Université de Montréal

Cognitive Training Optimization with a Closed-Loop System

par

Yannick Roy

École d'optométrie

Thèse présentée en vue de l'obtention du grade de
Philosophiæ Doctor (Ph.D.)
en Sciences de la vision

Orientation neurosciences de la vision et psychophysique

August 4, 2022

Université de Montréal

École d'optométrie

Cette thèse intitulée

Cognitive Training Optimization with a Closed-Loop System

présentée par

Yannick Roy

a été évaluée par un jury composé des personnes suivantes :

Matthieu Vanni
(président-rapporteur)

Jocelyn Faubert
(directeur de recherche)

Franco Lepore
(membre du jury)

Alan Evans
(examinateur externe)

Martin Arguin
(représentant du doyen de la FESP)

If the brain were so simple we could understand it,
we would be so simple we couldn't.

Lyall Watson

Résumé

Les interfaces cerveau-machine (ICMs) nous offrent un moyen de fermer la boucle entre notre cerveau et le monde de la technologie numérique. Cela ouvre la porte à une pléthore de nouvelles applications où nous utilisons directement le cerveau comme entrée. S'il est facile de voir le potentiel, il est moins facile de trouver la bonne application avec les bons corrélats neuronaux pour construire un tel système en boucle fermée. Ici, nous explorons une tâche de suivi d'objets multiples en 3D, dans un contexte d'entraînement cognitif (3D-MOT).

Notre capacité à suivre plusieurs objets dans un environnement dynamique nous permet d'effectuer des tâches quotidiennes telles que conduire, pratiquer des sports d'équipe et marcher dans un centre commercial achalandé. Malgré plus de trois décennies de littérature sur les tâches MOT, les mécanismes neuronaux sous-jacents restent mal compris. Ici, nous avons examiné les corrélats neuronaux via l'électroencéphalographie (EEG) et leurs changements au cours des trois phases d'une tâche de 3D-MOT, à savoir l'identification, le suivi et le rappel. Nous avons observé ce qui semble être un transfert entre l'attention et la de mémoire de travail lors du passage entre le suivi et le rappel. Nos résultats ont révélé une forte inhibition des fréquences delta et thêta de la région frontale lors du suivi, suivie d'une forte (ré)activation de ces mêmes fréquences lors du rappel. Nos résultats ont également montré une activité de retard contralatérale (CDA en anglais), une activité négative soutenue dans l'hémisphère contralatérale aux positions des éléments visuels à suivre.

Afin de déterminer si le CDA est un corrélat neuronal robuste pour les tâches de mémoire de travail visuelle, nous avons reproduit huit études liées au CDA avec un ensemble de données EEG accessible au public. Nous avons utilisé les données EEG brutes de ces huit études et les avons analysées avec le même pipeline de base pour extraire le CDA. Nous avons pu reproduire les résultats de chaque étude et montrer qu'avec un pipeline automatisé de base, nous pouvons extraire le CDA.

Récemment, l'apprentissage profond (deep learning / DL en anglais) s'est révélé très prometteur pour aider à donner un sens aux signaux EEG en raison de sa capacité à apprendre de bonnes représentations à partir des données brutes. La question à savoir si l'apprentissage profond présente vraiment un avantage par rapport aux approches plus traditionnelles reste une question ouverte. Afin de répondre à cette question, nous avons examiné 154 articles appliquant le DL à l'EEG, publiés entre janvier 2010 et juillet 2018, et couvrant différents domaines d'application tels que l'épilepsie, le sommeil, les interfaces cerveau-machine et la surveillance cognitive et affective.

Enfin, nous explorons la possibilité de fermer la boucle et de créer un ICM passif avec une tâche 3D-MOT. Nous classifions l'activité EEG pour prédire si une telle activité se produit pendant la phase de suivi ou de rappel de la tâche 3D-MOT. Nous avons également formé un classificateur pour les essais latéralisés afin de prédire si les cibles étaient présentées dans l'hémichamp gauche ou droit en utilisant l'activité EEG. Pour la classification de phase entre le suivi et le rappel, nous avons obtenu un 80% lors de l'entraînement d'un SVM sur plusieurs sujets en utilisant la puissance des bandes de fréquences thêta et delta des électrodes frontales.

Mots clés: Suivi d'objets multiples, MOT, CDA, EEG, Mémoire de travail, Attention, Apprentissage profond

Abstract

Brain-computer interfaces (BCIs) offer us a way to close the loop between our brain and the digital world of technology. It opens the door for a plethora of new applications where we use the brain directly as an input. While it is easy to see the disruptive potential, it is less so easy to find the right application with the right neural correlates to build such closed-loop system. Here we explore closing the loop during a cognitive training 3D multiple object tracking task (3D-MOT).

Our ability to track multiple objects in a dynamic environment enables us to perform everyday tasks such as driving, playing team sports, and walking in a crowded mall. Despite more than three decades of literature on MOT tasks, the underlying and intertwined neural mechanisms remain poorly understood. Here we looked at the electroencephalography (EEG) neural correlates and their changes across the three phases of a 3D-MOT task, namely identification, tracking and recall. We observed what seems to be a handoff between focused attention and working memory processes when going from tracking to recall. Our findings revealed a strong inhibition in delta and theta frequencies from the frontal region during tracking, followed by a strong (re)activation of these same frequencies during recall. Our results also showed contralateral delay activity (CDA), a sustained negativity over the hemisphere contralateral to the positions of visual items to be remembered.

In order to investigate if the CDA is a robust neural correlate for visual working memory (VWM) tasks, we reproduced eight CDA-related studies with a publicly

accessible EEG dataset. We used the raw EEG data from these eight studies and analysed all of them with the same basic pipeline to extract CDA. We were able to reproduce the results from all the studies and show that with a basic automated EEG pipeline we can extract a clear CDA signal.

Recently, deep learning (DL) has shown great promise in helping make sense of EEG signals due to its capacity to learn good feature representations from raw data. Whether DL truly presents advantages as compared to more traditional EEG processing approaches, however, remains an open question. In order to address such question, we reviewed 154 papers that apply DL to EEG, published between January 2010 and July 2018, and spanning different application domains such as epilepsy, sleep, brain-computer interfacing, and cognitive and affective monitoring.

Finally, we explore the potential for closing the loop and creating a passive BCI with a 3D-MOT task. We classify EEG activity to predict if such activity is happening during the tracking or the recall phase of the 3D-MOT task. We also trained a classifier for lateralized trials to predict if the targets were presented on the left or right hemifield using EEG brain activity. For the phase classification between tracking and recall, we obtained 80% accuracy when training a SVM across subjects using the theta and delta frequency band power from the frontal electrodes and 83% accuracy when training within subjects.

Keywords: Multiple-Object Tracking, MOT, Contralateral Delay Activity, CDA, EEG, Working Memory, Attention, Deep Learning

Contents

List of tables

List of figures

Liste des sigles et des abréviations

BCI	Brain-Computer Interface
CDA	Contralateral Delay Activity
DL	Deep Learning
EEG	Electroencephalography
EMG	Electromyography
EOG	Electroculography
ERP	Event-Related Potential
MEG	Magnetoencephalography
ML	Machine Learning
RSVP	Rapid Serial Visual Presentation
SNR	Signal-to-Noise Ratio

Remerciements

My mission to democratize and popularize the exciting field of *NeuroTechnology* extended way beyond my academic work within UdeM's walls and I couldn't have asked for a better supervisor than Prof. Jocelyn Faubert to support me throughout this amazing journey. Over the past years I've had the chance to start and spearhead NeuroTechX an international neurotechnology community that has inspired and mobilized several thousands of neurotech enthusiasts around the world through several initiatives, and I've also had the pleasure of cofounding NeuroTechX Services, a consulting and recruiting company, now working with the big names in the field providing various services. I've had the chance to help organize over 100 neurotech events of various sizes, be featured in more than 10 local radio and TV shows, be invited as a speaker to over 20 events and conferences, be invited to several round tables and most importantly, building an incredible network of highly talented and motivated individuals. I'm extremely grateful for the opportunity I had to lead an international neurotech community for the past several years while doing my research and being part of the rapid growth of the field. I'm now excited for the next chapter of my life and my next contributions to the amazing field of neurotechnology.

As I go on with my career, I will forever cherish the endless hours Prof. Faubert and I have spent talking about the brain, entrepreneurship and gaming. And for these

memories, I thank you Jocelyn.

-Yannick.

Introduction

The field of Brain-Computer Interface (BCI) started in the 1970s with the work of Jacques Vidal at the University of California, Los Angeles (UCLA) ([**9, 10**]) but only became popular in the early 2000s with Jonathan Wolpaw's paper on *Communication & Control* ([**11**]) which laid the foundation for the field as we know it today. Brain-computer interfaces stemmed from the need to help people with severe neuromuscular disorders, such as amyotrophic lateral sclerosis, brainstem stroke, and spinal cord injury. In such conditions, the body is somewhat unresponsive but brain functions remain, in most cases, intact. Therefore, the idea behind a brain-computer interface as its name implies, is to create a direct interface between the brain and a computer in order to bypass the body and leverage technology to help people communicate and have some control over their environment by controlling, for example, a robotic arm or a wheelchair. While this objective is still at the forefront of BCI research, the field has now evolved to wider ambitions extending to the general and healthy population as well.

0.1. Neurotechnology becoming mainstream

The quest to better understand the brain and cure neurological disorders is nothing new. While centuries of human experiments have enabled us to amass valuable knowledge on how the brain works, the recent technological advances in brain imaging have been a paradigm shift in how we study the brain and the speed at which we

gather evidences on brain functions. With tools such as functional magnetic resonance imaging (fMRI), functional near-infrared spectroscopy (fNIRS), positron emission tomography (PET), computed tomography (CT), magnetoencephalography (MEG), electroencephalography (EEG), and electrocorticography (ECoG), neuroscientists have been able to look at both the anatomical and functional aspect of the brain at unprecedented scale ([**1**]). While these advances have been very beneficial from both a medical and research standpoint, they have also paved the way to a cultural shift towards human enhancement and the ethical concerns that come with it [**29, 37, 4**].

In 2013, the US launched the Brain Research through Advancing Innovative Neurotechnologies® (BRAIN) Initiative under the Obama administration aiming to revolutionize our understanding of the human brain. Shortly after, other major brain initiatives were launched in different parts of the world, all committing significant budget allocations towards understanding the brain. For example, in 2013 as well the Human Brain Project was announced as a concerted effort in Europe. In 2014, Japan launched their Japan Brain/MINDS (Brain/Mapping by Innovative Neurotechnologies for Disease Studies) project. South Korea announced their ten-year brain-mapping project in 2016. The China Brain Project, a 15-year project, was approved in March 2016. Many other countries such as Canada, Australia, Cuba and others also launched their own brain initiatives [**36**]. In 2017, the International Brain Initiative (IBI) was established to coordinate efforts across all these existing and emerging brain initiatives ([**2**]). These years marked an unprecedented effort and budget allocation towards understanding the brain.

In 2016 and 2017, the neurotechnology field was shook again and brain-computer interfaces suddenly became a popular mainstream subject when back-to-back announcements were made by Elon Musk and Facebook, announcing that they were venturing in neurotechnology to create brain-computer interfaces enhancing human experiences. Other announcements from other groups were made during that period as well but nothing compares with the reach and influence that Facebook and Elon

Musk had and more importantly, the ethical debate it generated on public forums. Regardless of the opinions on these specific entities and projects, the capital and attention that these announcements generated is massively positive for the field as a whole. Future will tell if this was for better or for worse.

For a good review on neurotechnology covering the recent trends in research as well as industry and the societal impact of such technology, I recommend the book I co-authored with the NeuroTechX community *The Neurotech Primer: A Beginner's Guide to Everything Neurotechnology* ([**22**]).

0.2. Consumer EEG & BCI

There are two very distinct trends in the BCI field; the research & medical devices and the consumer devices. The research and medical devices are usually bucketed together as they both require the upmost quality, as opposed to the consumer devices usually sacrificing quality for form factor, manufacturing, price, and ease of use. For the BCIs as medical devices, they are usually separated in two categories; invasive and non-invasive. For consumer devices, there aren't any invasive devices as the regulation in most countries does not allow such procedure on healthy individuals. While there are different technologies and approaches to record brain activity and make use for a brain-computer interface, the BCI field is heavily dominated by electroencephalography (EEG) for non-invasive approaches and electrocorticography (ECoG) as well as microelectrode arrays for invasive approaches. This thesis however, focuses on non-invasive technology in a research context using research devices looking at neural correlates from a fundamental perspective but with the objective of potentially translating such research into the consumer space.

EEG is not a new technology, however, the early 2010s marked a pivotal moment in the evolution of the technology when four companies successfully disrupted a field that was previously reserved to the medical and research world given the price tag

of such equipment. Neurosky, Muse, Emotiv and OpenBCI paved the way for a new industry by offering low cost (sub $1000) EEG devices for people to use and develop applications using their software development kit (SDK). Until then, the market was dominated by a handful of EEG manufacturers selling devices in the tens of thousands of dollars. Neurosky was the first one to challenge the whole industry when they proposed their EEG device with 1 channel on the forehead for only $200. While the quality of such device can't be compared, these consumer EEG devices democratized the EEG technology and enabled a plethora of opportunities.

In the late 2010s and early 2020s, we have seen a pivot from making custom EEG headsets to rather try to embed these sensors into existing devices that we are already wearing such as headphones, glasses and headmounted display for augmented reality (AR), virtual reality (VR), and mixed reality (MR).

0.3. Towards Pervasive Passive and Reactive BCIs

In 2011, Thorsten Zander suggested to breakdown brain-computer interfaces into three distinct categories: (1) Active BCIs, (2) Passive BCIs, and (3) Reactive BCIs ([**12**]). Active BCIs refers to the interfaces where the user has to do a conscious effort to control it, like in a motor imagery paradigm where the user voluntarily thinks about moving their right or left arm to control a device ([**18, 34**]). Passive BCIs are those where the experience is seamless to the user. The user doesn't have to do anything specific while the BCI records and analyze the brain activity to modulate the task or the experience in real-time based on brain activity ([**5, 38**]). Reactive BCIs are those were a specific stimulus is presented while the BCI monitor the brain's reaction to that stimulus. The most popular use case of a reactive BCI is the P300 speller where a grid of letters, representing a keyboard, is shown on a screen and different rows and columns of letters are being randomly highlighted in rapid flashing sequences. When the letter that the user is paying attention to (i.e. trying to write) is highlighted, the

brain will naturally produce an evoked potential called the P300 ([**17, 38**]) that can be detected by the BCI and the letter be written on the screen allowing the user to type words and sentences.

Because non-invasive active BCIs are hard to implement given the quality of the signal, the mental fatigue, the exhaustion, and the frustation they cause, passive and reactive BCIs are most likely the first everyday life BCI implementations that we might see in the coming years.

0.4. Neural Correlates

Each brain imaging modality has its pros and cons. Electroencephalography (EEG), offers a very good temporal resolution as it measures the electrical activity of the brain, which is generated by the neurons firing. On the counterpart, EEG is a non-invasive technology recording from the scalp and therefore, reading a distorted signal prone to noise. Moreover, given that the potential difference (i.e. voltage measured) is calculated from two points on the scalp, the measurement obtained is the sum of all electrical signals being propagated and entangled together making it hard to interpret and offering a poor spatial resolution on where the activity is originating from.

The most popular analogy to describe EEG is using microphones in a stadium to infer what people are talking about. In this analogy the microphones represent the electrodes and the people represent the neuron communicating together. If we put microphones outside the stadium (non-invasive), we will be able to tell when there is a goal and we'll also be able to tell which team scored because these are synchronised signals with almost the whole stadium shouting in unison. On the other hand, we wouldn't be able to understand the various conversations people are having in smaller groups.

Despite such limitations, there are many different signals that can be used in a BCI context or to infer the underlying brain mechanisms at play. Just to name a few,

in the time domain, the most used signals are the evoked-related potentials (ERPs) which represent a reaction of the brain to a given stimulus. Hence their name: *evoked*. The most well known ERPs include the P300, a component elicited in the process of decision making and the N270, usually elicited in conflict or incongruity processing. In the frequency domain, the scientific community have separated the EEG spectrum in five main frequency bands: *delta*, *theta*, *alpha*, *beta* and *gamma*. Different groups use different limits for these bands, but usually they fall around the following breakdown: [1-3]Hz for Delta, [4-7]Hz for Theta, [8-12]Hz for Alpha, [13-30]Hz for Beta, and over 30Hz for Gamma. The role of these bands is still vaguely understood and many studies link the different bands with different cognitive processes. These bands are believed to be natural oscillatory speed at which the different brain networks communicate together ([**2**]). Another interesting EEG pattern linked to working memory is the contralateral delay activity (CDA) which will be explored in greater details in the next chapter.

0.5. Cognitive Training

One of the most disruptive concept in modern neuroscience is undoubtedly brain plasticity. It is now accepted among the scientific community that the adult brain can alter it structure and *re-wire* itself in response to environmental demands ([**9, 16**]). This fundamental concept has paved the way for the *cognitive training* subfield, where the idea is to harness brain plasticity by offering a controlled and targeted training environment. Despite hundreds of studies on the matter, the debate is still going on as if cognitive training really works and if participants in such training can derive value in their daily life activities. The main argument reside in the generalization of the training across different domains, also called *far transfer*. Simply put, it's one thing to train on a task and get better at it, but can it really help be better at other tasks as well. Systematic reviews have reached inconsistent conclusions ([**27**]) on the subject.

Since the publication from Bavelier and Green in 2003, showing that playing an action video game leads to a higher performance on complex visual tasks ([**13**]), many research groups and companies have tried to leverage video games as a tool for cognitive training. On paper, video games are the perfect vehicle for cognitive training because they are engaging, easy to make and easy to adapt to target specific brain functions. Several companies such as CogMed, Lumosity, BrainHQ, NeuroTracker, and Akili have commercialized brain training applications, often called *brain games*. While they all have different claims and level of scientific validation, Akili paved the way to not only for cognitive training but also for digital therapeutics by being the first company ever to obtain FDA approval for the marketing of a video game as a medical solution for ADHD in June 2020.

Adam Gazzaley, the scientist behind Akili and their video game as a digital therapy, with his colleague Jyoti Mishra, claimed in their 2015 paper that closed-loop systems will be the next evolution of cognitive training, using the term *closed-loop cognition* ([**21**]).

The question remains however as if video games can offer something more than just a means of entertainment but rather a platform to train the brain and enhance our cognitive abilities? And if so, how do we close the loop in real-time to improve the experience and create a direct relationship between the brain and the training task?

0.6. Multiple Object Tracking (MOT)

NeuroTracker™ is a commercially available 3D-MOT task currently used by a multitude of users in many countries around the world as a perceptual-cognitive training and assessment tool. It has been used and studied in various fields such as sport ([**1, 13, 26, 38**]), ESports ([**3**]), education ([**43**]), aviation ([**18**]) and military ([**45**]). 3D-MOT training has been demonstrated to enhance attention, working memory and visual information processing speed ([**34**]). Given the wide adoption of

the NeuroTracker™ and existing literature showing effective transfer, we opted for a modified version of the NeuroTracker™ for our study.

The multiple-object tracking (MOT) paradigm was originally developed by Pylyshyn & Storm in 1988 ([**6**]). Since then, many researchers have developed their own version of the task with slight variations to study different aspect of perception and cognition. In a MOT task, the participant is asked to follow number of targets (usually between 1 and 4 objects of interest) while ignoring distractors. In order to modulate task difficulty, parameters usually include the number of targets ([**35**]), speed ([**7, 19**]), and distance between objects ([**1, 41**]).

Most MOT studies are using a 2D-MOT, however, given that we live in a 3D environment we believe that a 3D-MOT makes for a more ecological task to study and train cognitive functions. The modified version of the NeuroTracker (i.e. 3D-MOT task) is a great candidate for a passive BCI based on attention and working memory neural correlates. Moreover, while the task remains simple as opposed to a fancy video game with too many confounding elements, the task still engages complex brain networks as would some real-life activities such as driving and playing a team sport.

Despite the increasing literature on multiple-object tracking and evidence of effective training showing transfer on real tasks in real environments ([**1**]), our understanding of the underlying neural mechanisms remains fuzzy. Over the last few decades, several cognitive models have been brought forward trying to explain how individual items are being encoded and deciphering the intertwined roles of attention and working memory in such tasks. Such models are discussed in chapter 2. Recent EEG studies try to disentangle working memory and attention during a 2D-MOT task. For example, Drew and colleagues tried to delineate the neural signatures of tracking spatial position and working memory during attentive tracking. They found that there was a unique contralateral negativity related to the process of monitoring target position during tracking which was absent when objects briefly stopped moving.

These results suggest that the process of tracking target locations elicits an electrophysiological response that is distinct and dissociable from neural responses of the number of targets being attended ([**10**]). In this research we seek to explore if we can classify via machine learning the neural signatures of the participants to differentiate the phases of the 3D-MOT task, the number of targets being tracked as well as the hemifield (i.e. side) in which they where presented.

0.7. Machine Learning or Deep Learning?

Machine learning plays a central role in a BCI application as that's where a decision will be made by the system from interpreting brain activity. While most brain-computer interface pipeline still use traditional machine learning approaches such as Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), or more complex approaches such as Reimannian geometry (which is still considered state-of-the-art for most BCI classification problems ([**103**])), the promises of AI and deep learning have generated a huge interest in the community. While the field lacks the data size that other fields such as image processing and natural language processing have access to, large neuro-related datasets have recently been released publicly to foster innovation in this area. In 2019, we published the first review on deep learning for EEG (see chapter 3), which was quickly appreciated by the community, where we shed light on the trends and tried to answer a few questions researchers interested in the field might have such as *how much data is enough data?*, *how deep should the model be?*, *should one use convolutional neural networks or recurrent neural networks?*, can one use the raw signal or should we extract features first?, does the model generalize across subjects or should one retrain a new model for each subject?. While a few questions remained unanswered we were able to extract the trends and to highlight practices providing the best results.

0.8. This Research

Given the current pace of innovation in neurotechnology and the different trends mentioned throughout this introduction, it is now clear that the quest to use the brain as an input for technology is here to stay and will spawn across many different sectors over the coming decades. This thesis seeks to explore new neural correlates related to attention and working memory that could be used to create a passive BCI to improve cognitive training tasks. Having a closed-loop system analyzing brain activity in real-time while the user is performing a cognitive training task, would allow for a more personalized experience leading to, we believe, a better training and transfer effect.

