Evaluation, University of Osnabrück, Germany.

[7]School of Artificial Intelligence, Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian,116024, China.

[8]Key Laboratory of Integrated Circuit and Biomedical Electronic System, Dalian University of Technology, Dalian, 116024, China.

Abstract

Objective: Scalp electroencephalogram (EEG) provides a substantial amount of data about information processing in the human brain. In the context of conventional event-related potential (ERP) analysis, it is typically assumed that individual trials for one subject share similar properties and stem from comparable neural sources. However, group-level ERP analysis methods (including cluster analysis) can miss important information about the relevant neural process due to a rough estimation of the brain activities of individual subjects while selecting a fixed time window for all the subjects.

Method: We designed a multi-set consensus clustering method to examine cognitive processes at the individual subject level. First, consensus clustering from diverse clustering methods was applied to single-trial EEG epochs of individual subjects. Next, the second level of consensus clustering was applied across the trials of each subject. Afterward, a modified time window determination is applied to identify the ERP of interest of individual subjects.

Results: The proposed method was applied to real EEG data from the active visual oddball task experiment to qualify the P3 component. Our findings disclosed that the estimated time windows for individual

* Corresponding author's addresse: timo.t.hamalainen@jyu.fi, Tel: +358-407726470 (T. Hämäläinen).

subjects can provide more precise ERP identification than considering a fixed time window for all subjects. Moreover, based on standardized measurement error and established bootstrap for single-trial EEG, our assessments revealed suitable stability in the calculated scores for the identified P3 component.

Significance: The new method provides a realistic and information-driven understanding of the single trials' contribution towards identifying the ERP of interest in the individual subjects.

Keywords: Single-Trial EEG, Time window, Multi-Set consensus clustering, Standardization, EEG microstates, Cognitive process.

1. Introduction

Electroencephalogram (EEG) is a non-invasive neuroimaging technique to record electrophysiological brain activity from multiple electrodes placed on the scalp. For decades, investigating a group-level EEG has been the golden standard for testing different experimental hypotheses. Moreover, the researchers have expected to qualify brain responses from the individual subjects and the trials, especially for the clinical experiments. However, due to the complexity and high noise level of raw EEG data, averaging EEG trials called the event-related potential (ERP) technique has been used to identify the ERP components, which, in turn, are associated with specific perceptual, motor, or cognitive processes. Averaging is justified with the assumption that single-trial EEG signals represent similar properties of the cognitive process in question, which can be identified from ERP. Although the ERP technique is popular because of the high signal-to-noise ratio (SNR), simplicity of statistical analysis, and further interpretation of the brain information processing reflected by different ERP components, it does not have full access to possibly valuable and meaningful information available at individual trials (Cohen & Cavanagh, 2011; Delorme et al., 2002). Additionally, studying the variability of single trials (Knuth et al., 2006) is of great importance in, e.g., clinical studies due to inhomogeneity among the individual subjects (i.e., the latencies of ERP components can be unidentical for control and patient groups). On the other hand, ERP identifies the time-locked response to stimulus-onset, reducing the physiological and recording noise contributions that are usually not time-locked.

Some sophisticated statistical techniques have been used to explore the ERP components from single-trial EEG, such as independent component analysis (ICA; Makeig et al., 1997), principal component analysis (PCA; Schölkopf et al., 1998), electrode-wise time-frequency analysis (TFA; Herrmann et al., 2014), and spectrum power analysis (Cong et al., 2015). Moreover, a vast majority of ERP studies have used ICA/PCA to extract shared ERP components from the concatenated ERP data of all subjects (Bugli &

Lambert, 2007; Calhoun et al., 2009; Dien et al., 2007). The underlying assumption for qualifying an ERP is that the ICs/PCs associated with the ERP of interest are fixed across all subjects (Makeig et al., 1997). However, this is often in conflict with the real ERP waveform or topography, as individuals sometimes show substantial variance either in temporal or spatial characteristics of activation. Therefore, to overcome this problem, some advanced methods from this class of solutions attempt to extract the temporal and spatial features of ERPs of interest from single-trial EEG or individual subjects (Cong et al., 2010; Huster et al., 2020; Rissling et al., 2014; Zhang et al., 2023).

Other studies have applied ICA on single-trial EEG (Delorme et al., 2002) aimed at identifying the brain response from trials by manually (subjectively) confirming the ERP component of interest. The above-mentioned ICA method has been used in the popular EEGLAB toolbox (Delorme & Makeig, 2004), supported by statistical bootstrap methods. The core challenge for the methods mentioned above was that the latency and phase of individual trials varied to some extent. To extract these variables of ERPs among subjects, temporal-PCA has been used to extract ERP components of interest from single trials EEG epochs of an individual and found that the number of PCs related to a specific ERP component varied across subjects (Zhang et al., 2023). These findings, in turn, reveal that the latency and phase actually vary across subjects. Nevertheless, alignment of brain response within the trials (i.e., based on adjusting stimulus and responses with the averaged response) and using ICA decomposition as a subjective selection of the components were used to solve the inconsistency problem in trials (Jung et al., 2001; Onton et al., 2006).

In recent decades, cluster analysis has emerged as a promising tool for modeling event-related and resting-state EEG for exploring brain activations. The idea of EEG cluster analysis was first described by Lehmann (Lehmann et al., 1987), introducing the "atom of thoughts" that refers to the quasi-stable electrical potential (also called EEG microstates) that remain unchanged (semi-stable) for milliseconds of time range like 80-100 ms (D'Croz-Baron et al., 2021). The two steps of this method are calculating canonical cluster maps (template maps) that express the high variance explained and reassigning the template maps to the time points based on spatial correlation (Khanna et al., 2014). Yet, two popular clustering techniques, namely, modified *k*-means (Pascual-Marqui et al., 1995) and atomize and agglomerate hierarchical clustering (AAHC; Murray et al., 2008), were commonly used in this category. This method (microstate analysis), however, ignores the polarity of the time point and considers the global field power (GFP) or GFP extreme of data points (i.e., the scalp electrodes' standard deviation) for clustering.

Other clustering methods such as the Gaussian mixture model for individual subjects (De Lucia et al., 2007b) and single-trial EEG (De Lucia et al., 2007a), and stimulus-related statistical information from single-trial responses (Tzovara, Murray, Plomp, et al., 2012) have been used for EEG data. On the other hand, it has also been shown that consensus clustering can result in consistent and reliable clustering outcomes for biological data (Abu-Jamous et al., 2015; Liu et al., 2017), especially in ERP identification from group-averaged ERP data (Mahini et al., 2020; Mahini et al., 2022). The remaining challenge with the clustering analysis of single-trial EEG is the existing high degree of inconsistency in the EEG data that may lead to uncertain or faulty clustering results. However, the extraction of ERPs from single-trial EEG for individual subjects has not been thoroughly investigated in previous studies.

This study introduces a multi-set consensus clustering pipeline for identifying the ERP of interest from individual subjects using single-trial EEG data (**Figure 1**). Initially, we assess all single trials by comparing their spatial characteristics with the isolated ERP components found in group-averaged ERP data. The idea of using consensus clustering for single-trial EEG epochs is to generate aggregated cluster maps from each single trial that most likely contain interesting ERP responses. Subsequently, we apply a second-level consensus clustering to identify robust cluster maps among the selected trials of each subject. The whole process involves consensus clustering at both the individual trial level and across trials, called multi-set consensus clustering. Following this, we employ a newly modified time window determination method to precisely explore the latency of the ERP of interest at the individual subject level. We aim to develop a pipeline that effectively explores evoked responses for each condition or group at the individual subject level. We anticipate that this method will reliably identify the consistent ERP of interest within the single-trial EEG data of individual subjects.

2. Materials and Methods

2.1 EEG Data

We used real EEG data from a previous study that employed an active visual oddball task to evaluate our proposed pipeline. Here, we briefly introduced the paradigm, participants, and preprocessing pipelines for which more details are accessible in the original paper (Kappenman et al., 2021). The P3 component was initially designed to assess 'stimulus evaluation times', focusing on response time duration rather than the component's latency (Luck et al., 2009). The present experiment (Kappenman et al., 2021) adapted the stimulus set, which consisted of letter stimuli (A, B, C, D, and E) in which one of these letters was designated as the target, while the others served as nontargets. Hence, following the prior study, the P3

wave was considered in the maximum positive peak around 300 to 600 ms (i.e., the recommended time window for P3), which was used as the experimental information about the ERP component.

The EEG data were recorded from 40 participants (25 female and 15 male) via 30 scalp electrodes adhering to the international 10/20 system from two conditions, namely, 'Rare' and 'Frequent.' The recorded signals were digitized at 1024 Hz resolution, then downsampled to 256 Hz for faster processing, and referenced offline to the average of P9 and P10. The experimenters extracted roughly 50 to 70 trials but less in some cases for each condition from each subject. Epochs were selected from 200 ms before the stimulus onset to 800 ms after the stimulus onset. DC noise was removed, and the high-pass and low-pass filters were meticulously applied at 0.1 and 20 Hz to minimize any influence on stimulus onset latency. Subsequently, ICA was applied to address component-related artifacts, including eyeblinks and eye movements, which were removed via assessment by visual inspection and topographic representation of the components. Statistical power analysis was performed on the Pz electrode (recommended by the experimenters) and from the selected trials (see Section 2.2.1).

2.2 Proposed Method

In the following, we describe the role of each stage in the designed pipeline. Equally, **Figure 1** demonstrates the developed pipeline for identifying the interesting ERP of the individual subject.

Figure 1

2.2.1 Trial Selection

We examined each trial in order to eliminate trials carrying no/low correlated responses with the identified component (described below) from the group average ERP data. To this aim, each trial was individually clustered (using consensus clustering) to examine the interesting ERP component. Each single-trial EEG epoch was considered a dataset for clustering in which the time points are observations, and the electrodes' potentials were used as features (e.g., dataset size: 256 time points × 28 electrodes). Hence, the interesting ERP pattern (topographical configuration of ERP) was masked in single-trial clustering results. This was done by measuring the spatial correlation between candidate cluster map(s), i.e., cluster maps with high inner similarity in the roughly expected experimental interval, and with the identified ERP from group averaged data (Mahini et al., 2022). Two sensitivity parameters were used for adjusting trial examination: the inner similarity (e.g., > 0.90) and spatial correlation, e.g., > 0.50 with template map (determined cluster maps for representing the ERP), which can be adjusted when no map is found. The clustering design details will be described in Section 2.2.2. Notably, the proposed method was considered to keep a

sufficient number of trials (e.g., > 50% of trials for each subject and condition at minimum) by applying a silent decrement on the spatial correlation threshold (if needed).

2.2.2 Multi-Set Consensus Clustering

We designed a consensus clustering from clustering methods implemented in our toolbox (Mahini et al., 2022) in two levels: clustering of trials' datasets and ensemble clustering from results of trials for each subject/condition. For better cluster modeling of individual datasets, the clustering method selection was used based on the M-N plot method (Abu-Jamous et al., 2014; Mahini et al., 2022) on each subject's temporal concatenated ERP dataset. The M-N plot investigates two criteria that are fixed across the subjects: the inner-similarity of samples is high enough (e.g., > 95), and the duration of the elicited ERP (from the individual) is large enough (e.g., > 50 ms). Although estimating the optimal number of clusters from individual subject ERP data can be more appropriate, this was determined by testing the inner similarity of the estimated time window from the group average ERP data following our previous work (Mahini et al., 2022) to keep simplicity. Thereby, the selected trials were clustered using multi-set consensus clustering. The cluster-based similarity partitioning algorithm (CSPA) consensus function (Karypis & Kumar, 1998; Nguyen & Caruana, 2007) was chosen based on hypergraph partitioning using the 'supra' test (Ghosh et al., 2002) to explore the best possible ensemble clustering solution for trial and subject levels consensus clustering. Using CSPA potentially can provide some sense of tolerance on the mentioned varieties of information distribution in the single trials in this design.

Mathematically, let us consider the consensus clustering problem of n samples, $X = \{x_1, x_2, \dots, x_n\}$ into K groups, where each group is represented by a centroid $\mu_k, k = \{1,2, \dots, K\}$ and $x_t \in R^F, t = \{1,2, \dots, n\}$ and F denotes the number of features (electrodes in the EEG scalp). A set of m clusterings $L^{(1,2,\dots,m)}$ is used for combining into a final clustering L. Therefore, the objective function for cluster ensemble from m clusterings, a consensus function, Γ can be defined as a function of $N^{n\times m} \to N^n$, which maps the clustering to a final set of clusters.

$$\Gamma: \{L^{(i)} | i \in \{1,2, \dots, m\}\} \to L, \tag{1}$$

Given a set of clusters $\{L^{(i)} | i \in \{1,2, \dots, m\}\}$, the goal is to explore the firmest clustering that shares the most information from all clusterings. Therefore, the optimal labeling from m clusterings can be defined as:

$$L^*_{tm} = argmax_{L\in L} \sum_{l=1}^{m} \Gamma^{(NMI)}(L_l), \tag{2}$$

where Γ denotes a similarity measurement, e.g., NMI (Meila, 2007), which measures mutual information between a set of m clusterings and ${L^*}_{tm}$ is an optimally combined clustering with maximum average similarity to all other clusterings L_l for the individual trial. Next, we combine the clustering results of trials using further trial-level consensus clustering. The clustering at this level provides mutual information from all the trials. The consensus function for the trials can be presented as follows:

$$L^{**}{}_c^p = argmax_{L \in L_T} \sum_{i=1}^{T_c^p} \Gamma(L_i) \tag{3}$$

where, T_c^p denotes the number of selected trials for participant *p* in condition *c*. $L^{**}{}_c^p$ denotes the result of ensemble clustering across the trials. Together, the optimal ensemble clustering across the trials for each subject *p* can be noted by:

$$L^{**}{}_c^p = argmax_{L \in L_{X,T}} \sum_{i=1}^{T_c^p} \sum_{j=1}^{r} \Gamma\left(L_j^i\right) \tag{4}$$

Considering the CSPA consensus function's mechanism based on exploring the most aggregated cluster sets across the input clustering results, this property can guarantee to assigning consecutive time points to a cluster map that shares similar information in most cluster sets from diverse clusterings.

2.2.3 Time Window Determination

Once clustering results were obtained from the individual subjects, a modified version of the time window determination of our previous work (Mahini et al., 2020) was applied for each subject. We modified the time window determination through two criteria in two steps: First, we detected the candidate cluster maps, i.e., the cluster maps with high inner similarity, e.g., > 0.95, in the experimentally interesting interval. The experimental parameters (e.g., expected rough time window for response, rough estimation of the duration, and the region of interest) were estimated based on previous literature using a similar experimental design (Kappenman & Luck, 2012). Next, among those selected candidate cluster maps, we select those maps with a better fit and higher spatial correlation with the template map of P3 (e.g., > 0.90 that can be changed if needed). Note that the time window determination was used in both trial levels to calculate statistical scores (see section 2.3.3) and the subject level for identifying P3 from the clustering result of the subject.

2.3 Statistical Analysis and Performance Metrics

2.3.1 Statistical Analysis

We used a repeated measures ANOVA with a within-subject factor stimulus (conditions: 'Rare' and 'Frequent') in the electrode side of Pz (the same interesting electrode site as the original study) to test the

null hypothesis of no significant difference between the conditions in the determined time windows from individual subjects. The mean amplitude was calculated in the estimated time windows from individual subjects to investigate the effect of stimulus on the P3 component. The statistical comparisons were made at the alpha of 0.05.

2.3.2 Inter-Trial and Inter-Subject Reproducibility Tests

Inter-trial/subject reproducibility measures the consistency and predictability of stimulus-locked response properties at the individual trial/subject level. Unlike repeatability, which measures the consistency of generated repetitive results, reproducibility is considered a measure of obtaining consistent results from different generators (e.g., trials, subjects) that are not necessarily identical. In this study, reproducibility refers to the consistency of calculated scores from the proposed pipeline. Therefore, we established two standardization analysis methods, analytical and bootstrapping measurements, to assess identifying the P3 component. For analytical scores measurement, we calculated the standard measurement error (SME) of estimated scores at two levels, single-trial EEG and individual subject ERP. A similar concept of the SME index was introduced by Luck et al. (Luck et al., 2021) for ERP to assess the score and data measurement quality. Following their study, scoring refers to the results of identifying the ERP of interest (e.g., time window properties, peak latency, inner similarity) from individual subjects/trials. In general, the $\widehat{SME}$ (estimated SME) from the n results given score can be calculated as:

$$\widehat{SME} = \frac{\widehat{SD}}{\sqrt{n}}, \tag{5}$$

where the $\widehat{SD}$ is the standard deviation (SD) of the scores, and n is the number of contributed scores. Note that the true value of SME is unknown; thus, its estimation is denoted as $\widehat{SME}$ in the following sections. More clearly, given n trials of one condition from one subject and calculated scores (e.g., peak amplitude) from each trial, the standard error from n trials can be calculated from Eq. 5.

Let us introduce the score items used in our measurement. At the single-trial EEG level, the spatial correlation is evaluated between the estimated ERP and the identified template map from cluster analysis of the group average ERP. Therefore, the yielded $\widehat{SME}$ across the scores (spatial correlation) indicates the spatial error of the results at the individual trial level. Furthermore, we evaluate the temporal reproducibility by assessing the consistency of the estimated time windows across trials. Noticeably, the temporal reproducibility of the estimated temporal properties of the scores carries considerable noise and variability, which is associated with the nature of single-trial EEG data. Likewise, the reproducibility of

the spatial and temporal properties of the estimated time windows was investigated for qualifying ERP at the individual subject level. In the following, we described the conducted bootstrap procedure.

2.3.3 Bootstrapping and Reliability Tests

We established a bootstrapping process on the calculated scores for testing the reproducibility of generating/processing the results and reliability. Bootstrapping provides an estimation of the standard error if the experiment can be repeated many times. The idea of performing the bootstrap procedure is that, given an experiment, we can simulate trials/results by generating an adequate number of trials many times (e.g., 1000 times trial generation) *with replacement* for each condition and each subject, rather than repeating the experiment many times. In our design, the bootstrap mechanism was applied to generate the single trials clustering results used in the designed multi-set consensus clustering. Therefore, given R repeats and scores, $\widehat{bSME}_s^c$ for subject $s = \{1,2,\dots,S\}$ is calculated as averaged squared errors as:

$$\widehat{bSME}_s^c = \sqrt{\frac{\sum_{r=1}^{R} \widehat{SME}_r{}^2}{R}}, \tag{6}$$

where the standard error ($\widehat{SME}_r$) for each of the repeats $r = \{1,2,\dots,R\}$ is calculated as:

$$\widehat{SME}_r = \frac{\widehat{SD}_r}{\sqrt{N_c^s}}, \tag{7}$$

and N_c^s denotes the number of trials for subject s in condition c in each generation of bootstrapping. Therefore, the scores from each generation can be calculated followed by obtaining the measurement error for all the individual subjects as aggregated error:

$$MS\left(\widehat{SME}\right) = \frac{\widehat{SME}_1^2 + \widehat{SME}_2^2 + \cdots + \widehat{SME}_S^2}{S}. \tag{8}$$

Furthermore, an additional parameter called total error $\widehat{Var}_{all}$ is calculated from the individual subjects $\widehat{Var}_{par}$ called true variance, and the measurement error (calculated from Eq. 8). This calculation can be illustrated as:

$$\widehat{Var}_{all} = \widehat{Var}_{par} + MS(\widehat{SME}). \tag{9}$$

Although this metric was not originally designed for single-trial EEG analysis, we adapted it to generate simulated clusterings obtained from individual trials during the bootstrap process. Indeed, we assume sufficient trials are available for the ensemble clustering. Moreover, this can enhance the complexity of applying consensus clustering since no cluster generation step is required for each trial in each iteration.

Consequently, we seamlessly integrated the scoring results of the trials with individual subject scores, ensuring robust evaluations. Therefore, the reliability of the measurement can be calculated as follows:

$$\widehat{Reliability} = 1 - \frac{MS(\widehat{SME})}{\widehat{Var}_{all}}. \tag{10}$$

Furthermore, we used Cronbach's alpha and standard error of measurement (SEM) to calculate the reliability, estimating the error in individual scores within the subjects. The Cronbach's alpha is calculated as:

$$\alpha = \frac{q}{q-1}\left(1 - \frac{\sum_{i=1}^{q} \hat{V}_i}{\hat{V}_{tot}}\right), \tag{11}$$

where, q is the number of items (the number of scoring tests) and $\hat{V}_i$ denotes the variance associated with each measure and $\hat{V}_{tot}$ is the variance associated with all the scores. The $\widehat{SEM}$ is then calculated as:

$$\widehat{SEM} = \widehat{SD} \times \sqrt{1-\alpha}. \tag{12}$$

3. Results

We first present the consensus clustering results and the spatial and temporal properties of the identified ERP for individual subjects. Then, further performance analysis and standardization results are illustrated from the established bootstrap.

Figure 2

3.1 Multi-Set Consensus Clustering Results and Temporal Properties

Two levels of consensus clustering were employed. The first level was applied to group average ERP data to identify the ERP of interest, which served as the target template map information. The second level involved multi-set consensus clustering at the single-trial EEG level. For clustering at the group level, four clustering methods, namely, *k*-means, self-organizing map (SOM), modified *k*-means (with polarity adjustment), *k*-medoids clustering (KMD), and Gaussian mixture model (GMM), were selected from the implemented clustering methods in our toolbox (Mahini et al., 2022). **Figure 2** shows the clustering results, determined time windows, and topographical maps for group average ERP data. Observing the results from the group averaged ERP reveals identifying P3 by cluster maps 4 for both conditions with a high inner similarity, indicating the stability in the time window of 307.81 to 596.88 ms and 335.16 to 471.88 ms in 'Rare' and 'Frequent' conditions, respectively. Notably, the identified P3 has been used as the reference to examine spatial properties of the single trials and spatial correlation scores.

Before cluster analysis of the trials, the proposed method selects a clustering method set for single trials of each subject (see Section 2.2.2). **Table 1** presents the selected clustering methods for the single-trial EEG epochs of each subject for the subjects' temporal concatenated ERP dataset. Noteworthy that determining the suitable set of clustering methods for individuals is associated with a more precise cluster analysis, at least at the individual level. However, we did not find suitable clustering methods set for a few subjects (e.g., subjects 13, 38, 40) due to the criteria for selection in which the cluster method set for group-averaged ERP was replaced as a general set of clustering methods. We showed the clustering results of individual subjects in **Figure 3,** which involves the ERP waveforms at the Pz site, estimated time windows (red-colored rectangles), and clustering results for each subject. The results reveal variations in estimated responses from individual subjects. For some subjects, like subject 39, we did not find a distinct P3, which can pose challenges to experimental hypotheses (i.e., expecting similar P3 from all the subjects). We will discuss this issue in more detail in Section 4.

Table 1

Figure 3

3.2 Spatial Properties of P3 in Individual Subjects

We measured the mean of topographical maps (see **Figure 4**) in the estimated time window of individual subjects to demonstrate the topographical representation of the P3 component. The results in **Figure 4** and the spatial correlation results in **Table 2** highlight that the topographical maps of most subjects declare a fair spatial correlation with the template maps (topographical maps from the group average ERP results). However, we did not find a reasonable correlation between the topography of P3 and the template map in some subjects, like subject 39, which can be due to existing overlapped components or the lack of the expected strong brain response in the recorded signals. **Table 2** reports the scoring results, including estimated time windows, the inner similarity of time windows, the mean amplitude (at the Pz electrode), and the spatial correlation between the mean map and the template map regarding qualifying the P3 component for the individual subjects.

The scoring results for the individual subjects disclose the following findings: i) A reasonable spatial correlation was obtained between most of the individual subjects and the template maps in both conditions, with an average of 0.74 and $\widehat{SD}$ of 0.13 in the 'Rare'and 0.64 and $\widehat{SD}$ of 0.28 in the 'Frequent' condition. Nevertheless, we noticed a low correlation with the template map in some subjects, such as subjects 9 and 36, that may indicate different or dimed brain responses. ii) A larger amplitude was observed in the 'Rare'

(i.e., average 13.44 μv and $\widehat{SD}$ of 6.34 μv) compared to the amplitude in the 'Frequent' condition (i.e., average 7.16 μv and $\widehat{SD}$ of 4.03 μv) in the majority of subjects. iii) Additionally, high inner similarity in both conditions was observed in most subjects, i.e., 0.91 and $\widehat{SD}$ of 0.05 in the 'Rare' and 0.92 and $\widehat{SD}$ of 0.03 in the 'Frequent' condition, disclosing the quality of representative cluster maps. iv) Eventually, the average latency properties of 'Rare' (start and end) from the subjects were 351.08 ms to 495.91 ms and $\widehat{SD}$ of 76.12 and 90.64 ms, respectively. Those parameters for the 'Frequent' condition were obtained in 366.01 to 498.62 ms with $\widehat{SD}$ of 94.32 and 100.99 ms for start and end, respectively. These results, in turn, reveal a suitable consistency across the subjects regarding the obtained scores and are complementary to the results from the group-averaged ERP data in the original study.

Figure 4

Table 2

3.3 Evaluation Metrics and Performance Results

Figure 5 illustrates the reproducibility scores, including *analytical scores* from trials and subjects, as well as *bootstrap scores* obtained through 1000 iterations of trial clustering *with replacement*. The same scoring items were investigated, including mean amplitude, inner similarity, time window properties, and correlation between the mean topography of P3 in individual subjects and the corresponding obtained template map from group average ERP data. Realizing that the $aSME$ results were derived from single trials of individual subjects and $bSMEs$ were obtained from the bootstrapping procedure. The SME of the scores reveals how the scores can change if the experiment (here, processing) is repeated. Observing **Figure 5** and **Table 2** disclose the measured consistency in the scoring results at individual and group levels, indicating the experiment conduction quality and signal processing performances.

Figure 5

In order to illustrate the difference between the obtained corresponding $\widehat{aSME}$ and $\widehat{bSME}$ scores, we conducted two-sample *t*-tests. The results showed a significant difference between the $\widehat{aSME}$ and $\widehat{bSME}$ scores in the 'Rare' condition for both the start (p-value < 0.0001) and the end (p-value < 0.0001) points of the estimated time windows. These values were greater (i.e., mean of 22.44 ms) in $\widehat{aSME}$ than in $\widehat{bSME}$ (12.12 ms) in the star points and were similarly higher in $\widehat{aSME}$ (25.77ms) than in $\widehat{bSME}$ (14.37 ms) scores for the endpoints in 'Rare.' However, we did not find a significant difference between the scores in the 'Frequent' condition for the start points (p-value = 0.125) and the end (p-value = 0.1346) points. We

provided further detailed results of $\widehat{aSME}$ and $\widehat{bSME}$ from the estimated time windows in **Table S1**. For the obtained mean amplitudes (see **Table S2**), we also found a significant difference (p-value < 0.0001) between the $\widehat{aSME}$ and $\widehat{bSME}$ scores in 'Rare' that was greater for $\widehat{aSME}$ (mean of scores 2.26 μv) than for v$\widehat{bSME}$ (1.27 μv). The difference was not significant (p-value = 0.330) in the 'Frequent' condition. Furthermore, the spatial correlation scores (see **Table S3**) revealed a significant difference (p-value < 0.0001) between the $\widehat{aSME}$ and $\widehat{bSME}$ of the scores, with a relatively larger mean of 0.042 for $\widehat{aSME}$ than that of 0.025 for $\widehat{bSME}$ scores in the 'Rare' condition. Nevertheless, there was no significant difference (p-value = 0.389) between $\widehat{aSME}$ and $\widehat{bSME}$ scores in the 'Frequent' condition.

Finally, the statistical analysis of obtained SMEs from the inner similarity (see **Table S4**) of estimated time windows revealed a significant difference (p-value < 0.0001) between the $\widehat{SMEs}$ in 'Rare' that was a larger mean of 0.014 for $\widehat{aSME}$ than that of 0.007 for $\widehat{bSME}$ scores. Likewise, we also found a significant difference (p-value = 0.020) between $\widehat{aSME}$ and $\widehat{bSME}$ in 'Frequent,' where a higher inner similarity error of 0.008 was obtained in $\widehat{aSME}$ than $\widehat{bSME}$ with an error of 0.006. Further reliability tests on the obtained scores in **Table 2** revealed Cronbach's alpha of 0.70 from the scores in two conditions, implying a suitable consistency of the scoring results between the subjects.

Table S1-S4 Supplementary Results

3.4 Statistical Analysis Results

The statistical analysis of the variances on the selected time windows from individual subjects revealed that the main effect of the stimulus was significant, where the average results from the bootstrapping revealed F(1,39)=74.69 and SD of 14.49, p-value <0.0001, η_p^2= 0.651 and SD of 0.044, indicating a large effect of the P3 component. Notably, a large positive potential in the central lobe region was observed in the 'Rare' compared to the 'Frequent' condition, confirming the previous findings in the original study. The statistical analysis results implied the potential of the proposed pipeline to obtain more informative evoked activity measurement from individual subjects instead of using a fixed time window for all subjects or conventional conducting point-by-point statistical analysis (multiple t-tests of whole time points to search for significant differences).

4. Discussion

Given the rising interest in the role of single trials in determining time window, we proposed a multi-set consensus clustering pipeline for cluster analyzing of single-trial EEG in order to identify the brain

response at individual subject level. To tackle this problem, we designed the consensus clustering at the single trials level and combined the obtained clusterings across the trials for each individual subject. The idea of combining the clusterings of trials is that by employing the consensus mechanism, our method extracts similar cognitive responses by identifying time points with consistent contributions across trials, leading to mutually aggregated clustering results and reducing the influence of noisy clusters. Furthermore, the performance analysis revealed a suitable reproducibility of the clustering results with respect to the obtained scores for identifying the ERP of interest. Additionally, the use of bootstrap and analytical techniques exposed the stability of the proposed pipeline, providing robust clustering and scoring results for evoked single-trial EEG epochs of individuals. As a result, the estimated time windows offer a realistic representation of individual subjects' cognitive responses, making them suitable for both group-level and individual analyses.

The proposed method differs from the conventional methods in two main aspects. Firstly, it investigates the spatial and temporal properties of the cognitive response from single-trial EEG to the individual subject. This was accomplished by exploring the mutual temporal information from the single trials and the inner similarity investigation (stable spatial configuration) while determining the time window. In contrast, conventional microstate analysis-based methods evaluate spatial properties to classify microstates into commonly four privileged dominant classes of maps for event-related and resting state EEG (Antonova et al., 2022; Michel & Koenig, 2018; Zappasodi et al., 2019). For ERP data, the microstate analysis method assigns the GFP points from individual subjects' ERP data into the template maps obtained from group average ERP data clustering (Murray et al., 2008; Ruggeri et al., 2019). Therefore, the temporal structure of the ERP of interest is statistically identified wherein specific topography is dominated (i.e., obtained using clustering of single trial data). Meanwhile, identifying the temporal occurrence of template maps has relied on statistical analysis (De Lucia et al., 2007a; Tzovara, Murray, Michel, et al., 2012; Tzovara, Murray, Plomp, et al., 2012).

Secondly, the proposed pipeline investigates a dynamic clustering configuration in the consensus clustering generation phase for each subject using the M-N plot-based clustering selection. However, a fixed set of clustering methods was considered for all subjects in conventional microstate studies and the consensus clustering method on EEG/ERP data (Koenig et al., 2014; Mahini et al., 2022; Ruggeri et al., 2019). Noteworthy that the proposed method may yield suboptimal clustering performance in low SNR data, as it could result in a large number of noisy clusters - a common challenge in clustering-based approaches (Mahini et al., 2023). To address this issue, we introduced a post-hoc processing step that can

be applied at different clustering levels, such as the initial clustering of single trials with different methods. This step involves identifying thin cluster maps with a small number of samples (e.g., <10 ms) and assigning them to neighboring cluster maps if they exhibit sufficiently high spatial correlation (e.g., > 0.90 between mean topography maps). Another consideration is the potential difficulty in identifying highly overlapped components due to the nature of the clustering methods. This issue arises because the real brain response may be mixed with other components, particularly during the group average ERP data processing, which entails averaging trials from all subjects. To mitigate this, our approach analyzes individual subjects' responses from real trials, providing a more realistic representation of their brain responses compared to the averaging process.

Regarding the stability of the proposed pipeline results, both the analytical and bootstrapping scores demonstrated a reasonable consistency among individual subjects' results across different test items, especially spatial correlation scores, aligning with the expectations of the experiment. However, a few subjects exhibited deviant results where the corresponding bSME was not necessarily lower than the aSME. Subjects 9 and 39, for instance, displayed relatively deviant results. These deviations could be attributed to two potential factors. Firstly, the obtained topographical maps may have had a low SNR and might not have been statistically reliable within the determined time windows. Secondly, the selected trials during the preprocessing phase may not have contained sufficiently strong ERP responses, potentially leading to the inclusion of trials with lower spatial correlation in an effort to retain a minimum number of individual trials.

In the context of this study, lower bSMEs were interpreted as indicating greater reproducibility in the clustering results of the selected trials and the obtained scores. The bootstrapping test indicated higher stability in all the examined score items compared to the corresponding analytical scores. To ensure the robustness of the designed bootstrapping process, we further repeated the bootstrapping process with varying numbers of iterations (e.g., 100, 1000, and 3000) for the same single-trial clusterings. Pairwise t-tests applied to the resulting bSMEs from these repetitions showed no significant differences (p-value ~ 1.00) between the corresponding scores, confirming the reliability of the bootstrapping test (Davison & Hinkley, 1997; Efron, 1992). However, increasing the number of repetitions did enhance stability, particularly in the scoring results for the 'Frequent' condition compared to the 'Rare' condition. It is noteworthy that we assume to include an adequate number of trials (e.g., a minimum of 50 trials for bootstrap generation) after the trial selection step. However, in cases where subjects had fewer selected

trials, like less than 10 trials, we adjusted the criteria (e.g., to a minimum of 20 trials) to ensure a reasonable number of trial selections.

The developed pipeline revealed spatially correlated brain activity with similar temporal properties (though not necessarily identical), supporting the principle of consistent brain responses across single trials and individual subjects. The designed reproducibility evaluation also illustrated the consistency of the obtained results, implying the reliability of the proposed cluster analysis if the random trials were generated iteratively. The statistical analysis highlighted a significant effect at the Pz electrode site, along with the identified time windows, exhibiting a larger positive potential in the 'Rare' condition compared to 'Frequent' in most subjects, confirming the experiment hypotheses. The MATLAB demo code for the proposed pipeline is available at https://github.com/remahini/Single_trial_EEG_MSCC, which can be easily modified by the researchers to test their hypothesis. Noteworthy that our method is not limited to identifying the standard P3 component, as demonstrated in this study; it holds the potential for identifying other ERP components from event-related single-trial EEG data. Moreover, the proposed method meets suitable confidence in exploring ERP of interest for individual subjects, which is essential for different individual subject investigations. However, more detailed studies and reliability tests are required to overcome potential risks and ethical concerns to be used in critical applications.

5. Conclusions

In conclusion, using the proposed data-driven approach for investigating single-trial EEG suggests that the evoked response in an individual subject can be reliably identified from single trials when examining single-trial EEG clustering. The proposed method addressed the complexity of identifying ERP of interest with single-trial EEG by hierarchically combining clustering information from single trials with minimum knowledge about the component of interest. Our early findings support existing spatially correlated cluster maps in a single trial of individual subject data associated with reliable estimations of the brain response. The proposed pipeline offers an unbiased means of identifying interesting potentials, enhancing the likelihood of uncovering real components. The future outlook of this project can be a promising tool for reliable investigations of individual subject brain activity, particularly in clinical applications that are still open research questions for individual subject single-trial EEG data analysis. Future developments may leverage multi-dimensional single-trial EEG processing, which offers a powerful tool to explore brain responses across/combined from various domains and perspectives through clustering analyses.

Authors' Contributions

R. M. conceptualized and conducted the study, including developing the algorithm, data processing, and writing and revising the paper. G. Z. contributed to data processing technical support and writing and reviewing the manuscript. T. P., R. D., and A. N. conceptualized the study, provided technical support, and revised the manuscript. F. C. and T. H. conceptualized and supervised the whole project, provided technical support, and wrote and revised the manuscript. All authors have read and agreed to publish the final version of the manuscript.

Corresponding author

Correspondence to T. Hämäläinen.

Supplementary Materials

The open-access demo code and testing data are available at https://github.com/remahini/Single_trial_EEG_MSCC.

Funding Declaration

The authors have no relevant financial or non-financial interests to disclose. The authors have no funding for this study.

Ethics declarations

Consent to Participate

This study does not include data collection from individual participants, and public data has been used.

Competing Interests

The authors declare no competing interests.

References

Abu-Jamous, B., Fa, R., Roberts, D. J., & Nandi, A. K. (2014, 4-9 May 2014). M-N scatter plots technique for evaluating varying-size clusters and setting the parameters of Bi-CoPaM and Uncles methods. 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), https://doi.org/10.1109/ICASSP.2014.6854902

Abu-Jamous, B., Fa, R., Roberts, D. J., & Nandi, A. K. (2015, Jun 4). UNCLES: method for the identification of genes differentially consistently co-expressed in a specific subset of datasets. *BMC Bioinformatics, 16*, Article 184. https://doi.org/10.1186/s12859-015-0614-0

Antonova, E., Holding, M., Suen, H. C., Sumich, A., Maex, R., & Nehaniv, C. (2022, 2022/06/01/). EEG microstates: Functional significance and short-term test-retest reliability. *Neuroimage: Reports, 2*(2), 100089. https://doi.org/10.1016/j.ynirp.2022.100089

Bugli, C., & Lambert, P. (2007). Comparison between Principal Component Analysis and Independent Component Analysis in Electroencephalograms Modelling. *Biometrical Journal, 49*(2), 312-327. https://doi.org/10.1002/bimj.200510285

Calhoun, V. D., Liu, J., & Adalı, T. (2009, 2009/03/01/). A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data. *Neuroimage, 45*(1, Supplement 1), S163-S172. https://doi.org/10.1016/j.neuroimage.2008.10.057

Cohen, M., & Cavanagh, J. F. (2011, 2011-February-28). Single-Trial Regression Elucidates the Role of Prefrontal Theta Oscillations in Response Conflict [Original Research]. *Frontiers in Psychology, 2*. https://doi.org/10.3389/fpsyg.2011.00030

Cong, F., Kalyakin, I., Huttunen-Scott, T., Li, H., Lyytinen, H., & Ristaniemi, T. (2010). SINGLE-TRIAL BASED INDEPENDENT COMPONENT ANALYSIS ON MISMATCH NEGATIVITY IN CHILDREN. *International Journal of Neural Systems, 20*(04), 279-292. https://doi.org/10.1142/s0129065710002413

Cong, F., Lin, Q.-H., Kuang, L.-D., Gong, X.-F., Astikainen, P., & Ristaniemi, T. (2015, 2015/06/15/). Tensor decomposition of EEG signals: A brief review. *Journal of Neuroscience Methods, 248*, 59-69. https://doi.org/10.1016/j.jneumeth.2015.03.018

D'Croz-Baron, D. F., Bréchet, L., Baker, M., & Karp, T. (2021, 2021/01/01). Auditory and Visual Tasks Influence the Temporal Dynamics of EEG Microstates During Post-encoding Rest. *Brain Topography, 34*(1), 19-28. https://doi.org/10.1007/s10548-020-00802-4

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap Methods and their Application*. Cambridge University Press. https://doi.org/10.1017/CBO9780511802843

De Lucia, M., Michel, C. M., Clarke, S., & Murray, M. M. (2007a, 2007). Single-trial topographic analysis of human EEG: A new `image' of event-related potentials. 6th International Special Topic Conference on Information Technology Applications in Biomedicine, https://doi.org/10.1109/itab.2007.4407353

De Lucia, M., Michel, C. M., Clarke, S., & Murray, M. M. J. I. J. o. B. (2007b). Single subject EEG analysis based on topographic information. *9*(3), 168-171.

Delorme, A., & Makeig, S. (2004, Mar). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis [Article]. *Journal of Neuroscience Methods, 134*(1), 9-21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Delorme, A., Makeig, S., Fabre-Thorpe, M., & Sejnowski, T. (2002). From single-trial EEG to brain area dynamics. *Neurocomputing, 44-46*, 1057-1064. https://doi.org/10.1016/s0925-2312(02)00415-0

Dien, J., Khoe, W., & Mangun, G. R. (2007). Evaluation of PCA and ICA of simulated ERPs: Promax vs. infomax rotations. *Human Brain Mapping, 28*(8), 742-763. https://doi.org/10.1002/hbm.20304

Efron, B. (1992). Bootstrap Methods: Another Look at the Jackknife. In S. Kotz & N. L. Johnson (Eds.), *Breakthroughs in Statistics: Methodology and Distribution* (pp. 569-593). Springer New York. https://doi.org/10.1007/978-1-4612-4380-9_41

Ghosh, J., Strehl, A., & Merugu, S. (2002). A consensus framework for integrating distributed clusterings under limited knowledge sharing. Proc. NSF Workshop on Next Generation Data Mining,

Herrmann, C. S., Rach, S., Vosskuhl, J., & Strueber, D. (2014, Jul). Time-Frequency Analysis of Event-Related Potentials: A Brief Tutorial. *Brain Topography, 27*(4), 438-450. https://doi.org/10.1007/s10548-013-0327-5

Huster, R. J., Messel, M. S., Thunberg, C., & Raud, L. (2020, 2020/11/01/). The P300 as marker of inhibitory control – Fact or fiction? *Cortex, 132*, 334-348. https://doi.org/10.1016/j.cortex.2020.05.021

Jung, T.-P., Makeig, S., Westerfield, M., Townsend, J., Courchesne, E., & Sejnowski, T. J. (2001). Analysis and visualization of single-trial event-related potentials. *Human Brain Mapping, 14*(3), 166-185. https://doi.org/10.1002/hbm.1050

Kappenman, E. S., Farrens, J. L., Zhang, W., Stewart, A. X., & Luck, S. J. (2021, 2021/01/15/). ERP CORE: An open resource for human event-related potential research. *Neuroimage, 225*, 117465. https://doi.org/10.1016/j.neuroimage.2020.117465

Kappenman, E. S., & Luck, S. J. (2012). Manipulation of orthogonal neural systems together in electrophysiological recordings: the MONSTER approach to simultaneous assessment of multiple neurocognitive dimensions. *Schizophrenia Bulletin, 38*(1), 92-102. https://doi.org/10.1093/schbul/sbr147

Karypis, G., & Kumar, V. (1998, 1998/01/10/). Multilevelk-way Partitioning Scheme for Irregular Graphs. *Journal of Parallel and Distributed Computing, 48*(1), 96-129. https://doi.org/10.1006/jpdc.1997.1404

Khanna, A., Pascual-Leone, A., & Farzan, F. (2014). Reliability of Resting-State Microstate Features in Electroencephalography. *PloS One, 9*(12), e114163. https://doi.org/10.1371/journal.pone.0114163

Knuth, K. H., Shah, A. S., Truccolo, W. A., Ding, M., Bressler, S. L., & Schroeder, C. E. (2006, 2006/05/01). Differentially Variable Component Analysis: Identifying Multiple Evoked Components Using Trial-to-Trial Variability. *Journal of Neurophysiology, 95*(5), 3257-3276. https://doi.org/10.1152/jn.00663.2005

Koenig, T., Stein, M., Grieder, M., & Kottlow, M. (2014, Jan). A Tutorial on Data-Driven Methods for Statistically Assessing ERP Topographies. *Brain Topography, 27*(1), 72-83. https://doi.org/10.1007/s10548-013-0310-1

Lehmann, D., Ozaki, H., & Pal, I. (1987). EEG alpha map series: brain micro-states by space-oriented adaptive segmentation. *67*(3), 271-288. https://doi.org/10.1016/0013-4694(87)90025-3

Liu, C., Abu-Jamous, B., Brattico, E., & Nandi, A. K. (2017). Towards Tunable Consensus Clustering for Studying Functional Brain Connectivity During Affective Processing. *International Journal of Neural Systems, 27*(02), 1650042. https://doi.org/10.1142/S0129065716500428

Luck, S. J., Kappenman, E. S., Fuller, R. L., Robinson, B., Summerfelt, A., & Gold, J. M. (2009, 2009/07/01). Impaired response selection in schizophrenia: Evidence from the P3 wave and the lateralized readiness potential. *Psychophysiology, 46*(4), 776-786. https://doi.org/https://doi.org/10.1111/j.1469-8986.2009.00817.x

Luck, S. J., Stewart, A. X., Simmons, A. M., & Rhemtulla, M. (2021). Standardized measurement error: A universal metric of data quality for averaged event-related potentials. *Psychophysiology, 58*(6), e13793. https://doi.org/10.1111/psyp.13793

Mahini, R., Li, F., Zarei, M., Nandi, A. K., Hämäläinen, T., & Cong, F. (2023, 2023/09/01/). Ensemble deep clustering analysis for time window determination of event-related potentials. *Biomedical Signal Processing and Control, 86*, 105202. https://doi.org/10.1016/j.bspc.2023.105202

Mahini, R., Li, Y., Ding, W., Fu, R., Ristaniemi, T., Nandi, A. K., Chen, G., & Cong, F. (2020, 2020-October-21). Determination of the Time Window of Event-Related Potential Using Multiple-Set Consensus Clustering [Methods]. *Frontiers in Neuroscience, 14*(1047). https://doi.org/10.3389/fnins.2020.521595

Mahini, R., Xu, P., Chen, G., Li, Y., Ding, W., Zhang, L., Qureshi, N. K., Hämäläinen, T., Nandi, A. K., & Cong, F. (2022). Optimal Number of Clusters by Measuring Similarity Among Topographies for Spatio-Temporal ERP Analysis. *Brain Topography*. https://doi.org/10.1007/s10548-022-00903-2

Makeig, S., Jung, T.-P., Bell, A. J., Ghahremani, D., & Sejnowski, T. J. (1997). Blind separation of auditory event-related brain responses into independent components. *Proceedings of the National Academy of Sciences, 94*(20), 10979-10984. https://doi.org/doi:10.1073/pnas.94.20.10979

Meila, M. (2007, May). Comparing clusterings - an information based distance. *Journal of Multivariate Analysis, 98*(5), 873-895. https://doi.org/10.1016/j.jmva.2006.11.013

Michel, C. M., & Koenig, T. (2018, 2018/10/15/). EEG microstates as a tool for studying the temporal dynamics of whole-brain neuronal networks: A review. *Neuroimage, 180*, 577-593. https://doi.org/10.1016/j.neuroimage.2017.11.062

Murray, M. M., Brunet, D., & Michel, C. M. (2008, Jun). Topographic ERP analyses: A step-by-step tutorial review. *Brain Topography, 20*(4), 249-264. https://doi.org/10.1007/s10548-008-0054-5

Nguyen, N., & Caruana, R. (2007, 28-31 Oct. 2007). Consensus Clusterings. Seventh IEEE International Conference on Data Mining (ICDM 2007), https://doi.org/10.1109/ICDM.2007.73

Onton, J., Westerfield, M., Townsend, J., & Makeig, S. (2006, 2006/01/01/). Imaging human EEG dynamics using independent component analysis. *Neuroscience and Biobehavioral Reviews, 30*(6), 808-822. https://doi.org/10.1016/j.neubiorev.2006.06.007

Pascual-Marqui, R. D., Michel, C. M., & Lehmann, D. J. I. T. o. B. E. (1995). Segmentation of brain electrical activity into microstates: model estimation and validation. *42*(7), 658-665. https://doi.org/10.1109/10.391164

Rissling, A. J., Miyakoshi, M., Sugar, C. A., Braff, D. L., Makeig, S., & Light, G. A. (2014, 2014/01/01/). Cortical substrates and functional correlates of auditory deviance processing deficits in schizophrenia. *NeuroImage: Clinical, 6*, 424-437. https://doi.org/10.1016/j.nicl.2014.09.006

Ruggeri, P., Meziane, H. B., Koenig, T., & Brandner, C. (2019, Mar 6). A fine-grained time course investigation of brain dynamics during conflict monitoring. *Scientific Reports, 9*, Article 3667. https://doi.org/10.1038/s41598-019-40277-3

Schölkopf, B., Smola, A., & Müller, K. (1998). Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation, 10*(5), 1299-1319. https://doi.org/10.1162/089976698300017467

Tzovara, A., Murray, M. M., Michel, C. M., & De Lucia, M. (2012, 2012/08/01). A Tutorial Review of Electrical Neuroimaging From Group-Average to Single-Trial Event-Related Potentials. *Developmental Neuropsychology, 37*(6), 518-544. https://doi.org/10.1080/87565641.2011.636851

Tzovara, A., Murray, M. M., Plomp, G., Herzog, M. H., Michel, C. M., & De Lucia, M. (2012, 2012/06/01/). Decoding stimulus-related information from single-trial EEG responses based on voltage topographies. *Pattern Recognition, 45*(6), 2109-2122. https://doi.org/10.1016/j.patcog.2011.04.007

Zappasodi, F., Perrucci, M. G., Saggino, A., Croce, P., Mercuri, P., Romanelli, R., Colom, R., & Ebisch, S. J. (2019). EEG microstates distinguish between cognitive components of fluid reasoning. *Neuroimage, 189*, 560-573. https://doi.org/10.1016/j.neuroimage.2019.01.067

Zhang, G., Li, X., Lu, Y., Tiihonen, T., Chang, Z., & Cong, F. (2023, 2023/02/01/). Single-trial-based temporal principal component analysis on extracting event-related potentials of interest for an individual subject. *Journal of Neuroscience Methods, 385*, 109768. https://doi.org/10.1016/j.jneumeth.2022.109768

Figure captions

Figure 1. Demonstration of the proposed pipeline for identifying the time window of interesting ERP of individual subjects using multi-trial consensus clustering. **i)** The clustering method selection procedure selects the appropriate clustering configuration to feed consensus clustering (CC) for individual subjects. **ii)** Initial clustering provides information on the existing target response by testing the spatial correlation between the grand average topography (selection criteria) and candidate cluster maps resulting from the initial clustering. **iii)** Clustering of the selected trials in two levels: individual trial clustering and then clustering of the results across the trials. **iv)** Exploring for the most appropriate time window, examining the candidate maps' inner similarity and spatial correlation. C= condition.