This research sits at the intersections of several fields such as brain-computer interfaces, neurotechnology, cognitive training, perception and cognition, machine learning and artificial intelligence. The next four chapters represents four articles either published or at least submitted in peer reviewed journals. In the first chapter, we conducted a reproducibility study on CDA to better understand if this represent a robust neural correlate for visual working memory and if it could potentially be used in a passive BCI. In order to do so, we used eight different EEG datasets from published studies with open access data and with the same pipeline we successfully extracted the CDA signal, showing robustness across different visual WM tasks, EEG recording devices, experimenters and subjects. In the second chapter, we conducted our own experiment on a 3D-MOT task exploring the CDA as well as the changes of neural oscillations across different phases of the task. In the third chapter, before using the insights from the two previous studies to develop a passive BCI, we look at recent developments in deep learning for EEG to better understand if deep learning approaches surpasses traditional machine learning approaches. To do so, we reviewed 154 papers using deep learning on EEG data. Finally, we bring it all that together to build a passive BCI for a 3D-MOT task in the forth chapter.

References

[1] Tony Abraham and Janine Feng. Evolution of brain imaging instrumentation. In *Seminars in nuclear medicine*, volume 41, pages 202–219. Elsevier, 2011.

[2] Amy Adams, Stephanie Albin, Katrin Amunts, Tasia Asakawa, Amy Bernard, Jan G Bjaalie, Khaled Chakli, James O Deshler, Yves De Koninck, Christoph J Ebell, et al. International brain initiative: an innovative framework for coordinated global brain research efforts. *Neuron*, 105(2):212–216, 2020.

[3] George A Alvarez and Steven L Franconeri. How many objects can you track?: Evidence for a resource-limited attentive tracking mechanism. *Journal of vision*, 7(13):14–14, 2007.

[4] Andrea Antal, Ivan Alekseichuk, M Bikson, J Brockmöller, André R Brunoni, Robert Chen, LG Cohen, G Dowthwaite, Jens Ellrich, A Flöel, et al. Low intensity transcranial electric stimulation: safety, ethical, legal regulatory and application guidelines. *Clinical Neurophysiology*, 128(9):1774–1809, 2017.

[5] Pietro Aricò, Gianluca Borghini, Gianluca Di Flumeri, Nicolina Sciaraffa, and Fabio Babiloni. Passive bci beyond the lab: current trends and future directions. *Physiological measurement*, 39(8):08TR02, 2018.

[6] Erol Başar, Canan Başar-Eroglu, Sirel Karakaş, and Martin Schürmann. Gamma, alpha, delta, and theta oscillations govern cognitive processes. *International journal of psychophysiology*, 39(2-3):241–248, 2001.

[7] Julie Justine Benoit, Eugenie Roudaia, Taylor Johnson, Trevor Love, and Jocelyn Faubert. The neuropsychological profile of professional action video game players. *PeerJ*, 8:e10211, 2020.

[8] Wei-Ying Chen, Piers D Howe, and Alex O Holcombe. Resource demands of object tracking and differential allocation of the resource. *Attention, Perception, & Psychophysics*, 75(4):710–725, 2013.

[9] Bogdan Draganski, Christian Gaser, Volker Busch, Gerhard Schuierer, Ulrich Bogdahn, and Arne May. Changes in grey matter induced by training. *Nature*, 427(6972):311–312, 2004.

[10] Trafton Drew, Todd S Horowitz, Jeremy M Wolfe, and Edward K Vogel. Delineating the neural signatures of tracking spatial position and working memory during attentive tracking. *Journal of Neuroscience*, 31(2):659–668, 2011.

[11] Jocelyn Faubert. Professional athletes have extraordinary skills for rapidly learning complex and neutral dynamic visual scenes. *Scientific reports*, 3(1):1–3, 2013.

[12] Jocelyn Faubert and Lee Sidebottom. Perceptual-cognitive training of athletes. *Journal of Clinical Sport Psychology*, 6(1):85–102, 2012.

[13] C Shawn Green and Daphne Bavelier. Action video game modifies visual selective attention. *Nature*, 423(6939):534–537, 2003.

[14] Jaclyn Hoke, Christopher Reuter, Thomas Romeas, Maxime Montariol, Thomas Schnell, and Jocelyn Faubert. Perceptual-cognitive & physiological assessment of training effectiveness. In *Interservice/Industry Training, Simulation, and Education Conference, Orlando, FL*, 2017.

[15] Lucica Iordanescu, Marcia Grabowecky, and Satoru Suzuki. Demand-based dynamic distribution of attention and monitoring of velocities during multiple-object tracking. *Journal of vision*, 9(4):1–1, 2009.

[16] Julia Karbach and Torsten Schubert. Training-induced cognitive and neural plasticity. *Frontiers in Human Neuroscience*, 7:48, 2013.

[17] Dean J Krusienski, Eric W Sellers, Dennis J McFarland, Theresa M Vaughan, and Jonathan R Wolpaw. Toward enhanced p300 speller performance. *Journal of neuroscience methods*, 167(1):15–21, 2008.

[18] Mikhail A Lebedev and Miguel AL Nicolelis. Brain–machine interfaces: past, present and future. *TRENDS in Neurosciences*, 29(9):536–546, 2006.

[19] Fabien Lotte, Laurent Bougrain, Andrzej Cichocki, Maureen Clerc, Marco Congedo, Alain Rakotomamonjy, and Florian Yger. A review of classification algorithms for eeg-based brain–computer interfaces: a 10 year update. *Journal of neural engineering*, 15(3):031005, 2018.

[20] Gerald T Mangine, Jay R Hoffman, Adam J Wells, Adam M Gonzalez, Joseph P Rogowski, Jeremy R Townsend, Adam R Jajtner, Kyle S Beyer, Jonathan D Bohner, Gabriel J Pruna, et al. Visual tracking speed is related to basketball-specific measures of performance in nba players. *The Journal of Strength & Conditioning Research*, 28(9):2406–2414, 2014.

[21] Jyoti Mishra and Adam Gazzaley. Closed-loop cognition: the next frontier arrives. *Trends in cognitive sciences*, 19(5):242–243, 2015.

[22] NeuroTechX. *The Neurotech Primer: A Beginner's Guide to Everything Neurotechnology*. 2021.

[23] Brendan Parsons, Tara Magill, Alexandra Boucher, Monica Zhang, Katrine Zogbo, Sarah Bérubé, Olivier Scheffer, Mario Beauregard, and Jocelyn Faubert. Enhancing cognitive function using perceptual-cognitive training. *Clinical EEG and neuroscience*, 47(1):37–47, 2016.

[24] Zenon Pylyshyn. The role of location indexes in spatial perception: A sketch of the finst spatial-index model. *Cognition*, 32(1):65–97, 1989.

[25] Zenon W Pylyshyn and Ron W Storm. Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial vision*, 3(3):179–197, 1988.

[26] Thomas Romeas, Antoine Guldner, and Jocelyn Faubert. 3d-multiple object tracking training task improves passing decision-making accuracy in soccer players. *Psychology of Sport and Exercise*, 22:1–9, 2016.

[27] Giovanni Sala and Fernand Gobet. Cognitive training does not enhance general cognition. *Trends in cognitive sciences*, 23(1):9–20, 2019.

[28] Won Mok Shim, George A Alvarez, and Yuhong V Jiang. Spatial separation between targets constrains maintenance of attention on multiple objects. *Psychonomic bulletin & review*, 15(2):390–397, 2008.

[29] Wessel Teunisse, Sandra Youssef, and Markus Schmidt. Human enhancement through the lens of experimental and speculative neurotechnologies. *Human Behavior and Emerging Technologies*, 1(4):361–372, 2019.

[30] Domenico Tullo, Jocelyn Faubert, and Armando Bertone. Examining the benefits of training attention with multiple object-tracking for individuals diagnosed with a neurodevelopmental condition: A cross-over, cognitive training study. *Journal of Vision*, 18(10):1021–1021, 2018.

[31] Oshin Vartanian, Lori Coady, and Kristen Blackler. 3d multiple object tracking boosts working memory span: Implications for cognitive training in military populations. *Military Psychology*, 28(5):353–360, 2016.

[32] Jacques J Vidal. Toward direct brain-computer communication. *Annual review of Biophysics and Bioengineering*, 2(1):157–180, 1973.

[33] Jacques J Vidal. Real-time detection of brain events in eeg. *Proceedings of the IEEE*, 65(5):633–641, 1977.

[34] Piotr Wierzgała, Dariusz Zapała, Grzegorz M Wojcik, and Jolanta Masiak. Most popular signal processing methods in motor-imagery bci: a review and meta-analysis. *Frontiers in neuroinformatics*, 12:78, 2018.

[35] Jonathan R Wolpaw, Niels Birbaumer, Dennis J McFarland, Gert Pfurtscheller, and Theresa M Vaughan. Brain–computer interfaces for communication and control. *Clinical neurophysiology*, 113(6):767–791, 2002.

[36] Rafael Yuste and Cori Bargmann. Toward a global brain initiative. *Cell*, 168(6):956–959, 2017.

[37] Rafael Yuste, Sara Goering, Guoqiang Bi, Jose M Carmena, Adrian Carter, Joseph J Fins, Phoebe Friesen, Jack Gallant, Jane E Huggins, Judy Illes, et al. Four ethical priorities for neurotechnologies and ai. *Nature*, 551(7679):159–163, 2017.

[38] Thorsten O Zander, Jonas Brönstrup, Romy Lorenz, and Laurens R Krol. Towards bci-based implicit control in human–computer interaction. In *Advances in Physiological Computing*, pages 67–90. Springer, 2014.

[39] Thorsten O Zander and Christian Kothe. Towards passive brain–computer interfaces: applying brain–computer interface technology to human–machine systems in general. *Journal of neural engineering*, 8(2):025005, 2011.

First Article.

Is the Contralateral Delay Activity (CDA) a robust neural correlate for Visual Working Memory (VWM) tasks? A reproducibility study.

by

Yannick Roy[1], and Jocelyn Faubert[1]

([1]) Université de Montréal

This article was submitted in Psychophysiology.

Résumé. La mémoire de travail visuelle nous permet d'activement conserver et de manipuler l'information visuel nous entourant. Bien que les mécanismes sous-jacents à la mémoire de travail visuelle ne soient pas bien compris, l'activité contra latérale (contralateral delay activity ou CDA en anglais) est souvent utilisée pour les étudier. Il s'agit d'une activité négative soutenue dans l'hémisphère du côté opposé aux objets visuels présentés. Afin d'étudier si le CDA est une activité cérébrale robuste et fiable pour les tâches de mémoire de travail visuelle, nous avons reproduit huit études dans le domaine avec des données EEG accessible et les avons analysées avec la même séquence d'analyse rudimentaire. Nous avons été capable de reproduire les résultats de toutes les études et de montrer qu'une séquence automatisée de base permet d'extraire le signal CDA. Dans cette étude, nous partageons les tendances observées à travers les études reproduites et soulevons quelques questions sur le déclin du CDA ainsi que le CDA durant la phase de rappel, qui surprenamment n'a pas été adressé dans aucune de ces huit études. Finalement, nous proposons des recommandations sur la reproducibilité basés sur notre expérience et les difficultés rencontrées durant la reproduction de ces études.
Mots clés : CDA, EEG, Mémoire de travail, Activité contra latérale

Abstract. Visual working memory (VWM) allows us to actively store, update and manipulate visual information surrounding us. While the underlying neural mechanisms of VWM remain unclear, contralateral delay activity (CDA), a sustained negativity over the hemisphere contralateral to the positions of visual items to be remembered, is often used to study VWM. To investigate if the CDA is a robust neural correlate for VWM tasks, we reproduced eight CDA-related studies with a publicly accessible EEG dataset. We used the raw EEG data from these eight studies and analysed all of them with the same basic pipeline to extract CDA. We were able to reproduce the results from all the studies and show that with a basic automated EEG pipeline we can extract a clear CDA signal. We share insights from the trends observed across the studies and raise some questions about the CDA decay and the CDA during the recall phase, which surprisingly, none of the eight studies did address. Finally, we also provide reproducibility recommendations based on our experience and challenges in reproducing these studies.
Keywords: EEG, CDA, Contralateral Delay Activity, Working Memory

1. Introduction

Visual working memory (VWM) allows us to actively store, update and manipulate visual information surrounding us. Acting as a mental buffer for visual information, VWM has been an active area of research for several decades. While behavioural studies have shown that working memory (WM), of which VWM is a subset, has a limited capacity of only a few items, usually ranging between 3 to 5 ([**8, 5**]), the underlying neural mechanisms remain vague. One ongoing challenge in the field is dissociating WM from attention ([**8, 27, 3, 26**]). There is no clear consensus yet as how separated or intertwined these two mechanisms really are.

Given the crucial role of VWM in our everyday life, much prior work has tried to understand the neural correlates underlying its functioning via electroencephalogram (EEG). One such VWM neural measurement being studied is the contralateral delay activity (CDA) ([**46, 24**]). CDA is a sustained negativity over the hemisphere contralateral to the positions of the items to be remembered. A prevailing view of CDA is that it is modulated by the number of items held in WM reaching a plateau at around three or four items. The CDA has been shown to be linked with the number of items held in WM ([**46, 44, 25**]).

In order to study CDA, the most common task is a change detection task. The sequence of such task typically looks something like the following: the participant is cued with an array of items, varying in numbers but usually between one and eight, balanced on both sides of the visual field. The items then disappear, forcing the participant to hold relevant information in mind and after a short period of time, usually one or two seconds, the participant is quizzed on the visual representation held in WM. For example, a new array of items can be presented and the participant is asked if there has been a change versus the initial items. The participant will answer with a keyboard or a mouse and the results will be logged with different levels of

granularity depending on what the study is interested in. There are many variants of such tasks.

Many well-known event-related potentials (ERPs) such as the P300 and N270, benefit from a large volume of studies and replications and are well understood ([**30, 31, 11, 21**]). CDA, however, doesn't benefit from such volume of evidence yet. Our primary goal with this reproducibility study was to answer the following question: *Is the Contralateral Delay Activity (CDA) a robust and consistent neural correlate for Visual Working Memory (VWM) tasks?*. We wanted to know if CDA is a consistent measure across different tasks, subjects and EEG recording devices, or if it requires a lot of manual cleaning and handcrafting of the data to obtain it. To investigate the robustness of the CDA EEG signal, we looked for CDA-related EEG datasets available online and tried to reproduce their results using a basic independent automated pipeline with no human intervention on the data to extract the CDA. Given that *robustness* can be interpreted in different ways, we should mention that we use the same definition as in the Framework for Open and Reproducible Research Teaching (FORRT): *The persistence of support for a hypothesis under perturbations of the methodological/analytical pipeline. In other words, applying different methods/analysis pipelines to examine if the same conclusion is supported under different analytical conditions* ([**28**]).

Lastly, before we dive in, we should also define the terms *reproducibility* and *replicability* given that researchers from different fields, and even from within the same field, often use them interchangeably. Here we will use the same definitions as in the *Recommendations to Funding Agencies for Supporting Reproducible Research* by the American Statistical Association ([**7**]):

Reproducibility: *A study is reproducible if you can take the original data and the computer code used to analyze the data and reproduce all of the numerical findings from the study. This may initially sound like a trivial task but experience has shown that it's not always easy to achieve this seemingly minimal standard.*

Replicability: *This is the act of repeating an entire study, independently of the original investigator without the use of original data (but generally using the same methods).*

Simply put, one can replicate a study or an effect (outcome of a study) but reproduce results (data analyses). A third term, *repeatability*, is also used although less often, referring to the same group repeating the same experiment with the same analysis and obtaining the same results.

This current work focuses on reproducibility and not on replicability nor repeatability as we did not collect new data but only used publicly available datasets. Moreover, given the nature of the studies and the data we are handling, we could be even more specific using the terminology *computational reproducibility*, as defined in the FORRT ([**28**]). This computational reproducibility study also intersects with a review given that we extract common trends among studies and explore them from a computational point of view. Note that we initially reproduced the result either with the original code or with a re-written version of it before using the same pipeline for all studies to assess the robustness.

Other fields such as artificial intelligence (AI) have benefited tremendously from good reproducibility frameworks, standards and overall culture. The accelerated pace of innovation and breakthroughs in AI were mainly enabled by accessibility of both data and code. These best practices of sharing both data and code in a reproducible manner aren't, unfortunately, the default behavior in psycho- and neuro-related fields. Here, we share how a few groups of researchers have made their data and code available and we hope to inspire others to follow that path.

2. Method

After looking through CDA-related literature, searching for available datasets and asking (i.e. emailing) a few researchers in the field if they were aware of open

access CDA-related EEG datasets, we ended up with eight recent CDA-related studies published between 2018 and 2020 with different task paradigms. We do not claim that the list of studies included in this review represents an exhaustive list of CDA papers with available EEG datasets, however, we believe that these eight datasets represent a good sample of the CDA literature as they used different VWM tasks exploring things such as different set sizes (from 1 to 6 targets), adding new targets after the initial array, retro-cueing, adding targets bilaterally, adding interruptions, halving the targets to create more items to track, and all that using different shapes and colors as stimuli across studies. The search of CDA-related literature was done on Google Scholar using the following search terms in different orders and combinations: *CDA*, *Contralateral Delay Activity*, and *EEG*. From the bibliography section of these articles, a few additional relevant studies were added to the list. Open Science Framework (OSF) with the keywords previously listed was also used to search for additional studies and datasets. It is important to note that six out of the eight studies included here have Edward Awh and Edward Vogel as co-authors, highlighting a lack of independent studies on CDA with openly available EEG data.

In this section, we detail the studies we've reproduced and provide the resulting CDA signal figure, which matches the CDA figure from the original study. It is important to note that some of the reproduced studies were also looking at different neural correlates, however, for simplicity and readability we focus only on the CDA relevant part of the study. In our analysis, we used all the available data, but did not include any data files that were clearly marked as not being used (for various reasons). All the studies were reproduced using MNE-Python ([**15**]) despite the original code of most studies being in MATLAB. All figures in this section were generated using the following pipeline: (1) load raw data, (2) rereference and downsample based on the original study's preprocessing steps, (3) filter the data between 1 and 30 Hz, (4) epoch the data, (5) automated removal of bad trials (i.e. epochs) and interpolation of bad channels via *autoreject* ([**3**]), (6) obtain each subject's CDA for each condition by

subtracting the averaged ipsilateral electrodes to the averaged contralateral electrodes (i.e. contra minus ipsi), (7) average the CDA for each condition across subjects.

Table 2 shows the high-level information of the datasets to provide the reader an idea about the number of subjects and trials for each study.

Year	**First Author**	**Title**
2020	Tobias Feldmann-Wüstefeld	Neural measures of working memory in a bilateral change detection task
2020	Nicole Hakim	Perturbing neural representations of working memory with task-irrelevant interruption
2019	Mario Villena-Gonzalez	Data from brain activity during visual working memory replicates the correlation between contralateral delay activity and memory capacity
2019	Haley Balaban	Neural evidence for an object-based pointer system underlying working memory
2019	Eren Gunseli	EEG dynamics reveal a dissociation between storage and selective attention within working memory
2019	Nicole Hakim	Dissecting the Neural Focus of Attention Reveals Distinct Processes for Spatial Attention and Object-Based Storage in Visual Working Memory
2018	Kristen Adam	Contralateral delay activity tracks fluctuations in working memory performance
2018	Tobias Feldmann-Wüstefeld	Contralateral Delay Activity Indexes Working Memory Storage, not the Current Focus of Spatial Attention

Table 1. Reproduced Studies

Dataset	Task	Subjects	Trials	Target(s)
FW2020	Change Detection Task	21	1560	2,4,6
H2020 - Exp1	Change Detection Task	22	2400	4
H2020 - Exp2	Change Detection Task	20	1920	4
B2019 - Exp1	(Bilateral) Change Detection Task	16	840	2,4
B2019 - Exp2	(Bilateral) Change Detection Task	16	660	2,4
B2019 - Exp3	(Bilateral) Change Detection Task	16	840	2,4
H2019	Change Detection Task	97	1600	2,4
G2019	Orientation Retro-Cued Task	30	500	1, 3*
VG2019	Change Detection Task	23	96	1,2,4
FW2018 - Exp1	(Sequential) Change Detection Task	23	960	1,2,3,4
FW2018 - Exp2	(Sequential) Change Detection Task	20	960	1,2,3,4
A2018 - Exp1	Lateralized Whole-Report Task	31	650	1,3,6
A2018 - Exp1	Lateralized Whole-Report Task	48	540	1,3,6

Table 2. Datasets - Details. If a study contains more than one experiment of interest, the different experiments have been listed with "- Exp #". The number of trials represents the total theoretical number of trials per participant according to the design of the study.

2.1. Feldmann-Wüstefeld et al., 2020

In *Neural measures of working memory in a bilateral change detection task* ([**12**]), Feldmann-Wüstefeld and colleagues used a novel change detection task in which both the CDA and the negative slow wave (NSW) can be measured at the same time. They presented memory items bilaterally with different set sizes in both hemifields inducing an imbalance or “net load” as they called it. Their results showed that the NSW increased with set size, whereas the CDA increased with net load. There were three

different set sizes participants had to remember: two, four or six targets. In Figure 1, we can see the five combinations of targets they used (*2:0, 3:1, 4:2, 4:0, 5:1*). With their nomenclature *2:0* means 2 targets in one hemifield and 0 target in the other hemifield for a net load of 2. *4:2* represents 6 targets total, 4 in one hemifield and 2 in the other for a net load of 2. As we can see, the highest CDA is obtained with *4:0* for a net load of 4. Interestingly, the *5:1* condition shows a higher CDA than 3:1 and 4:2 however lower than 2:0 indicating that having targets in both hemifields reduces the overall CDA amplitude. On the graph, t=0s is when the memory display appeared on the screen with the targets, then they stayed visible for 500ms after which the participant had to remember the targets for 1 second before the probe display appeared for the participant to provide their answer as to confirm if the item in the probe display was indeed part of the targets or not.