Figure 2. Consensus clustering results on group-averaged ERP data and the identified P3 component from grand mean data. The spatial property of the elicited P3 is used as the template map for selecting trials and comparing individual subjects' results.

Figure 3. Illustration of the clustering results for individual subjects and estimated time window for each condition (indicated by the red rectangle) for identifying the P3 component. The waveform has been shown from the 'Pz' electrode in single trial data.

Figure 4. Illustration of the mean topographical maps (in the determined time windows) for P3 calculated from the ERP dataset of subjects. The two bigger topographical maps (in the results top panel) illustrate the ground truth maps calculated from the grand mean ERP data.

Figure 5. Estimated analytical standard measurement error ($\widehat{aSME}$) and bootstrapping SME ($\widehat{bSME}$) of the calculated scores for each condition. **A)** $\widehat{SMEs}$ of calculated scores of the temporal properties of time windows in single trials of subjects and the bootstrapping procedure with 1000 iterations. **B)** The $\widehat{aSME}$ and $\widehat{bSME}$ of the calculated amplitude scores in the determined time windows. **C)** $\widehat{SMEs}$ of spatial correlation scores, i.e., between the obtained maps (in the estimated time window) and the template map. **D)** The $\widehat{SMEs}$ of inner similarity scores determined from individual subjects in the determined time windows.

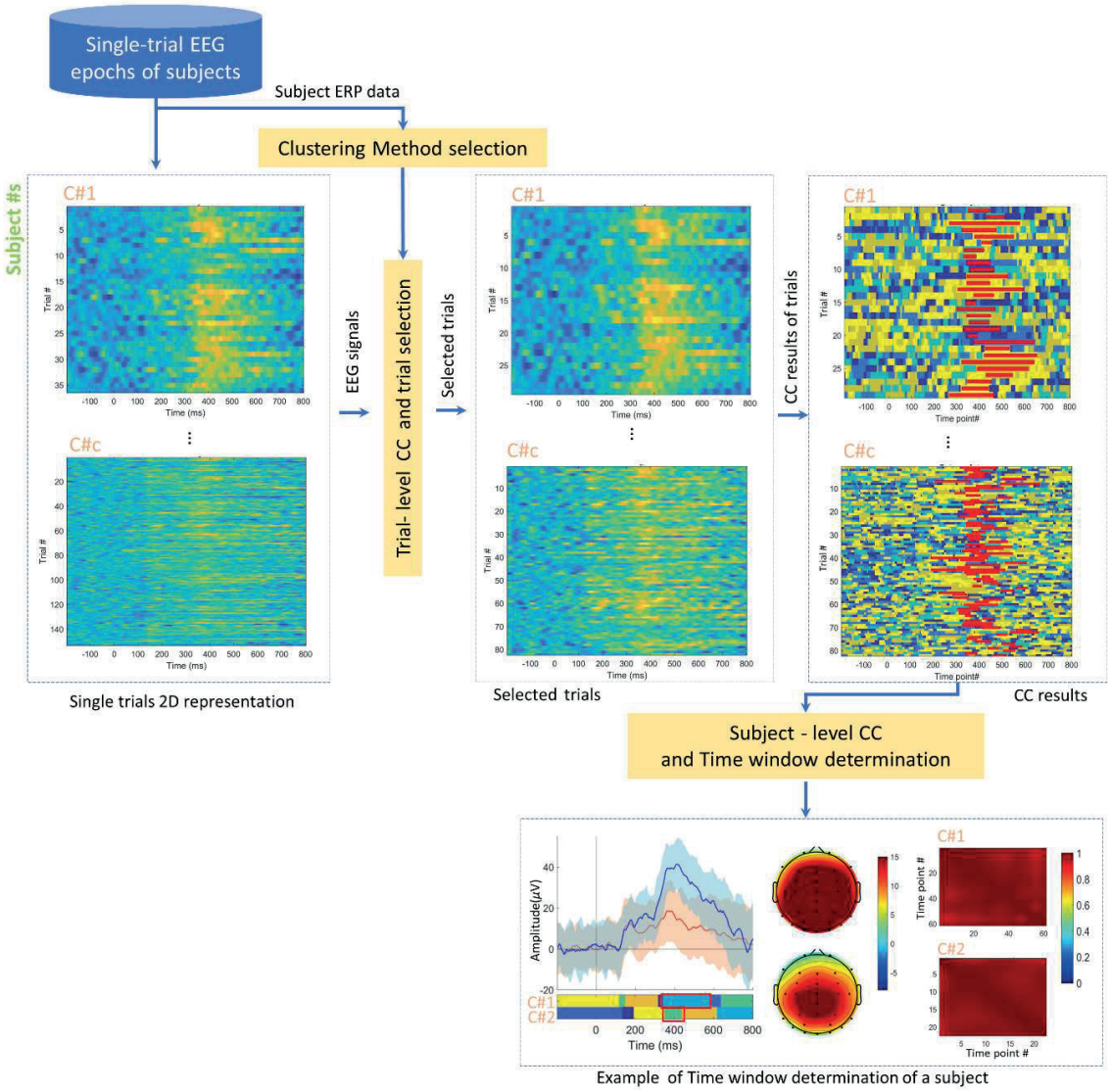


Figure 1. Demonstration of the proposed pipeline for identifying the time window of interesting ERP of individual subjects using multi-trial consensus clustering. **i)** The clustering method selection procedure selects the appropriate clustering configuration to feed consensus clustering (CC) for individual subjects. **ii)** Initial clustering provides information on the existing target response by testing the spatial correlation between the grand average topography (selection criteria) and candidate cluster maps resulting from the initial clustering. **iii)** Clustering of the selected trials in two levels: individual trial clustering and then clustering of the results across the trials. **iv)** Exploring for the most appropriate time window, examining the candidate maps' inner similarity and spatial correlation. C= condition.

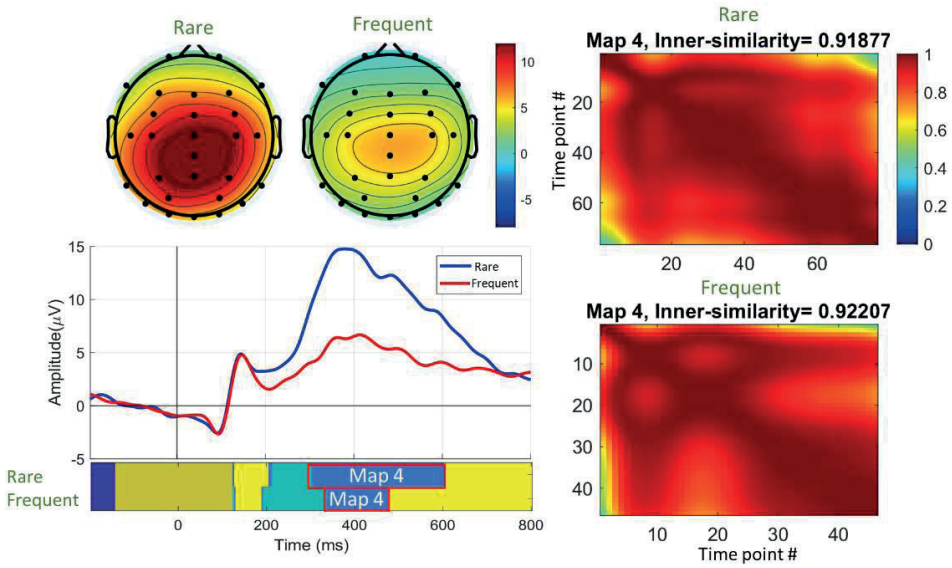


Figure 2. Consensus clustering results on group-averaged ERP data and the identified P3 component from grand mean data. The spatial property of the elicited P3 is used as the template map for selecting trials and comparing individual subjects' results.

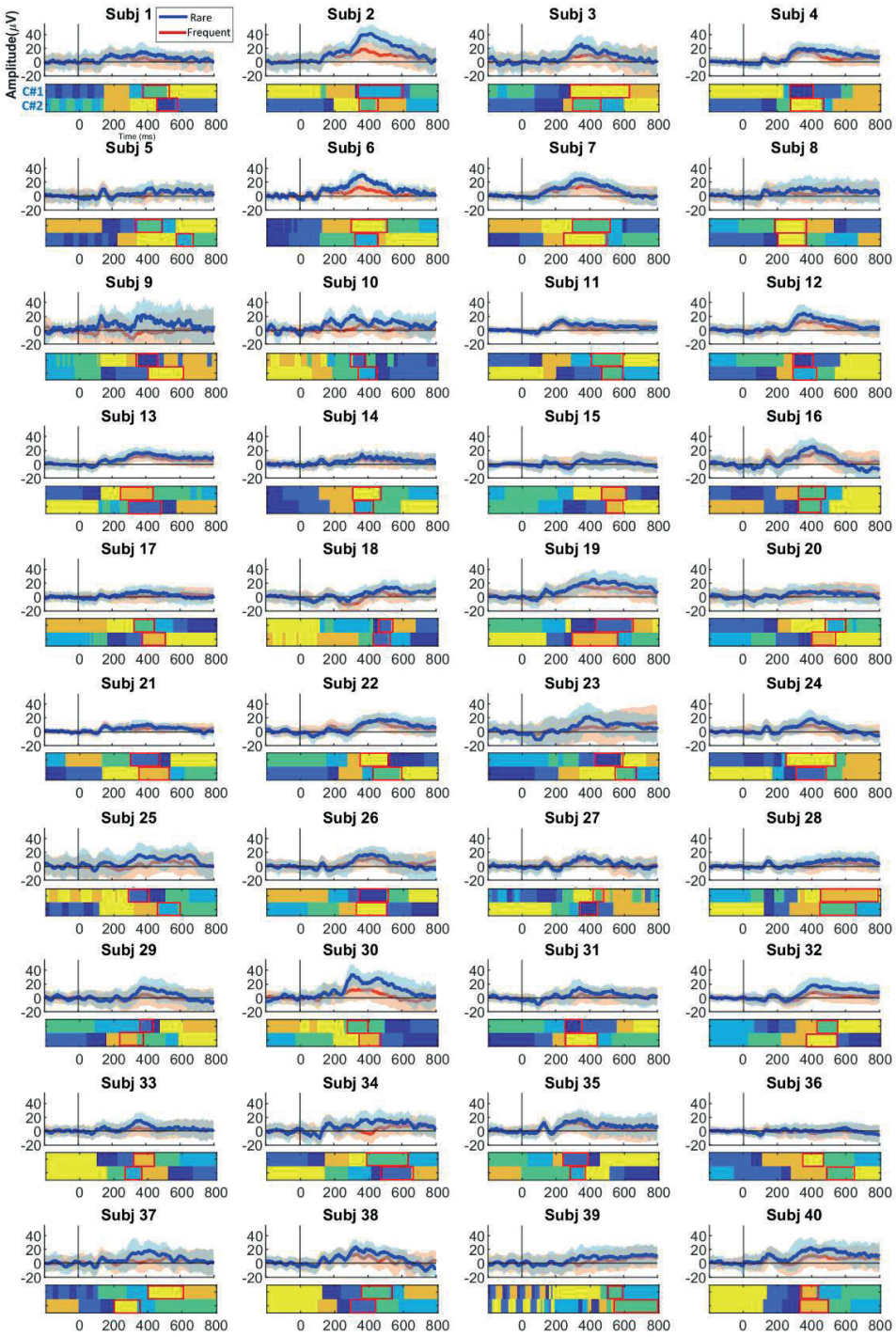


Figure 3. Illustration of the clustering results for individual subjects and estimated time window for each condition (indicated by the red rectangle) for identifying the P3 component. The waveform has been shown from the 'Pz' electrode in single trial data.

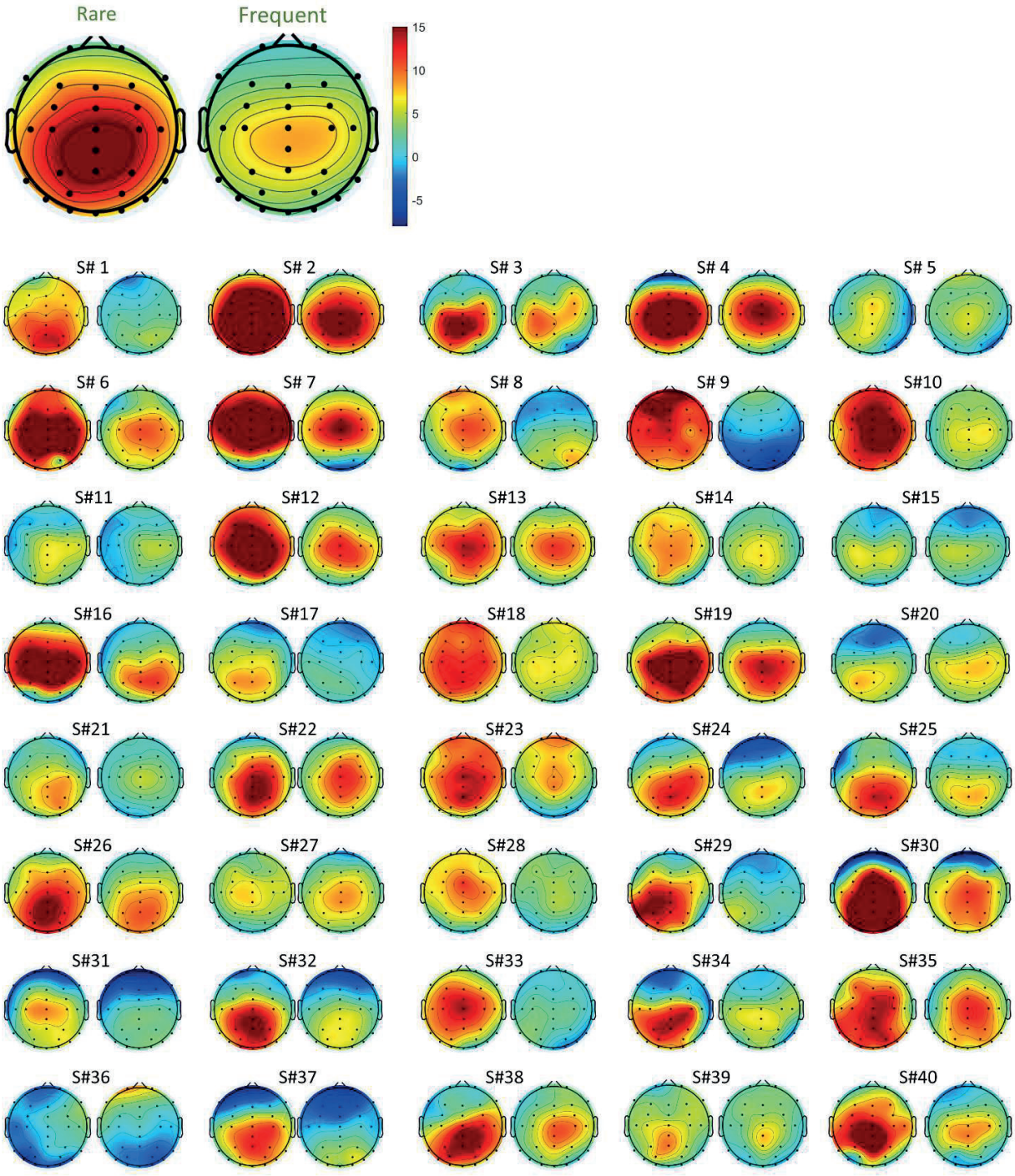


Figure 4. Illustration of the mean topographical maps (in the determined time windows) for P3 calculated from the ERP dataset of subjects. The two bigger topographical maps (in the results top panel) illustrate the ground truth maps calculated from the grand mean ERP data.

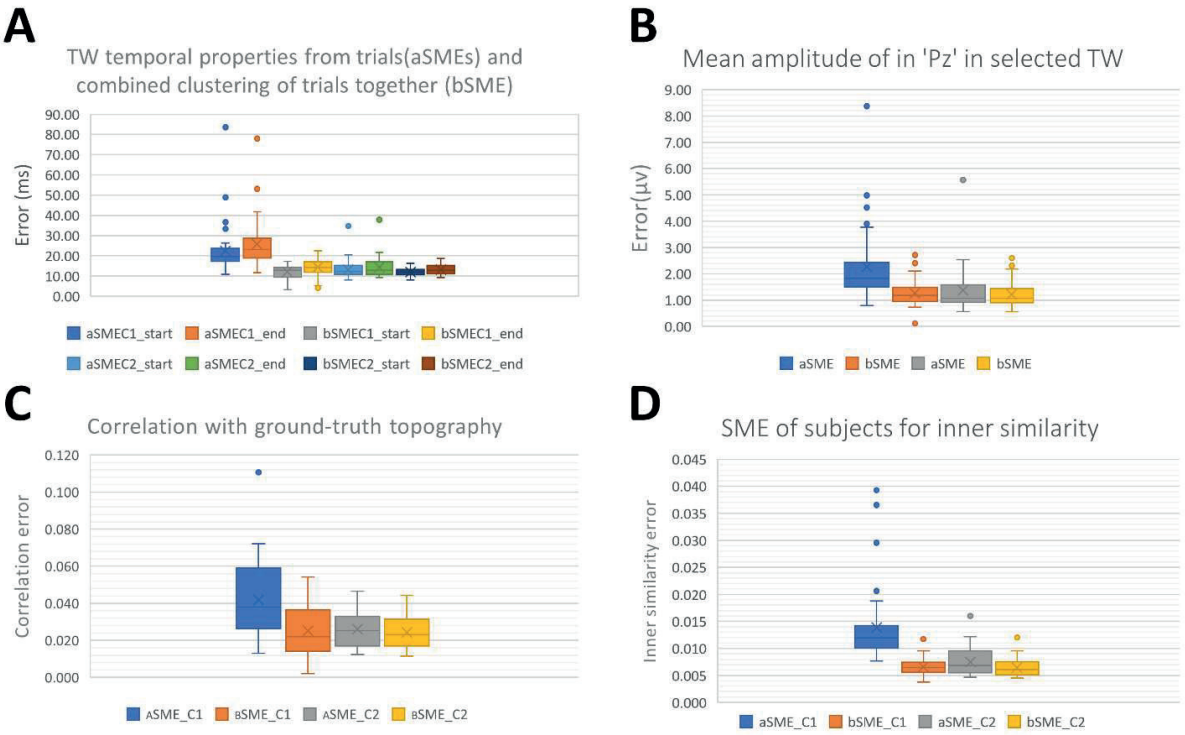


Figure 5. Estimated analytical standard measurement error ($\widehat{aSME}$) and bootstrapping SME ($\widehat{bSME}$) of the calculated scores for each condition. **A)** $\widehat{SMEs}$ of calculated scores of the temporal properties of time windows in single trials of subjects and the bootstrapping procedure with 1000 iterations. **B)** The $\widehat{aSME}$ and $\widehat{bSME}$ of the calculated amplitude scores in the determined time windows. **C)** $\widehat{SMEs}$ of spatial correlation scores, i.e., between the obtained maps (in the estimated time window) and the template map. **D)** The $\widehat{SMEs}$ of inner similarity scores determined from individual subjects in the determined time windows.

Table captions

Table 1. Illustration of the selected clustering methods for the individual subjects, evaluating the results of the M-N plot test on individuals' temporal concatenated ERP data. Note that the replacement list is used when no suitable method is selected or an individual clustering method is chosen. KM= *k*-means, HC = hierarchical clustering, SOM= self-organizing map, DSPC = diffusion map spectral clustering, MKMS = modified *k*-means, SPC = spectral clustering, KMD = *k*-Medoids clustering, and GMM = Gaussian mixture model.

Table 2. Illustration of the calculated scores obtained from the determined time windows of individual subjects using the proposed pipeline. The scores involve the temporal properties of the estimated time windows (start and end), inner similarity, amplitude in the Pz electrode, and correlation of mean topography with ground truth topography. TW = time window, Innsim = inner similarity, Amp = mean amplitude, Corr = spatial correlation.

Table S1. The calculated $\widehat{aSME}$ and $\widehat{bSME}$ scores for the determined time windows for the individual subjects. The $\widehat{aSME}$ is obtained from processing all the trials, and the $\widehat{bSME}$ is calculated from the bootstrapping by selecting (with replacement) at least 50 trials for each repeat.

Table S2. The obtained $\widehat{aSME}$ and $\widehat{bSME}$ scores from the calculated mean amplitudes in the estimated time windows for the individual subjects.

Table S3. Illustration of the $\widehat{aSME}$ and $\widehat{bSME}$ scores of the individual subjects, calculated from the spatial correlation between the mean topography in the estimated time windows and the ground truth topography.

Table S4. Calculated $\widehat{aSME}$ and $\widehat{bSME}$ scores from the inner similarity of time points in the estimated time windows for the individual subjects.

Table 1. Illustration of the selected clustering methods for the individual subjects, evaluating the results of the M-N plot test on individuals' temporal concatenated ERP data. Note that the replacement list is used when no suitable method is selected or an individual clustering method is chosen. KM= *k*-means, HC = hierarchical clustering, SOM= self-organizing map, DSPC = diffusion map spectral clustering, MKMS = modified *k*-means, SPC = spectral clustering, KMD = *k*-Medoids clustering, and GMM = Gaussian mixture model.

Subj_ID	Selected methods (Method-code)	Replaced List (Method-code)
S1	KM-1, SOM-4, DSPC-5, SPC-8, KMD-9, GMM-10	-
S2	KM-1, SOM-4, DSPC-5, MKM-6, KMD-9, GMM-10	-
S3	KM-1, HC-2, SOM-4, DSPC-5, SPC-8, GMM-10	-
S4	KM-1, HC-2, SOM-4, DSPC-5, MKM-6, GMM-10	-
S5	SOM-4, DSPC-5, MKM-6, SPC-8	-
S6	KM-1, HC-2, SOM-4, DSPC-5, MKM-6, SPC-8, KMD-9, GMM-10	-
S7	KM-1, HC-2, DSPC-5, MKM-6, SPC-8, GMM-10	-
S8	KM-1, SOM-4, DSPC-5, MKM-6, SPC-8, KMD-9	-
S9	KM-1, SOM-4, MKM-6, KMD-9, GMM-10	-
S10	KM-1, HC-2, DSPC-5, MKM-6, KMD-9, GMM-10	-
S11	KM-1, HC-2, SOM-4, MKM-6, SPC-8, KMD-9, GMM-10	-
S12	DSPC-5, MKM-6, SPC-8, KMD-9, GMM-10	-
S13	No Method determined	KM-1, SOM-4, MKM-6, KMD-9
S14	DSPC-5, MKM-6, SPC-8, KMD-9, GMM-10	-
S15	KM-1, HC-2, SOM-4, DSPC-5, MKM-6, SPC-8, KMD-9, GMM-10	-
S16	HC-2, SOM-4, DSPC-5, MKM-6, SPC-8, KMD-9, GMM-10	-
S17	KM-1, SOM-4, DSPC-5, MKM-6, KMD-9, GMM-10	-
S18	KM-1, SOM-4, MKM-6, SPC-8, KMD-9	-
S19	DSPC-5, GMM-10	-
S20	KM-1, HC-2, SOM-4, DSPC-5, MKM-6, KMD-9, GMM-10	-
S21	KM-1, MKM-6, SPC-8, KMD-9, GMM-10	-
S22	KM-1, SOM-4, DSPC-5, MKM-6, SPC-8, KMD-9, GMM-10	-
S23	SOM-4, MKM-6, KMD-9	-
S24	KM-1, HC-2, SOM-4, DSPC-5, MKM-6, KMD-9, GMM-10	-
S25	KM-1, HC-2, SOM-4, DSPC-5, MKM-6, KMD-9, GMM-10	-
S26	KM-1, SOM-4, DSPC-5, MKM-6, KMD-9, GMM-10	-
S27	KM-1, HC-2, SOM-4, DSPC-5, GMM-10	-
S28	KM-1, HC-2, SOM-4, DSPC-5, MKM-6, GMM-10	-
S29	KM-1, SOM-4, DSPC-5, MKM-6, KMD-9,GMM-10	-
S30	KM-1, HC-2, SOM-4, DSPC-5, MKM-6, KMD-9, GMM-10	-
S31	KM-1, SOM-4, DSPC-5, MKM-6, KMD-9, GMM-10	-
S32	KM-1, SOM-4, DSPC-5, MKM-6, KMD-9, GMM-10	-
S33	HC-2, SOM-4, DSPC-5, KMD-9, GMM-10	-
S34	KM-1, HC-2, DSPC-5, SPC-8 GMM-10	-
S35	KM-1, HC-2, SOM-4, DSPC-5, MKM-6, KMD-9, GMM-10	-
S36	KM-1, HC-2, SOM-4, MKM-6, KMD-9, GMM-10	-
S37	KM-1, HC-2, MKM-6, SPC-8, KMD-9	-
S38	One method (GMM-10)	KM-1, SOM-4, MKM-6, KMD-9
S39	KM-1, HC-2, SOM-4, MKM-6, KMD-9, GMM-10	-
S40	One method (DSPC-5)	KM-1, SOM-4, MKM-6, KMD-9

Table 2. Illustration of the calculated scores obtained from the determined time windows of individual subjects using the proposed pipeline. The scores involve the temporal properties of the estimated time windows (start and end), inner similarity, amplitude in the Pz electrode, and correlation of mean topography with ground truth topography. TW = time window, Innsim = inner similarity, Amp = mean amplitude, Corr = spatial correlation.

	Rare					Frequent				
Subj-ID	**TW start(ms)**	**TW end(ms)**	**Innsim**	**Amp(μv)**	**Corr**	**TW start(ms)**	**TW end(ms)**	**Innsim**	**Amp(μv)**	**Corr**
S1	382.03	507.03	0.92	11.12	0.71	456.25	550.00	0.93	2.04	0.53
S2	342.97	577.34	0.98	33.37	0.97	350.78	432.81	0.98	16.50	0.95
S3	288.28	612.50	0.87	15.78	0.91	303.91	444.53	0.87	8.86	0.71
S4	280.47	397.66	0.96	17.79	0.89	280.47	452.34	0.91	13.48	0.82
S5	342.97	467.97	0.86	6.42	0.63	573.44	647.66	0.95	5.82	0.38
S6	303.91	487.50	0.89	23.23	0.84	319.53	444.53	0.92	9.98	0.88
S7	296.09	503.13	0.97	20.29	0.52	249.22	479.69	0.92	12.47	0.65
S8	198.44	346.88	0.83	7.52	0.69	202.34	350.78	0.88	4.12	0.34
S9	342.97	444.53	0.84	18.48	0.47	405.47	604.69	0.90	-1.73	-0.12
S10	303.91	366.41	0.86	17.06	0.73	346.88	428.91	0.92	6.20	0.75
S11	421.09	577.34	0.89	6.68	0.74	471.88	573.44	0.89	3.56	0.50
S12	300.00	389.84	0.97	21.60	0.76	296.09	413.28	0.96	11.66	0.89
S13	241.41	417.19	0.93	13.27	0.78	292.19	464.06	0.94	10.89	0.85
S14	311.72	452.34	0.85	9.39	0.64	319.53	409.38	0.93	6.94	0.85
S15	467.97	596.88	0.87	5.56	0.82	499.22	577.34	0.95	4.71	0.89
S16	327.34	471.88	0.95	20.70	0.74	237.50	319.53	0.91	8.96	0.72
S17	323.44	428.91	0.96	7.35	0.78	374.22	491.41	0.90	1.42	0.56
S18	467.97	526.56	0.90	12.58	0.63	428.91	503.13	0.89	6.24	0.51
S19	436.72	635.94	0.96	20.22	0.91	300.00	546.09	0.93	13.88	0.96
S20	495.31	581.25	0.85	5.36	0.73	409.38	522.66	0.88	7.52	0.88
S21	303.91	452.34	0.89	8.37	0.80	346.88	510.94	0.89	4.61	0.72
S22	350.78	499.22	0.92	14.78	0.92	428.91	573.44	0.96	11.32	0.89
S23	436.72	573.44	0.91	12.24	0.68	553.91	647.66	0.96	9.18	0.28
S24	257.03	522.66	0.86	11.63	0.89	315.63	471.88	0.93	7.39	0.78
S25	292.19	385.94	0.92	11.60	0.66	467.97	565.63	0.88	6.63	0.76
S26	346.88	495.31	0.91	15.60	0.81	319.53	491.41	0.94	9.43	0.64
S27	467.97	514.84	0.84	7.19	0.62	346.88	428.91	0.90	8.98	0.95
S28	452.34	772.66	0.90	8.83	0.49	452.34	643.75	0.93	4.60	0.44
S29	362.50	413.28	0.99	14.10	0.77	339.06	499.22	0.87	2.47	0.56
S30	257.03	479.69	0.92	25.06	0.81	350.78	452.34	0.94	10.91	0.80
S31	260.94	335.16	0.94	7.93	0.77	264.84	428.91	0.95	3.35	0.61
S32	440.63	542.19	0.98	15.44	0.77	378.13	534.38	0.98	6.36	0.66
S33	323.44	425.00	0.93	12.39	0.75	276.56	350.78	0.89	2.59	0.58
S34	389.84	620.31	0.96	13.84	0.77	479.69	639.84	0.87	6.72	0.77
S35	245.31	385.94	0.85	14.49	0.77	288.28	358.59	0.90	11.52	0.79
S36	350.78	460.16	0.89	0.14	0.32	491.41	635.94	0.97	0.05	-0.38
S37	405.47	589.06	0.95	11.75	0.80	210.16	331.25	0.91	1.65	0.27
S38	362.50	510.94	0.92	16.37	0.77	296.09	428.91	0.83	10.40	0.84
S39	510.94	573.44	0.88	8.70	0.82	550.00	796.09	0.86	7.72	0.61
S40	346.88	421.09	0.98	19.80	0.85	335.16	483.59	0.93	9.48	0.91
Mean	**351.08**	**495.91**	**0.91**	**13.44**	**0.74**	**366.01**	**498.62**	**0.92**	**7.16**	**0.64**
$\widehat{SD}$	**76.12**	**90.64**	**0.05**	**6.34**	**0.13**	**94.32**	**100.99**	**0.03**	**4.03**	**0.28**

Table S1. The calculated $\widehat{aSME}$ and $\widehat{bSME}$ scores for the determined time windows for the individual subjects. The $\widehat{aSME}$ is obtained from processing all the trials, and the $\widehat{bSME}$ is calculated from the bootstrapping (with replacement).

	Rare				Frequent			
Subject ID	$\widehat{aSME}$ **start(ms)**	$\widehat{aSME}$ **end(ms)**	$\widehat{bSME}$ **start(ms)**	$\widehat{bSME}$ **end(ms)**	$\widehat{aSME}$ **start(ms)**	$\widehat{aSME}$ **end(ms)**	$\widehat{bSME}$ **start(ms)**	$\widehat{bSME}$ **end(ms)**
S1	23.00	28.27	14.29	17.87	15.85	18.01	14.12	16.06
S2	10.86	13.36	8.55	10.77	9.55	10.52	10.50	11.24
S3	26.32	22.93	13.12	14.35	12.49	11.42	12.65	11.64
S4	13.34	20.73	9.63	14.16	11.33	10.24	11.26	10.24
S5	19.58	17.57	15.91	14.30	15.60	15.90	15.48	15.75
S6	83.61	77.94	16.78	15.64	11.69	11.68	11.54	11.66
S7	25.41	23.38	15.22	13.97	10.38	11.09	10.94	10.96
S8	33.85	41.69	17.27	21.28	20.51	21.47	13.90	14.57
S9	16.26	25.94	3.25	5.19	34.71	37.83	16.25	17.70
S10	48.83	29.30	6.90	4.14	15.29	14.04	12.62	11.59
S11	17.04	19.30	12.83	14.47	15.98	18.95	15.80	18.81
S12	19.18	25.48	8.20	14.75	11.37	11.77	11.32	11.86
S13	16.42	17.35	11.21	9.00	8.49	9.16	8.53	9.21
S14	36.64	38.55	13.94	13.81	13.16	17.13	12.12	14.43
S15	20.72	22.79	11.73	12.85	11.99	12.83	12.05	12.88
S16	18.01	28.75	11.70	18.62	8.07	9.87	8.04	9.83
S17	20.99	23.88	14.22	16.19	11.85	13.42	14.98	15.76
S18	18.95	18.96	14.90	15.27	17.17	21.75	11.34	14.50
S19	20.15	27.39	13.32	17.60	10.75	11.24	10.25	11.38
S20	24.60	22.16	15.16	13.66	9.61	10.79	9.57	10.73
S21	18.61	24.85	13.94	18.56	12.39	14.46	12.33	14.43
S22	18.41	34.55	10.83	20.97	8.40	9.34	8.38	9.33
S23	20.46	26.12	12.62	16.09	13.23	16.01	13.14	15.89
S24	18.70	16.43	13.25	11.60	11.33	10.20	11.49	10.12
S25	23.98	29.99	14.33	17.67	14.57	17.18	13.21	15.81
S26	14.41	18.22	9.59	10.17	12.52	12.35	12.78	12.76
S27	14.02	13.63	8.37	8.16	10.77	9.96	10.73	9.91
S28	23.14	37.21	9.83	15.79	15.18	16.29	12.73	13.63
S29	12.94	18.96	8.14	12.02	18.25	17.09	15.27	14.25
S30	21.34	19.96	11.26	11.75	15.11	16.76	13.43	15.13
S31	20.72	21.91	13.44	14.24	12.17	12.09	12.08	11.96
S32	17.59	23.10	13.10	16.80	8.61	10.18	8.55	10.13
S33	11.32	11.63	9.18	9.41	9.51	12.52	9.47	12.48
S34	14.41	23.87	9.55	20.28	13.08	14.03	13.88	15.07
S35	33.34	53.09	14.18	22.62	12.44	19.33	10.60	16.47
S36	20.47	20.03	15.06	15.32	15.25	14.20	12.75	11.88
S37	17.84	18.85	9.56	11.90	11.83	13.02	11.86	13.57
S38	18.45	19.19	12.93	12.99	10.80	10.22	10.76	10.18
S39	19.93	15.01	16.81	13.47	11.07	12.71	12.50	15.45
S40	23.81	38.32	10.64	17.17	16.80	20.71	12.94	15.91
Mean	**22.44**	**25.77**	**12.12**	**14.37**	**13.23**	**14.44**	**12.05**	**13.13**
$\widehat{SD}$	**12.24**	**11.96**	**3.01**	**3.98**	**4.51**	**5.18**	**2.02**	**2.48**
RMS($\widehat{SME}$)	**25.49**	**28.35**	**12.49**	**14.91**	**13.96**	**15.32**	**12.22**	**13.36**

Table S2. The obtained $\widehat{aSME}$ and $\widehat{bSME}$ scores from the calculated mean amplitudes in the estimated time windows for the individual subjects.

	Rare		Frequent	
Subj ID	$\widehat{aSME}$(μv)	$\widehat{bSME}$(μv)	$\widehat{aSME}$(μv)	$\widehat{bSME}$(μv)
S1	2.44	1.51	1.30	1.16
S2	1.22	0.97	0.96	1.06
S3	2.21	1.06	1.39	1.51
S4	1.05	0.73	0.73	0.73
S5	1.62	1.33	1.49	1.47
S6	4.52	0.91	1.09	1.05
S7	1.83	1.09	0.97	0.98
S8	3.76	1.90	1.74	1.18
S9	8.37	1.67	5.56	2.61
S10	0.79	0.11	1.54	1.27
S11	1.31	0.98	0.92	0.91
S12	1.36	0.78	0.80	0.80
S13	1.17	0.74	0.56	0.56
S14	1.62	0.80	1.42	1.14
S15	2.31	1.31	0.80	0.81
S16	2.61	1.69	1.09	1.08
S17	2.10	1.42	1.07	1.14
S18	1.80	1.49	1.80	1.13
S19	1.79	1.45	0.92	0.91
S20	1.45	0.89	0.88	0.87
S21	1.26	0.94	0.91	0.90
S22	1.62	0.90	1.01	1.01
S23	3.90	2.42	2.35	2.33
S24	1.95	1.37	0.94	0.92
S25	2.39	1.59	1.58	1.68
S26	1.93	1.09	0.89	0.97
S27	1.79	1.07	0.99	0.98
S28	1.71	0.73	0.93	0.78
S29	2.40	1.47	2.54	2.12
S30	2.26	1.02	1.96	1.80
S31	1.83	1.18	0.90	0.89
S32	1.52	1.20	0.92	0.92
S33	1.50	1.22	0.98	0.98
S34	1.83	1.32	1.07	1.15
S35	4.98	2.11	2.29	1.94
S36	1.30	1.10	1.19	0.99
S37	2.50	1.72	1.40	1.40
S38	1.58	0.99	1.34	1.33
S39	2.98	2.73	1.72	2.20
S40	3.95	1.76	1.95	1.53
Mean	**2.26**	**1.27**	**1.37**	**1.23**
$\widehat{SD}$	**1.37**	**0.49**	**0.83**	**0.47**
RMS($\widehat{SME}$)	**2.64**	**1.36**	**1.60**	**1.32**

Table S3. Illustration of the $\widehat{aSME}$ and $\widehat{bSME}$ scores of the individual subjects, calculated from the spatial correlation between the mean topography in the estimated time windows and the ground truth topography.

	Rare		Frequent	
Subj ID	$\widehat{aSME}$	$\widehat{bSME}$	$\widehat{aSME}$	$\widehat{bSME}$
S1	0.059	0.038	0.031	0.027
S2	0.013	0.010	0.018	0.020
S3	0.026	0.014	0.018	0.027
S4	0.013	0.009	0.025	0.025
S5	0.061	0.049	0.046	0.044
S6	0.062	0.012	0.028	0.028
S7	0.030	0.018	0.015	0.017
S8	0.049	0.025	0.017	0.011
S9	0.066	0.013	0.029	0.014
S10	0.014	0.002	0.016	0.013
S11	0.039	0.029	0.032	0.032
S12	0.033	0.019	0.015	0.015
S13	0.025	0.014	0.018	0.018
S14	0.043	0.027	0.017	0.016
S15	0.072	0.040	0.017	0.017
S16	0.055	0.035	0.015	0.015
S17	0.059	0.040	0.033	0.021
S18	0.047	0.037	0.042	0.016
S19	0.026	0.017	0.012	0.012
S20	0.032	0.020	0.024	0.024
S21	0.063	0.047	0.037	0.036
S22	0.026	0.013	0.031	0.031
S23	0.042	0.026	0.025	0.025
S24	0.035	0.024	0.017	0.018
S25	0.024	0.018	0.013	0.021
S26	0.024	0.015	0.014	0.020
S27	0.066	0.039	0.032	0.031
S28	0.040	0.017	0.022	0.019
S29	0.029	0.018	0.038	0.032
S30	0.023	0.007	0.036	0.032
S31	0.028	0.018	0.023	0.023
S32	0.037	0.029	0.023	0.022
S33	0.041	0.033	0.031	0.031
S34	0.046	0.028	0.028	0.030
S35	0.111	0.047	0.038	0.032
S36	0.062	0.052	0.046	0.039
S37	0.026	0.012	0.030	0.030
S38	0.034	0.026	0.034	0.033
S39	0.065	0.054	0.036	0.041
S40	0.027	0.012	0.016	0.013
Mean	**0.042**	**0.025**	**0.026**	**0.024**
$\widehat{SD}$	**0.020**	**0.014**	**0.010**	**0.009**
RMS($\widehat{SME}$)	**0.046**	**0.028**	**0.028**	**0.026**

Table S4. Calculated $\widehat{aSME}$ and $\widehat{bSME}$ scores from the inner similarity of time points in the estimated time windows for the individual subjects.

	Rare		Frequent	
Subj ID	$\widehat{aSME}$	$\widehat{bSME}$	$\widehat{aSME}$	$\widehat{bSME}$
S1	0.015	0.006	0.010	0.007
S2	0.010	0.006	0.005	0.005
S3	0.012	0.010	0.009	0.008
S4	0.013	0.007	0.005	0.005
S5	0.008	0.007	0.010	0.012
S6	0.039	0.008	0.006	0.006
S7	0.010	0.005	0.006	0.005
S8	0.014	0.007	0.010	0.005
S9	0.030	0.006	0.016	0.010
S10	0.037	0.004	0.009	0.008
S11	0.009	0.006	0.007	0.006
S12	0.012	0.005	0.006	0.006
S13	0.011	0.007	0.005	0.005
S14	0.019	0.012	0.010	0.008
S15	0.014	0.010	0.006	0.006
S16	0.010	0.006	0.005	0.005
S17	0.010	0.006	0.007	0.008
S18	0.010	0.006	0.012	0.008
S19	0.012	0.008	0.006	0.006
S20	0.013	0.009	0.005	0.005
S21	0.011	0.008	0.006	0.006
S22	0.012	0.005	0.005	0.005
S23	0.012	0.007	0.007	0.005
S24	0.010	0.006	0.006	0.005
S25	0.012	0.008	0.007	0.006
S26	0.012	0.005	0.007	0.005
S27	0.012	0.006	0.006	0.006
S28	0.018	0.004	0.010	0.007
S29	0.013	0.004	0.009	0.006
S30	0.015	0.006	0.008	0.007
S31	0.012	0.007	0.005	0.005
S32	0.009	0.005	0.005	0.005
S33	0.009	0.007	0.005	0.005
S34	0.009	0.007	0.007	0.008
S35	0.019	0.006	0.010	0.007
S36	0.011	0.006	0.011	0.008
S37	0.013	0.008	0.005	0.005
S38	0.010	0.004	0.007	0.007
S39	0.009	0.007	0.008	0.009
S40	0.021	0.008	0.010	0.006
Mean	**0.014**	**0.007**	**0.008**	**0.006**
$\widehat{SD}$	**0.007**	**0.002**	**0.002**	**0.002**
RMS($\widehat{SME}$)	**0.015**	**0.007**	**0.008**	**0.007**

IV

ENSEMBLE DEEP CLUSTERING ANALYSIS FOR TIME WINDOW DETERMINATION OF EVENT-RELATED POTENTIALS

by

Reza Mahini, Fan Li, Mahdi Zarei, Asoke K. Nandi, Timo Hämäläinen, Fengyu Cong
2023

Biomedical Signal Processing and Control, 86, 105202

https://doi.org/10.1016/j.bspc.2023.105202

Biomedical Signal Processing and Control 86 (2023) 105202

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Ensemble deep clustering analysis for time window determination of event-related potentials

Reza Mahini [a], Fan Li [b], Mahdi Zarei [f], Asoke K. Nandi [c], Timo Hämäläinen [a,*], Fengyu Cong [a,b,d,e,*]

[a] Faculty of Information Technology, University of Jyväskylä, Jyväskylä, Finland
[b] School of Biomedical Engineering, Faculty of Electronic and Electrical Engineering, Dalian University of Technology, China
[c] Department of Electronic and Electrical Engineering, Brunel University London, Uxbridge UB8 3PH, UK
[d] School of Artificial Intelligence, Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian, China
[e] Key Laboratory of Integrated Circuit and Biomedical Electronic System, Liaoning Province. Dalian University of Technology, Dalian, China
[f] Department of Bioengineering and Therapeutic Sciences, and Programs in Biological Sciences and Human Genetics, University of California, San Francisco, CA 94158, USA

ARTICLE INFO

Keywords:
Event-related potentials
Time window
Deep clustering
Ensemble learning
Consensus clustering
ERP microstates

ABSTRACT

Objective: Cluster analysis of spatio-temporal event-related potential (ERP) data is a promising tool for exploring the measurement time window of ERPs. However, even after preprocessing, the remaining noise can result in uncertain cluster maps followed by unreliable time windows while clustering via conventional clustering methods.
Methods: We designed an ensemble deep clustering pipeline to determine a reliable time window for the ERP of interest from temporal concatenated grand average ERP data. The proposed pipeline includes semi-supervised deep clustering methods initialized by consensus clustering and unsupervised deep clustering methods with end-to-end architectures. Ensemble clustering from those deep clusterings was used by the designed adaptive time window determination to estimate the time window.
Results: After applying simulated and real ERP data, our method successfully obtained the time window for identifying the P3 components (as the interest of both ERP studies) while additional noise (e.g., adding 20 dB to −5 dB white Gaussian noise) was added to the prepared data.
Conclusion: Compared to the state-of-the-art clustering methods, a superior clustering performance was yielded from both ERP data. Furthermore, more stable and precise time windows were elicited as the noise increased.
Significance: Our study provides a complementary understanding of identifying the cognitive process using deep clustering analysis to the existing studies. Our finding suggests that deep clustering can be used to identify the ERP of interest when the data is imperfect after preprocessing.

1. Introduction

Event-related potentials (ERPs) data is a rich source of information about the cognitive process in the human brain. Information processing units are known as ERP components (i.e., particularly emerge as the ERP peaks). Qualifying ERP of interest for measuring the cognitive process is the key element for reporting results of processing ERP data and testing the research hypothesis. The conventional method for identifying an ERP is to measure the ERP's peak latency or the latency's mean amplitude in the time window measurement interval [30]. The conventional method for selecting the time window is primarily performed via visual inspection for the prominent peak amplitude or obtaining significant differences between the conditions/groups [20,21]. Another popular method is moving time intervals commonly used in different resolutions to find a large effect size [41,50,64]. This method, however, can report the effect size obtained from high-frequency noise as a biased result. The problem with such measurements is that if the underlying assumption for selecting the time window is invalid, analyzing the peak latency, i.e., aiming to detect a larger effect size, can be misleading or result in a problematic estimation of the brain response.

Regardless of the experiment design, various uncertainties can be investigated while processing ERP data. First, there is no available

* Corresponding authors.
E-mail addresses: timo.t.hamalainen@jyu.fi (T. Hämäläinen), cong@dlut.edu.cn (F. Cong).

https://doi.org/10.1016/j.bspc.2023.105202
Received 28 October 2022; Received in revised form 5 May 2023; Accepted 21 June 2023
Available online 2 July 2023

standard aggregated parameter setting for preprocessing methods. For instance, artifact rejection via visual inspection [10] or artifact rejection employing independent component analysis (ICA) [34] and electroencephalography (EEG) referencing methods [66] are commonly uncertain. However, some advanced preprocessing methods implemented in popular software, e.g., FieldTrip [45] and systematic artifact removal methods [16], can somewhat improve data quality. Next, data inconsistency can occur by EEG recording conditions and devices (if more than one), the participants' cortical differences, and the response delay inconsistency in the trials and individual subjects. This can also be associated with the involved number of trials [7]. However, the ERP process (averaging the trials) by itself can somewhat reduce the issue of incoherency between the trials. Finally, new recording technology and devices may require new and more robust analysis method designs.