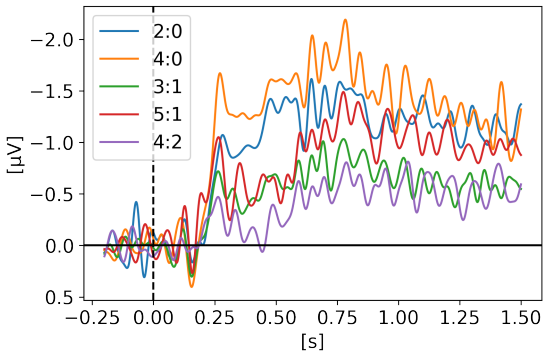


Fig. 1. Reproduced results from Feldmann-Wüstefeld et al., 2020, using our simple comparative pipeline.

2.2. Hakim et al., 2020

In *Perturbing neural representations of working memory with task-irrelevant interruption* ([**18**]), Hakim and colleagues investigated the impact of task-irrelevant interruptions on neural representations of working memory across two experiments

looking at both the CDA and lateralized alpha power. What they found is that after interruption, the CDA amplitude momentarily sustained but was gone by the end of the trial. On the other hand, lateralized alpha power, which has been used as an effective tool for discerning which visual hemifield is being attended, was immediately influenced by the interrupters but recovered by the end of the trial. Hakim and colleagues suggested that dissociable neural processes contribute to the maintenance of working memory information and that brief irrelevant onsets disrupt two distinct online aspects of working memory and also that task-irrelevant interruption could motivate the transfer of information from active to passive storage, explaining the reason why the CDA drops significantly after interruptions, even on trials with good performance (i.e. the participant remembered the targets correctly). In their 2nd experiment, they go further and test if the expectation of interruption changes the CDA. The full 1.65s epoch is displayed but not the response period which took place after 1.65s (see 2). The CDA was obtained by using only PO7 and PO8 electrodes.

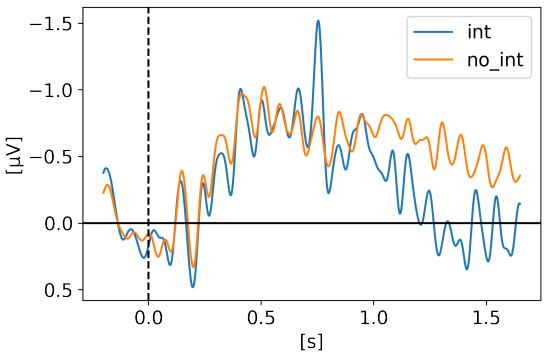


Fig. 2. Reproduced results from Hakim et al., 2020 - Experiment 1, using our simple comparative pipeline.

2.3. Balaban et al., 2019

In *Neural evidence for an object-based pointer system underlying working memory* ([**4**]), Balaban and colleagues argued that to update our representation of the environment, our VWM depends on a pointer system such that each representation is stably and uniquely mapped to a specific stimulus. Therefore, without these pointers, our VWM representations are inaccessible. Via three experiments, they examined whether the pointers are allocated in an object-based, featural, or spatial manner. Their results showed that the separation of an object in two halves invalidated the pointers. It happened in a shape task, where the separation changed both the objects and the task-relevant features, but also in a color task, where the separation destroyed the objects while leaving the task-relevant features intact. They suggested that objects, and not task-relevant features, underlie the pointer system. Two of their three experiments are displayed in Figure 3 and Figure 4, while the third experiment is available in the supplementary material to reduce the length of this manuscript and enhance readability.

For Experiment 1, t=0s is when the memory display appeared on the screen with the moving objects which either separated in halves after 400ms or either continued moving as a whole for another 600ms after which they stopped moving for 300ms and disappeared. After 900ms of retention with an empty screen displaying only a fixation cross in the middle, the participant had to give an answer as if the objects being displayed are the same or different than the initial ones. In one condition the shapes were the relevant features (i.e. have some of the shapes changed?) and in a second condition the colors were the relevant features (i.e. have some of the colors changed?). On the graph, the cognitive impact of the separation that happened for half the trials (at 400ms) is clearly visible in the CDA signal around 200ms as highlighted by the grayed region. The full 2.2s epoch is displayed on Figure 3 but not the response period

which took place after 2.2s. The CDA was obtained by using only PO7-PO8, P7-P8 and PO3-PO4 electrode pairs.

For Experiment 2 (Figure 4), the epochs were shorter in time and only the colors were the relevant features. The targets were all moving squares that could either separate after 400ms or continue moving as a whole. Some trials had two targets and some trials had four targets. In both experiments we see that after the separation, the CDA amplitude increases, as the number of targets to track has now increased.

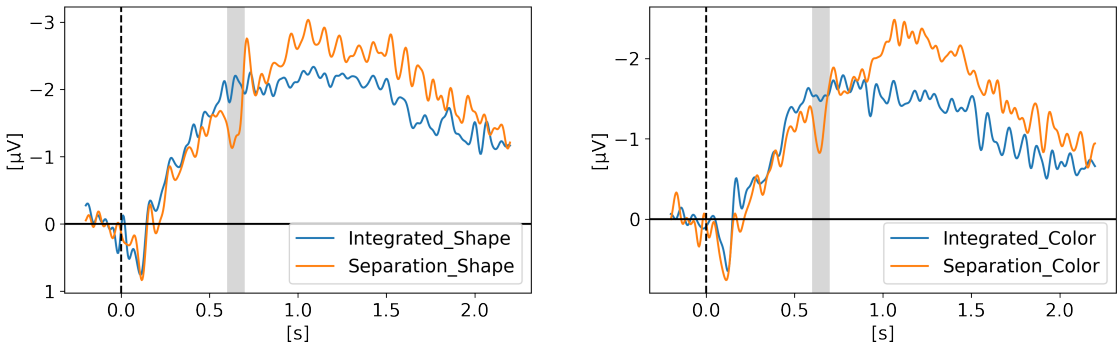


Fig. 3. Reproduced results from Balaban et al., 2019 - Experiment 1, using our simple comparative pipeline.

2.4. Gunseli et al., 2019

In *EEG dynamics reveal a dissociation between storage and selective attention within working memory* ([**4**]) Gunseli and colleagues tested the hypothesis that within WM, selectively attending to an item and stopping storing other items are independent mechanisms. In order to make participant drop items from WM, they used a retro-cue to indicate which of the items were the target(s). As opposed to identifying the targets at the beginning of the trial like in most WM studies, here they showed 3 items (bars with different orientations) and then, 1s later they showed a retro-cue indicating

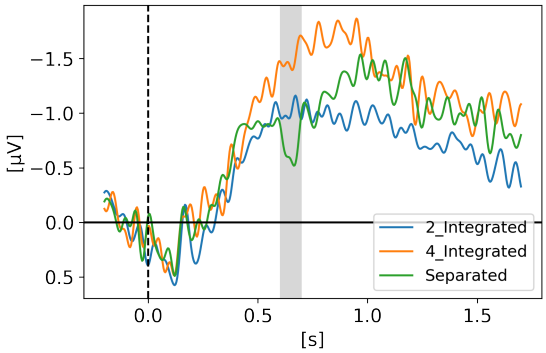


Fig. 4. Reproduced results from Balaban et al., 2019 - Experiment 2, using our simple comparative pipeline.

which item is most likely to be tested, or probed, after the retention phase. Their hypothesis was that if the retro-cue is reliable, the participant would drop the other item(s) creating an imbalance between hemispheres and therefore increasing the CDA signal. Whereas if the retro-cue is not reliable the participants would not drop the item(s) and therefore resulting in a smaller CDA signal. On Figure 5, we see that indeed the CDA is of higher amplitude for trials where the retro-cue is valid 80% of the time in comparison to the other condition where the retro-cue is valid only 50% of the time. The right graph was generated using their MATLAB preprocessed data files (.mat) and the left graph shows our reproduced version from their raw EEG data, with a lowpass filter at 6Hz, as they mention in their paper to leave the alpha band out of the CDA signal. Unfortunately, the preprocessing code used to generate the MATLAB preprocessed files was not available and it seems like these files benefited from additional manual cleaning because when we use a higher filter (e.g. 20Hz or 30Hz) with the same pipeline as other studies, both signals end up with a similar amplitude and the effect isn't visible anymore because of a high variability on both signals (50% vs 80%). In their paper they mentioned that using a higher filter (e.g. 40Hz) didn't change the results of the statistical analysis (relative to the 6Hz filter).

We did not perform any statistical analysis, however, we were only able to obtain a visible CDA difference when plotting the grand average with heavy filtering (6Hz), which helps reduce the variability. Only the PO7-PO8, P7-P8, and O1-O2 electrode pairs were used to generate the CDA.

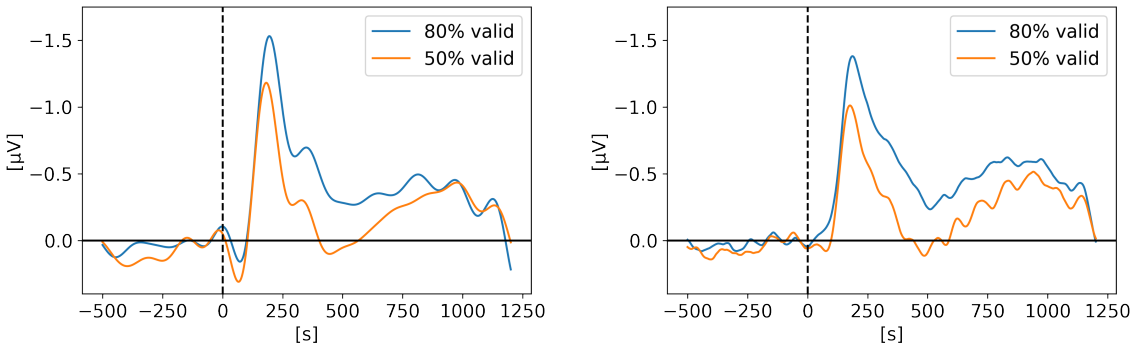


Fig. 5. Reproduced results from Gunseli et al., 2019, using our simple comparative pipeline. (a) CDA calculated from raw EEG files. (b) CDA obtained from preprocessed MATLAB files.

2.5. Villena-Gonzalez et al., 2019

In *Data from brain activity during visual working memory replicates the correlation between contralateral delay activity and memory capacity* ([**34**]), Villena-Gonzalez and colleagues replicated the results from Vogel, 2004 showing that the amplitude of the CDA correlates with the number of items held in WM using a change detection task with set sizes of one, two and four. Moreover, they also looked at the individual performances and showed that participants with higher WM capacity (i.e. better performance on the task) also had a higher CDA amplitude. Figure 6 shows a clear difference between one target and two or four targets, however, we see similar amplitudes for two and four targets. Unfortunately, we were not able to reproduce

their results showing a clear difference of amplitude between two and four targets. In their paper, they had a significant higher amplitude for four targets, which we failed to reproduce after a few attempts at with different automated preprocessing pipelines. Our version doesn't invalidate their conclusion as we also see a higher CDA amplitude the larger the set size, however the effect between two and four isn't as strong as in their findings. This might indicate that the results could have benefited from extra manual cleaning. The following electrode pairs were used to obtain the CDA: TP7-TP8, CP5-CP6, CP3-CP4, CP1-CP2, P1-P2, P3-P4, P5-P6, P7-P8', P9-P10, PO7-PO8, PO3-PO4, and O1-O2.

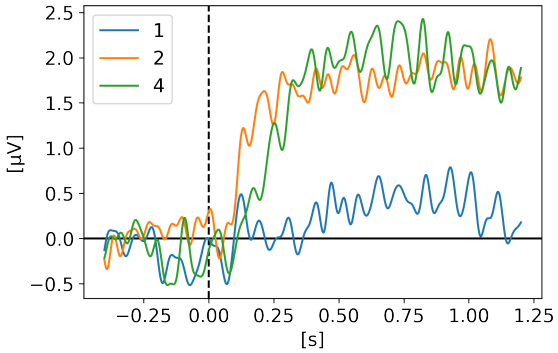


Fig. 6. Reproduced results from Villena-Gonzalez et al., 2019, using our simple comparative pipeline.

2.6. Hakim et al., 2019

In *Dissecting the Neural Focus of Attention Reveals Distinct Processes for Spatial Attention and Object-Based Storage in Visual Working Memory* ([**17**]), Hakim and colleagues showed that the focus of attention in WM is not a monolithic construct but rather involves at least two neurally separable processes: (a) attention to regions in space and (b) representations of objects that occupy the attended regions. On Figure 7, the CDA is clearly visible, showing a slightly higher amplitude for a set size

of 4 targets vs 2. The full 1.45s epoch of the change detection task is displayed but excludes the response period which took place right after. It is important to note that the graph represent the aggregation of 4 sub-experiments with slight variations on the task. The CDA was obtained by using only O1-O2, PO3-PO4, PO7-PO8, P3-P4, and P7-P8 electrode pairs.

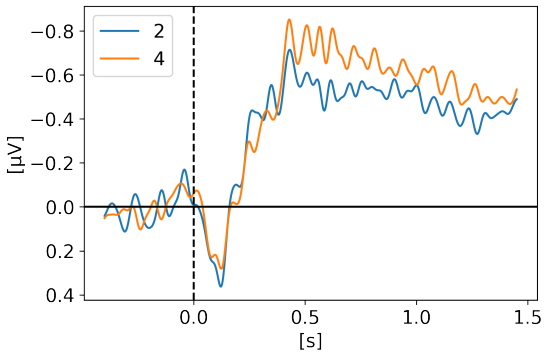


Fig. 7. Reproduced results from Hakim et al., 2019, using our simple comparative pipeline.

2.7. Feldmann-Wüstefeld et al., 2018

In *Contralateral Delay Activity Indexes Working Memory Storage, not the Current Focus of Spatial Attention* ([**13**]), Feldmann-Wüstefeld and colleagues seek to explore the recent hypothesis from Berggren and Eimer suggesting that the CDA tracks the current focus of spatial attention as opposed to working memory storage ([**6**]). Figure 8 shows the CDA results of both their experiments in which they displayed four targets among distractors via two sequential memory displays in a change detection task. The first set of targets is shown at t=0s for 200ms then disappears for 500ms after which a second set of targets and distractors appear for 200ms and then disappear for another 500ms after which the probe display is shown for the participant to confirm if the items currently displayed are the same as the targets. The whole 1.4s epoch is

displayed on Figure 8, leaving the response period out of that figure. A total of four targets were always shown to the participant. The first memory display (i.e. the first set of targets and distractors to be shown) could either have 1, 2 or 3 targets and the second memory display, 700ms later, could add 3, 2 or 1 targets for a combined total of 4. The targets could be either added in the same hemifield or the different (i.e. opposite) hemifield. The expected result is a higher CDA when the targets are added in the same hemifield as this will increase the "net load" (term they will later use in their 2020 paper) and a lower CDA amplitude when added in the opposite hemifield because it would then decrease the net load. Experiment 1, on the upper row, shows such expected results somewhat perfectly as on the top left graph we see all three CDA signals reaching the same amplitude after both sets of targets are added and equals to four items in the same hemifield. As expected, the top right graph shows a slight decrease on the CDA amplitude for *3+1 diff* but more interestingly, a CDA of pretty much zero for the *2+2 diff* condition where two targets were shown in both hemifields cancelling out the CDA signal. For the *1+3 diff* condition, we see the CDA flipping side after the new set of three targets is shown on the opposite side. Experiment 2 was similar to Experiment 1 except for the probe display when the participant provides the answer. In Experiment 1 the probe display showed the items at the same spatial location that they were displayed in either the memory display 1 (at t=0ms) or 2 (at t=700ms). However, in Experiment 2 the probe display was modeled after Berggren and Eimer ([**6**]) experiment such that the items were spatially translated and interleaved. The authors suggested that given that the retention period of Experiment 1 and 2 were identical (i.e. the 1.4s displayed on the graph) and that only the probe display was different, the differences observed between both experiment can only be explained by different memory strategies. They therefore suggested that because the mental representation in Experiment 2 is more difficult as the items are not displayed "as is" but translated on the probe display, the participants most likely transferred items of display 1 (i.e. first set of targets) from working memory (WM)

into long-term memory (LTM) therefore explaining why in Experiment 2 the CDA seems more affected by the second set of items rather than equally affected from the first and second set of targets as in Experiment 1. A clear example of that difference between experiments is the *2+2 diff* condition where the CDA is pretty much zero in Experiment 1 but biased towards the second set of targets in Experiment 2. Only the PO7-PO8 electrode pair was used for the CDA signal.

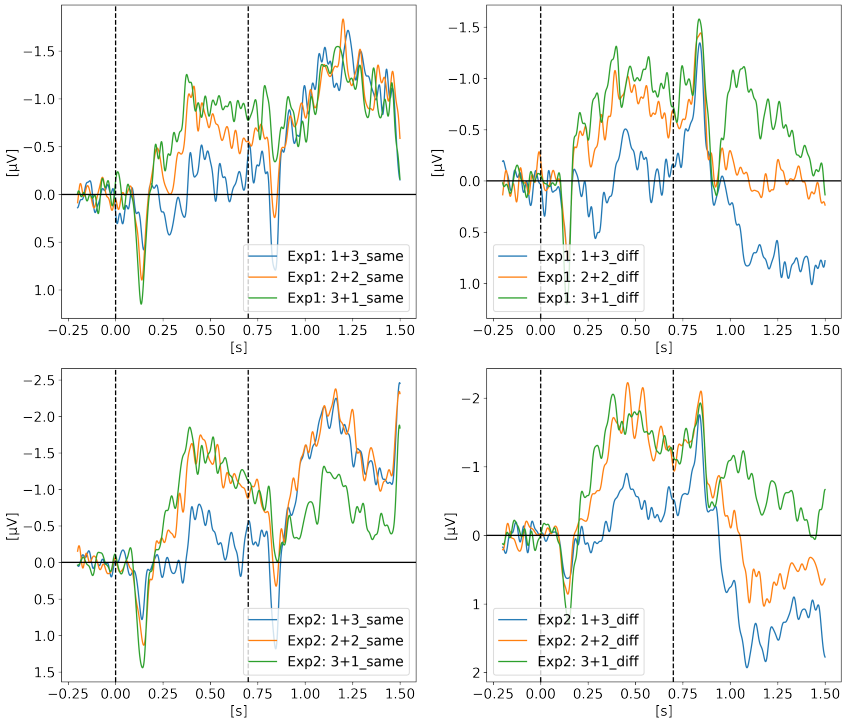


Fig. 8. Reproduced results from Feldmann-Wüstefeld et al., 2018, using our simple comparative pipeline.

2.8. Adam et al., 2018

In *Contralateral delay activity tracks fluctuations in working memory performance* ([**1**]), Adam and colleagues looked at the relationship between the CDA amplitude and working memory performance. Their hypothesis was that if working memory

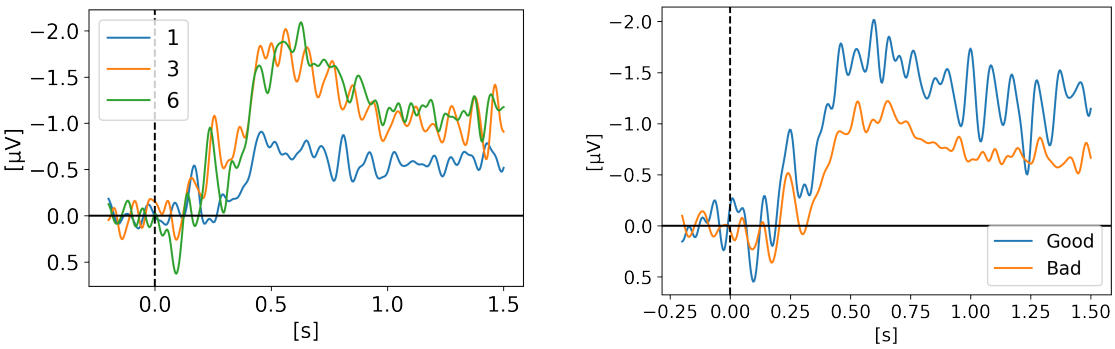


Fig. 9. Reproduced results from Adam et al., 2018, using our simple comparative pipeline. Experiment 1 (left) & 2 (right)

failures are due to decision-based errors and retrieval failures, CDA amplitude would not differentiate good and poor performance trials when load is held constant. If failures arise during storage (i.e. retention phase), then CDA amplitude should track both working memory load and trial-by-trial performance. Their Experiment 1, shown on Figure 9, showed that the CDA amplitude increased with set size but plateaued at three targets showing a similar amplitude for three or six targets. In their Experiment 2, they kept the set size at 6 and compared the good trials (accuracy of 3 or more targets identified correctly out of 6) vs bad trials (accuracy of 2 targets or less identified correctly). Figure 9 (on the right) shows that indeed the amplitude of the CDA correlated with the performance. The O1-O2, OL-OR, P3-P4, PO3-PO4, T5-T6 electrode pairs were used for the CDA.

3. Results

Before analysing the CDA and drawing conclusions on the underlying cognitive functions, we should understand how to best obtain a clean CDA signal in the first place. In this section, we will first discuss what channels and reference(s) various

groups have used to obtained their CDA. Second we will discuss some trends we have observed across the reviewed studies.

Note that we did not look at any eye-tracking data and assumed that the subjects respected the instructions of fixating the middle of the screen. Many studies had the eye-tracking data available, however we did not use it nor have we excluded any trials based on eye-movement. This does certainly impact the results we obtained when compared to the original curves of the authors in their paper. Given that such data was not available for all datasets, we decided to not consider it at all.

3.1. EEG Channels

The contralateral delay activity (CDA), as its name implies, is a difference in the activity between the left and right hemisphere. The signal is obtained by subtracting one or more contralateral channels to equivalent ipsilateral channel(s) to the attended side. Unsworth et al., mentioned in their 2015 paper that it is now standard procedure for measuring the CDA to use posterior parietal, lateral occipital and posterior temporal electrodes (PO3, PO4, T5, T6, OL, and OR), citing the work of Vogel & Machizawa, 2004 and Vogel et al., 2005. Looking at the electrodes used in the reproduced studies from 2018 to 2020, there seems to be a slight change towards favouring PO7/PO8 as the best electrode pair. Table 3 shows the selected channels used by the reviewed studies to obtain the CDA. Adam, 2018, used the recommendations from Unsworth, 2015, without the PO7/8 pair, while all the others included at least the PO7/8 pair. Feldmann-Wüstefeld, 2018 & 2020, and Hakim, 2020, used only that pair for their CDA results and did not average with any other pairs. Interestingly, in her 2019 study, Hakim had use PO7/8 but also P7/8, P3/4, PO3/4, and O1/2, which were not included in the 2020 analysis. Villena-Gonzalez, 2019 used the most electrode pairs of the reviewed studies, including parietal, occipital, temporal and central sites. It is worth noting that the PO7/8 is not an electrode pair in the standard 32-electrode

10-20 placement. This most likely explains why Adam, 2018 did not include this pair given that they use a system with 20 electrodes with 10-20 locations.