In recent decades, cluster analysis has emerged as a technical solution used in different aspects of EEG/ERP studies. For example, using fuzzy clustering for class discrimination of evoked potential (EP) waveforms [35] and clustering of principal component analysis (PCA) results [14] for decreasing the effect of noise in the ERP waveforms were introduced. The more conservative methods for ERP identification, reviewed by Kallionpää et al. [18], showed that the cluster-based nonparametric testing method (in the popular FieldTrip software) relies on the temporally adjacent time point (as a cluster) if a significant effect size is identified. Furthermore, spatial clustering, initially introduced by Pascual-Marqui [47], revealed that the brain state called a 'microstate' can be explained as a quasi-stable spatial configuration [27]. Hence, the topographical configuration for a neural response does not change in milliseconds, e.g., 80–120 ms [68].

Two popular cluster analyzing approaches were used for processing spatio-temporal ERP data. First, microstate analysis that assigns the ERP microstates, i.e., represented by global field power (GFP) or GFP maxima, to the template map obtained from clustering the grand average ERP [23,42,62]. The template maps are cluster maps with a high explained variance (e.g., 70% of variance). The post-hoc processing is required, such as smoothing and refining processes based on spatial correlation evaluation if the data is noisy [38]. This method, however, ignores the polarity of time points when assigning clusters. Despite using GFP and the winner-takes-all strategy in determining template maps in the microstates analysis, as argued in some research [11,55], the second group takes whole time points and polarity into account for clustering of spatio-temporal ERP. Recently, we discussed qualifying ERP of interest using consensus clustering as a reliable method for ERP data in different resolutions [31,32,33]. However, cluster analysis of noisy data can result in many noisy clusters and loss of the main components due to being sensitive to the data quality if inappropriate clustering is applied.

Considering the uncertainty of the data, deep learning powered by deep neural networks (DNNs) achieved tremendous success using multiple hidden layers, particularly for EEG data with different designs [3,9,56,67] and our previous work for sleep staging [28]. Roy et al. [52] reviewed a wide range of DNNs used to analyze EEG data. Deep clustering, by definition, is introduced as a method encouraging DNNs to learn cluster-oriented feature representation and clustering. Therefore, DNNs with an embedded clustering module are used with the aim of transforming data points into cluster-friendly representations [2,51]. Yet, two popular strategies have been introduced for deep clustering [2,39]. First, a two-step process in which the DNN is trained to learn initialized labels investigating non-clustering loss (i.e., only the DNN's loss is considered). Then, a clustering method (e.g., *k*-means) is applied to the transformed data in the latent space (i.e., cluster-friendly representation). Another approach uses a jointly training DNN and clustering to optimize clustering and the DNN's weights simultaneously, in which the deep clustering improves the labeling obtained from the clustering layer/module. Deep clustering for brain imaging has been used in some recent studies [49,56]. However, there is little discussion about unsupervised identifying ERP components in the literature.

In this study, we investigate the determinants of the time window of the ERP of interest from spatio-temporal ERP data with various additional noises. We design an ensemble clustering pipeline from two groups of deep clustering methods. Semi-supervised deep clustering methods have been used in which DNN models are trained to learn the labeling that is calculated by state-of-the-art consensus clustering. Unsupervised deep clustering methods are designed via end-to-end DNN architectures for learning the input signal (i.e., with a joint clustering, depending on design). A newly updated time window determination method has been used to qualify ERPs of interest from ensemble clustering results. On the other hand, DNNs are powerful tools for learning nonlinear properties of neuroimaging data and are tolerant to noise and fault [13,36]. This motivates us to apply DNN to learn the most efficient features compared to conventional handcrafted features (with extensive domain expertise). We applied our method to two different ERP data for qualifying P3 components. We demonstrate that the proposed pipeline reliably estimates the time window of the ERP of interest when there is noise in the data after preprocessing.

2. Materials and methods

This section describes the ERP datasets used for testing our method, the proposed method in detail, and the assessment performance metrics.

2.1. ERP data

In order to assess the proposed method, we employed two ERP data, simulated and real data. For the simulated data [31], we test the proposed method against our prior knowledge, i.e., about the spatial and temporal properties of pre-defined ERP components when more noise is added to the data. Likewise, for the real ERP data, we test our method for qualifying the ERP of interest in the prior study [19] when the existing noise in the data increases.

2.1.1. Simulated ERP data

The simulated data was conducted using the 'DipoleSimulator 3.3.0.2′ software from BESA Research (https://www.besa.de). To this end, first, we defined dipoles to generate six pre-defined components (i.e., P1, N1, P2, N2, P3, and N4) and the corresponding data from each component with a simulated scalp containing 65 electrodes for 20 subjects (one group). Once data was generated, the sampling rate of generated ERP data was 214 Hz and epoched from 100 ms pre-stimulus onset to 600 ms post-stimulus via the software (i.e., the dataset size is 65×150). The signal was interpolated to 429 Hz (i.e., dataset size is 65×300) to increase the resolution of the data to provide potentially more isolation accuracy of ERP components. This was done via electrode-wise increasing the original sample rate to a higher rate by inserting zeros into the signal and applying a finite impulse response (FIR) digital interpolation low-pass filter [44,58] to expand the signal. We showed the simulated ERP components' properties (spatial and temporal) and the combined waveform in Fig. 1 for reference.

In order to provide individual ERP data of the subjects, a random resampling interpolation method was applied by increasing the duration of the component with a maximum of 11.5 ms (5 time points $\times$ 2.3 ms), resulting in a new signal (for each dipole). Then a further random shift was performed for each dataset within $\pm$ 4.6 ms. Finally, a combined ERP dataset from two conditions (i.e., 'Cond1′ and 'Cond2′) was prepared using MATLAB code. Noteworthy that an additional strength (subjectively) is applied to some of the components' waveforms (e.g., N2 and P3) to provide a significant difference between conditions. The P3 component in this data refers to the positive response from 266 to 357 ms post-stimulus. Statistical amplitude power differences were measured at CPz/Cz electrode sites.

2.1.2. Real ERP data

The proposed method was applied to the real ERP data, i.e., the active visual oddball task study published by Kappenman et al. [19], to

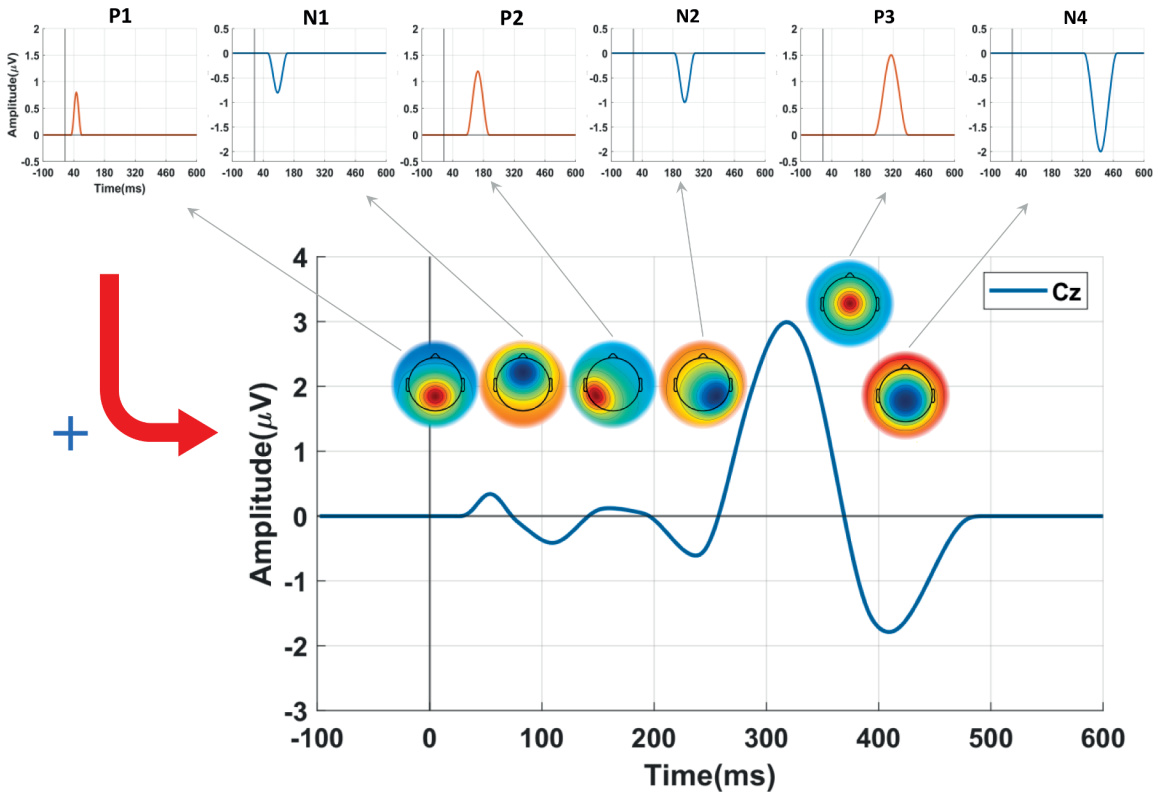


Fig. 1. Illustration of the simulated ERP components, corresponding topographical maps, and the combined waveform (in Cz electrode). The corresponding topographic maps of the pre-defined components are shown with the ERP waveform.

qualify the P3 component. The P3 component refers to the maximum positive peak around 300 to 600 ms (i.e., the rough time window for P3 from the prior study). The EEG data was recorded from 40 participants (i.e., 25 females and 15 males with a mean year of age = 21.5) with 30 scalp electrodes in the international 10/20 system from two conditions, 'Rare' and 'Frequent' letters. The recorded signals were digitized at 1024 Hz (sampling rate), downsampled to 256 Hz for faster processing, and referenced offline the average of P9 and P10. The location electrodes were excluded from processing (i.e., only 28 electrodes were considered). The signals were high-pass and low-pass filtered at 0.1 Hz and 20 Hz and epoched from 200 ms pre-stimulus onset to 800 ms post-stimulus onset by the experimenters. The experimenters extracted approximately 50 to 70 trials for each condition from each subject. The electrode Pz (state-of-the-art electrode) was considered for statistical power analysis following the prior research.

2.2. Our proposed method

Fig. 2 illustrates the proposed method in four steps, data preparation, consensus clustering, ensemble deep clustering, and time window determination. A more detailed explanation of each step and their corresponding role are described as follows:

2.2.1. Data preparation

The temporal concatenation for ERP data [42] was applied to the ERP data from individual subjects. Concatenating was employed along with the conditions for each individual subject of the group. Hence, given N time points from each condition and F scalp electrode (i.e., each condition data size is $F \times N$), the temporal concatenated data is the size of $F \times (N \times C)$ for each individual where C denotes the number of conditions. Then group-wise averaging was performed to be used in clustering analysis. Thus, the temporal concatenated grand average ERP dataset (from two conditions) is the size of 65×600 for the real data, and for the simulated data (from two conditions) is the size of 28×512. Fig. 2**A** demonstrates the temporal concatenating for subjects and the grand average calculation prepared for feeding individual clustering methods. Less formally, the samples for cluster analysis are the time points, and the primary features are the recorded voltage from the scalp electrodes. In order to assess the proposed method, an additional white Gaussian noise (e.g., 20, 10, 5, 0, −5 dB) as a whole (all electrodes' data) is applied to the prepared grand averaged ERP using the MATLAB function *awgn*. This will be in contrast to the assumption that averaging signal from trials/subjects removes noise from the signal for carrying the most powerful ERP responses.

2.2.2. Consensus clustering

The most popular clustering methods for neuroimaging were employed aimed to initialize consensus clustering, including polarity-independent, i.e., after a polarity adjustment to avoid the risk of putting samples with different polarity in the same cluster, and polarity-invariant methods. The clustering methods for the generation phase were selected to provide an appropriate consensus clustering configuration from our toolbox [33], employing the M-N plot method [1]. This approach selects the clustering methods in which the inner-similarity and duration of the obtained time window for a given ERP are appropriate from various clustering options (e.g., repetitive runs on 2 to 20 clusters). Therefore, for the simulated data, *k*-means [48] and hierarchical clustering [60] with correlation similarity function, spectral clustering [43] with *k*-means with Euclidean similarity, and modified *k*-means [46] were selected. Likewise, for the real data, *k*-means and hierarchical clustering with correlation similarity, fuzzy *c*-means (FCM) [6], self-organizing maps (SOM) [24], spectral clustering, and modified *k*-means were selected.

Moreover, the optimal number of clusters was determined by the

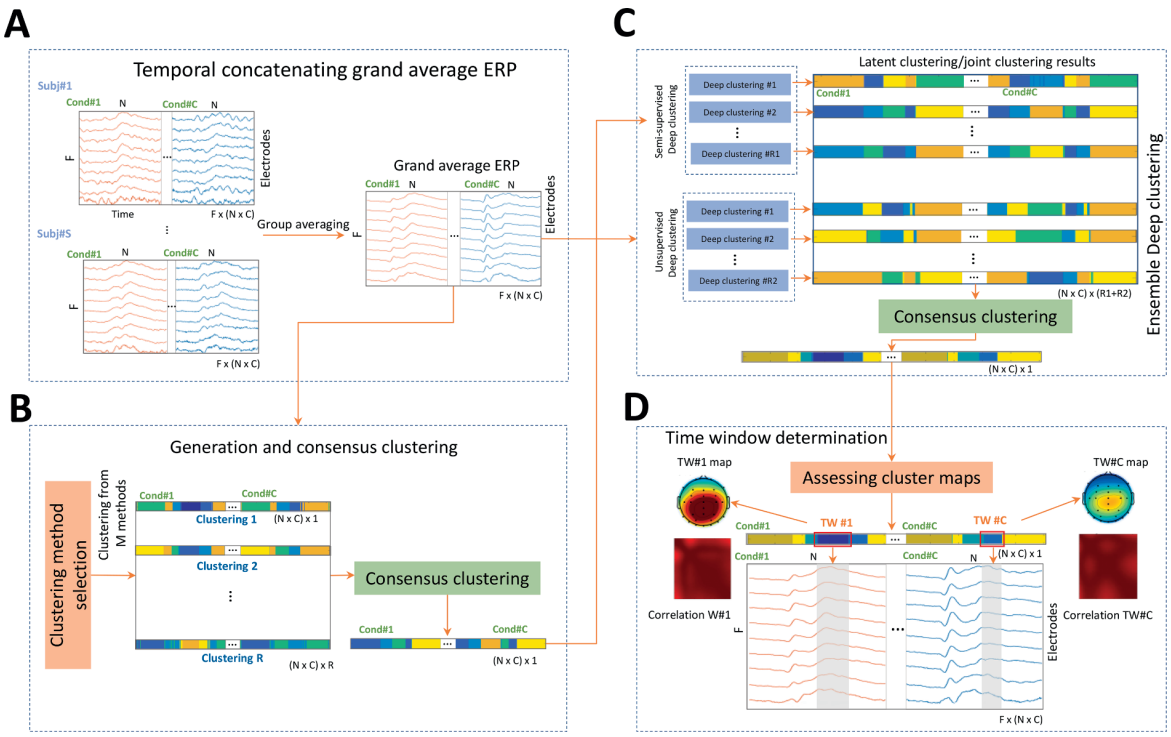


Fig. 2. Proposed pipeline for determining the time window (TW) of the event-related potential (ERP) of interest. **A.** Temporal concatenating data from the C conditions for S subjects and calculating grand average ERP dataset size of $F \times (N \times C)$, where F is the number of electrodes and N denotes the number of time points for each condition dataset. **B.** Selection of the clustering algorithm and generation phase of consensus clustering to initialize semi-supervised deep clustering methods. **C.** Ensemble deep clustering from *R1* semi-supervised and *R2* supervised deep clusterings. **D.** Time window determination from the ensemble clustering result. Subj = subject and Cond = condition.

pipeline following our prior work [33] in which the independent repetitive consensus clustering (e.g., up to 100 runs) is performed on the clustering options (e.g., from 2 to 15 clusters) and the inner-similarity of identified time windows is calculated. Then, the optimal number of clusters is estimated in a clustering option where the qualified time windows reach a high inner-similarity (e.g., > 0.95) and become stable. The inner-similarity of a cluster map can be defined as the mean of correlation coefficients between topographical maps for each of two different time points (except self-correlation). As a result, the optimal number of clusters for the simulated and real data were obtained in 6 and 5 clusters, respectively. Hence, once the clustering results are obtained, we apply cluster-based similarity partitioning (CSPA) that is based on hypergraph clustering [57]to calculate the final clustering result.

Mathematically, let us consider the clustering problem of N samples, $X = \{x_1, x_2, ..., x_N\}$ into K groups, where each group is represented by a centroid μ_k, $k = \{1, 2, \cdots, K\}$ and $x_t \in \mathbb{R}^F$, $t = \{1, 2, \cdots, N\}$ and F denotes the number of features, i.e., the electrodes in the EEG scalp. A set of R clusterings $L^{(1,2,\cdots,R)}$ is used for combining into a result clustering L. Therefore, the objective function for cluster ensemble from R clusterings, a consensus function Γ can be defined as:

$$\Gamma : \left\{L^{(i)} \middle| i \in \{1, 2, \cdots, R\}\right\} \rightarrow L \tag{1}$$

which is a function of $\mathbb{N}^{N\times R} \rightarrow \mathbb{N}^N$ that maps clusterings to a final set of clusters. Given a set of clusterings $\{L^{(i)} | i \in \{1, 2, \cdots, R\}\}$, the goal is to explore the clustering result that shares the most information from all clusterings. The mutual information between two clustering results like L_i, L_j is denoted by $I(L_i,L_j)$,and $H(L_i)$ denotes the entropy of L_i. Hence, the normalized mutual information (NMI), i.e., in the range between 0 and 1, between L_i, L_j using geometric mean can be denoted by:

$$NMI(L_i, L_j) = \frac{I(L_i, L_j)}{\sqrt{H(L_i)H(L_j)}} \tag{2}$$

$$I(L_i, L_j) \leq \min\left(H(L_i)H(L_j)\right), \tag{3}$$

$$H(\widehat{L}) = \sum_{a=1}^{K} N_a \log \frac{N_a}{N} \tag{4}$$

where N_a refers to the number of samples in the cluster C_a according to $\widehat{L}$. Thus, for two clustering results L_i, L_j, the mutual information is calculated as:

$$\Gamma^{(NMI)}(L_i, L_j) = \frac{\sum_{a=1}^{K}\sum_{b=1}^{K} N_{a,b} \log\left(\frac{N.N_{a,b}}{N_a N_b}\right)}{\sqrt{\left(\sum_{a=1}^{K} N_a \log \frac{N_a}{N}\right)\left(\sum_{b=1}^{K} N_b \log \frac{N_b}{N}\right)}} \tag{5}$$

where N_a, N_b present the number of samples in the clusters C_a, C_b according to L_i, L_j, respectively. $N_{a,b}$ refers to the number of samples in cluster a according to C_a as well as in cluster b according to C_b. Thus, the mutual information between r clusterings (Λ) can be defined as the average NMI (ANMI):

$$\Gamma^{(ANMI)}(\Lambda, \widehat{L}) = \frac{1}{R}\sum_{l=1}^{R} \Gamma^{(NMI)}(\widehat{L}, L_l) \tag{6}$$

Therefore, the optimal labeling from r clusterings can be simply defined as:

$$L^* = argmax_{L\in\mathbb{L}} \sum_{l=1}^{R} \Gamma^{(NMI)}(L_l) \tag{7}$$

where Γ denotes a similarity measurement (e.g., NMI), which measures mutual information between a set of R clusterings and L^* is an optimal

combined clustering with maximum average similarity to all other clusterings L_l. Note that the L^* (consensus clustering labeling) has the same size with individual labelings L_l. Notably, the applied CSPA consensus function can calculate the clustering result from non-heterogeneous labeling (i.e., different number of clusters or including missing labels) [57]. As a result, once the clustering labels are assigned via clustering methods (generation phase), the consensus function explores the maximum aggregation between the clusterings.

2.2.3. Deep clustering

Two groups of deep clustering methods were designed. First, the semi-supervised methods, i.e., initialized with a consensus clustering result, were applied to the prepared ERP data (i.e., including additional noise) to obtain cluster-friendly transformed data by learning the K class of clusters. The second group was designed based on the end-to-end autoencoder (AE)-based unsupervised deep clusterings to learn the most powerful features of the data for clustering into K groups. Depending on the deep clustering design, the clustering module was embedded as a layer or linked to be fed by the transformed dataset independently. The following describes general mathematical logic for both groups of designed deep clustering methods.

Let X be the prepared data, e.g., size of N × F from N time points and F electrodes, and $Y = \{y_1, y_2, ..., y_N\}$ denotes the labels obtained via consensus clustering in the dataset. The transforming function can be defined as $S_\varnothing : X \rightarrow Y$, which maps each time point $x_t = \{e_1, e_2, \cdots, e_F\}$ (i.e., a topography map) associated with a label y_t, $t \in 1, 2, \cdots, N$, where $\varnothing$ are the learnable parameters by the network. The role of the deep clustering method is to assign the input space to K clusters $L = \{C_1, C_2, \cdots, C_K\}$, where $C_k = \{x_t | y_t = k, \forall t \in 1, 2, \cdots, N\}$. Therefore, X and Y are defined as input and output spaces, in which input space is transformed with a nonlinear mapping $f_\theta : X \rightarrow Z$ where θ are learnable parameters and Z is embedded feature space, $Z \in \mathbb{R}^K$. Then a parametrized classifier such as g_ω is used to predict the correct labels on top of the features $f_\theta(x_t)$, where the classifier and mapping parameters ω and θ are jointly learned by optimizing the following problem:

$$min_{\theta,\omega} \frac{1}{N} \sum_{t=1}^{N} Loss_{net}(g_\omega(f_\theta(x_t)), y_t) \tag{8}$$

where $Loss_{net}$ is the multinomial logistic loss, also known as the negative log-*softmax* function.

Regardless of the type of applied layers in the semi-supervised methods, the DNN is encouraged to minimize $Loss_{net}$ in order to optimize the prediction of labels (see Eq. (8)). Hence, the input for semi-supervised deep clustering methods is the prepared grand average ERP data order of $\mathbb{R}^{F \times (N \times C)}$, and the output is the order of $\mathbb{N}^K$ (K notes). Thus, the cluster-friendly transformed data (after training) is the size of $(N \times C) \times K$. Next, we apply a stabilized clustering [31] using k-means for fine-tuning and obtaining the clustering result.

For unsupervised deep clustering methods, the DNN optimizes the network knowledge about input signal jointly with a clustering module, i.e., depending on the design, the clustering module can be connected to the bottleneck layer. The *Loss* function usually is the combination of the network and clustering losses, denoted as follow:

$$Loss = Loss_{net} + \gamma Loss_{cl}, \tag{9}$$

where $Loss_{cl}$ denotes the clustering (embedded) loss. γ is a hyper-parameter, which is used to balance the two costs in jointly learning deep clustering method. Note that $Loss_{net}$ for unsupervised methods is defined depending on the learning method and DNNs' structure. Therefore, the reconstruction loss can be easily defined as:

$$min_{\theta_1,\omega_1} L_{rec} = min \frac{1}{N} \sum_{t=1}^{N} \| x_t - g_{\omega_1}(f_{\theta_1}(x_t)) \|^2 \tag{10}$$

where the network is composed of two groups of layers corresponding to the encoder $f_{\theta_1}(.)$ and decoder $g_{\omega_1}()$ with a bottleneck layer(s). The input of the connected clustering module is the encoder's output from the bottleneck layer as the cluster-friendly data. In this definition, the transformed data size for clustering is $(N \times C) \times K$.

2.2.3.1. Design of studied deep clustering methods. The configuration problem of consensus clustering is considered as finding a balance between the selected clustering methods (i.e., called exploration in machine learning) to obtain optimal/sub-optimal combination (i.e., exploitation). In this study, we provided the M-N plot [1 33] to pre-test (see Fig. 3) the studied methods against the different datasets (regarding the noise levels) to avoid trivial results from the individual methods. As a result, the clustering method with a higher risk of obtaining worse candidate cluster maps (cluster maps in the critical area) with an insufficient number of time points and unstable (considering noise in the data) can be eliminated. Although some of the methods achieved less

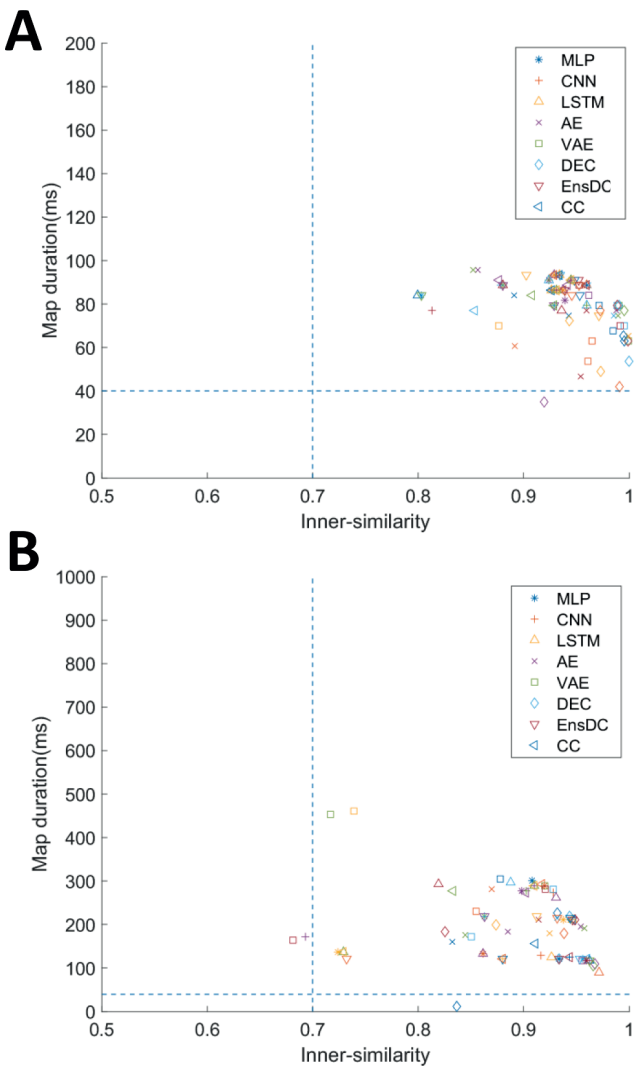


Fig. 3. Illustration of the M–N plot pre-test of the studied methods on the prepared simulated data (**A**) and the real data (**B**) with the additional noises. All the studied deep clustering designs identified cluster maps with high inner similarity (particularly noisy data resulted in lower inner similarity) from both ER data. MLP = fully connected multi-layer perceptron, CNN = one dimensional convolutional deep neural network, LSTM = long short-term memory, AE = autoencoder, VAE = variational autoencoder, DEC = deep embedded clustering, EnsDC = ensemble deep clustering, and CC = consensus clustering.

compact or even fewer samples in candidate cluster maps, i.e., in some of the tests, this does not affect the proposed method's performance as there are no trivial results from the proposed method (Ens_DC) in Fig. 3**A** and Fig. 3**B**. This can also estimate whether those clustering methods are appropriate for given ERP data with additional noise.

We designed six standard deep clustering models from the popular deep clustering designs for sequential data after evaluating them. Therefore, from the semi-supervised methods, we designed: a fully-connected multilevel perceptron (FC-MLP) [39] DNN to learn the essential features of ERP data; a one-dimensional convolutional neural network (1D CNN) [9] to learn complex features of the prepared ERP data; and long short-term memory (LSTM) [26] to learn sequential features of data. We employed a stabilized clustering module via consensus clustering [31] to cluster the transformed data from each DNN.

For the unsupervised group, we designed: an end-to-end AE deep network [40] to use the capability of AE for learning ERP features; variational autoencoder (VAE) [22] for learning input space distribution (e.g., Gaussian distribution) in the latent space; and deep embedded clustering (DEC) [12] to simultaneously learn feature representations and optimize cluster assignments (soft assignment). Likewise, in the first group, a stabilized clustering module was used for clustering transformed data from the encoder's output of AE and VAE based deep clustering methods. Table 1 illustrates the designed blocks of deep clustering models for prepared ERP datasets.

For the DNNs' hyperparameters, we have used the sparse categorical cross-entropy loss function as the net loss and a common *adam* optimizer for supervised and *RMSProp* optimizer for the semi-supervised group with default hyperparameters. Furthermore, the other hyperparameters of all the DNNs models, such as the number of units, the number of layers, the learning rate (e.g., 0.001), batch size (e.g., 150), and the number of iterations (e.g., 100 iterations) were determined by tuning the network using a coarse grid search [5]. Finally, 5-fold cross-validation with mentioned optimizers has been applied to 80 percent of the data for training and validation and 20 percent for the test evaluation. All DNNs were built-in using Keras deep learning libraries.

2.2.3.2. Ensemble deep clustering. Among different strategies for ensemble clustering [8,54], we combined the results of individual methods to calculate the ensemble result from non-heterogeneous elements (i.e., various deep DNNs strategies). Once the results from deep clustering models are obtained, the deep clustering results are fed into consensus clustering (CSPA consensus function) for exploring the most aggregate clustering result. Hence, the consensus clustering at the deep clustering level combines the labeling results from semi-supervised deep clustering methods, AE and VAE from the supervised group, and the clustering result of labeling optimization from DEC (see Fig. 2**C**).

Mathematically, following the same principle to calculate the mutual information in Eq. (7) (in Section 2.2.2), the ensemble clustering can be described as:

$$\widehat{L}^{*} = argmax_{L\in\mathbb{L}} \sum_{l=1}^{\widehat{R}} \Gamma(\widehat{L}_l) \tag{11}$$

where $\widehat{L}^{*}$ is the result of consensus clustering from $\widehat{R} = R1 + R2$ deep clustering methods (i.e., including $R1$semi-supervised clustering and $R2$ unsupervised deep clustering methods) and $\widehat{L}_l$ represents the results from all deep clustering methods. We used the CSPA consensus function, which has suitable tolerance for selecting the number of clusters and the combination of unstable clusters [57]. Therefore, the labeling result from the mentioned three semi-supervised and three unsupervised deep clusterings (i.e., size of 600 × 6 for the simulated data and 512 × 5 for

Table 1
Illustration of designed deep clustering models applied to ERP datasets where N is the number of time points, F is the number of electrodes, and C is the number of conditions. FC_MLP = fully connected multi-layer perceptron, 1D_CNN = one dimensional convolutional deep neural network, LSTM = long short-term memory, AE = autoencoder, VAE = variational autoencoder, DEC = deep embedded clustering, Relu = rectified linear unit, tanh = hyperbolic tangent function, LB = consensus clustering results for feeding the semi-supervised methods. P = current estimation of clustering labels, Q = the previous estimation of the labels, and KL = Kullback-Leibler divergence.

	Semi-supervised			**Unsupervised**		
Deep clustering	**FC_MLP**	**1D_CNN**	**LSTM**	**AE**	**VAE**	**DEC**
Input	(N × C) × F, LB	(N × C) × F, LB	(N × C) × F, LB	(N × C) × F	(N × C) × F	(N × C) × F
Layer 1	FC (64, Relu)	1D_Conv (64, Relu, input), kernel = 1	Lstm (64, Relu)	FC (256, tanh)	FC (125, tanh)	FC (64, Relu)
Layer 2	Batch normalization	1D_Conv (64, Relu), kernel = 1	Batch normalization	Batch normalization	Batch normalization	Batch normalization
Layer 3	Dropout 5%	Max_Pooling_1D	Dropout 5%	Dropout 2%	Dropout 5%	Dropout 5%
Layer 4	FC (512, Relu)	Batch normalization	Lstm (128, Relu)	FC (512, tanh)	FC (256, tanh)	FC (256, Relu)
Layer 5	Batch normalization	Dropout 5%	Batch normalization	Batch normalization	Batch normalization	Batch normalization
Layer 6	Dropout 5%	1D_Conv (256, Relu), kernel = 1	Dropout 5%	Dropout 5%	Dropout 5%	Dropout 5%
Layer 7	FC (256, Relu)	1D_Conv (256, Relu), kernel = 1	FC (128, Relu)	FC (K, Softmax)	FC (256, tanh)	FC (256, tanh)
Layer 8	Batch normalization	Batch normalization	Batch normalization	**Clustering**	Batch normalization	Batch normalization
Layer 9	Dropout 5%	Dropout 5%	Dropout 5%	FC (512 tanh)	Dropout 5%	Dropout 5%
Layer 10	FC (128, Relu)	1D_Conv (64, Relu), kernel = 1	FC (64, Relu)	Batch normalization	Z(z_mean(K), z_log_var(K))	**Clustering Layer (KL-divergence (P,Q))**
Layer 11	Batch normalization	1D_Conv (64, Relu), kernel = 1	Batch normalization	Dropout 5%	Lambda (sampling)	FC (256, Relu)
Layer 12	Dropout 5%	Global_average_pooling_1D	Dropout 5%	FC (256, tanh)	**Clustering**	Batch normalization
Layer 13	FC (K, Softmax)	Batch normalization	FC (K, Softmax)	Batch normalization	FC (256, tanh)	Dropout 5%
Layer 14	**Clustering**	FC (K, Softmax)	**Clustering**	Dropout 5%	Batch normalization	FC (256, Relu)
Layer 15		**Clustering**			Dropout 5%	Batch normalization
Layer 16					FC (256, tanh)	Dropout 5%
Layer 17					Batch normalization	FC (64, Relu)
Layer 18					Dropout 5%	Batch normalization
Layer 19					FC (128, tanh)	Dropout 5%
Layer 20					Batch normalization	
Layer 21					Dropout 5%	

the real data) was achieved in a firm aggregate labeling of 600 and 512 time points in the concatenated simulated and the real ERP data, respectively.

2.2.4. Time window determination

We modified the previously designed time window determination method [31] to provide more flexibility in the inner similarity and duration thresholds of candidate cluster maps. It should be noted that the time window determination method requires experimental information about the ERP of interest. This means that considering the experimental design (e.g., visual and auditory) and participant group (e.g., age, sex, and health level), a rough expectation (at least based on stimulation) of some neurological brain response (e.g., attention, memory, and mismatch components) is approachable. Therefore, the stimulation onset/offset time, the target response, and the electrode site are expected. The adaptive time window adjusts the inner similarity threshold (e.g., $0.7 \leq$ minimum inner-similarity ≤ 0.95) and the consecutive number of time points in the candidate cluster maps (e.g., 30 ms $\leq$ minimum number of time points $\leq$ 50 ms) while needed. In other words, the time window determination method starts from the highest possible inner similarity with sufficient duration and applies a silent change (e.g., 0.003 for inner similarity and 2 ms for the duration of the map, which can be adjusted when needed) if no suitable representative map is found.

2.3. Performance analysis

2.3.1. Evaluation metrics

We applied the popular performance evaluation metrics, namely, accuracy (ACC) [65], NMI [57], and adjusted rand index (ARI) [37], to assess the performance of clustering methods. Hence, given the known clustering L (i.e., ground-truth) and the clustering result L', the accuracy index can be defined as:

$$ACC(L, L^{'}) = max \frac{\sum_{i=1}^{N} 1\{L(i) = m(L^{'}(i))\}}{N}, \tag{12}$$

where m provides overall possible one-to-one mappings between clusters and labels using the Hungarian algorithm [25]. It is, however, not possible to get the same label for the given similar clusters from multiple clustering methods (i.e., different labels might be generated for the same cluster in multiple runs or various methods). Therefore, the Rand index (RI), as a suitable index to compare clustering results, can be defined as:

$$\mathscr{R}(L, L^{'}) = \frac{TP + TN}{TP + TN + FP + FN}, \tag{13}$$

where *TP*, *TN*, *FP*, and *FN* signify true positive, true negative, false-positive, and false-negative rates, respectively. By calculating the expectation of *R*, i.e., $E[R]$, the adjusted rand index (*ARI*) is calculated as:

$$ARI(L, L^{'}) = \frac{R(L, L^{'}) - E[\mathscr{R}]}{1 - E[\mathscr{R}]}. \tag{14}$$

Besides, mutual information provides a suitable concept of the shared information between a pair of clusterings as an asymmetric measure to quantify the statistical information shared between two distributions [59], which we have defined in subsection 2.2.2. Another reasonable index called adjusted mutual information (AMI) for NMI [63] is used by calculating the NMI expectation as the following:

$$AMI(L, L^{'}) = \frac{I(L, L^{'}) - E\{I(L, L^{'})\}}{\max\{H(L), H(L^{'})\} - E\{I(L, L^{'})\}} \tag{15}$$

The ground truth clustering for the comparison using those metrics mentioned above is the clustering results on the prepared grand average ERP data using state-of-the-art consensus clustering when no additional noise is applied. The clustering performance of the studied clustering methods is assessed from the data of the different additional noise.

2.3.2. Statistical analysis

We provided a standard analysis of variances to determine whether the identified P3 effect (from each ERP data) is statistically significant. For the simulated data, statistical analysis was carried out via a repeated measures analysis of variance (rmANOVA) with a within-subject factor: *Task* (conditions: 'Cond1′ and 'Cond2′) in two pre-defined electrode sites CPz and Cz. This was performed by measuring the mean voltage of the P3 amplitude in the determined time window. The effect of the *Task* was tested against the null hypothesis of existing no significant difference between the conditions from those selected electrodes and the estimated time windows. Likewise, statistical power analysis for real data was carried out via a rmANOVA with a within-subject factor: *Task* (conditions: 'Rare' and 'Frequent') by measuring the mean amplitude of P3 on the priory selected electrode site over the parietal region (electrode: Pz). We tested the effect of the *Task* for the hypothesis that a significant difference exists between the 'Rare' and 'Frequent' conditions at the selected electrode site and time windows. Statistical comparisons were made at p-values of $p < 0.05$ for both data.

3. Results and evaluations

The summarized results of applying the proposed pipeline to two ERP data are illustrated, including the performance of each DNN, clustering results, estimated time windows, and statistical analysis results from different noise levels. Furthermore, we present the performance results based on the defined metrics.

3.1. Performance of the studied DNNs

Table 2 and Table 3 show the training performance of the studied DNNs on the test datasets from the simulated and real data, respectively. Observing the results in Table 2 and Table 3 discloses that for semi-supervised DNNs, the DNNs are able to learn the ERP data and the labeling depending on the DNNs' structures with high accuracy, even for noisy data. The unsupervised DNN models, on the other hand, have learned the input space properties with relatively worse *Loss* rates than the semi-supervised methods. Together, from the results of both datasets, the designed DNNs successfully trained on the prepared ERP data from additional noises.

3.2. Clustering results and time windows

Fig. 4 shows the clustering results (randomly selected) from the proposed and state-of-the-art methods (consensus clustering) in the simulated data, i.e., when no additional noise exists and the maximum reasonable noise is added (e.g., −5 dB). We excluded results from datasets with additive noise between them to keep the figure readable. The qualified cluster maps for identifying the interesting ERP were marked in gray color for both Fig. 4 and Fig. 5. Observing Fig. 4**A,** i.e., results of consensus clustering in the prepared simulated data without noise, shows that the time windows for the P3 component are isolated with cluster maps 5 (colored gray areas) from 268.67 to 355.00 ms for 'Cond1′ and 268.67 to 362.00 ms (ground-truth) for 'Cond2′. Similarly, in Fig. 4**B** (i.e., the proposed method results), those time windows have been elicited by maps 4 in the identical time windows, i.e., in 268.67 to 355.00 ms and 268.67 to 362.00 ms for 'Cond1′ and 'Cond2′, respectively. P3 was isolated for highly noisy simulated data by maps 6 from 273.33 to 350.33 ms for 'Cond1′ and 271.00 to 355.00 ms for 'Cond2′ using consensus clustering (see Fig. 4**C**). The proposed method extracted P3 by maps 4 from 273.33 to 352.67 ms and 268.67 to 357.33 ms for 'Cond1′ and 'Cond2′ (see Fig. 4**D**) for the noisy simulated data. Noticeably, a larger peak was observed (in the determined time windows) in 'Cond1′ than in 'Cond2′ when no noise was added and from the maxima noisy datasets, from the clustering results via two methods.

Observing Fig. 5 (a randomly selected result), for the real data with no additional noise, P3 was isolated by map 1 and map 2 from 303.90 to

Table 2
The studied DNNs' performances (on the test dataset) in the simulated data while additional noise is included on the original signal (i.e., from 20 dB to −5 dB). acc = accuracy, SNR = signal-to-noise ratio. The SNR value denotes the additive white Gaussian noise in the prepared ERP signal.

Method	No noise added		SNR = 20 dB		SNR = 10 dB		SNR = 5 dB		SNR = 0 dB		SNR = -5 dB	
	loss	acc	loss	acc	loss	acc	loss	acc	loss	acc	loss	acc
FC_MLP	0.002	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000
1D CNN	0.002	1.000	0.000	1.000	0.000	1.000	0.001	1.000	0.000	1.000	0.001	1.000
LSTM	0.002	1.000	0.003	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000
AE	0.002	–	0.004	–	0.003	–	0.004	–	0.003	–	0.003	–
VAE	0.016	–	0.011	–	0.032	–	0.037	–	0.040	–	0.042	–
DEC	0.046	–	0.056	–	0.050	–	0.050	–	0.053	–	0.060	–

Table 3
The studied DNNs' performances on the test dataset for the real data when the additive noise increases from 20 dB to −5 dB.

Method	No noise added		SNR = 20 dB		SNR = 10 dB		SNR = 5 dB		SNR = 0 dB		SNR = -5 dB	
	loss	acc	loss	acc	loss	acc	loss	acc	loss	acc	loss	acc
FC_MLP	0.003	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000
1D CNN	0.001	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000
LSTM	0.004	1.000	0.001	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000
AE	0.012		0.015	–	0.016	–	0.021		0.026		0.029	–
VAE	0.022	–	0.035	–	0.035	–	0.044		0.040		0.050	–
DEC	0.016	–	0.024	–	0.019	–	0.034		0.042		0.040	–

514.84 ms and 342.97 to 464.06 ms (ground-truth), in condition (1) and condition (2), respectively, using the consensus clustering (Fig. 5**A**). Those time windows for the P3 component were elicited by map 1 and map 2 (Fig. 5**B**), in 303.90 to 514.84 ms and 342.97 to 460.16 ms for conditions (1) and (2), respectively, using the proposed method. While a high noise (SNR = -5 dB) was added in the real data, P3 was isolated by maps 1 from 331.25 to 589.06 ms and 342.97 to 428.91 ms in conditions (1) and (2), respectively, using consensus clustering. Whereas those time windows were elicited by map 1 and map 2 (colored gray areas), from 296.09 to 514.84 ms and 354.69 to 475.78 ms, respectively, using the proposed method. Together, the clustering results for both ERP data (the simulated and real) with different amounts of additive noise revealed that the clusterings include noisier clusters, particularly where no strong response exists (e.g., pre-stimulus onset). Observably, the proposed method seems to provide a more robust clustering result (Fig. 4**D** and Fig. 5**D**) than consensus clustering (Fig. 4**C** and Fig. 5**C**).

We provided detailed results of the estimated time windows (start, end, and duration) of the P3 identification from the studied method in different noise levels in Table 4 and Table 5**.** For simulated data (see Table 4), all the studied methods identified P3 response with some degree of accuracy. However, consensus clustering obtained better time window accuracy, especially in real data when adding non-intensive noise. Semi-supervised deep clustering obtained suitable identification due to supervising by consensus clustering. The proposed method obtained more accurate and stable results among different methods. Likewise, for the real data (see Table 5), the time window of the P3 response was identified from the clustering results of all the methods studied. Notably, the proposed method achieved more accuracy and stability than other studied methods.

Furthermore, Table 6 illustrates the standard deviation error of the estimated time windows from the different clustering methods in the examined datasets with additional noises. Together, the time window determination and the stability evaluation results reveal that all the deep clustering methods successfully identified P3 from different datasets. Our method performed better in the real data than in the simulated data, which was relatively better than other methods.

To provide more evidential results and test the spatial properties of qualifying the isolated ERPs of interest, we examined the spatial correlation between the mean topography maps in the ground-truth time window and the time windows from the studied methods. We included the spatial correlation test results in Table 7 and Table 8 in the simulated and the real data, respectively. The results showed a high spatial correlation between the identified ERP and the ground truth P3 topographical map from the majority of the studied methods. However, a silently less correlation was obtained in low SNR due to the noise effect on topography configuration.

3.3. Evaluation and statistical analysis results

Fig. 6**A** and Fig. 6**B** show the clustering performance based on the performance metrics (ACC, ARI, and AMI) for simulated and real data, respectively. We have included the performance of the studied deep clustering methods for a better comparison and understanding of their role in the proposed clustering. Observing Fig. 6 exposes a suitable performance from the studied deep clustering methods and, consequently, the ensemble deep clustering (proposed method). On the other hand, consensus clustering results reveal a suitable performance with comparatively better stability while the noise ratio is changed. Observing Fig. 6**A** indicates that, except for the clean data, the proposed method provides a more confident performance than the studied clustering methods, especially consensus clustering in the simulated data. Likewise, Fig. 6**B** shows that the proposed method obtained remarkable and stable results while the data noise varied in the real data. Noticeably, except for the ground truth in the simulated data and corresponding semi-supervised methods' performance, the proposed method discloses a relatively superior and stable performance for both datasets.

For the studied methods, the statistical analysis results of the elicited P3 effects from the estimated time windows (see Table 4 and Table 5) were illustrated in Table 9 and Table 10 for the simulated and real ERP data, respectively. For the simulated data, our results revealed a large P3 effect and a significant difference between the conditions ($F(1,19) = 81317$, p -value < 0.0001, $\eta_p^2 = 1.000$) in the region of the interest (the central area) and the obtained time windows when no additional noise exists. However, the obtained large effect and calculated highly significant difference between the conditions can be seen to be overestimated. This can be because of the occurred alignment in the subjects' responses in the peak/mean amplitude (i.e., due to being calculated in the same ratio with silent changes) from the simulation mechanism when there is no additional noise. Hence, a larger response was identified in 'Cond1′ than in 'Cond2′, which was expected following the simulation mechanism. Observing Table 9 reveals a silent decrease in the obtained effect size while the data are noisier. Nevertheless, regardless of the noise

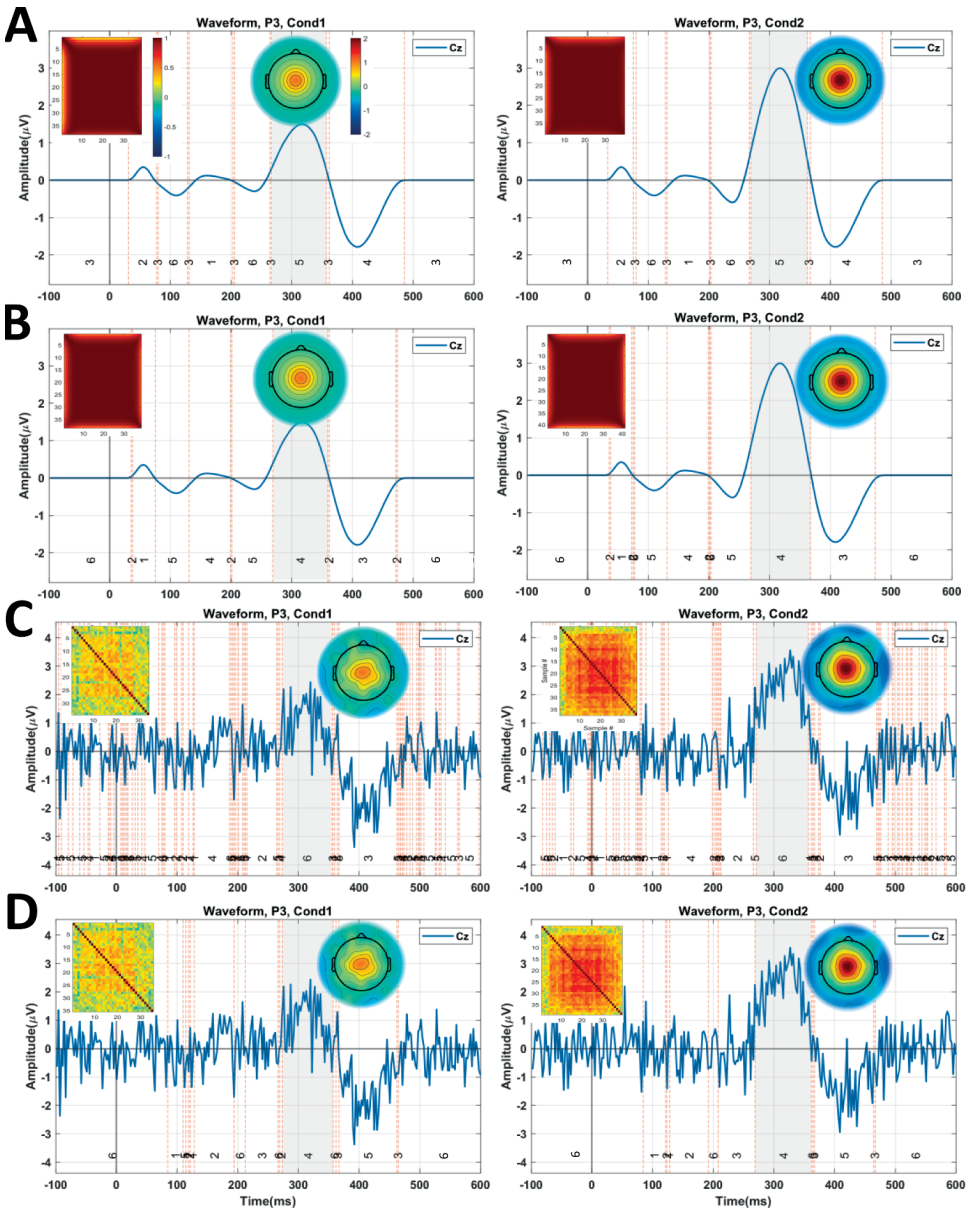


Fig. 4. Illustration of clustering results and selected time windows (colored gray areas), including the corresponding topographies and correlation between the time points for identifying the P3 component in the simulated data. **A.** Isolated time windows with maps 5 (cluster maps 5) in Cond1 and Cond2 from consensus clustering result when no additional noise is added. **B.** Identified time windows by maps 4 in both conditions from the proposed method clustering result when no additional noise is added. **C.** Identified time windows with maps 6 (in both conditions) from consensus clustering in maximum additional noise of −5 dB. **D.** Isolated time windows with maps 4 (in both conditions) of the proposed method when the additional noise is −5 dB. The numbers for each segment present the associated cluster map's number. Cond1 = condition (1), Cond2 = condition (2).

level, a large effect size was obtained from the majority of the studied clustering methods in the estimated time windows.