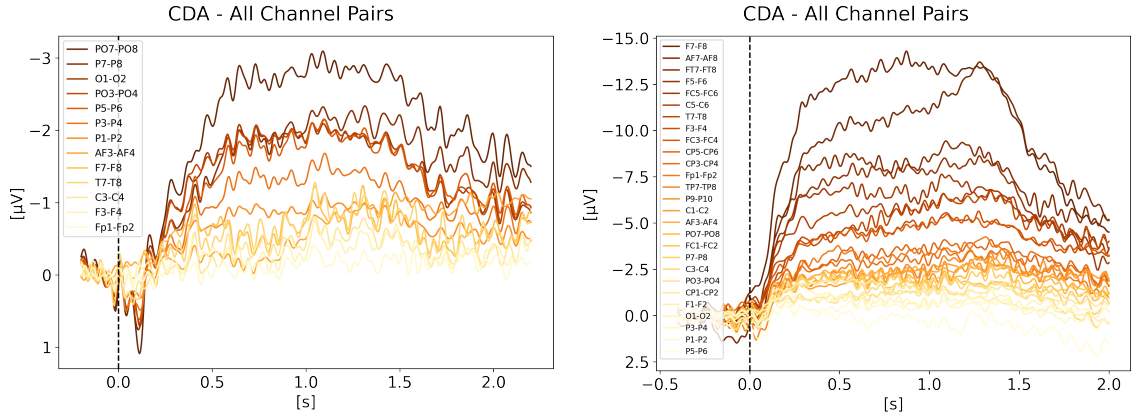
To confirm our empirical estimation of PO7/8 being the best electrode pair for CDA, we calculated the effect size for each available pair of electrodes by comparing contralateral vs. ipsilateral activity for each study. All the datasets had the PO7/8 pair available but Adam 2018. Aside from Villena 2019, the largest effect size was always observed from a parietal electrode pair. PO7/8, PO3/4 and P7/8 dominated the top three electrode pairs in most studies. PO7/8 had the largest or second largest effect size for 5 of the 8 studies. The breakdown is available on the repository online.

Study	**Channels**
Feldmann-Westerfel, 2018	PO7/8
Feldmann-Westerfel, 2020	PO7/8
Hakim, 2020	PO7/8
Gunseli, 2019	PO7/8, P7/8, O1/2
Balaban, 2019	PO7/8, P7/8, PO3/4
Adam, 2018	O1/2, OL/R, P3/4, PO3/4, T5/6
Hakim, 2019	O1/2, PO3/4, PO7/8, P3/4, P7/8
Villena-Gonzalez, 2019	TP7/8, CP5/6, CP3/4, CP1/2, P1/2, P3/4, P5/6, P7/8, P9/10, PO7/8, PO3/4, O1/2

Table 3. Channel pairs used for the CDA.

In order to better understand the signal shape and amplitude coming from the various electrode pairs, we looked at each pair for each condition of each study. Regardless of what the study actually used for their analysis, we used all channel pairs available in the raw data. We included here only the breakdown for Balaban 2019 and Villena 2019 on Figure 10. All the other studies followed the trend of Balaban 2019

showing a stronger CDA from the parietal sites with PO7/8 being the best candidate (i.e. strongest CDA amplitude) across studies. Villena 2019, is the only one showing different results with a very strong frontal lateralized activity (see Figure 10(b)). We initially thought these results were odd, until we analyzed our own data recorded for another project while writing this review, for which we also observed frontal activity being way stronger than parietal activity. Our task is a 3D-MOT task (see [**13**]).



(a) (b)

Fig. 10. CDA Channel Pairs. (a) Shows the CDA from Balaban et al., 2019 for all available channel pairs. The CDA is the grand average across subjects from all trials with a good performance for the condition integrated shape in experiment 1. (b) Shows the CDA from Villena-Gonzalez et al., 2019 for all available channel pairs. The CDA is the grand average across subjects from all trials with a good performance for the condition with 2 items.

Finally, the approach for averaging the channel pairs also varied. Some groups subtracted the pairs first and then averaged across pairs, while others averaged one side first and then subtracted to the average on other side. The order in which the averaging and subtracting happens doesn't really matter if no other operation is done on the signal and if they are all referenced to the same reference. For example, if

we are trying to use the pairs PO7-PO8, P7-P8 and PO3-PO4, we could either start by subtracting each pair individually (e.g. signal from PO8 minus signal from PO7) and then averaging the 3 signals obtained, giving us the resulting CDA or we could average one side together (PO7, P7, and PO3) and subtract it from the average signal from the other side (PO8, P8, and PO4). Both approaches would give the same CDA for a given trial.

3.2. EEG Reference(s)

The EEG signal being an electric potential difference between two electrodes, namely the electrode of interest and the reference, it is no surprise that the reference plays an important role in EEG studies. Changing the reference can drastically change the signal obtained and therefore influence the conclusion of a study. The ideal reference would be a neutral point with no electrical activity to which we could measure a difference of potential being only the activity of interest. Unfortunately, no such point exists on the body, leaving the problem of finding a good reference unsolved or open to interpretation. Best practices in EEG studies include (1) using the average of left and right mastoids or (2) to re-reference the signal offline to the average of all channels or (3) using one or multiple electrode(s) on the midline such as Cz or (4) using a mathematical reference such as the reference electrode standardization technique (REST) or a Laplacian approach ([**37**]). Here in the reviewed studies, all of them used to the average of left and right mastoids except for Feldmann-Wusterfeld 2018 and 2020 which used the average of all electrodes.

Given that CDA is a difference between left and right hemispheres, the reference isn't as important as in most evoked-related potential (ERP) studies. The choice of reference will impact the visual inspection and the cleaning of the data, however, when it comes to obtaining the CDA itself, because we are subtracting one electrode to the other, the reference gets somewhat cancelled and therefore does not impact the

resulting CDA signal as much. For example, if the contralateral electrode is PO8(-Ref) and the ipsilateral electrode is PO7(-Ref), then CDA (contralateral - ipsilateral) is (PO8 - Ref) - (PO7 - Ref) which is equivalent to PO8 - PO7 as the Ref cancels out. It is therefore still important to consider the reference for cleaning the signal but to keep in mind that the CDA itself isn't as affected as much by the reference as typical, non-latelarized, ERPs. This actually makes the CDA signal even more robust to noise, assuming that most EEG noise would be visible on both channels, and be cancelled out during the subtraction. A noise that would survive that subtraction would be a noise seen only by one electrode or one 'side' of the head.

3.3. CDA Decay

If the amplitude of the CDA correlates with the number of items being tracked, one could expect the amplitude to remain somewhat stable for the whole retention or tracking duration. However, in all the replicated studies we can observe the amplitude of the CDA declining way before the end of the retention phase even when the participants did not lose track of the item(s).

In later section 3.6 we look at the CDA when subjects lost track of one of more item(s) (i.e. bad trials) but all figures previously showed with the reproduced results and discussed in the methods section were generated from *good performances* only, meaning that the subjects did remember all the items for the whole duration of the trial and provided good answers at the end. As we can see on Figures 1, 3, 4, 5, and 7 the CDA is starting to decrease way before the end of the trials. In Hakim, 2020, they added interruptions trying to interfere with WM processes and CDA indeed dropped significantly shortly after such interruptions. And yet, despite the important drop in amplitude the participants still provided correct answers. While none of the eight studies addressed directly what seems to be a natural CDA decay, some suggested (e.g. Hakim, 2020; Feldmann-Wüstefeld, 2018) that the items could be transferred

from short term memory to long term memory therefore affecting the CDA amplitude observed here.

3.4. Recall

One interesting phenomenon that most studies did not mention in their paper is the recall, or response, period (i.e. when the participant is providing the answer). Interestingly, in most studies, we are observing a re-increase in the amplitude of the EEG signal, sharing similarities with the CDA observed during the identification and retention (or tracking) periods. One potential reason most studies don't look at that period is because eye movements are generally not controlled in this phase, introducing artifacts in the EEG signal. Moreover, the induced artifacts are most likely lateralized and not normally distributed as the participants might be fixating and moving their eyes to the side of interest, complicating the dissociation between eye-movements and cognitive processes in the EEG signal.

Figures 11, 12, 13, and 14 show the CDA graphs but with a longer epoch this time, including two seconds after the end of the retention phase. The right-most dashed line represents when the participant was asked to provide an answer. The figures all show a re-increase in the signal amplitude also followed by a decay of the signal as discussed in the last section. Only two studies did not show such re-increase of a CDA-like signal during recall: Villena-Gonzalez, 2019 and Gunseli, 2019. The graphs are available online in our repository with the analysis.

In order to investigate if the re-increase of the signal is mainly driven by eye-movements or by cognitive processes, we looked at different channel pairs with the assumption that the frontal pairs, especially the ones further from the midline, would be the most affected by eye-movements and such artifacts would then leak to central and parietal channels. We show the breakdown of channel pairs for Balaban, 2019 Experiment 1 on Figure 15, Feldmann-Wüstefeld, 2018 Experiment 1 on Figure 16,

and Hakim, 2019 on Figure 17. The different channel pairs available in the dataset, not only the one used for the CDA, were plotted for the whole trial and extending over the response period. What we observe is that indeed there seems to be a mixture of cognitive processes and eye-movement artifacts. For Balaban, 2019, we see the re-increase in amplitude in the same parietal channels while the frontal channels remain unaffected during the early part of recall, suggesting that the underlying activity is not only from eye-movements but similar to the one during the retention period eliciting the CDA. For Feldmann-Wüstefeld, 2018, on the 2nd and 3rd graph of Figure 16, we see a big change in amplitude for the F7-F8 pair during recall, which is most likely driven by eye-movements rather than cognitive processes. However, on the 1st graph, for the *1+3 same* condition, the F7-F8 pair isn't being impacted as much and have a similar amplitude as the parietal channels, suggesting that the activity observed from the parietal channels is mostly cognitive while being slightly affected by eye-movement artifacts. It isn't very likely that the eye-movement artifacts would affect more the parietal channels than the frontal ones. In Hakim, 2019 on Figure 17, we observe a very strong increase of amplitude on the F7-F8 pair suggesting that the response period is highly contaminated by eye-movement artifacts making it quite difficult to draw any conclusion on the re-increase in amplitude in the parietal channels as it could simply be driven by the eye-movement artifacts.

Here we provide some exploratory graphs, however, it is too early to draw any strong conclusions on the nature of the CDA-like amplitude re-increase during recall, without a deeper analysis of eye-tracking and EOG data which was outside of the scope of this reproducibility study.

3.5. CDA Amplitude vs Number of Items

Looking at the reproduced results, it seems fair to conclude that indeed the amplitude of the CDA correlates with the number of items tracked by the participants

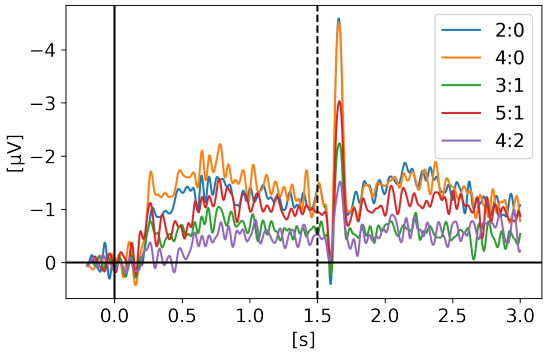


Fig. 11. Feldmann-Wüstefeld, 2020 - Recall/Probe. Same CDA as on Figure 1 but with a longer epoch, showing a CDA re-increase during recall (t > 1.5s).

up to a plateau of 3 to 4 items. This correlation and plateau has been discussed in previous CDA studies (e.g. [**46, 44, 25**]). Figure 6 from Villena-Gonzalez, 2019, data shows a clear difference in CDA amplitude between one, two and four items. Figure 9 from Adam, 2018, data shows a clear difference between one and three or one and six but a very similar amplitude for three and six items, aligned with some sort of CDA amplitude plateau around three or four items. Figure 7 from Hakim, 2019, data shows a small difference between two and four items.

The work from Feldmann-Wüstefeld, 2018 (Figure 8) shows that even after the initial identification and tracking phases, the CDA can be increased by adding items on the same side or decreased by adding items on the opposite side. Similarly, the work from Balaban, 2019 (Figures 3 and 4) shows that when a whole item splits in two separate parts that need to be tracked, the CDA increases. Feldmann-Wüstefeld, 2020 (Figure 1) used a bilateral task with different loads on both sides creating a net load of either two or four items and what their work shows is that different combinations of net load of two or four produce different CDA amplitudes. The more objects on the opposite side, the lower the CDA amplitude despite the same net load. However, the results are aligned with the CDA theory saying that the more objects the higher

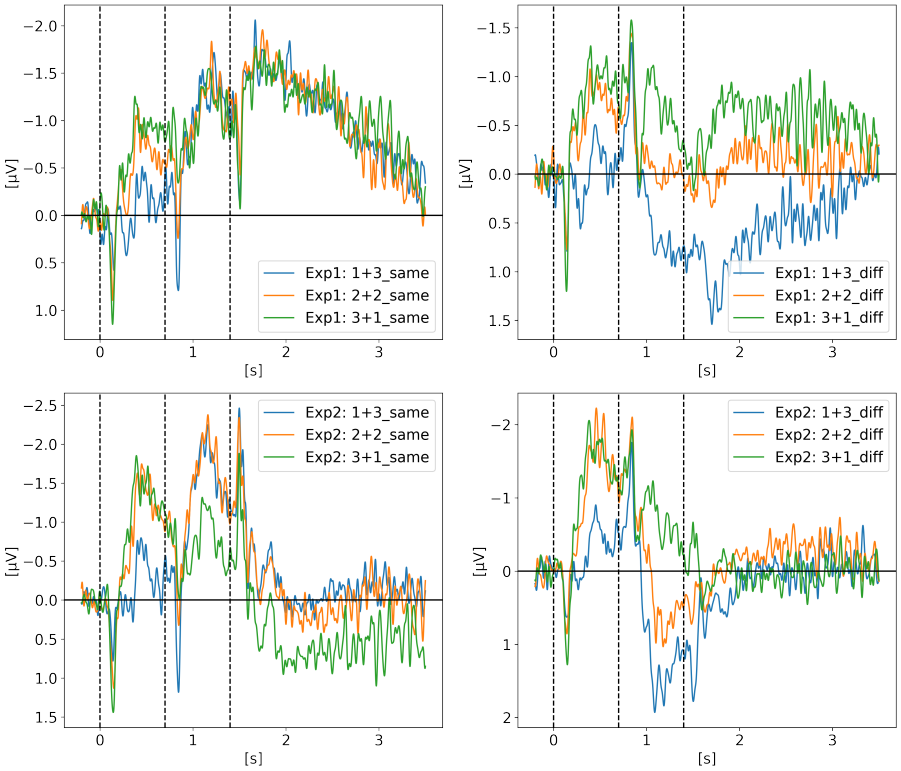


Fig. 12. Feldmann-Wüstefeld, 2018 - Recall/Probe. Same CDA as on Figure 8 but with a longer epoch, showing a signal amplitude re-increase during recall (t > 1.5s) for Experiment 1.

the amplitude. The 4:0 condition shows a higher amplitude than the 2:0 condition. The 5:1 shows a higher amplitude than the 3:1.

3.6. CDA Amplitude vs Individual Performance

If the CDA amplitude correlates with the number of items held in memory, the CDA should be indicative of the performance of a specific trial. This obviously holds true only if the working memory failures occur during the retention phase and not during the recall phase to provide the answer. If such memory fault occurs during the initial identification phase, the CDA would not drop per se but instead reach a lower

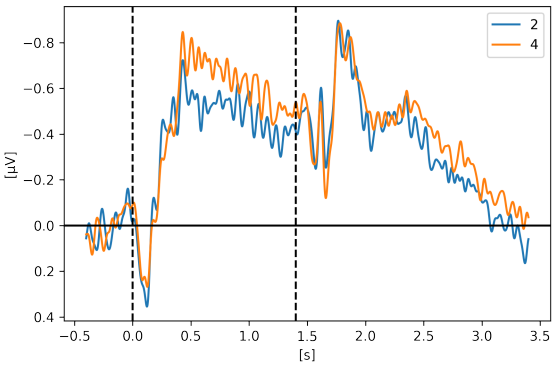


Fig. 13. Hakim, 2019 - Recall/Probe. Same CDA as on Figure 7 but with a longer epoch, showing a CDA re-increase during recall (t > 1.4s).

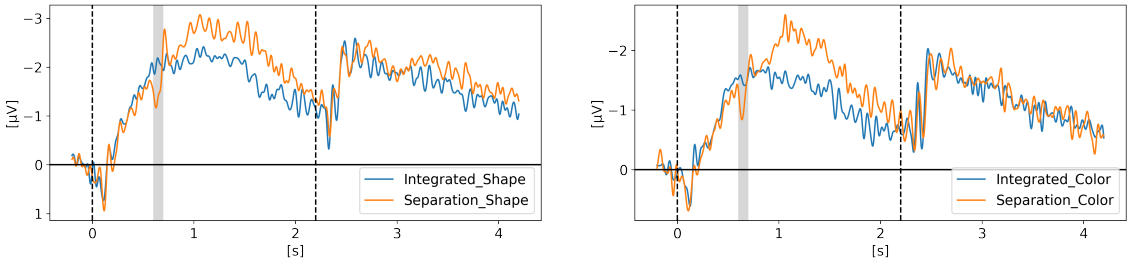


Fig. 14. Balaban, 2019 - Recall/Probe. Same CDA as on Figure 3 but with a longer epoch, showing a CDA re-increase during recall (t > 2.2s)

amplitude than the expected peak amplitude should the participant have tracked the correct amount of items. However, if the participant mistakenly identifies the wrong target (for example in a multiple-object tracking task) and confidently holds and tracks the correct number of items, despite some of them being wrong, this kind of mistake would result in the same CDA amplitude as if the right items were kept in memory. While the current state of CDA literature lacks the trial-by-trial analysis to evaluate the role CDA plays in performance, several studies have looked at the individual differences in CDA amplitude vs working memory performance or capacity.

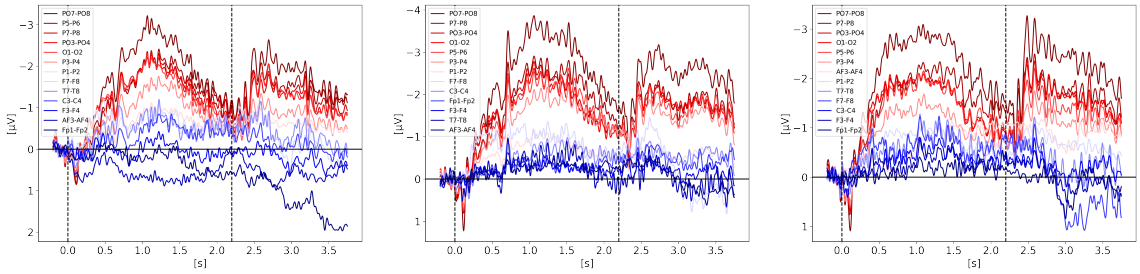


Fig. 15. Balaban, 2019 - Recall/Probe. Same CDA as on Figure 3 but with a longer epoch, showing a CDA re-increase during recall (t > 2.2s). Three conditions are shown, from left to right: *Separation Color, Separation Shape, Integrated Shape.*

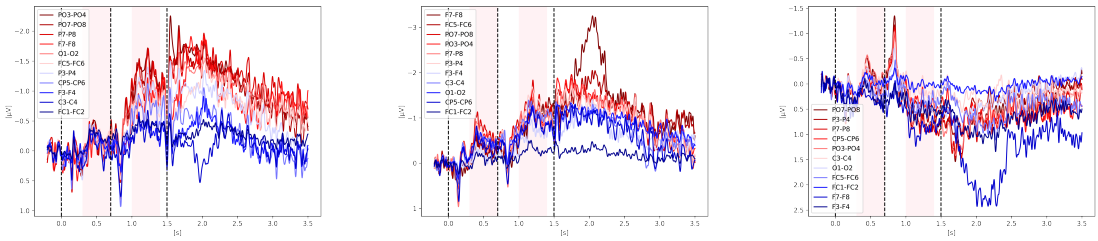


Fig. 16. Feldmann-Wüstefeld, 2018 - Recall/Probe. Same CDA as on Figure 8 but with a longer epoch, showing a signal amplitude re-increase during recall (t > 1.5s) for Experiment 1. Three conditions are shown, from left to right: *1+3 Same, 2+2 Same, 1+3 Diff.*

For example, Unsworth and colleagues in 2015 ([**44**]) replicated the work from Vogel & Machizawa, 2004 ([**46**]) showing a correlation between CDA and performance on a change detection task where high working memory individuals had larger CDA (r=0.30; Unsworth, 2015) and concluded their study saying that CDA is a reliable and valid individual measure of working memory that predicts behavioral performance on visual working memory tasks. Of the studies reviewed here, only three looked at the direct correlation between individual working memory performance and CDA amplitude. Most studies looked at the relationship between task conditions and performance,

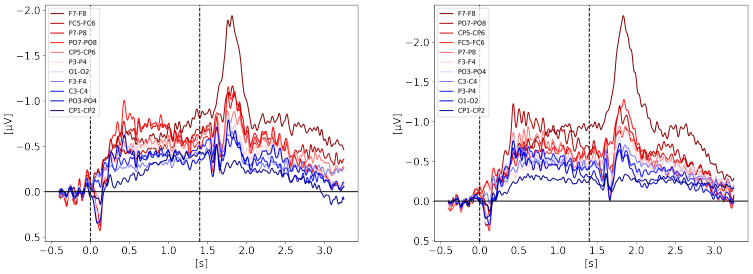


Fig. 17. Hakim, 2019 - Recall/Probe. Same CDA as on Figure 7 but with a longer epoch, showing a CDA re-increase during recall (t > 1.4s). Two conditions are shown, *set size = 2* on the left and *set size = 4* on the right.

as well as the relationship between task conditions and CDA amplitude, however, only three looked at the direct correlation between performance and CDA amplitude (Adam, 2018; Feldmann-Wüstefeld, 2020; Villena-Gonzalez, 2019). Among these three studies, working memory performance, or capacity, was evaluated differently based on the task and the granularity in the answers provided by the participants.

In Adam, 2018, they used a whole report task with a set size of six items and noted the participants' answer accuracy between zero and six targets identified correctly. With this granularity, they looked at the correlation between CDA amplitude and accuracy (from 0/6 to 6/6, steps of 1) and showed individual differences in working memory performance with a correlation of r=-.26 and p=.028. They also noted that the magnitude of the effect is relatively small, but consistent with previously observed effects in the literature, citing Unsworth, 2015 work. With their EEG and behavioural data we were able to reproduce a similar result with a slightly higher and more significant correlation (r=-0.34, p=0.001). It is important to note that the classes are quite unbalanced as there are very few trials with an accuracy score of 0/6 (all wrong) or 6/6 (all good) and two thirds (66%) of the trials have an accuracy of either 2/6 or 3/6.

In Feldmann-Wüstefeld, 2020, they used the K score from [**8**] as a performance measure for each subject, where $K = N \times$ *(hit rate — false alarm rate)* and reported a correlation r=-0.43, p=0.026 in their study. With their EEG and behavioural data we were able to reproduce a similar result with a slightly higher correlation (r=-0.53, p=0.016).

Villena-Gonzalez 2019 used the formula from [**29**], *K = S x ((H-F)/(1-F))*, where H is the observed hit rate, F the false alarm rate and S is the higher set size (maximum number of to-be-remembered items). In their study, instead of looking at the correlation with WMC and CDA amplitude, they looked at the increase in CDA amplitude between set size of two and set size of four items and reported r=0.448; p=0.0159 showing that participants with high WMC (i.e. better performance) showed larger amplitude increase in CDA between two and four items, compared with participants with low WMC. Unfortunately, their behaviour files were not available with their EEG data, preventing us from calculating the WMC with the same formula. In the EEG files, a binary trigger identifying good and bad answers was available so we were able to calculate the accuracy for each participant. Unfortunately, when correlating the difference in amplitude between ss=2 and ss=4 with accuracy for each participant we found no significant correlation and therefore were unable to reproduce the effect of CDA on individual performance. It is worth noting that according to a recent study by Ngiam and colleagues, a substantial number of subjects and trials is required to detect a significant CDA difference between set sizes (approximately 400 trials with 25 subjects for approximately 80% statistical power) ([**30**]). Therefore, according to Ngiam's study, the Villena-Gonzalez design is quite underpowered.

The figures of the correlations, statistical analyses and distributions are available in the supplementary material.

3.7. Subjects Variability

Given the weak but consistent correlation between CDA amplitude and individual performance, we looked at the individual level to explore the variability on CDA across participants. On figures 19, 18, 21, and 20 in supplementary material, we did plot the CDA amplitude of the *top 5* and *bottom 5* participants of four different studies reviewed here (Adam, 2018; Villena-Gonzalez, 2019; Balaban, 2019; Feldmann-Wüstefeld, 2020), including three different conditions for each. These figures are exploratory and no further statistics were performed, however, we believe it gives more perspective on the CDA shape and its variability across subjects and across studies. The blue graphs on left represents the CDA of the 5 participants with the best performance on the task (top 5) and the orange graphs represents the CDA for the 5 participants with the worst performance (bottom 5). The top left graphs in blue on the first row represent the best participant, performance wise, and the bottom right graphs in orange the last row represent the participant with the lowest performance score. By visually looking at the graphs, it would be difficult to decipher a clear trend of CDA amplitude on performance, explaining the weak correlation. It is important to note here that the y axis (CDA amplitude in microvolt) was not fixed since the peak-to-peak amplitude varies quite significantly between participants and finding a once-size-fits-all range ends up hiding the shape of the CDA, which is what we seek to showcase here. We invite the reader to pay a close attention to the value on the y axis before drawing any conclusion.

4. Discussion

As shown on previous figures, all the reproduced studies showed a clear CDA across different VWM paradigms. While it is now fair to say that the CDA amplitude correlates with the number of items in WM, the shape and amplitude of the CDA requires more investigation to be better understood. For example, in Balaban, 2019,

when the items are separated, the CDA increases and reaches a higher peak than for integrated shape as expected if the CDA amplitude correlates with the number of items in working memory. However, by the end of the retention phase, just before being probed, the two conditions have pretty much the same CDA amplitude, which would normally indicate that at that point in time there is the same number of items in working memory, which is not the case. Moreover, in Feldmann-Wüsterfeld 2018 Experiment 1, we see a cumulative effect of the CDA amplitude, whether from external stimuli or during recall. As mentioned before, the CDA decay can either come from an amplitude increase in the ispilateral electrodes, reducing the difference between contra minus ipsi, or from a decrease of the amplitude in the contralateral electrodes. A combination of both is also possible. Fukuda, Woodman, and Vogel suggested that the waning CDA amplitude over time is actually a result of a selective increase in the negativity of the ipsilateral electrodes, while activity in the contralateral electrodes appears mostly unchanged ([**14**]). If the decay is caused by a decrease of amplitude in the contralateral electrodes, it would be coherent with WM models where oscillatory processes are keeping the information 'alive' via cell assembly firing together synchronizing lower frequencies in the theta and alpha bands (4 Hz to 12 Hz) with higher frequencies (40 Hz). The cell assembly passively decays, requiring regular updates before it decays too much and the information is lost ([**20, 22**]).

While we did not investigate further the CDA-like signal during recall, we thought it would be relevant to highlight it in this review as this period was not looked at in any of the reproduced studies and could help shed light on the neural correlates involved in VMW and underlying the CDA. For example, in their 2018 study, Feldmann-Wüstefeld and colleagues mentioned that the difference between their Experiment 1 and Experiment 2 is the memory strategy the participants might have used. Interestingly, when we compare the recall/response period of Experiment 1 we see the CDA amplitude re-increasing, however in Experiment 2 we do not see the CDA re-increasing, as if indeed a different recall strategy was used. Interestingly, the CDA-like signal during

recall seems to have the same amplitude for all conditions which is not the case during the retention phase (see Figures 14, 13, 11, and 12 top-left corner). If the observed EEG signal represents cognitive processes and not simply eye-movement artifacts, a possible explanation could be that during the initial identification of the items, the indexing happens one by one in a serial fashion, increasing the CDA in a cumulative way for each item being indexed, explaining the different amplitudes for different number of items. During recall, the update or refreshing of working memory is internally induced and could bring back all the items at once, hence showing the same CDA amplitude for all conditions. Moreover, the rate of ascend of the signal is significantly higher during recall, as if the items were brought back in a more parallel fashion to memory. Here we shared our early speculative ideas, however, given that we did not look at eye-tracking data nor exhaustively remove trials based on EOG and eye-movements during the response period, a more throughout examination would be required before drawing any strong conclusions on the meaning of that CDA-like signal.

As mentioned earlier, this sample of studies does not represent an exhaustive list of CDA studies with publicly available EEG datasets and this review could suffer from a bias given that many of the authors of selected studies are present and/or previous colleagues and collaborators.

There are several barriers for researchers to make their research reproducible, however, one that can be alleviated is the lack of knowledge and awareness about the available tools and best practices for reproducibility. In this study, we thought we would share what others have done and how they've done it, with the hope of helping other fellow researchers to do the same. Hopefully this study will reduce the fear of the extra steps required to make your study reproducible as well as highlight the value of taking these extra steps enabling one's work to generate a greater impact on the field.

Moreover, the FAIR principles ([**193**]) and the Brain Imaging Data Structure (BIDS) ([**55**]) both provide guidelines and standards on how to acquire, organize and share brain-related data and code. As the amount of recorded brain data keep increasing around the world and becomes more openly accessible it is important to have best practices to reduce the friction and wasted time. In addition to reproducibility aiming at validating or invalidating scientific evidences, data mining leveraging new approaches such as artificial intelligence (AI) and deep learning (DL) can benefit from having access well documented and openly accessible brain datasets ([**7**]). Trainees are also benefiting tremendously from reproducible experiments. Unintuitively however, most young trainees venturing in a new field think that it will be quick and easy to reproduce someone else's work and then take it from there to either modify or improve it. Unfortunately, they quickly realize that despite being theoretically easy, it isn't. There are always a multitude of complications from different operating system (OS) compatibility issues to versions of software libraries to missing parts of the code or lack of documentation making it hard to understand the code and its logic. A better reproducibility culture and best practices can help reduce significantly such friction and waste of time.

Reproducibility recommendations. (1) EEG preprocessing; in most published papers, the preprocessing steps are well described, however, many studies have a "visual inspection" part to remove some of the contaminated data. This subjective step can make reproducing the results difficult and we encourage researchers to also share the preprocessed EEG files of each participants in order to be able to compare where the results start diverging when failing to reproduce the results. This obviously increases the size of data being stored and shared but adds a lot of value. (2) Excluded subjects; in their documentation file (e.g. README.md), researchers should make clear which subjects have been excluded from the analysis. Usually, in the published paper there is a mention like "3 subjects were removed because of ...", while this informs the reader about the number of subjects that were included, if the raw data of these subjects are

included in the data folder, it puts the burden on someone trying to reproduce the study to find the excluded ones. We've encountered four ways to dealing with this issue while reproducing the studies for this review. (a) Not sharing the raw data of excluded subjects. In most cases we would not recommend this as there might be useful parts in the data. It really depends on the reason for excluding the participants from the study. (b) Making a specific *Excluded* folder and inserting the excluded subjects' data in it. (c) Identifying the excluded participant in the documentation file. Sometimes that information is available somewhere in the code as a *if* statement like *if fname == "participant_X"*, however this information should be brought forward and featured in the main documentation file. (d) Identifying the excluded participants in the published paper. This works well when one or two participants are excluded with a mention like "participant #12 was excluded because of ..." but doesn't scale well for multiple participants. (3) Document triggers/events; the triggers/events allowing to epoch the data properly should be clearly identified. Such information is usually a mix of hardware triggers from the EEG file and information from a behavioural file (e.g. csv file). For example, in most studies the side and the set size of each trial is available in the EEG file (saved as a signal with the EEG hardware) but the performance of the trial is available in a behavioural file. Such structure should be clearly described the main documentation file. (4) Performance scoring code; there are many ways to assess performance (e.g. K Score) and it would be beneficial to share the corresponding code to avoid confusion. In most studies, the formula is described in the paper but the corresponding code is not available. (5) Online repository platform; we recommend using OSF to share the data. *OSF is a free, open platform to support your research and enable collaboration*, as stated on their website. There is a limit of 50GB for the free tier which might not be enough for most EEG studies, however the cost of additional storage capacity is fairly low and has a one-time payment which works well with grants funding. All of the reviewed studies, but one, used OSF to share both the data and the code. Villena-Gonzalez used Mendeley Data instead. Another option is

OpenNeuro, *a free and open platform for validating and sharing BIDS-compliant MRI, PET, MEG, EEG, and iEEG data*, as one can read on their website. (6) README file; we recommend having a README.md or .pdf file in the root folder explaining the crucial information about the data with reproducibility in mind. We provide as good example from Adam 2018 below. We took a print screen of the beginning of the file. Their README file provides a clear explanation of the files and the data contained in these files saving precious time of investigating to figure it out. Beware of the copy/paste across experiments! Their first sentence shows a good example of a copy/paste that wasn't edited. In this case, the information isn't misleading and the mistake is pretty obvious but in other contexts it can be very counterproductive. (7) Raw data; researchers should include the raw data from the EEG recording device and not an exported version from EEGLab or else. Even if the data hasn't been modified and has been exported as is, it creates doubt and questions for no added value. The preprocessed data will be exported from tools like EEGLab or .mat files, however the raw data files should be the one coming from the recording device directly.

Finally, our objective with this review, aside from better understanding the cognitive mechanisms at play during a VWM task and the key role of CDA, was to evaluate the CDA reliability as a potential candidate for brain-computer interfaces (BCIs). Many other ERPs and neural correlates are being used in brain-computer interface paradigms, however, to our knowledge, the CDA has never been used in a BCI context. We have, however, seen a recent interest in using machine learning for CDA classification with regards to the number of items held in WM (e.g. [**2**]. We believe that within certain contexts, CDA could be used to enhance passive BCIs ([**12**]).

5. Conclusion

With this study we were able to look at CDA across different VWM tasks, from different groups of subjects, recorded by different groups of researchers using different EEG equipment. By using the same simple independent pipeline for all studies, we have shown that CDA is a robust neural correlate of visual working memory. As for its exact role and meaning, more research is required.

Moreover, having access to all these datasets allowed us to look beyond the usual numerical CDA mean amplitude over a window of interest but to also observe two phenomenon that are understudied and underdocumented. First, the decay happening shortly after the CDA peaks while the participant still has to maintain the information. Second, the CDA-like signal during the response period. For the latter, it can be difficult to control efficiently in order to dissociate eye-movements and other artifacts from cognitive processes in the EEG signal during the response period, however, we believe it deserves more investigation as it could help shed more light on the CDA and working memory.

Finally, all the code and figures generated for this review is available online on our repository. Many more figures that we did not include in the manuscript to keep it as short and concise as possible for better readability, are available on the repository.

Code and analysis: *https://github.com/royyannick/cda-reprod*

6. Funding

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC-RDC) (reference number: RDPJ 514052-17) and an NSERC Discovery fund.

7. Conflict of Interest

The authors declare that there is no conflict of interest.

8. Supplementary Material

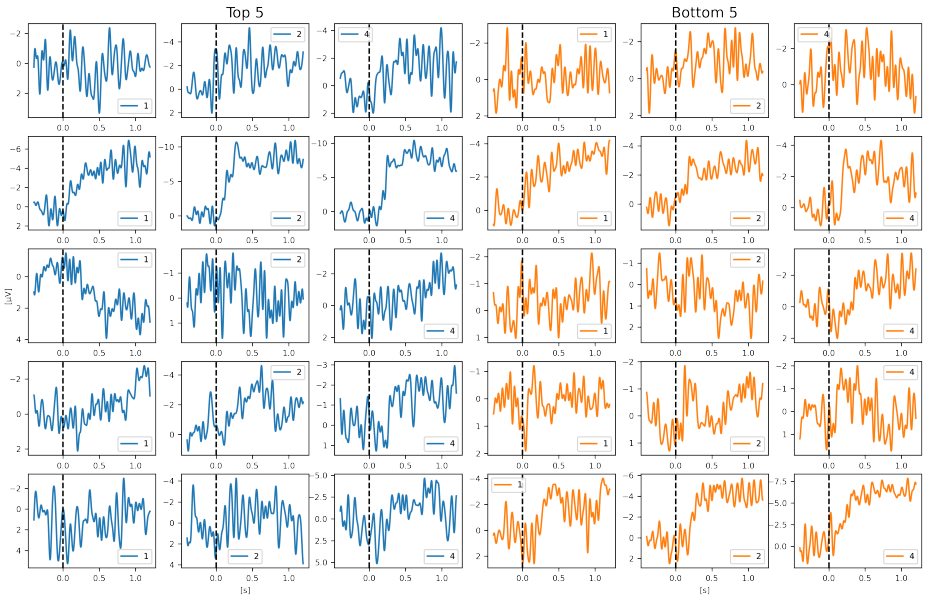


Fig. 18. Top 5 and Bottom 5 from Villena-Gonzalez, 2019. The graphs show the CDA averaged across trials for 3 different conditions for a total of 10 participants. The *Top 5*, shown in blue, represents the 5 participants with the highest performance while the *Bottom 5*, in orange, represents the 5 participants with the lowest performance.

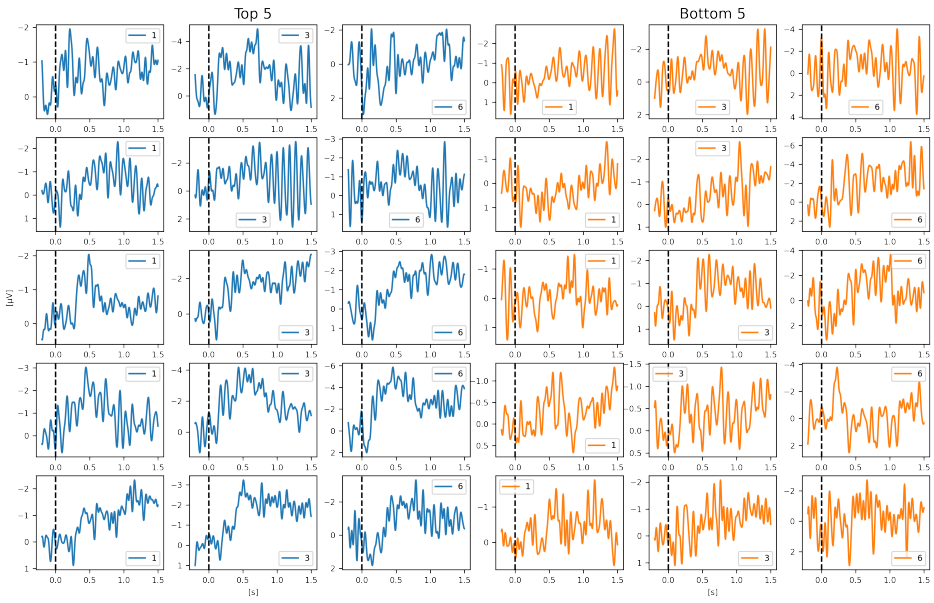


Fig. 19. Top 5 and Bottom 5 from Adam, 2018. The graphs show the CDA averaged across trials for 3 different conditions for a total of 10 participants. The *Top 5*, shown in blue, represents the 5 participants with the highest performance while the *Bottom 5*, in orange, represents the 5 participants with the lowest performance.

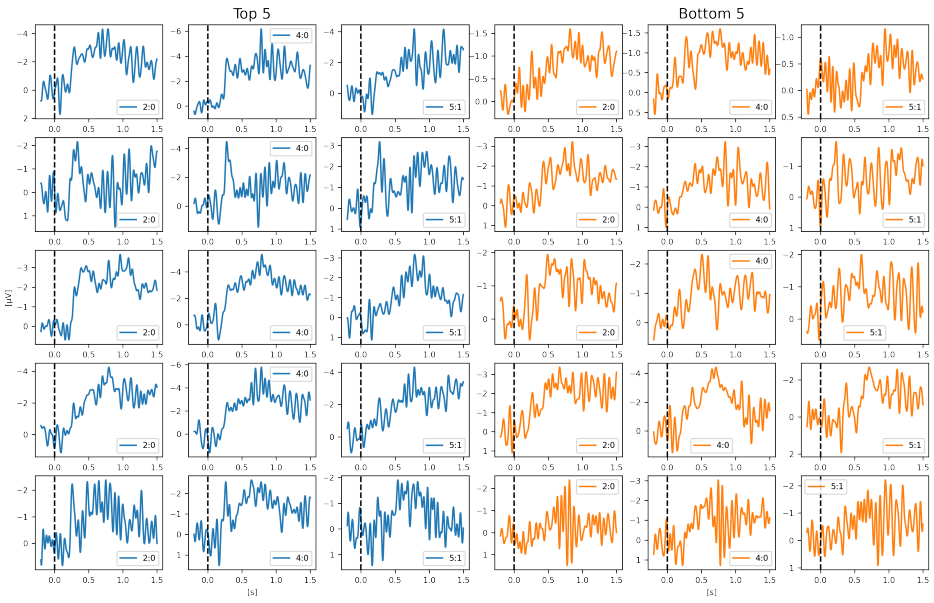


Fig. 20. Top 5 and Bottom 5 from Feldmann-Wusterfel 2020. The graphs show the CDA averaged across trials for 3 different conditions for a total of 10 participants. The *Top 5*, shown in blue, represents the 5 participants with the highest performance while the *Bottom 5*, in orange, represents the 5 participants with the lowest performance.