For real data, our results confirmed the previous findings on the main effect of the *Task*, indicating existing large effect size and significant difference ($F(1,39) = 121.18$, p-value < 0.0001, $\eta_p^2 = 0.76$) between conditions that was larger amplitude in the 'Rare' condition (target) than'Frequent' condition (non-target) in the central lobe of the occipital region. Table 10 shows the results of the statistical power analysis from different noise levels for the studied methods. Similar to the simulated data, all the studied methods identified a significant effect of P3 in the estimated time windows.

4. Discussion

This study presents the ensemble deep clustering pipeline to reliably determine the time window of the ERP of interest when existing noise on the ERP data is unknown after preprocessing. To tackle the problem, we designed the ensemble clusterings from multiple deep clustering methods, including semi-supervised and unsupervised, to explore the most aggregated clusters in the data. To this end, we clustered the weighted data from the trained DNNs (in the latent space), namely, FC_MLP, LSTM, 1DCNN, AE, and VAE, for fine-tuning. Then, those clusterings and DEC's results were combined using the consensus function to calculate the final clustering result. Finally, the modified time window determination was used for estimating the ERP of interest from the candidate cluster maps in each condition/group. The proposed pipeline was built on three methodologies, cluster analysis of spatio-temporal ERP, deep learning as a powerful and noise-tolerant tool, and ensemble learning.

The idea of ensemble clustering in this study is that the scalp EEG data recorded from the same or different devices, multiple-subject (e.g., age, brain size, healthy level) from different conditions/groups carries various artifacts even after preprocessing affects the quality of data. On the other hand, considering the fact that even the popular clustering algorithms could fail spectacularly for certain datasets that do not match the corresponding modeling assumptions (Acharya and Ghosh, 2011),

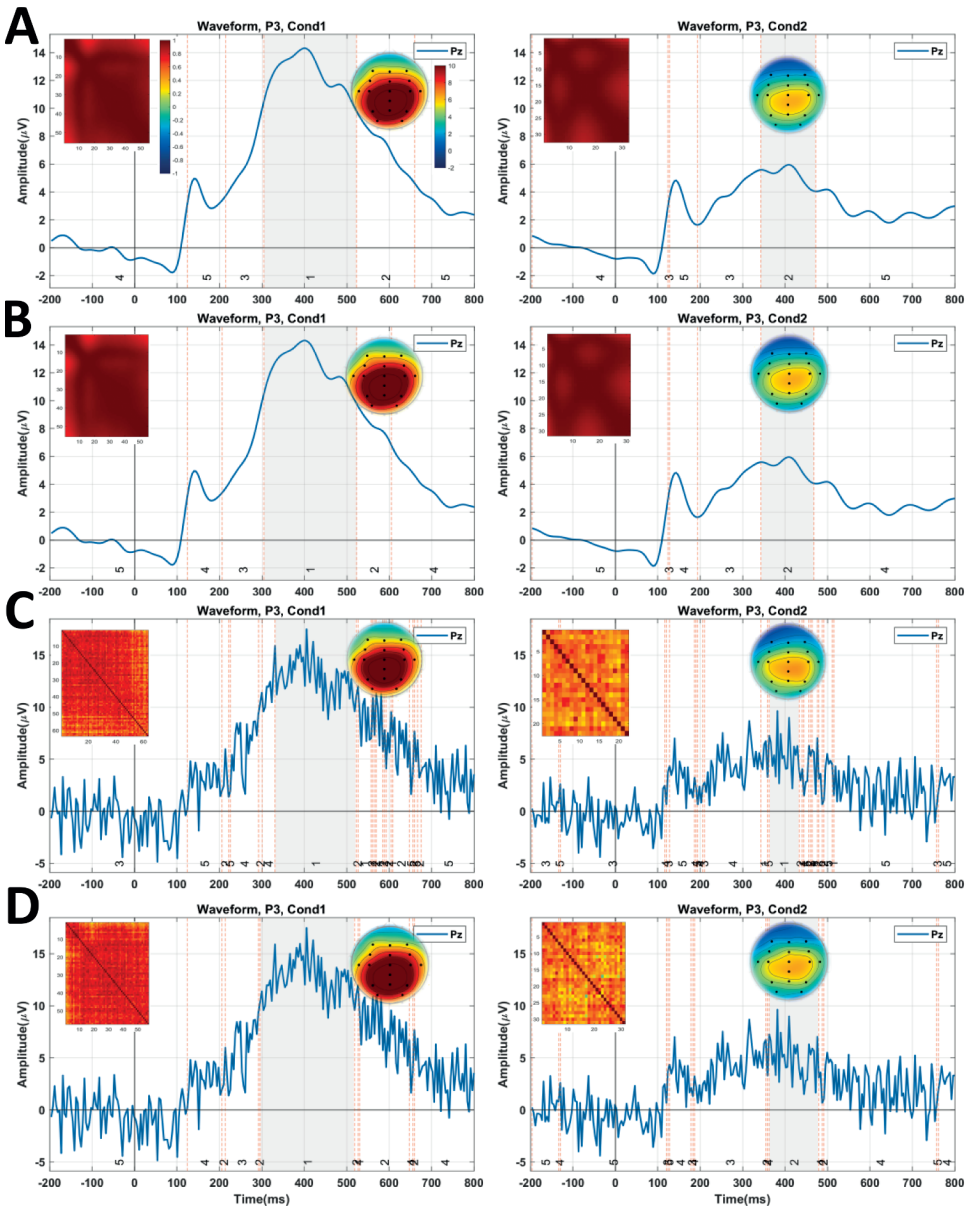


Fig. 5. Illustration of clustering results and selected time windows (colored gray areas) from the proposed and consensus clustering methods for identifying the P3 component in the real data. The results of each condition include the corresponding topographies and the correlation between the time points for the determined time window. **A.** Isolated time windows by maps 1 and 2 for Cond1 and Cond2, respectively, using consensus clustering when no additional noise is applied. **B.** Identified time windows by maps 1 and 2 in Cond1 and Cond2, respectively, using the proposed method when no additional noise is applied. **C.** Identified time windows with maps 1 (in both conditions) by consensus clustering in maximum additional noise of −5 dB. **D.** Isolated time windows with maps 1 and 2 Cond1 and Cond2 by the proposed method when the additional noise is −5 dB, respectively.

using a more reliable clustering method is suggested. In addition, clustering noisy or unbalanced EEG/ERP tensors considering the only spatial properties can result in unreliable cluster maps (i.e., for qualifying the ERP components) since numerous small peaks can be recognized as brain responses [38]. As a result, although available clustering techniques besides ICA/PCA provided a more reliable decomposition of ERP of interest, more challenging data can lead to a problematic result (e.g., determination of divided component, missing ERP). These problems can be more severe if inappropriate preprocessing is performed.

One important issue with ensemble learning methods is the configuration consistency for such a combination. Although this mechanism eliminates the contribution of trivial results, it cannot guarantee the optimization of ensemble clustering. Considering the fact that there is no straightforward solution for the configuration of ensemble clustering [61] and existing a large variety of deep clustering designs [51], we provided an M-N plot pre-test of the studied methods against the different datasets (in terms of noise level) to avoid using deep clustering methods with trivial results. Noting that we avoided testing sophisticated deep clustering designs to keep our design implementable and understandable at this stage. Our early findings revealed three important characteristics of using DNNs for training ERP data. First, the studied DNNs are powerful learners in learning ERP data even when data is considerably noisy. Next, the studied individual deep clustering methods result in clustering in which the interesting components in two datasets and other few components (e.g., N4 in the simulated data) can be identified using the time window determination method. Finally, the ensemble deep clustering provides stable performance compared to other methods associated with the proposed ensemble deep clustering tolerance to artifacts (particularly with noise) without compromising the performance.

The advantages of the proposed method compared to conventional methods are: **i)** using the minimum amount of knowledge in the designed deep clustering methods; **ii)** exploring a firm clustering model for spatio-temporal ERP data by ensemble deep clustering results; **iii)** designing the adaptive time window determination, considering the spatial and temporal properties, from noisy data; **iv)** obtaining the

Table 4
The temporal properties of the estimated time windows (start, end, and duration) through the proposed method, the consensus clustering, and the studied deep clustering methods to qualify the P3 component in the prepared simulated data with different additive noises. The bold marks represent the significant results. Ens_DC = ensemble deep clustering (proposed method), CC = consensus clustering, Cond1 = condition (1), and Cond2 = condition (2).

Method	Properties(ms)	No noise		SNR = 20 dB		SNR = 10 dB		SNR = 5 dB		SNR = 0 dB		SNR = -5dB	
		Cond1	Cond2	Cond1	Cond2	Cond1	Cond2	Cond1	Cond2	Cond1	Cond2	Cond1	Cond2
Ens_DC	Start	**268.67**	**268.67**	**268.67**	**268.67**	**268.67**	**268.67**	**268.67**	**268.67**	**273.33**	**268.67**	**273.33**	**268.67**
	End	**355.00**	**362.00**	**355.00**	**362.00**	**355.00**	**362.00**	**355.00**	**359.67**	**352.67**	**362.00**	**352.67**	**357.33**
	Duration	**86.33**	**93.33**	**86.33**	**93.33**	**86.33**	**93.33**	**86.33**	**91.00**	**79.33**	**93.33**	79.33	**88.67**
CC	Start	**268.67**	**268.67**	**268.67**	**268.67**	**268.67**	**268.67**	273.33	**268.67**	**273.33**	271.00	**273.33**	**268.67**
	End	**355.00**	**362.00**	**355.00**	**362.00**	**355.00**	**362.00**	**355.00**	**359.67**	**352.67**	357.33	350.33	**357.33**
	Duration	**86.33**	**93.33**	**86.33**	**93.33**	**86.33**	**93.33**	81.67	**91.00**	**79.33**	86.33	77.00	**88.67**
MLP_FC	Start	266.33	**268.67**	**268.67**	**268.67**	**268.67**	**268.67**	273.33	**268.67**	**273.33**	**268.67**	**273.33**	**268.67**
	End	352.67	357.33	**355.00**	**362.00**	**355.00**	**362.00**	**355.00**	**359.67**	**352.67**	359.67	**357.33**	**357.33**
	Duration	**86.33**	88.67	**86.33**	**93.33**	**86.33**	**93.33**	81.67	**91.00**	**79.33**	91.00	**84.00**	**88.67**
1DCNN	Start	266.33	**268.67**	**268.67**	**268.67**	**268.67**	**268.67**	273.33	**268.67**	**273.33**	**268.67**	280.33	**268.67**
	End	352.67	357.33	**355.00**	**362.00**	**355.00**	**362.00**	352.67	**359.67**	**352.67**	359.67	**357.33**	**357.33**
	Duration	**86.33**	88.67	**86.33**	**93.33**	**86.33**	**93.33**	79.33	**91.00**	**79.33**	91.00	77.00	**88.67**
LSTM	Start	266.33	**268.67**	**268.67**	**268.67**	**268.67**	**268.67**	273.33	**268.67**	275.67	**268.67**	**273.33**	**268.67**
	End	352.67	357.33	**355.00**	**362.00**	**355.00**	**362.00**	352.67	**359.67**	**352.67**	359.67	**357.33**	**357.33**
	Duration	**86.33**	88.67	**86.33**	**93.33**	**86.33**	**93.33**	79.33	**91.00**	77.00	91.00	**84.00**	**88.67**
AE	Start	273.33	287.33	273.33	271.00	282.67	**268.67**	273.33	271.00	**273.33**	**268.67**	**273.33**	**268.67**
	End	350.33	341.00	348.00	357.33	343.33	352.67	348.00	355.00	348.00	357.33	345.67	**357.33**
	Duration	77.00	53.67	74.67	86.33	60.67	84.00	74.67	84.00	74.67	88.67	72.33	**88.67**
VAE	Start	280.33	273.33	275.67	273.33	285.00	271.00	278.00	266.33	280.33	273.33	275.67	266.33
	End	343.33	350.33	345.67	352.67	341.00	352.67	345.67	357.33	343.33	352.67	**352.67**	**357.33**
	Duration	63.00	77.00	70.00	79.33	56.00	81.67	67.67	**91.00**	63.00	79.33	77.00	**91.00**
DEC	Start	287.33	275.67	280.33	273.33	287.33	275.67	278.00	**268.67**	290.00	273.33	294.33	278.00
	End	341.00	352.67	343.33	352.67	338.67	350.33	345.67	357.33	341.00	350.33	329.33	350.33
	Duration	53.67	77.00	63.00	79.33	51.33	74.67	67.67	88.67	51.00	77.00	35.00	72.33

Table 5
The temporal properties of estimated time windows via the proposed method, consensus clustering, and the studied deep clustering methods in the real data to qualify the P3 component for different additional noise. The bold marks are the significant results.

Method	Properties(ms)	No noise		SNR = 20 dB		SNR = 10 dB		SNR = 5 dB		SNR = 0 dB		SNR = -5dB	
		Cond1	Cond2	Cond1	Cond2	Cond1	Cond2	Cond1	Cond2	Cond1	Cond2	Cond1	Cond2
Ens_DC	Start	**303.91**	**342.97**	**303.91**	339.06	300.00	**342.97**	**300.00**	**346.88**	**303.91**	**342.97**	296.09	354.69
	End	**514.84**	460.16	**514.84**	**460.16**	**514.84**	**464.06**	**518.75**	**467.97**	522.66	**464.06**	**514.84**	**475.78**
	Duration	210.94	117.19	**210.94**	**121.09**	**214.84**	**121.09**	**218.75**	**121.09**	**218.75**	**121.09**	**218.75**	**121.09**
CC	Start	**303.91**	**342.97**	**303.91**	**342.97**	303.91	**342.97**	315.63	**346.88**	**303.91**	**342.97**	331.25	**342.97**
	End	**514.84**	**464.06**	**514.84**	491.41	592.97	**464.06**	592.97	**467.97**	600.78	479.69	589.06	428.91
	Duration	**211.94**	**122.09**	**210.94**	148.44	289.06	**121.09**	277.34	**121.09**	296.88	136.72	257.81	85.94
MLP_FC	Start	292.19	335.16	**303.91**	**342.97**	**303.91**	**342.97**	311.72	**346.88**	331.25	**342.97**	307.81	346.88
	End	596.88	460.16	565.63	444.53	604.69	503.13	581.25	**467.97**	608.59	456.25	510.94	440.63
	Duration	304.69	125.00	261.72	101.56	300.78	160.16	269.53	**121.09**	277.34	113.28	203.13	93.75
1D CNN	Start	292.19	335.16	**303.91**	335.16	**303.91**	339.06	**300.00**	335.16	315.63	**342.97**	288.28	**342.97**
	End	592.97	460.16	565.63	432.81	592.97	**464.06**	592.97	**467.97**	600.78	483.59	510.94	440.63
	Duration	300.78	125.00	261.72	97.66	289.06	125.00	292.97	132.81	285.16	140.63	222.66	97.66
LSTM	Start	296.09	335.16	**303.91**	**342.97**	292.19	339.06	**300.00**	**339.06**	315.63	**342.97**	307.81	**342.97**
	End	592.97	460.16	596.88	460.16	592.97	**464.06**	577.34	**467.97**	608.59	479.69	510.94	440.63
	Duration	296.88	125.00	292.97	117.19	300.78	125.00	277.34	128.91	292.97	136.72	203.13	97.66
AE	Start	307.81	346.88	311.72	**342.97**	300.00	335.16	**300.00**	**339.06**	**303.91**	339.06	307.81	346.88
	End	503.13	487.50	503.13	526.56	**514.84**	510.94	**518.75**	503.13	526.56	460.16	507.03	507.03
	Duration	195.31	140.63	191.41	183.59	**214.84**	175.78	**218.75**	164.06	222.66	**121.09**	199.22	160.16
VAE	Start	284.38	335.16	311.72	331.25	300.00	**342.97**	**300.00**	**346.88**	**303.91**	**342.97**	**303.91**	350.78
	End	538.28	514.84	612.50	514.84	596.88	518.75	**518.75**	518.75	542.19	514.84	600.78	440.63
	Duration	253.91	179.69	300.78	183.59	296.88	175.78	**218.75**	175.78	238.28	171.88	296.88	89.84
DEC	Start	303.91	346.88	307.81	346.88	**303.91**	**342.97**	315.63	331.25	311.72	401.56	311.72	311.72
	End	522.66	456.25	518.75	452.34	530.47	**464.06**	495.31	514.84	**514.84**	491.20	507.03	510.94
	Duration	218.75	109.38	**210.94**	105.47	226.56	**121.09**	179.69	183.59	203.13	89.64	195.31	199.22

relatively stable clustering accuracy and time windows testing on the different intensities of additive noise. Our method, however, is limited in some aspects: the highly overlapped components are challenging to the proposed method as it is for newly developed methods and previously developed approaches. Together, our results show that the proposed method provides a new approach to improve our understanding of the discoverable nature of ERP from noisy data and determine a more reliable time window of ERP.

Another issue with the deep clustering methods is initializing DNN with no ground-truth classification/labeling exists. Commonly, *k*-means [17] is used for initializing and tuning of deep clusterings [2]. However, the random optimized results of the *k*-means-based tuning can affect the learning in the DNNs. A similar issue occurs when initializing the unsupervised deep clustering with a trivial clustering such as *k*-means, the Gaussian mixture model (GMM) [29], and hierarchical clustering [15]. To tackle this issue, we fed the semi-supervised methods and DEC with consensus clustering. The drawback to semi-supervised is that this initialization cannot guarantee to obtain the best labeling results. However, it encourages the network to learn the most powerful features of ERP data. Unsupervised methods can appropriately learn the

Table 6
Standard deviation error (SD) between the estimated time windows by the proposed method, consensus clustering, and the studied deep clustering methods in both ERP data when the different noise strengths are added. The proposed method (Ens_DC) has achieved better stability in estimating the time window, especially in the real data.

Method	Properties (ms)	Simulated data		Real data	
		Cond1	Cond2	Cond1	Cond2
Ens_DC	Start	**2.41**	**0.00**	**3.19**	5.38
	End	**1.20**	**1.95**	**3.27**	**5.88**
	Duration	**3.61**	**1.95**	**3.84**	**1.59**
CC	Start	2.56	0.95	11.23	**1.59**
	End	1.95	2.29	41.02	21.07
	Duration	4.11	2.95	37.91	21.05
MLP_FC	Start	3.10	**0.00**	12.94	4.28
	End	1.76	2.09	36.51	22.47
	Duration	2.95	2.09	36.73	23.27
1D CNN	Start	5.04	**0.00**	9.70	3.84
	End	1.91	2.09	34.11	18.61
	Duration	4.34	2.09	29.00	18.10
LSTM	Start	3.61	**0.00**	8.44	3.19
	End	1.91	2.09	35.26	12.78
	Duration	4.09	2.09	37.22	13.40
AE	Start	3.81	7.33	4.73	4.73
	End	2.41	6.38	9.46	22.91
	Duration	5.90	13.50	13.30	23.13
VAE	Start	3.54	3.43	9.05	7.27
	End	4.02	2.86	39.66	31.00
	Duration	7.18	6.20	35.35	35.94
DEC	Start	6.08	3.19	4.73	29.95
	End	5.67	2.73	12.35	27.80
	Duration	11.35	5.67	16.88	45.31

important features of the data with roughly less accuracy than semi-supervised methods.

From the cognitive process perspective, the proposed method resulted in interpretable and reasonable findings based on prior studies. Notably, the statistical analysis results on the artificial data or real data can be crucial when there is a large effect of ERP component(s) due to obtaining uninterpretable statistical differences (e.g., obtaining extremely significant p-value between the parameters). Although this might be considered a better performance of the new methods, further sophisticated statistical analysis of temporal and sensory parameters can provide more information on the effect of interesting ERP. In the simulated data, the determined time windows from the cluster maps (even noisy conditions) were qualified by our previous findings [31] and the pre-defined components' properties. Additionally, the P3 component was reliably identified in both target and non-target conditions, which is interpretable with the purpose of the experiment and findings from the prior study [19]. Noteworthy, the generated responses from different groups and conditions could differ according to the neurological and experimental mechanisms [4,53]. Therefore, the cluster maps from the same potential can emerge in different temporal and spatial properties, e.g., in the ERP study by Koenig et al. [23]. The reason to study the P3 components is that the results are comparable with the ground truth results (in the simulated data) and interpretable for the real data. Therefore, the proposed method is not limited to identifying P3; it can be applied to identifying some other ERP components, such as N2, P2, and N4, in the simulated data. N4, for example, is identified by maps 3 (for 'Cond1′ and 'Cond2′) in Fig. 4**B** and maps 5 in Fig. 4**D**. Similarly, in real data, the identification of the P1 component by maps 4 in Fig. 5**B** and Fig. 5**D** can be discussed using the proposed method.

Considering the likelihood of obtaining imperfect clustering results, even using state-of-the-art clustering methods (including deep clustering), our method provides a confident result and time window determination for testing the researchers' hypotheses. Our early findings showed that combining different deep clustering methods can be useful for processing ERP data. One important advantage of ensemble learning is that different combinations of clustering methods, including deep clustering, are possible in our pipeline even with an unknown number of clusters. This study provides a positive message about using deep clustering methods for processing ERP data. We have provided a GitHub repository (https://github.com/remahini/Deep-Clustering) for the deep clustering methods used in this study, which can be used as a toolbox by changing the input and initializing parameters to be used by the researchers.

Table 7
Spatial correlation between the mean topography map in the ground truth time windows of P3 and the mean topography maps in the obtained time windows acquired by the proposed method, consensus clustering, and the studied deep clustering methods in the simulated data with different additional noises.

Method	No noise		SNR = 20 dB		SNR = 10 dB		SNR = 5 dB		SNR = 0 dB		SNR = -5dB	
	Cond1	Cond2	Cond1	Cond2	Cond1	Cond2	Cond1	Cond2	Cond1	Cond2	Cond1	Cond2
Ens_DC	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.998	0.999	0.998	0.997	0.982
CC	1.000	1.000	1.000	0.998	0.995	1.000	0.994	0.998	0.994	0.998	0.994	0.982
FC_MLP	0.996	0.999	0.998	0.999	0.994	0.996	0.995	0.998	0.986	0.996	0.998	0.986
1D CNN	0.997	0.999	0.998	0.995	0.995	0.999	0.996	0.998	0.991	0.998	0.996	0.987
LSTM	0.997	0.999	0.995	1.000	0.997	0.999	0.997	0.998	0.990	0.998	0.998	0.987
AE	1.000	0.998	1.000	0.991	1.000	0.997	1.000	0.997	0.999	0.996	0.998	0.979
VAE	1.000	0.996	0.991	0.997	0.995	0.993	1.000	0.993	0.998	0.993	0.995	0.989
DEC	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.997	0.999	0.949	0.998	0.992

Table 8
Spatial correlation of mean topography map in the ground-truth time window and mean topography maps in the identified time windows of the studied methods for P3 in the real data with different additional noises.