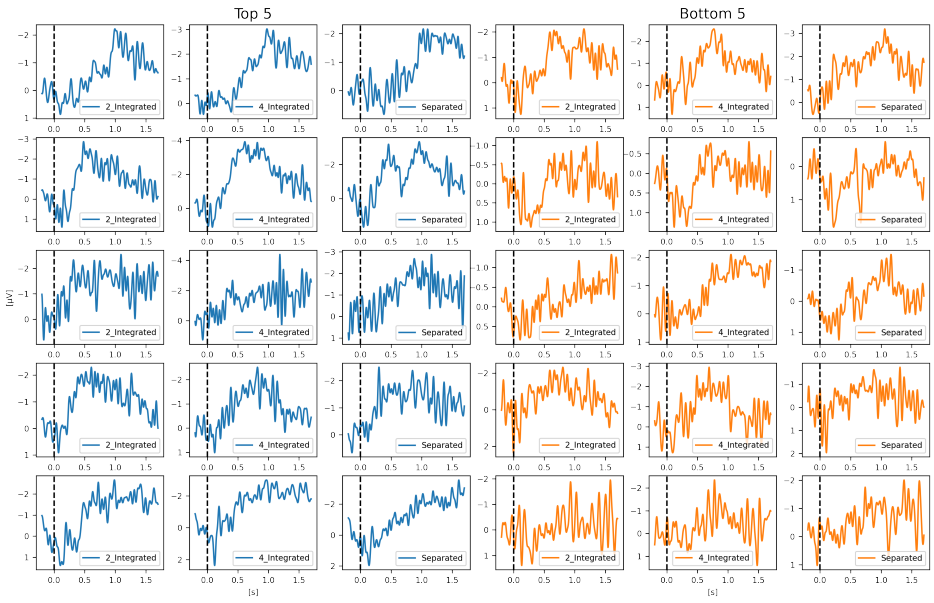


Fig. 21. Top 5 and Bottom 5 from Balaban 2019 Exp. 2. The graphs show the CDA averaged across trials for 3 different conditions for a total of 10 participants. The *Top 5*, shown in blue, represents the 5 participants with the highest performance while the *Bottom 5*, in orange, represents the 5 participants with the lowest performance.

```
--------------------------------------------------
Corr K-score vs CDA Amp (Both G/B): (0.5015810171230934, 0.02052519919753655)
--------------------------------------------------

Text(0, 0.5, 'CDA mean amp')
```

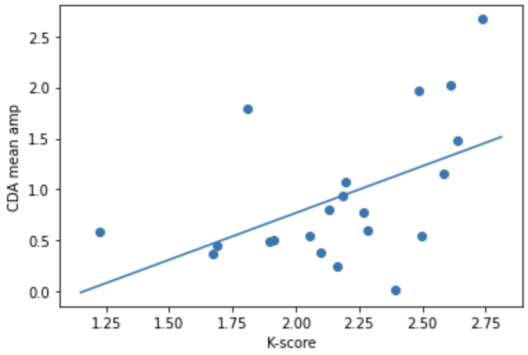


Fig. 22. CDA Amplitude vs Performance - Feldmann-Wüstefeld, 2020

```
--------------------------------------------------
Corr Acc vs CDA Amp: (-0.34455401248972156, 0.0018743257307676277)
With 79 subjects.
--------------------------------------------------

Text(0.5, 0, 'Mean Accuracy (SS6)')
```

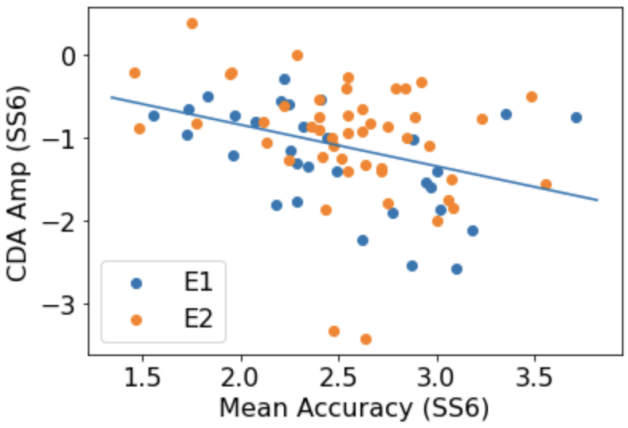


Fig. 23. CDA Amplitude vs Performance - Adam, 2018

Getting started

- The folders required to replicate the data in the paper are “rawBehavior”, “rawEEG”, and “rawEyeData” and “scripts”. Simply copy them on a local hard drive.
- Then, to run same analyses as in paper, execute “RUNALL.m” in folder “scripts”. This will use raw data and go through all processing steps calling a variety of functions. The end result are figures, and exported data files etc.
- To change settings, edit “preprocSettings.m” (pre-processing) “artRejectEye.m” (artifact rejection), or the functions starting with “plot” (e.g., to analyse different electrodes) or “Classify” (to change the multivariate classifier settings).

Overview of folders

- AV: averaged ERP data, sorted by conditions
- class: Multivariate classifier data
- export: exported data
- preproc: pre-processed and artifact-rejected data
- rawBehavior: behavioral result files
- rawEEG: EEG result files (.eeg, .vhmdr, .vmrk)
- rawEyeData: eye tracking result files
- runFiles: Matlab/PsychToolbox experimental run files
- scripts: Matlab code used to analyse data
- stats: SPSS files

Fig. 24. Getting Started from Feldmann-Wüstefeld, 2020

README for Data - Experiment 2

Behavior files

These .mat files contain the behavioral data recorded for Experiment 1.

List of files

`1_ColorK.mat` : Change detection K

`1_discreteWR_bilat_ss6.mat` : Whole-report task (during EEG data collection)

List of useful variables

Change detection

```
     stim.setSize: [36×4 double] % 36 trials x 4 blocks Set Size
      stim.change: [36×4 double] # Change (1 = Change, 0 = No Change)
    stim.response: [36×4 double] # Response (90 = Same, 191 = Different)
    stim.accuracy: [36×4 double] # Accuracy (1 = Correct, 0 = Incorrect)
          stim.rt: [36×4 double] # Response Time
            prefs: Various preferences and settings
                p: Various preferences and settings
```

Whole-report

- `data.setSize: [30×18 double]` : 30 trials x 30 blocks Set Size
- `data.retentionInterval: [30×18 double]` : Length of retention interval
- `data.screenSide: [30×18 double]` : Side of the Scren (1 = Left, 2 = Right)
- `data.acc: [30×6×18 double]` : 30 trials x 6 responses x 30 blocks (Response Accuracy, 1 = correct, 0 = incorrect)
- `data.trialAcc: [30×30 double]` : Total number correct on each trial (0 to 6)
- `data.rt: [30×6×18 double]` : Response time for each response (cumulative time within each trial

Fig. 25. README from Adam, 2018

References

[1] Kirsten CS Adam, Matthew K Robison, and Edward K Vogel. Contralateral delay activity tracks fluctuations in working memory performance. *Journal of Cognitive Neuroscience*, 30(9):1229–1240, 2018.

[2] Kirsten CS Adam, Edward K Vogel, and Edward Awh. Multivariate analysis reveals a generalizable human electrophysiological signature of working memory load. *Psychophysiology*, 57(12):e13691, 2020.

[3] Edward Awh, Edward K Vogel, and S-H Oh. Interactions between attention and working memory. *Neuroscience*, 139(1):201–208, 2006.

[4] Halely Balaban, Trafton Drew, and Roy Luria. Neural evidence for an object-based pointer system underlying working memory. *cortex*, 119:362–372, 2019.

[5] Paul M Bays and Masud Husain. Dynamic shifts of limited working memory resources in human vision. *Science*, 321(5890):851–854, 2008.

[6] Nick Berggren and Martin Eimer. Does contralateral delay activity reflect working memory storage or the current focus of spatial attention within visual working memory? *Journal of Cognitive Neuroscience*, 28(12):2003–2020, 2016.

[7] Karl Broman, Mine Cetinkaya-Rundel, Amy Nussbaum, Christopher Paciorek, Roger Peng, Daniel Turek, and Hadley Wickham. Recommendations to funding agencies for supporting reproducible research. In *American Statistical Association*, volume 2, 2017.

[8] Nelson Cowan. *Attention and memory: An integrated framework*. Oxford University Press, 1998.

[9] Nelson Cowan. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and brain sciences*, 24(1):87–114, 2001.

[10] Jocelyn Faubert. Professional athletes have extraordinary skills for rapidly learning complex and neutral dynamic visual scenes. *Scientific reports*, 3(1):1–3, 2013.

[11] Reza Fazel-Rezai, Brendan Z Allison, Christoph Guger, Eric W Sellers, Sonja C Kleih, and Andrea Kübler. P300 brain computer interface: current challenges and emerging trends. *Frontiers in neuroengineering*, page 14, 2012.

[12] Tobias Feldmann-Wüstefeld. Neural measures of working memory in a bilateral change detection task. *Psychophysiology*, 58(1):e13683, 2021.

[13] Tobias Feldmann-Wüstefeld, Edward K Vogel, and Edward Awh. Contralateral delay activity indexes working memory storage, not the current focus of spatial attention. *Journal of cognitive neuroscience*, 30(8):1185–1196, 2018.

[14] Keisuke Fukuda, Geoffrey F Woodman, and Edward K Vogel. Individual differences in visual working memory capacity: Contributions of attentional control to storage. *Mechanisms of sensory working memory: Attention and perfomance XXV*, 105, 2015.

[15] Krzysztof J Gorgolewski, Tibor Auer, Vince D Calhoun, R Cameron Craddock, Samir Das, Eugene P Duff, Guillaume Flandin, Satrajit S Ghosh, Tristan Glatard, Yaroslav O Halchenko, et al. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific data*, 3(1):1–9, 2016.

[16] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, et al. Meg and eeg data analysis with mne-python. *Frontiers in neuroscience*, 7:267, 2013.

[17] Nicole Hakim, Kirsten CS Adam, Eren Gunseli, Edward Awh, and Edward K Vogel. Dissecting the neural focus of attention reveals distinct processes for spatial attention and object-based storage in visual working memory. *Psychological Science*, 30(4):526–540, 2019.

[18] Nicole Hakim, Tobias Feldmann-Wüstefeld, Edward Awh, and Edward K Vogel. Perturbing neural representations of working memory with task-irrelevant interruption. *Journal of cognitive neuroscience*, 32(3):558–569, 2020.

[19] Mainak Jas, Denis A Engemann, Yousra Bekhti, Federico Raimondo, and Alexandre Gramfort. Autoreject: Automated artifact rejection for meg and eeg data. *NeuroImage*, 159:417–429, 2017.

[20] John E Lisman and Marco AP Idiart. Storage of 7±2 short-term memories in oscillatory subcycles. *Science*, 267(5203):1512–1515, 1995.

[21] Steven J Luck. *An introduction to the event-related potential technique*. MIT press, 2014.

[22] Steven J Luck and Edward K Vogel. Visual working memory capacity: from psychophysics and neurobiology to individual differences. *Trends in cognitive sciences*, 17(8):391–400, 2013.

[23] Roy Luria, Halely Balaban, Edward Awh, and Edward K Vogel. The contralateral delay activity as a neural measure of visual working memory. *Neuroscience & Biobehavioral Reviews*, 62:100–108, 2016.

[24] Andrew W McCollough, Maro G Machizawa, and Edward K Vogel. Electrophysiological measures of maintaining representations in visual working memory. *Cortex*, 43(1):77–94, 2007.

[25] William XQ Ngiam, Kirsten CS Adam, Colin Quirk, Edward K Vogel, and Edward Awh. Estimating the statistical power to detect set-size effects in contralateral delay activity. *Psychophysiology*, 58(5):e13791, 2021.

[26] Klaus Oberauer. Access to information in working memory: exploring the focus of attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(3):411, 2002.

[27] Christian NL Olivers, Judith Peters, Roos Houtkamp, and Pieter R Roelfsema. Different states in visual working memory: When it guides attention and when it does not. *Trends in cognitive sciences*, 15(7):327–334, 2011.

[28] Sam Parsons, Flávio Azevedo, Mahmoud M Elsherif, Samuel Guay, Owen N Shahim, Gisela H Govaart, Emma Norris, Aoife O'mahony, Adam J Parker, Ana Todorovic, et al. A community-sourced glossary of open scholarship terms. *Nature human behaviour*, 6(3):312–318, 2022.

[29] Harold Pashler. Familiarity and visual change detection. *Perception & psychophysics*, 44(4):369–378, 1988.

[30] John Polich and Albert Kok. Cognitive and biological determinants of p300: an integrative review. *Biological psychology*, 41(2):103–146, 1995.

[31] Bruno Rossion, Isabel Gauthier, M J Tarr, P Despland, Raymond Bruyer, S Linotte, and Marc Crommelinck. The n170 occipito-temporal component is delayed and enhanced to inverted faces but not to inverted objects: an electrophysiological account of face-specific processes in the human brain. *Neuroreport*, 11(1):69–72, 2000.

[32] Yannick Roy, Hubert Banville, Isabela Albuquerque, Alexandre Gramfort, Tiago H Falk, and Jocelyn Faubert. Deep learning-based electroencephalography analysis: a systematic review. *Journal of neural engineering*, 16(5):051001, 2019.

[33] Nash Unsworth, Keisuke Fukuda, Edward Awh, and Edward K Vogel. Working memory delay activity predicts individual differences in cognitive abilities. *Journal of Cognitive Neuroscience*, 27(5):853–865, 2015.

[34] Mario Villena-González, Iván Rubio-Venegas, and Vladimir López. Data from brain activity during visual working memory replicates the correlation between contralateral delay activity and memory capacity. *Data in brief*, 28:105042, 2020.

[35] Edward K Vogel and Maro G Machizawa. Neural activity predicts individual differences in visual working memory capacity. *Nature*, 428(6984):748–751, 2004.

[36] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9, 2016.

[37] Dezhong Yao, Yun Qin, Shiang Hu, Li Dong, Maria L Bringas Vega, and Pedro A Valdés Sosa. Which reference should we use for eeg and erp practice? *Brain topography*, 32(4):530–549, 2019.

[38] Thorsten O Zander and Christian Kothe. Towards passive brain–computer interfaces: applying brain–computer interface technology to human–machine systems in general. *Journal of neural engineering*, 8(2):025005, 2011.

Second Article.

Significant changes in neural oscillations during different phases of three-dimensional multiple object tracking task (3D-MOT).

by

Yannick Roy[1], and Jocelyn Faubert[1]

([1]) Université de Montréal

This article was submitted in PLOS One.

Résumé. Notre habileté à suivre plusieurs objets dans un environnement dynamique nous permet de performer des tâches quotidiennes comme la conduite automobile, pratiquer des sports d'équipe, ou marcher dans un centre commercial achalandé. Malgré plus de trois décennie de littérature sur des tâches de suivi d'objets multiples (MOT en anglais), les mécanismes neuronaux sous-jacent demeurent très peu compris. Ici, nous regardons l'activité cérébrale via l'électroencéphalographie (EEG) et ses changements durant les trois phases d'une tâche 3D-MOT, soit l'identification des cibles, le suivi et le rappel. Nous avons enregistré l'activité EEG de 24 participants pendant le 3D-MOT avec soit 1, 2 ou 3 cibles ainsi que certains essais latéralisés à droite ou à gauche. Nous avons observé ce qu'il semble être un transfert entre les processus d'attention soutenue et de mémoire de travail lorsque le participant passe de la phase de suivi au rappel.
Mots clés : Suivi d'objets multiple, EEG, Mémoire de travail, Attention

Abstract. Our ability to track multiple objects in a dynamic environment enables us to perform everyday tasks such as driving, playing team sports, and walking in a crowded mall. Despite more than three decades of literature on multiple object tracking (MOT) tasks, the underlying and intertwined neural mechanisms remain poorly understood. Here we looked at the electroencephalography (EEG) neural correlates and their changes across the three phases of a 3D-MOT task, namely identification, tracking and recall. We recorded the EEG activity of 24 participants while they were performing a 3D-MOT task with either 1, 2 or 3 targets where some trials were lateralized and some where not. We observed what seems to be a handoff between focused attention and working memory processes when going from tracking to recall. Our findings revealed a strong inhibition in delta and theta frequencies from the frontal region during tracking, followed by a strong (re)activation of these same frequencies during recall. Our results also showed contralateral delay activity (CDA) for the lateralized trials, in both the identification and recall phases but not during tracking.
Keywords: Multiple-Object Tracking, MOT, EEG, Working Memory, Attention

1. Introduction

Our ability to track multiple moving objects simultaneously in a dynamic environment enables us to perform everyday tasks such as driving, playing team sports, and walking in a crowded mall. In such tasks, it is required to manage internal representations of relevant information in order to predict future spatial positions of surrounding objects and optimize decision making accordingly. A professional athlete being able to make a perfect pass to a teammate in a high speed sport while avoiding players from the other team, is a great example of the brain's remarkable ability to track multiple objects both in space and time. In order to study this ability in a laboratory setting, researchers often use a variant of the multiple-object tracking (MOT) task developed by Pylyshyn & Storm in 1988 ([**6**]). In typical MOT tasks we find two categories of visual objects: targets (objects of interest) and distractors (objects to ignore) both sharing identical visual properties. In order to modulate task difficulty, parameters usually include the number of targets ([**35**]), speed ([**7, 19**]), and distance between objects ([**1, 41**]).

Multiple-object tracking is an active area of research in humans but also in computer vision as we are observing an increasing demand for technology for automated tracking of vehicles and people in various contexts ([**29, 9, 32**]). Humans' visual system has inspired current neural network architectures driving most of the artificial intelligence (AI) field as we know it today ([**37**]) and recent neural network architectures are trying to take advantage of higher brain mechanisms such as attention ([**24**]). Therefore, better understanding the underlying mechanisms of the brain's ability to track multiple objects in time and space could also benefit AI-related fields.

Despite the increasing literature on multiple-object tracking and evidence of effective training showing transfer on real tasks in real environments ([**1**]), our understanding of the underlying neural mechanisms across the different phases and their transitions, remains fuzzy. Several studies have shown that our capacity to keep

individual items in working memory is limited to 3 or 4 ([**8, 46**]) and over the last few decades, several cognitive models have been brought forward trying to explain how individual items are being encoded and deciphering the intertwined roles of attention and working memory in such tasks. Older proposals suggested that early individuation of objects (up to 3-4) does not require attention mechanisms ([**42**]), however, recent research indicates otherwise suggesting that simultaneous indexing of items relies on attention mechanisms ([**5**]) before involving subsequent mechanisms such as visual WM to encode the individuated objects in greater details. Another proposal include multifocal spatial attention, where attention can be split and work in parallel for multiple targets ([**6**]). On the other hand, Oksama and Hyöna suggested that rather than having a fixed-capacity parallel mechanism, the tracking performance would be better explained by a serial model where the maintenance of moving objects requires continuous serial refreshing of identity-location bindings ([**31**]). Moreover, a previous behavioural study also showed that one cognitive strategy is to process the targets as one illusory object by mentally creating connections between the targets to make, for example, a geometrical shape ([**48**]).

The confounding aspects of attention versus working memory and location-based versus object-based tracking remains an active area of research and recent EEG studies were performed trying to provide more insights by disentangling them. For example, Drew and colleagues tried to delineate the neural signatures of tracking spatial position and working memory during attentive tracking. They found that there was a unique contralateral negativity related to the process of monitoring target position during tracking which was absent when objects briefly stopped moving. These results suggest that the process of tracking target locations elicits an electrophysiological response that is distinct and dissociable from neural responses of the number of targets being attended ([**10**]). Also, Merkel and colleagues looked at the spectral properties of the electrophysiological signal, mainly in the gamma range, during tracking to find a difference between location-based and object-based maintenance of visual information

([**28**]). Their results suggest that object-based tracking is supported by enhanced encoding during the initial presentation of the targets and location-based tracking is characterized by the sustained maintenance of the individual targets during the entire tracking period, in that same processing neural network. In a previous study Merkel and colleagues also showed that neural networks involved in both tracking processes (object-based and location-based) are at least partly overlapping ([**27**]).

The experiment described in this manuscript is part of a larger study where we seek to develop a passive brain-computer interface (BCI) composed of an EEG closed-loop system for a cognitively demanding task. For this experiment and this manuscript, we hypothesize that by looking at the EEG activity during the 3D-MOT task, we can identify different neural mechanisms at play during the three different phases of the task, namely (1) the identification phase, (2) the tracking phase and (3) the recall phase. If our hypothesis holds true and we observe a significant difference in EEG activity across the three phases of the 3D-MOT, a subsequent manuscript will explore how to leverage such differences in a BCI context. As part of the larger study, all subjects participated to three different tasks: MOT task, N-Back task, and flight simulator task. In this publication we share only our findings related to cognitive processes during the MOT task, and we don't cover the other tasks nor the BCI system which will all be discussed in depth in another publication.

Most MOT studies are conducted using a 2D-MOT task on a computer screen. However, we live in a 3D environment and different cognitive processes might be at play in a 3D environment. NeuroTracker^{TM} is a commercially available 3D-MOT task currently used by a multitude of users in many countries around the world as a perceptual-cognitive training and assessment tool. It has been used and studied in various fields such as sport ([**1, 13, 26, 38**]), ESports ([**3**]), education ([**43**]), aviation ([**18**]) and military ([**45**]). 3D-MOT training has been demonstrated to enhance attention, working memory and visual information processing speed ([**34**]). Given

the wide adoption of the NeuroTracker™ and existing literature showing effective transfer, we opted for a modified version of the NeuroTracker™ for our study.

According to existing literature and the nature of the task, different neural correlates linked to working memory, attention, workload and visual processes should be at play during 3D-MOT. In the time domain, event-related potentials (ERPs) should be observed at the beginning of each phase of the task given the sudden change in visual information displayed. For lateralized trials, where the targets are displayed and moving only in one hemifield, we expect to see lateralized activity. In the frequency domain, we are expecting different changes in frequency bands linked to working memory (e.g. Theta), attention (e.g. Alpha), workload (e.g. Beta), and visual processes (e.g. Gamma) during the different phases of the task.

2. Materials and Methods

2.1. Participants

Twenty-four participants (thirteen females) aged between 21 and 41 years of age (M=29.3, SD=4.9) took part in this study. The participants were healthy university students of various ethnicity from different universities in Montreal. All participant self-reported normal or corrected-to-normal visual acuity and passed the Randot Stereotest for stereo vision. The study was reviewed and approved by the Université de Montréal ethics committee for health research (Comité d'éthique de la recherche en santé; CERES #2018-334). Recommended ethics procedures and guidelines were followed, and informed consent was obtained from all participants. The three hour long session included the 3D-MOT task discussed here but also included the recording of a N-Back task and both tasks were part of a larger research project where the subjects also participated in a second session, on a different day, for a flight simulator task. All subjects received a monetary compensation for their participation to the two sessions covering also their transportation to the different facilities. The other

components of the larger research project (i.e. N-Back and flight simulator) are not discussed in this manuscript and will be published separately.

2.2. Task

Figure 56 shows the five different phases of the 3D-MOT task developed with the Unity engine. (A) *presentation phase* where 8 yellow spheres are shown in a 3D volume space for 2 seconds, (B) *indexing phase* where one, two or three spheres (targets) change colour (to red) and are highlighted (halo) for 2 seconds, (C) *tracking (or movement) phase* where the targets indexed in phase 2 return to their original colour (yellow) and 1 second later start moving for 8 seconds crisscrossing and bouncing off of each other and the virtual 3D volume cube walls, (D) *recall phase* where the spheres stop moving and the observer is prompted to identify the spheres originally indexed in phase 2. Each sphere is labelled with a number between 1 to 8. After identifying the targets, the observer is asked to provide a confidence level for each answer (either 0%, 25%, 75% or 100% confident). And finally, (E) *feedback phase* where the correct targets are clearly identified on the screen. The whole trial takes around 15s (2s + 2s + 9s + [1-4]s) depending on how long the participant takes to provide the answers. A video of the task is available online on the repository provided.