Method	No noise		SNR = 20 dB		SNR = 10 dB		SNR = 5 dB		SNR = 0 dB		SNR = -5 dB	
	Cond1	Cond2	Cond1	Cond2	Cond1	Cond2	Cond1	Cond2	Cond1	Cond2	Cond1	Cond2
Ens_DC	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.987	0.995	0.998	0.999
CC	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.987	0.995	0.959	0.992
MLP_FC	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	1.000	0.998	0.999
1dCNN	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	1.000	0.998	0.999
LSTM	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	1.000	0.998	0.999
AE	0.999	1.000	1.000	1.000	0.999	1.000	1.000	1.000	0.999	1.000	0.997	0.999
VAE	0.999	1.000	0.999	1.000	0.999	1.000	0.999	1.000	0.999	0.999	0.998	0.999
DEC	0.999	0.999	0.999	1.000	0.999	0.999	0.999	1.000	0.998	0.999	0.995	0.999

Fig. 6. The performance assessment results, in comparison to the ground truth clustering, for the studied clustering methods from the simulated (left panel) and the real ERP (right panel) data with different additive noise levels. **A.** the accuracy (ACC), adjusted rand index (ARI), and normalized mutual information (NMI) comparison results for clustering results in the simulated data. **B.** the performance metrics (ACC, ARI, and NMI) assessment results for the real data. Noticeably, the proposed clustering provides relatively stable and superior results (except when no noise is added) from both applied data.

Table 9
Illustration of the statistical analysis results of the identifying P3 effect from the measured mean amplitude in the estimated time windows and the Cz and CPz electrode sites from the studied clustering methods on the simulated ERP data at different noise levels.η_p^2 = Partial Eta Squared.

Noise	Method	Ens_DC	CC	FC_MLP	1D CNN	LSTM	AE	VAE	DEC
No noise	F(1,19)	81,317	81,317	199,263	199,263	199,263	235,023	135,063	80,934
	p-value	5.72E-36	5.72E-36	1.15E-39	1.15E-39	1.15E-39	2.39E-40	4.62E-38	5.98E-36
	η_p^2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
SNR = 20 dB	F(1,19)	68,237	68,237	68,237	68,237	68,237	62,941	118,338	75,717
	p-value	3.02E-35	3.02E-35	3.02E-35	3.02E-35	3.02E-35	6.52E-35	1.62E-37	1.13E-35
	η_p^2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
SNR = 10 dB	F(1,19)	27,965	27,965	27,965	27,965	27,965	15,805	15,532	22,446
	p-value	1.44E-31	1.44E-31	1.44E-31	1.44E-31	1.44E-31	3.25E-29	3.83E-29	1.16E-30
	η_p^2	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999
SNR = 5 dB	F(1,19)	16,285	14,151	14,151	12,442	12,442	13,330	9748	10,789
	p-value	2.45E-29	9.27E-29	9.27E-29	3.15E-28	3.15E-28	1.64E-28	3.18E-27	1.22E-27
	η_p^2	0.999	0.999	0.999	0.998	0.998	0.999	0.998	0.998
SNR = 0 dB	F(1,19)	1975	2784	2268	2268	2121	2339	2924	2892
	p-value	1.15E-20	4.50E-22	3.11E-21	3.11E-21	5.86E-21	2.33E-21	2.83E-22	3.15E-22
	η_p^2	0.990	0.993	0.992	0.992	0.991	0.992	0.994	0.993
SNR = -5dB	F(1,19)	883	856	989	1001	989	753	832	608
	p-value	2.15E-17	2.89E-17	7.51E-18	6.73E-18	7.51E-18	9.50E-17	3.74E-17	6.92E-16
	η_p^2	0.979	0.978	0.981	0.981	0.981	0.975	0.978	0.970

Table 10
Illustration of the statistical analysis results and the identified P3 effect measured by mean amplitude in the estimated time windows and Pz electrode from the studied clustering methods for the real ERP data at different noise levels.

Noise	Method	Ens_DC	CC	FC_MLP	1D CNN	LSTM	AE	VAE	DEC
No noise	F(1,39)	121.18	122.01	91.96	93.59	94.21	129.01	121.77	119.80
	p-value	1.57E-13	1.42E-13	8.27E-12	6.49E-12	5.91E-12	6.16E-14	1.46E-13	1.87E-13
	η_p^2	0.76	0.76	0.70	0.71	0.71	0.77	0.76	0.75
SNR = 20 dB	F(1,39)	121.02	127.47	100.69	98.39	92.41	134.86	106.20	120.56
	p-value	1.61E-13	7.38E-14	2.32E-12	3.22E-12	7.73E-12	3.15E-14	1.08E-12	1.70E-13
	η_p^2	0.76	0.77	0.72	0.72	0.70	0.78	0.73	0.76
SNR = 10 dB	F(1,39)	120.12	95.17	103.72	96.16	94.90	128.98	111.35	117.88
	p-value	1.79E-13	5.13E-12	1.52E-12	4.44E-12	5.34E-12	6.18E-14	5.46E-13	2.37E-13
	η_p^2	0.75	0.71	0.73	0.71	0.71	0.77	0.74	0.75
SNR = 5 dB	F(1,39)	121.73	99.19	104.64	98.46	105.05	128.78	130.45	135.42
	p-value	1.47E-13	2.87E-12	1.34E-12	3.19E-12	1.27E-12	6.33E-14	5.21E-14	2.95E-14
	η_p^2	0.76	0.72	0.73	0.72	0.73	0.77	0.77	0.78
SNR = 0 dB	F(1,39)	121.20	98.54	80.31	102.62	93.91	118.44	128.88	122.95
	p-value	1.57E-13	3.15E-12	5.19E-11	1.77E-12	6.18E-12	2.21E-13	6.26E-14	1.27E-13
	η_p^2	0.76	0.72	0.67	0.72	0.71	0.75	0.77	0.76
SNR = -5dB	F(1,39)	124.33	82.42	122.85	114.88	121.12	132.33	88.62	132.81
	p-value	1.07E-13	3.67E-11	1.28E-13	3.46E-13	1.59E-13	4.19E-14	1.38E-11	3.97E-14
	η_p^2	0.76	0.68	0.76	0.75	0.76	0.77	0.69	0.77

5. Conclusions

This research proposed an ensemble deep clustering methodology for qualifying ERP of interest from grand averaged spatio-temporal ERP data. The proposed method has been successfully applied to the simulated ERP and the real ERP data to assess and compare previous findings. Our findings suggested that the time window of ERP can be identified using ensemble deep clustering while a considerable amount of noise exists after preprocessing. Compared to the state-of-the-art clustering methods, the proposed method obtained superior results in terms of the temporal properties of the time windows and clustering performance. The robust clustering performance of the proposed method discloses its confidential properties for use in ERP data. Yet, studying the ensemble deep clustering in the subject, single-trial, and electrode resolution is an open question. Our further outline is to modify the current design to more sophisticated data, e.g., single-trial EEG data.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This study would like to remember Prof. Tapani Ristaniemi, who was involved in this study and passed away in 2020, for his great help to all the authors, especially Fengyu Cong, Asoke K. Nandi, Timo Hämäläinen, and Reza Mahini.

References

[1] B. Abu-Jamous, R. Fa, D.J. Roberts, A.K. Nandi, M-N scatter plots technique for evaluating varying-size clusters and setting the parameters of Bi-CoPaM and Uncles methods, in: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, https://doi.org/10.1109/ICASSP.2014.6854902.
[2] E. Aljalbout, V. Golkov, Y. Siddiqui, D.J.a.p.a. Cremers. Clustering with Deep Learning: Taxonomy and New Methods, 2018. https://doi.org/arXiv:1801.07648v2.
[3] P. Bashivan, I. Rish, M. Yeasin, N. Codella, Learning representations from EEG with deep recurrent-convolutional neural networks, 2015. arXiv preprint arXiv:1511.06448. https://doi.org/10.48550/arXiv.1511.06448.
[4] C. Berchio, A.-L. Küng, S. Kumar, P. Cordera, A.G. Dayer, J.-M. Aubry, C.M. Michel, C. Piguet, Eye-gaze processing in the broader bipolar phenotype revealed by electrical neuroimaging, Psychiat. Res.: Neuroimag. 291 (2019) 42–51, https://doi.org/10.1016/j.pscychresns.2019.07.007.
[5] J. Bergstra, Y. Bengio, Random search for hyper-parameter optimization, J. Mach. Learn. Res. 13 (2) (2012).
[6] J.C. Bezdek, Pattern recognition with fuzzy objective function algorithms, 1981. https://doi.org/10.1007/978-1-4757-0450-1.
[7] M.A. Boudewyn, S.J. Luck, J.L. Farrens, E.S. Kappenman, How many trials does it take to get a significant ERP effect? It depends. 55(6) (2018) e13049. https://doi.org/10.1111/psyp.13049.
[8] Y. Cao, T.A. Geddes, J.Y.H. Yang, P. Yang, Ensemble deep learning in bioinformatics, Nat. Mach. Intell. 2 (9) (2020) 500–508, https://doi.org/10.1038/s42256-020-0217-y.
[9] H. Cecotti, A. Graser, Convolutional neural networks for P300 detection with application to brain-computer interfaces, IEEE Trans. Patt. Anal. Mach. Intell. 33 (3) (2011) 433–445, https://doi.org/10.1109/TPAMI.2010.125.
[10] A. Delorme, S. Makeig, Mar). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis [Article], J. Neurosci. Methods 134 (1) (2004) 9–21, https://doi.org/10.1016/j.jneumeth.2003.10.009.
[11] M. Dinov, R. Leech, Modeling uncertainties in EEG microstates: analysis of real and imagined motor movements using probabilistic clustering-driven training of probabilistic neural networks, [Methods]. 11 (534) (2017), https://doi.org/10.3389/fnhum.2017.00534.
[12] K.G. Dizaji, A. Herandi, C. Deng, W. Cai, H. Huang, Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization, in: Computer Vision (ICCV), 2017 IEEE International Conference on, 2017, https://doi.org/10.1109/ICCV.2017.612.
[13] V. Duddu, N. Rajesh Pillai, D.V. Rao, V.E. Balas, Fault tolerance of neural networks in adversarial settings, J. Intell Fuzzy Syst 38 (5) (2020) 5897–5907, https://doi.org/10.3233/jifs-179677.
[14] A.B. Geva, H. Pratt, Unsupervised clustering of evoked potentials by waveform, 1994, Med Biol. Eng. Comput. 32 (5) (1994) 543–550, https://doi.org/10.1007/BF02515313.
[15] X. Guo, X. Liu, E. Zhu, X. Zhu, M. Li, X. Xu, J. Yin, Adaptive self-paced deep clustering with data augmentation, IEEE Trans. Knowl. Data Eng. 1–1 (2019), https://doi.org/10.1109/tkde.2019.2911833.
[16] M.K. Islam, A. Rastegarnia, Z. Yang, Methods for artifact detection and removal from scalp EEG: a review, 2016/11/01/), Neurophysiol. Clinique/Clinical Neurophysiol. 46 (4) (2016) 287–305, https://doi.org/10.1016/j.neucli.2016.07.002.
[17] A.K. Jain, Data clustering: 50 years beyond K-means, Pattern Recognit. Lett. 31 (8) (2010) 651–666, https://doi.org/10.1016/j.patrec.2009.09.011.
[18] R.E. Kallionpää, H. Pesonen, A. Scheinin, N. Sandman, R. Laitio, H. Scheinin, A. Revonsuo, K. Valli, Single-subject analysis of N400 event-related potential component with five different methods, Int. J. Psychophysiol. 144 (2019) 14–24, https://doi.org/10.1016/j.ijpsycho.2019.06.012.
[19] E.S. Kappenman, J.L. Farrens, W. Zhang, A.X. Stewart, S.J. Luck, ERP CORE: an open resource for human event-related potential research, 2021/01/15/), Neuroimage 225 (2021), 117465, https://doi.org/10.1016/j.neuroimage.2020.117465.
[20] E.S. Kappenman, S.J. Luck, ERP components: The ups and downs of brainwave recordings (2012) 3-30. https://doi.org/10.1093/oxfordhb/9780195374148.013.0014.
[21] A. Kiesel, J. Miller, P. Jolicœur, B. Brisson, Measurement of ERP latency differences: a comparison of single-participant and jackknife-based scoring

methods. 45(2) (2008) 250-274. https://doi.org/10.1111/j.1469-8986.2007.00618.x.
[22] D. Kingma, M. Welling. Auto-encoding variational bayes, ArXiv: 13126114. The 2nd International Conference on Learning Representations, 2013. https://doi.org/10.48550/arXiv.1312.6114.
[23] T. Koenig, M. Stein, M. Grieder, M. Kottlow, A tutorial on data-driven methods for statistically assessing ERP topographies, Brain Topography 27 (1) (2014) 72–83, https://doi.org/10.1007/s10548-013-0310-1.
[24] T. Kohonen, THE SELF-ORGANIZING MAP, Proc IEEE 78 (9) (1990) 1464–1480, https://doi.org/10.1109/5.58325.
[25] H.W.J.N.r.l.q. Kuhn, The Hungarian method for the assignment problem. *2*(1-2) (1955) 83-97.
[26] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444. https://www.nature.com/articles/nature14539.pdf.
[27] D. Lehmann, Brain Electric Microstates and Cognition: The Atoms of Thought. In: E. R. John, T. Harmony, L. S. Prichep, M. Valdés-Sosa, & P. A. Valdés-Sosa (Eds.), Machinery of the Mind: Data, Theory, and Speculations About Higher Brain Function (pp. 209-224). Birkhäuser Boston, 1990. https://doi.org/10.1007/978-1-4757-1083-0_10.
[28] F. Li, R. Yan, R. Mahini, L. Wei, Z. Wang, K. Mathiak, R. Liu, F. Cong, End-to-end sleep staging using convolutional neural network in raw single-channel EEG, Biomed. Signal Process. Control 63 (2021), 102203, https://doi.org/10.1016/j.bspc.2020.102203.
[29] K.L. Lim, X. Jiang, C. Yi, Deep clustering with variational autoencoder, IEEE Signal Process. Lett. 27 (2020) 231–235, https://doi.org/10.1109/LSP.2020.2965328.
[30] S.J. Luck, An introduction to the event-related potential technique, (Second edition ed.)., MIT press. (MIT press), 2014.
[31] R. Mahini, Y. Li, W. Ding, R. Fu, T. Ristaniemi, A.K. Nandi, G. Chen, F. Cong, Determination of the time window of event-related potential using multiple-set consensus clustering [Methods], 2020-October-21, Front. Neurosci. 14 (1047) (2020), https://doi.org/10.3389/fnins.2020.521595.
[32] R. Mahini, P. Xu, G. Chen, Y. Li, W. Ding, L. Zhang, N.K. Qureshi, T. Hämäläinen, A.K. Nandi, F. Cong, Correction: optimal number of clusters by measuring similarity among topographies for spatio-temporal ERP analysis, 2022/11/01), Brain Topography 35 (5) (2022) 558, https://doi.org/10.1007/s10548-022-00918-9.
[33] R. Mahini, P. Xu, G. Chen, Y. Li, W. Ding, L. Zhang, N.K. Qureshi, T. Hämäläinen, A.K. Nandi, F. Cong, Optimal number of clusters by measuring similarity among topographies for spatio-temporal ERP analysis. Brain Topography (2022b). https://doi.org/10.1007/s10548-022-00903-2.
[34] S. Makeig, A. Bell, T.-P. Jung, T.J. Sejnowski, Independent component analysis of electroencephalographic data, Adv. Neural Inform. Process. Syst. 8 (1995).
[35] P. Masulli, F. Masulli, S. Rovetta, A. Lintas, A.E.P. Villa, Fuzzy clustering for exploratory analysis of EEG event-related potentials, IEEE Trans. Fuzzy Syst. 28 (1) (2020) 28–38, https://doi.org/10.1109/TFUZZ.2019.2910499.
[36] M.C. Medeiros, M. McAleer, D. Slottje, V. Ramos, J. Rey-Maquieira, An alternative approach to estimating demand: neural network regression with conditional volatility for high frequency air passenger arrivals, 2008, J. Economet. 147 (2) (2008) 372–383, https://doi.org/10.1016/j.jeconom.2008.09.018.
[37] M. Meila, Comparing clusterings – an information based distance, J. Multivariate Anal. 98 (5) (2007) 873–895, https://doi.org/10.1016/j.jmva.2006.11.013.
[38] C.M. Michel, T. Koenig, EEG microstates as a tool for studying the temporal dynamics of whole-brain neuronal networks: a review, Neuroimage 180 (2018) 577–593, https://doi.org/10.1016/j.neuroimage.2017.11.062.
[39] E. Min, X. Guo, Q. Liu, G. Zhang, J. Cui, J.J.I.A. Long, A survey of clustering with deep learning: from the perspective of network architecture 6 (2018) 39501-39514. https://doi.org/10.1109/ACCESS.2018.2855437.
[40] N. Mrabah, N.M. Khan, R. Ksantini, Z. Lachiri, Deep clustering with a dynamic autoencoder: from reconstruction towards centroids construction, Neural Networks 130 (2020) 206–228, https://doi.org/10.1016/j.neunet.2020.07.005.
[41] Y. Mu, S. Han, Neural oscillations involved in self-referential processing, Neuroimage 53 (2) (2010) 757–768, https://doi.org/10.1016/j.neuroimage.2010.07.008.
[42] M.M. Murray, D. Brunet, C.M. Michel, Topographic ERP analyses: a step-by-step tutorial review, Brain Topography 20 (4) (2008) 249–264, https://doi.org/10.1007/s10548-008-0054-5.
[43] A.Y. Ng, M.I. Jordan, Y. Weiss, On spectral clustering: Analysis and an algorithm. Advances in Neural Information Processing Systems, 2002.
[44] G. Oetken, T. Parks, H. Schussler, New results in the design of digital interpolators, IEEE Trans. Acoust. Speech Signal Process. 23 (3) (1975) 301–309, https://doi.org/10.1109/tassp.1975.1162686.
[45] R. Oostenveld, P. Fries, E. Maris, J.-M. Schoffelen, FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data, Comput. Intell. Neurosci. 156869 (2011). https://doi.org/10.1155/2011/156869.
[46] R.D. Pascual-Marqui, C.M. Michel, D. Lehmann, Segmentation of brain electrical activity into microstates: model estimation and validation, IEEE Trans. Biomed. Eng. 42 (7) (1995) 658–665, https://doi.org/10.1109/10.391164.
[47] R.D. Pascual-Marqui, C.M. Michel, D.J.I.T.o.B.E. Lehmann, Segmentation of brain electrical activity into microstates: model estimation and validation. 42(7) (1995) 658-665. https://doi.org/10.1109/10.391164.
[48] J.M. Pena, J.A. Lozano, P.J.P.r.l. Larranaga, An empirical comparison of four initialization methods for the k-means algorithm. 20(10) (1999) 1027-1040. https://doi.org/10.1016/S0167-8655(99)00069-0.
[49] S.M. Peterson, R.P. Rao, B.W. Brunton, Learning neural decoders without labels using multiple data streams, J. Neural Eng. 19(4) (2022) 046032. https://doi.org/DOI 10.1088/1741-2552/ac857c.
[50] Y. Qi, F. Luo, W. Zhang, Y. Wang, J. Chang, D.J. Woodward, A.C. Chen, J. Han, Sliding-window technique for the analysis of cerebral evoked potentials, Beijing Da Xue Xue Bao Yi Xue Ban 35 (3) (2003) 231–235.
[51] Y. Ren, J. Pu, Z. Yang, J. Xu, G. Li, X. Pu, P.S. Yu, L. He, Deep clustering: a comprehensive survey, 2022. arXiv preprint arXiv:2210.04142. https://doi.org/10.48550/arXiv.2210.04142.
[52] Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T.H. Falk, J. Faubert, Deep learning-based electroencephalography analysis: a systematic review, J. Neural Eng.. (2019). https://orcid.org/0000-0003-4408-5221.
[53] P. Ruggeri, H.B. Meziane, T. Koenig, C. Brandner, A fine-grained time course investigation of brain dynamics during conflict monitoring, Article 3667, Scientific Reports 9 (2019), https://doi.org/10.1038/s41598-019-40277-3.
[54] O. Sagi, L. Rokach, Ensemble learning: a survey, WIREs Data Min. Knowl. Disc. 8 (4) (2018) e1249.
[55] S.B. Shaw, K. Dhindsa, J.P. Reilly, S. Becker, Capturing the forest but missing the trees: microstates inadequate for characterizing shorter-scale EEG dynamics, Neural Computat. 31 (11) (2019) 2177–2211, https://doi.org/10.1162/neco_a_01229.
[56] A. Sikka, H. Jamalabadi, M. Krylova, S. Alizadeh, J.N. van der Meer, L. Danyeli, M. Deliano, P. Vicheva, T. Hahn, T. Koenig, D.R. Bathula, M. Walter, Investigating the temporal dynamics of electroencephalogram (EEG) microstates using recurrent neural networks, Human Brain Mapping (2020), https://doi.org/10.1002/hbm.24949.
[57] A. Strehl, J. Ghosh, Cluster ensembles- a knowledge reuse framework for combining multiple partitions, J. Mach. Learn. Res. 3(3) (2003) 583-617. https://doi.org/10.1162/153244303321897735.
[58] L. Tan, J. Jiang, Chapter 11 - Multirate Digital Signal Processing, Oversampling of Analog-to-Digital Conversion, and Undersampling of Bandpass Signals, in: L. Tan, J. Jiang (Eds.), Digital Signal Processing, Third Edition, Academic Press, 2019, pp. 529–590, https://doi.org/10.1016/B978-0-12-815071-9.00011-7.
[59] M.C. Thomas, A.T. Joy. Elements of information theory, Wiley-Interscience, 2006.
[60] R. Tibshirani, G. Walther, Cluster validation by prediction strength, 2005/09/01, J. Computat. Graph. Statist. 14 (3) (2005) 511–528, https://doi.org/10.1198/106186005X59243.
[61] A. Topchy, A.K. Jain, W. Punch, Clustering ensembles: models of consensus and weak partitions, IEEE Trans. Pattern Anal. Mach. Intell. 27 (12) (2005) 1866–1881, https://doi.org/10.1109/TPAMI.2005.237.
[62] A. Tzovara, M.M. Murray, C.M. Michel, M. De Lucia, A tutorial review of electrical neuroimaging from group-average to single-trial event-related potentials, 2012/08/01, Dev. Neuropsychol. 37 (6) (2012) 518–544, https://doi.org/10.1080/87565641.2011.636851.
[63] N.X. Vinh, J. Epps, J.J.T.J.o.M.L.R. Bailey, Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance. 11 (2010) 2837-2854.
[64] A.J. Wills, A. Lavric, Y. Hemmings, E. Surrey, Attention, predictive learning, and the inverse base-rate effect: evidence from event-related potentials, Neuroimage 87 (2014) 61–71, https://doi.org/10.1016/j.neuroimage.2013.10.060.
[65] J. Xie, R. Girshick, A. Farhadi, Unsupervised deep embedding for clustering analysis. International Conference on Machine Learning, 2016.
[66] D. Yao, Y. Qin, S. Hu, L. Dong, M.L. Bringas Vega, P.A. Valdés Sosa, Which reference should we use for EEG and ERP practice?, 2019/07/01, Brain Topography 32 (4) (2019) 530–549, https://doi.org/10.1007/s10548-019-00707-x.
[67] P. Zhang, X. Wang, W. Zhang, J. Chen, Learning spatial–spectral–temporal EEG features with recurrent 3D convolutional neural networks for cross-task mental workload assessment, IEEE Trans. Neural Syst. Rehabilitat. Eng. 27 (1) (2019) 31–42, https://doi.org/10.1109/TNSRE.2018.2884641.
[68] A. Khanna, A. Pascual-Leone, F. Farzan, Reliability of Resting-State Microstate Features in Electroencephalography, PloS One 9 (12) (2014) e114163, https://doi.org/10.1371/journal.pone.0114163.