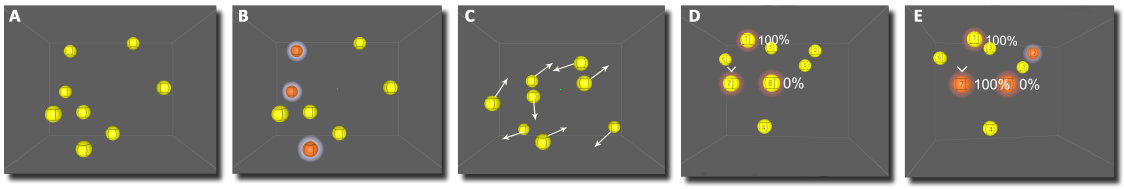


Fig. 26. 3D-MOT Task Sequence. (A) All spheres appear on screen. (B) Targets are highlighted in red for 2 seconds. (3) All the spheres are moving for 8 seconds. (D) Participant must identify the targets and provide a confidence level. (E) Feedback is provided to the participant showing the correct answers.

The participant was seated 1.75m and centered from the 65" 3D TV screen (Panasonic TC-P65VT60) wearing the EEG cap as well as active shutter 3D glasses (Panasonic TY-ER3D5MA). The 3D-MOT virtual cube was 35 degrees of visual angle (dva) in size and the spheres were 2.5 dva. The 3D glasses were carefully slid under the EEG cap in a way to minimally disrupt the EEG signal. Additional gel was added to some temporal electrodes for some participants to compensate for the gap created by the 3D glass legs. The participants had a keyboard on their lap to provide their confidence level after giving the answers orally to the instructor, seated 2m behind the participant with a keyboard.

Each participant started with a training block of 20 trials, without the EEG equipment, to familiarize themselves with the task and for us to obtain an individualised speed threshold. They were instructed to keep their eyes on the green fixation dot in the middle of the screen and to let their covert attention track the targets. An adaptive staircase algorithm modulated the speed of each new trial based on previous performances to find the individual speed where the participant gets a performance of 50% (i.e. they get half the trials right and half wrong) ([**1**]). To get a *good* trial the participant needs to identify correctly all the targets. For the training block, only the condition with 3 targets was used. After the 20 trial training block, the resulting speed was used for the remaining of the experiment.

The experiment consisted of 4 blocks of 21 trials with 2 conditions: side and set size. The speed was kept constant, based on the speed obtained at the end of the training block. In total, 30 trials were presented in the left hemifield, 10 for each set size (1,2,3), 30 trials were presented in the right hemifield, 10 for each set size (1,2,3), and 24 trials were not lateralized and the targets could freely cross from left to right and vice-versa, 8 for each set size (1,2,3). It is important to note that for lateralized trials only the targets stayed on one specific hemifield while the other distractors where moving freely with no restriction. Once the targets stopped moving, a number between 1 and 8 appeared on each of the spheres and the participants had to provide

their answer by saying the number of the target(s) out loud for the instructor to enter the answers. After the answers were entered, a visual cue on top of each of the selected spheres appeared for the participant to provide a confidence level for each target. The participant used the arrow on the keyboard to provide the confidence level. Up arrow means 100% confident (the instruction provided to the participant: "I was able to track it."). Down arrow means 0% ("I lost it, it's a random guess."). Right arrow means 75% confident ("I'm somewhat confident. I think I tracked it but I might have switch target during an overlap/occlusion."). Left arrow means 25% confident ("It's mostly a guess. I got confused."). The rational behind this discrete scoring of confidence was to clearly dissociate between a random guess and a fully confident answer (0% vs 100%). For the in-between, we decided to avoid having only one additional option as we felt like most participants might default to that option whether they are quite confident but do have a small doubt or if they are mostly guessing it, making this option hard to use for further analysis. Adding more granularity than four options for the confidence level would have only added a cognitive load with no additional benefit. Therefore the compromise with four discrete options was chosen. For the analysis presented in this manuscript, we regrouped 75% and 100% as being a confident answer and therefore a *good trial* is a trial where the participant identified all the correct targets and indicated a confidence level of 75% or 100%, otherwise it is labeled as a bad trial.

2.3. EEG Acquisition

The electroencephalogram (EEG), electrocardiogram (ECG), and electrooculogram (EOG) were recorded using the Biosemi ActiveTwo system (Biosemi, Amsterdam, Netherlands) with 71 Ag–Ag/Cl electrodes positioned at 64 standard International 10/20 System sites (EEG), left and right mastoids for offline EEG re-reference, 1cm lateral to the external canthi for horizontal EOG (HEOG), left and right ribs plus right collarbone for ECG. The HEOG was used for eye movements to confirm that

the participant was tracking the targets with covert and not overt attention (i.e. not moving their eyes). Electrophysiological signals were digitized at 2048Hz.

2.4. EEG Analysis

EEG offline analysis was performed in MNE-Python ([**15**]) an open-source Python package for neurophysiological data. Both our EEG data and code are available online. Here we will breakdown the analysis in two parts, first in the time domain and second, in the frequency domain to look at the oscillation variations over time and power spectrum across conditions. Four conditions were examined across different brain regions: side (left, right), set size (1, 2, 3), performance (good, bad), 3D-MOT phase (indexing, tracking, recall). Note that because of low amount of trials with bad performances, we looked at the data internally but do not draw any conclusion in this paper about performance. It is important to note that after looking at our results we realized that we unfortunately had a 150ms offset in our Unity code between the trigger (i.e. EEG event) and the color change (from yellow to red and red back to yellow) appearing on screen. The offset is fixed and due to a Unity fixed update delay we forgot to remove for the color change. This offset does not apply for when the movement stops and ends. What it means is that the events at t=0s, and at t=2s, are in fact visible on screen only at t=0.15s and t=2.15s.

Preprocessing. First, the EEG channels were re-referenced to the left and right mastoids. Second, independent component analysis (ICA) was used to remove eye blinks and eye movement artifacts. Third, the EEG data was epoched in [-1, 15]s windows where t=0s represents the stimuli/trigger of the spheres being highlighted in red. Fourth, AutoReject ([**3**]) was used to automatically remove bad trials and correct bad channels. After looking at the clean EEG data and separately analyzing EOG channels for eye movements during tracking to see if people were really tracking with

covert attention instead of overt attention, 4 subjects were removed and 20 subjects remained for the analysis.

Time domain. First, the non-lateralized occipital grand average signal was obtained over O1, O2 and Oz to confirm visual ERPs. Since the 3D-MOT is a visual task with sudden visual changes between the different phases, a visual/occipital ERP should be observed accordingly. Second, for the lateralized activity, we looked at different clusters of electrodes, namely *frontal*, *central*, *parietal*, *temporal* and *occipital*. For readability the electrodes used in each cluster are listed in Table 4 in supplementary material. For the frontal cluster, all channels with the letter 'F' were included. For the central cluster, all the electrodes with the letter 'C' were included. For the parietal cluster, all the channels with the letter 'P' were included. For the temporal cluster, all the channels with the letter 'T' were included. For the occipital cluster, all the channels with the letter 'O' were included. For lateralized activity of both left and right trials, we averaged the left channels (i.e. channels with odd numbers) and subtracted the average of the right channels resulting in the difference between the two hemispheres. The rationale behind this analysis is to confirm that indeed we see more activity in one hemisphere than the other for lateralized trials. The gross activity observed here should encompass specific neural signatures such as ERPs and contralateral delay activity (CDA). The CDA is a sustained negativity over the hemisphere contralateral to the positions of the items to be remembered. The CDA has been shown to be linked with the number of items held in WM ([**44, 25, 39**]) and previous studies have shown a correlation between CDA amplitude and the number of targets during a 2D-MOT task ([**11, 10**]). We therefore looked at the effect of the number of targets on the CDA by averaging of the channels contralateral to the hemifield where the targets were presented minus the average of the ipsilateral channels. The midline channels (ending with the letter z) were not included in the CDA nor the lateralized analysis. The CDA, as defined in the literature, should be strongest in the parietal region. However, we also calculated the same activity (i.e. contra minus ipsi) for the different clusters

mentioned above. For both lateralized activity and CDA we used only the trials with a good performance (i.e. the participant identified all the targets correctly with high level of confidence) for a total of N=982 trials, an average of 49 trials per subject. A three-way repeated measures analysis of variance (ANOVA) was performed with the number of targets (1, 2, and 3), phases (id, tacking, and recall), and clusters (frontal, central, parietal, temporal, and occipital) as independent variables and the mean CDA amplitude over a 1s time window as the dependent variable. For the identification (id) phase, the time window was from 0.5s to 1.5s, for the tracking we selected the 5s to 6s and for the recall, we used 11.5s to 12.5s.

Frequency domain. To obtain a more detailed representation of neural oscillation changes over time, Event-Related Spectrum Perturbation (ERSP) graphs were used. The time-frequency decomposition was computed using Morlet wavelets for frequencies between 1 and 50Hz with varying cycles of half the frequency. The ERSP maps were then obtained by getting the log ratio of the power relative to the baseline power. For the time-frequency analysis, the baseline was selected as -1s to 0s prior to the targets being colored in red (t=0.15s). Instead of using raw power, the log ratio has the advantage of normalizing the power across participants. To investigate the potential role of different brain regions we used the midline channels from frontal to occipital to give a representation in space of the time-frequency activity. For this analysis we included only the trials with a good performance. With the same time-frequency decomposition, we analyzed a 1s window of average power as a dependent variable using a repeated measure three-way ANOVA, with Phase (Identification [0.3, 1.3]s, Tracking [5, 6]s, and Recall [11.5, 12.5]s), Set Size (1, 2, and 3), and Frequency Band (delta [1-3]Hz, theta [4-7]Hz, alpha [8-12]Hz, beta [13-30]Hz, gamma [31-50]Hz) as independent variables. The time windows across the phases were selected to capture mostly top-down cognitive processes.

3. Results

3.1. Time Domain

Non-lateralized occipital activity in O1, O2 and Oz (see Figure 27) confirmed the visual ERPs when drastic visual changes occured, such as the target spheres changing color from yellow to red at t=0.15s, then reverting back to yellow at t=2.15s, start moving at t=3s and stop moving at t=11s. Noteworthy, the occipital ERP of the initial color change from yellow to red is significantly stronger than when the targets revert back to yellow (t-test on 200ms window; $p < 0.001$) and the strongest ERP occurs when the targets start moving. On the other hand, the ERP elicited when spheres stop moving isn't as sharp as the other ones, most likely due to the time it takes for individuals to realize that the spheres have indeed stopped moving. Moreover, we can also see a sustained cognitive activity following the ERPs at t=0.15s and t=11s, respectively during identification of targets and recall.

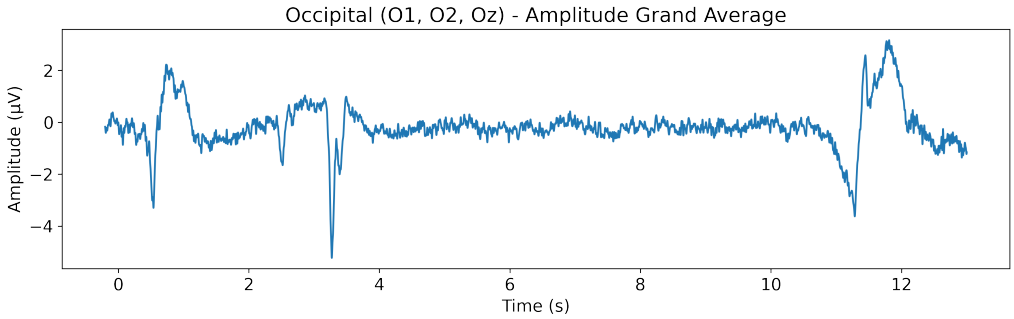


Fig. 27. Non-lateralized occipital ERPs (channels: O1, O2, Oz). Grand average over all participants for all conditions and performances.

As expected, we obtained a clear lateralization of activity in left vs right trials during both identification and recall phases as shown on Figure 28. However, we did not observe lateralized activity during tracking despite the targets staying in only one hemifield for the full duration of the trial. Zooming in on the identification and recall

phases, we observe that the same brain regions (clusters) are activated the strongest, namely the frontal and temporal regions, during both identification and recall (see Figure 29). This could be indicative of similar cognitive functions during identification and recall phases. Note that automated scales were used to preserve a clear shape of the signal, however the amplitude, or strength of the signal, varies across clusters.

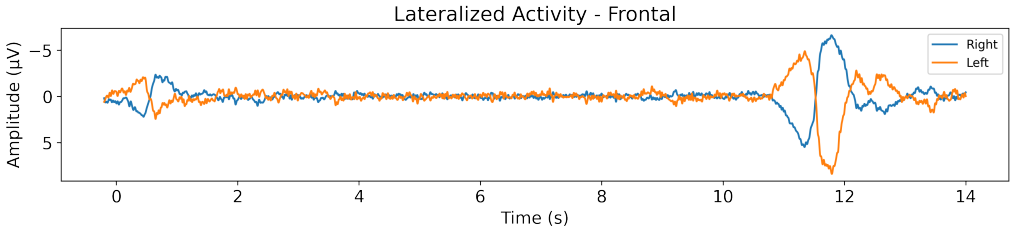


Fig. 28. Lateralized activity in the frontal region for whole sequence. Left vs right trials.

Figure 30 shows the lateralized amplitude for the frontal cluster for the whole duration of the trial and on Figure 31 we zoom in on the identification and recall phases across EEG channel clusters. The lateralized activity observed here has a similar shape as what we see in the literature for the CDA, peaking around 500-600ms post-stimuli and slowly decaying for another ∼500ms but is expected to be stronger in posterior parietal regions as opposed to frontal regions ([**46, 44**]). The initial three-way repeated measures ANOVA with the number of targets (1, 2, and 3), phases (id, tacking, and recall), and clusters (frontal, central, parietal, temporal, occipital) as independent variables and the mean CDA amplitude as the dependent variable, revealed a significant effect for the number of targets ($F_{(2,34)} = 3.59$, $p = .039$), a strong significant effect for the phase ($F_{(2,34)} = 20.17$, $p < .0001$), a strong significant effect for the cluster ($F_{(4,68)} = 32.85$, $p < .0001$), a near significant effect for the interaction between the targets and phases ($F_{(4,68)} = 2.46$, $p = .053$), a non significant effect for the interaction between targets and clusters ($F_{(8,136)} = 1.56$, $p = .14$), a strong significant effect for the interaction between phases and clusters ($F_{(8,136)} = 18.3$, $p <$

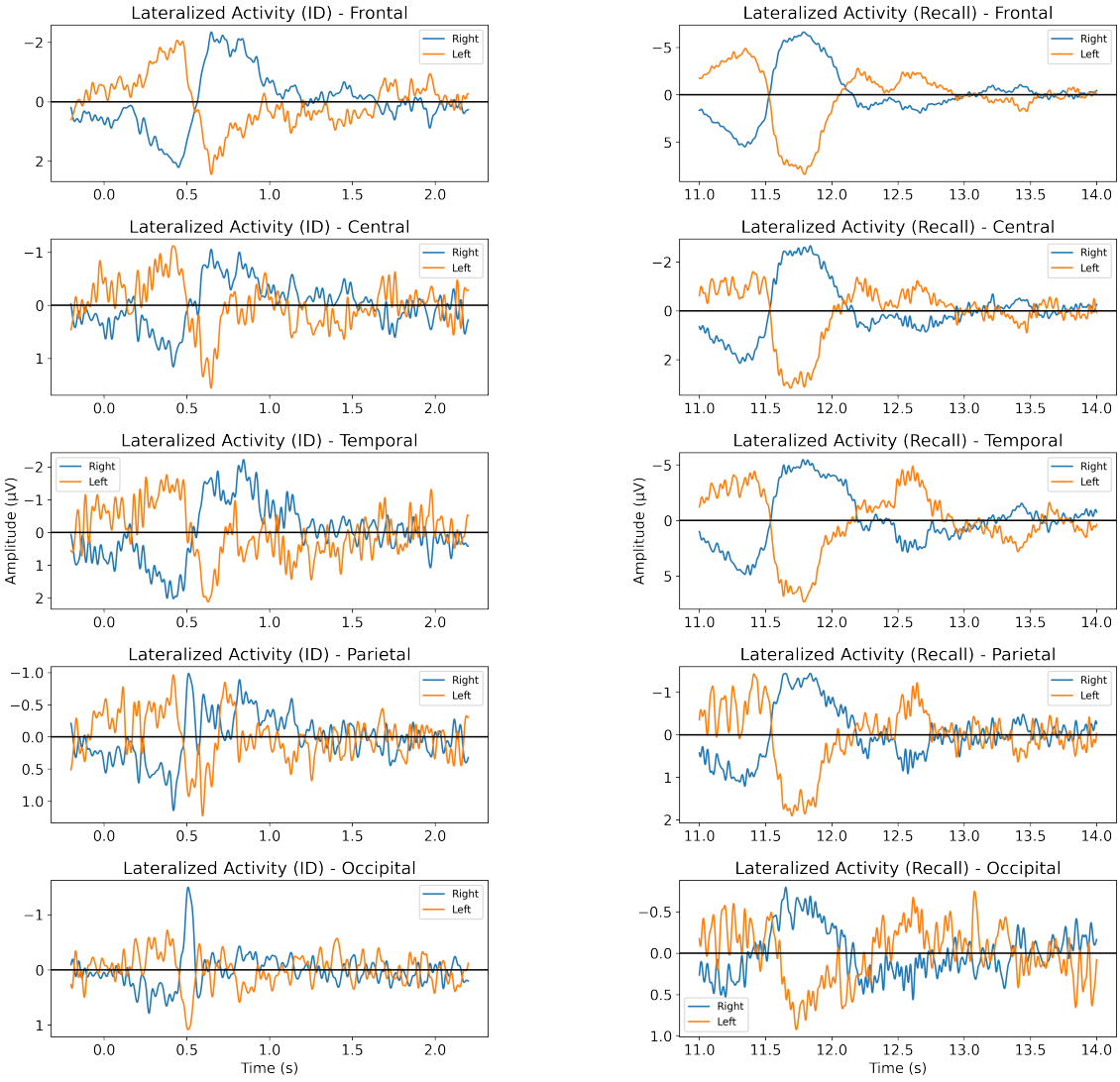


Fig. 29. Lateralized activity from the different brain regions for identification (first column) and recall (second column) phases. Left vs right trials.

.0001), and a trending but non significant effect for the interaction between the three ($F_{(16,272)} = 1.6$, $p = .067$). Post hoc one-way ANOVAs looking only at the number of the targets, independently for each phase and cluster, revealed no significant effect on the mean CDA amplitude during the identification phase nor during tracking, in any

of the clusters. During recall, however, the effect was significant in the parietal cluster ($F_{(2,34)} = 8.15$, $p < .005$), the temporal cluster ($F_{(2,34)} = 7.76$, $p < .005$), the central cluster ($F_{(2,34)} = 5.59$, $p < .01$), the occipital cluster ($F_{(2,34)} = 6.93$, $p < .005$) and trending but non significant in the frontal cluster ($F_{(2,34)} = 2.6$, $p = .088$). Note that the post hoc ANOVA values are provided as is and were not corrected for multiple comparisons. A total of 15 post hoc ANOVAs were performed (5 clusters x 3 phases). After obtaining and writing the results, We also performed the same analysis with a 0.5s window to see if it would change the results but it yielded similar results.

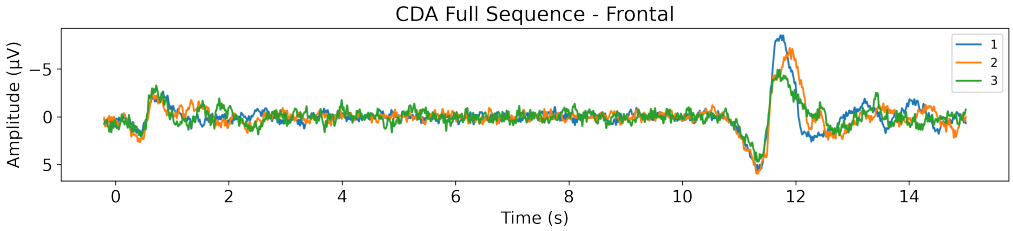


Fig. 30. CDA in the frontal region for the full sequence. Trials with 1, 2 and 3 targets.

3.2. Frequency Domain

The ERSP for the whole trial is shown on Figure 32 which includes 1 second prior to the targets being identified in red and up to 15s post identification, with frequencies ranging from 1 to 50Hz. It represents the grand average across all participants, all conditions, and all channels. Before breaking down the time-frequency analysis to look at specific elements, we can easily distinguish the different phases of the MOT task with drastic changes in neural oscillations across these phases. It is important to note the range of the color scale of the ERSP maps presented in it section as in EEG studies blue usually means a desynchronization (i.e. negative value) and red a synchronization (i.e. positive value). Here, we purposefully used a none symmetrical

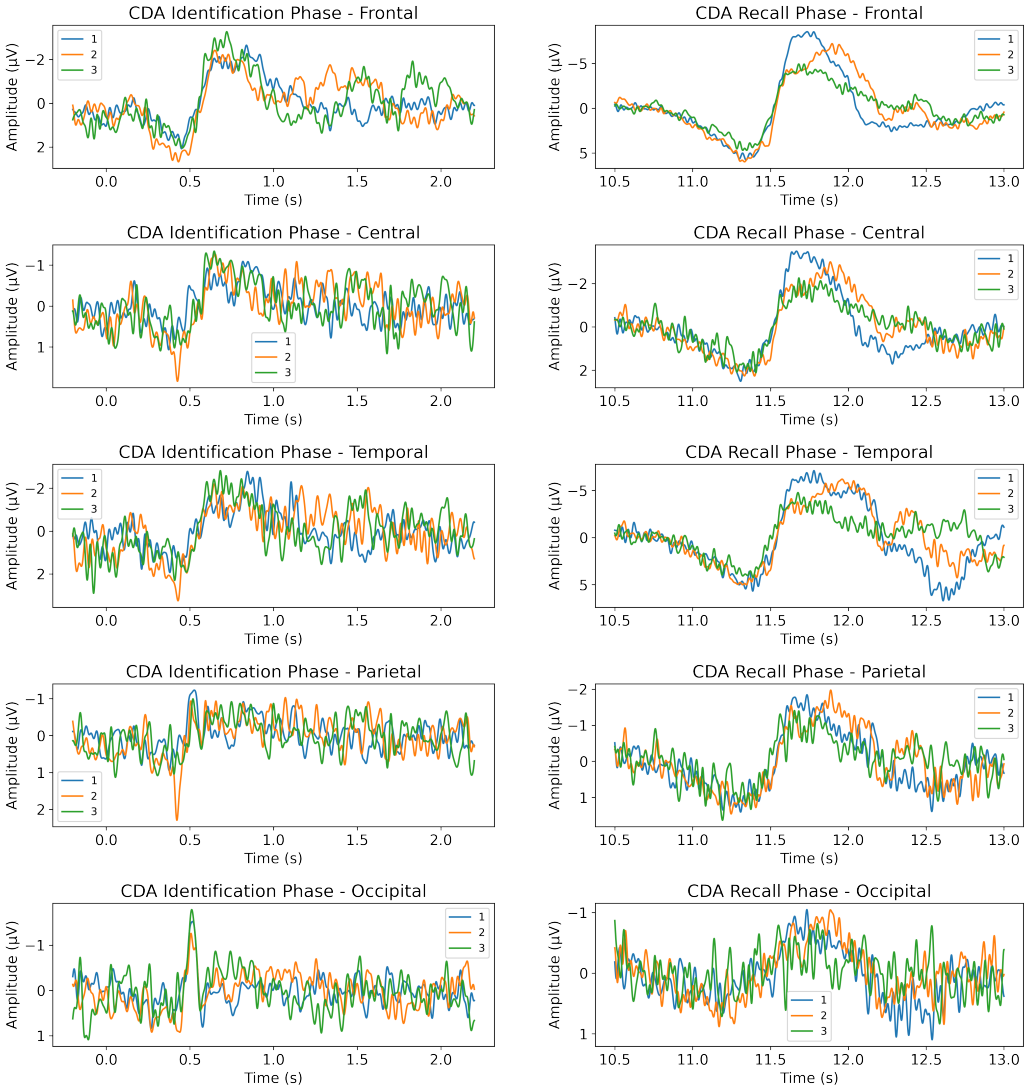


Fig. 31. Lateralized activity from the different brain regions for identification (first column) and recall (second column) phases. Trials with 1, 2 and 3 targets.

color scale to accentuate the differences observed given that most of the values are negative.

On the spectral map (Figure 32) we can identify, as expected, a perturbation near t=0.15s when the spheres turn red, creating a narrow perturbation corresponding to

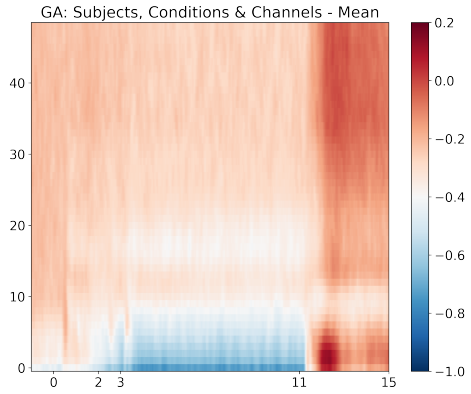


Fig. 32. ERSP: Grand Average (GA) across all subjects, all conditions and all channels. The color represents the log ratio of the power at each instant with average power of the baseline [-1,0]s for that same frequency, averaged across subjects.

the ERP in the time domain related to the sudden visual change. Also visible are perturbations from the ERP near t=2.15s when they turn back yellow and from the ERP near t=3s when they suddenly start moving. During tracking, we observe a strong inhibition of lower frequencies until alpha band (∼8Hz) where the inhibition isn't as strong, then the inhibition seems to continue in the ∼14-18Hz range after which there is a change in inhibition intensity for high beta and gamma. More strikingly, we observe a clear cognitive switch as soon as the spheres stop moving and the participant is asked to answer (at t=11s). Suddenly, the lower frequencies are re-activated and alpha slightly reduced. After t=13s, the values can't really be interpreted given the variability in speed to answer from trial to trial, across participants and the number of targets.

Looking at the spatial distribution over the midline channels on Figure 33, we observe that during tracking (t=[3-11]s), there is a strong delta and theta inhibition in the frontal regions. During recall (t=[11-13]s), we observe a sudden re-activation of these lower frequencies in the delta and theta bands from frontal to occipital regions.

During recall, we also observe a strong activation of high-beta to low-gamma in the occipital region.

To better understand the effect across participants, both the mean and the median values were calculated for each time-frequency points and what we observe on Figure 33 is that the mean is not being distorted by outliers and that the distribution across trials and participants is somewhat symmetrical. A paired t-test was done for each time-frequency point comparing with the mean power of the baseline and the time-frequency points with p >= .05 were grayed-out on Figure 34. For readability we show the statistical analysis of only three channels (Fpz, Cz, Oz) as the six others were showing similar overall trends. We did not perform any multiple comparison correction, so given that we have performed tens of thousands of comparisons (i.e. for each time-frequency point independently) we are exposed to familywise type 1 error, meaning that the graphs presented here with the statistical masks are most likely showing more significant effects then there really are. Performing multiple comparison corrections like false discovery rate (FDR) on such a high amount of comparisons would obviously result in the opposite and hide all effects (type 2 error). Different approaches have been suggested to deal with ERSP significance such as cluster-level statistical permutation tests, however, given the mean and median of the log ratio with the baseline being both similar and biologically sound in time and frequency, correcting for the comparisons would most likely not invalidate the general trends we are observing here. Also noteworthy, as we can see on the ERSP graphs, the edge between alpha and theta bands around 8Hz and the edge between alpha and beta bands around 12Hz are blurry and comes out as non-statistically significant (vs baseline). This is most likely due to the individual differences in frequency bands ([**17**]).

The analysis for the effect of the number of targets showed a significant difference only during recall. ERSP values for 1, 2 and 3 targets were obtained and paired t-tests were performed for condition 1 vs 3 targets on the whole trial for each time-frequency

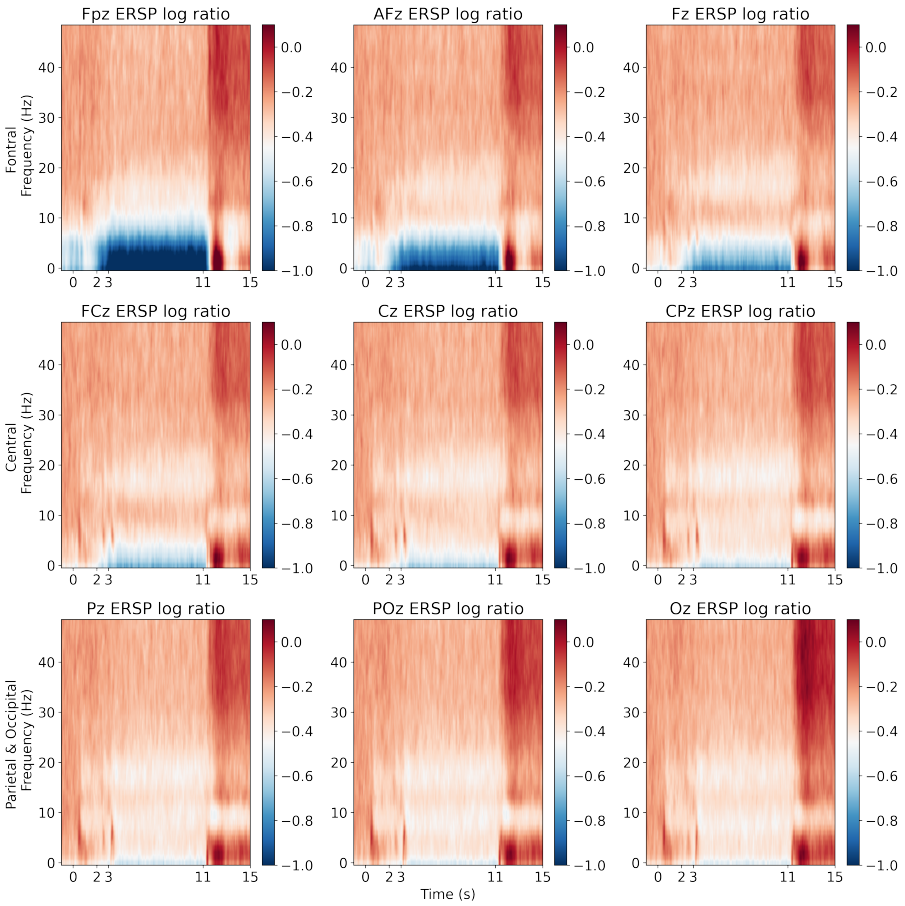


Fig. 33. ERSP: Midline electrodes, frontal to occipital. The color represents the log ratio of the power at each instant with average power of the baseline [-1,0]s for that same frequency, averaged across subjects.

point. The statistical mask shown on the right on Figure 35 grays out the points for which the p-value was higher or equal to 0.05, and therefore, not significantly different from 1 to 3 targets. The mask is plotted on top of the resulting ERSP map obtained by subtracting the ERSP values of 1 target to the ERSP values of 3 targets to highlight the difference in power between the two. The statistical tests for 1 vs 2 and 2 vs 3 targets were not performed because we assume some sort of linearity between 1, 2 and 3 targets and therefore their effect would be somewhere between

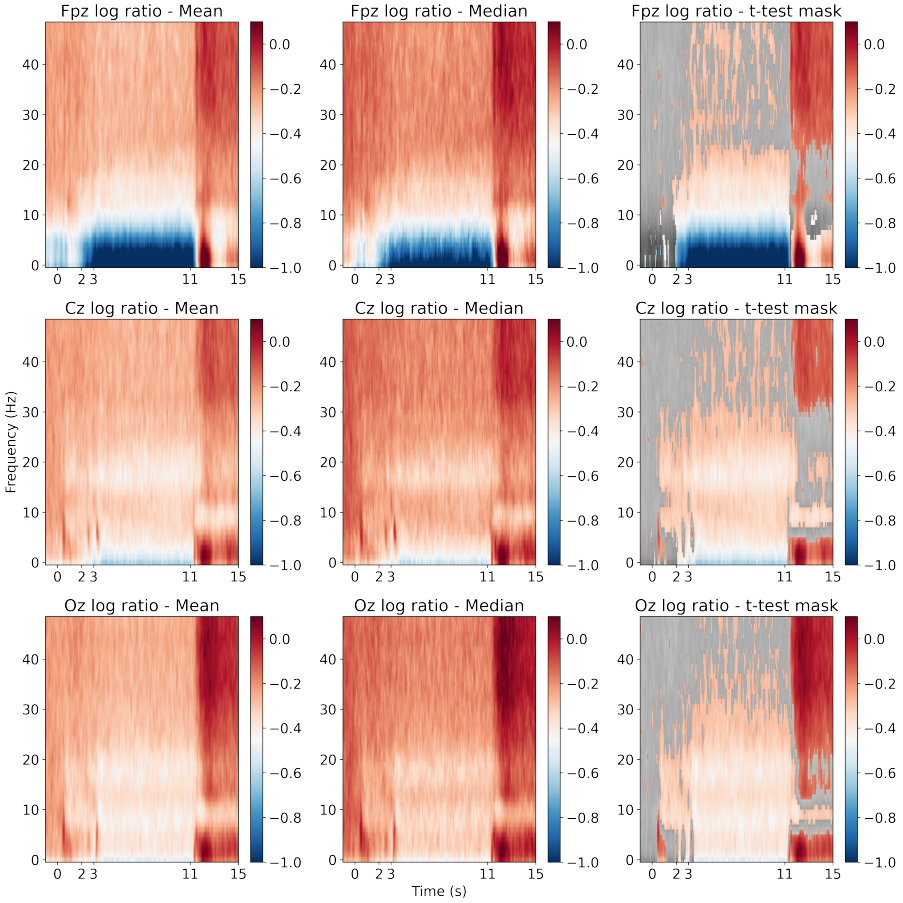


Fig. 34. ERSP: Fpz, Cz, and Oz electrodes. The color represents the log ratio of the power at each instant with average power of the baseline [-1,0]s for that same frequency. The first column is the mean across subjects. The second column is the median across subjects. The third column is the mean across subjects with a gray mask where the p-value of a t-test >=.05 (i.e. gray means not significantly different than baseline).

what we observe for 1 and 3 targets. For readability only the frontal and parietal clusters are displayed. Frontal because that's where we observed the strongest CDA in the time domain and parietal because that's where we initially expected the strongest difference.

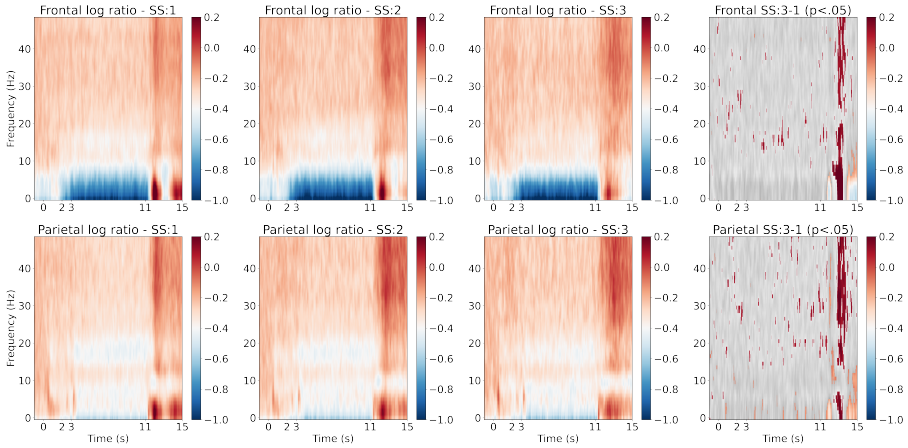


Fig. 35. ERSP: Frontal and parietal regions for 1, 2, and 3 targets. The color represents the log ratio of the power at each instant with average power of the baseline [-1,0]s for that same frequency. The first column is the mean across subjects. The last column is the ERSP with 3 targets minus the ERSP with 1 target with a gray mask where the p-value of a t-test between the two >=.05 (i.e. gray means no significant difference between set size of 1 target vs 3 targets.)

Finally, as expected given the results from the ERSP maps showed and explained above, the repeated measure three-way ANOVA on raw power revealed a significant effect for Phase ($F_{(2,36)} = 12.13$, $p < .0001$), for Frequency Band ($F_{(4,72)} = 50.4$, $p < .0001$), and smaller yet significant effect for Set Size ($F_{(2,36)} = 4.6$, $p = .016$). The interaction between Phase and Frequency Band was significant ($F_{(8,144)} = 11.58$, $p < .0001$), the interaction between Phase and Set Size was significant ($F_{(4,72)} = 6.36$, $p < .0005$), the interaction between Frequency Band and Set Size was significant ($F_{(8,144)} = 5.24$, $p < .0001$) and finally, the interaction between Phase, Set Size and Frequency Band was also significant ($F_{(16,288)} = 6.71$, $p < .0005$). The statistical analysis was performed on both the raw power (results mentioned above) and the power ratio (ERSP maps on figures). We reported the results from the raw power as most studies use the raw power directly. Using the power ratio instead of raw power

provides a more normalized value across subjects, and we obtained similar results. The differences include the Set Size effect that came out slightly over the p = .05 significance threshold, with ($F_{(2,38)}$ = 2.92, p = .066), the interaction between Set Size and Phase was trending towards significance ($F_{(4,76)}$ = 2.46, p = .0521) and the interaction between Set Size, Phase and Frequency Band was significant ($F_{(16,304)}$ = 2.88, p < .0005).

4. Discussion

Several studies have been published on EEG activity during MOT tasks (e.g. [**11, 4, 28**]), usually focusing on either the indexing or tracking phase but very little, if at all, on the recall phase. Here, we took a more holistic view to look at electrophysiological changes over the whole sequence (i.e. trial).

CDA. During the identification and tracking phases, we did observe lateralized activity as expected. However, the highest amplitude were seen in frontal and temporal regions and not in the parietal region as expected from the recent CDA literature. Also, the amplitude in the parietal region did not increase with the number of targets during identification and tracking which is in contradiction with the recent CDA literature ([**46, 11, 44**]) and our own CDA review paper ([**39**]). Moreover, during recall, we actually observed the opposite effect where the amplitude in certain regions was actually higher with less targets and the activity lasted longer in time with more targets. Given the clear lateralized ERPs and activity observed in the results (e.g. Figure 28) the design of the task clearly had laterialized targets. Therefore, to explain our results being different from what we expected based on the literature, we rule out a bad task design and point out four ways in which our 3D-MOT task is different than most VWM tasks used to study CDA. First, as opposed to change detection tasks where the participant has to remember all the items at once, here the participants could have recalled the items one by one, sustaining a longer CDA over time. Second,

most VWM tasks in CDA studies have very short trials of about 1 to 2 seconds long. Third, in most CDA related tasks, the objects are temporarily removed from the screen forcing the participant to have an internal representation of the objects. Here the objects are always visible for the participant and perhaps engage a different cognitive strategy. Forth, most tasks are done in 2D without stereoscopy. These confounding differences could make the participant use a different cognitive strategy relying more on attention mechanisms than working memory during tracking.

In order to better understand the results we obtained, two things have to be disentangled. First, is WM required for the 3D-MOT task and if so, are the targets held in memory for the whole duration of tracking or only held in memory during indexing and/or recall? Second, is the lateralized activity observed here the same as the CDA referred to in the literature and believed to play a role in WM, or is it a different, yet lateralized, activity? If the answer is yes to both we should have observed a CDA during tracking and the CDA amplitude should have varied in amplitude based on the number of targets. Our results diverge from [**11**] where they showed a clear CDA with different amplitudes during tracking in a different and shorter 2D-MOT task. As we can see in our results, the activity is clearly lateralized and has a CDA-like shape, in both the identification and recall phases but not during tracking which leads us to believe that there might be a cognitive switch between memory and attention, where the targets might not be held in working memory during tracking but rather only tracked with attention mechanisms. During recall, there is no doubt that the participant has to leverage working memory to provide the answers and even more so, the confidence level for each target. Interestingly, during the identification phase for the lateralized trials, the activity is nicely symmetrical early post-stimulus and then a change occurs and the following 500ms are not as symmetrical. We hypothesize that the period between 100-400ms engages more bottom-up brain mechanisms as a response to the stimuli and the following 500-600ms engages more

top-down mechanisms to which there seems to be a difference between left and right hemispheres.

Eye movements. In order to rule out the possibility that remaining eye movement artifacts be the main driver of the frontal EEG activity observed in the results, we calculated the correlation of the EOG channel pair with the lateralized signal obtained from the frontal cluster. The Pearson product-moment correlation coefficients was calculated for each participant and then averaged to obtain the group level correlation coefficient. A weak to moderate correlation is expected at the minimum, as nearby electrophysiological channels share information. For the EOG channels, we obtained a correlation coefficient of 0.765 which, when compared with the other channel pairs (F5-F6 r=0.875; F7-F8 r=0.845; AF7-AF8 r=0.823; F3-F4 r=0.798; FT7-FT8 r=0.788; AF3-AF4 r=0.769; F1-F2 r=0.739; Fp1-Fp2 r=0.581), is among the weakest correlation with the lateralized CDA-like activity presented in the results. This supports the notion that the signal obtained isn't driven by eye movements but rather cognitive processes.

Frequency bands. During tracking, there is a strong inhibition of both delta and theta frequencies, followed by a significant reactivation of these same frequencies during recall. For such a strong switch between inhibition and activation, we ruled out the hypothesis that it could be muscle activity or artifacts based on releasing the tension after being focused during tracking, because if that was the case, we would see the opposite effect on Figure 35 and the activity would spread longer with only 1 target as they would start releasing the tension and moving their body faster. However here we see a stronger but shorter activity in delta and theta during recall for 1 targets, less strong and slightly longer for 2 targets, and the weakest power but longer spread over time. This is aligned with the CDA plots in the time domain (see Figure 31). We also ruled out the motor and speech brain-related activity from providing the answers, as such activity would be more localized and not clearly visible on all clusters as seen here.

The activation of lower frequencies (delta and theta) during recall is distributed spatially but the strong inhibition during tracking is frontal. We hypothesize that such inhibition might be coming from a top-down cognitive mechanism to ignore task distractors and prevent them from being encoded during tracking. Also, the 3D-MOT task requires the participant's full attention, because many occlusions and contacts between the objects (targets and distractors alike) are happening and one tiny lapse in attention can make the participant fail the trial. Therefore, it is of upmost importance for our top-down mechanisms to protect our attention from task-related, environment-related as well as internal distractions. Theta has been linked to memory, cognitive control and attention systems ([**2, 21, 12**]) and as observed here, has a key role in the MOT task during tracking and recall. Theta and gamma phase coupling has also been accumulating supportive evidence for playing a key role in visual processes involving working memory ([**40, 22, 23**]). This theta-gamma coupling might explain why during recall we observe also a strong gamma activation at the same time as the theta activation, strongest in the occipital region. We haven't done any phase coupling nor connectivity analysis to confirm the link between these frequencies but it is in our future plans to look at phase coupling, source localization, and connectivity.

As for delta, the literature for its role in cognitive functions, aside from sleep studies, isn't as extensive as for the other frequency bands but it has been linked to similar attention mechanisms than theta ([**16**]). Alpha, one of the most prominent rhythm in the human brain, also plays a key role in attention, especially for inhibition of distractors. A previous study providing online feedback on a 3D-MOT task based on real time alpha peak frequency has helped improve performance on the task ([**33**]). While it is clear that alpha is playing a role, it remains unclear exactly how, as studies have shown both an increase ([**47**]) as well as a decrease ([**28**]) in alpha during object tracking. Based on the previous evidences, looking at the grand average over the whole trial and then averaging all the trials together might be the wrong way of looking at alpha as it might be involved in a more granular fashion during a trial.

Calculating the grand average might hide the subtle within-trial changes of alpha. On the ERPS graphs, we see that power changes are clearly happening around, and at, alpha frequencies however the effect isn't as strong as some other frequencies.

Given the clear EEG pattern we observed in the frequency domain for the whole trial (∼15s), we wanted to compare with another MOT task and run a similar frequency analysis to see if we'd see the same pattern of activity. Thanks to Nicholas S. Bland and colleagues, who shared their EEG data from their 2020 study ([**4**]), we were able to generate an ERSP graph and we found a striking resemblance. Bland and colleagues used a 2D-MOT (our task is a 3D-MOT), their stimuli were 2D circles with no filling (i.e. rings), their trials had either 2 or 4 targets presented either between-hemifield moving freely left and right but not crossing the middle part vertically or within-hemifield moving freely up and down but not crossing the middle part horizontally. At the start of the trial they presented all the circles in white, then highlighted the targets in blue for 2 seconds (like our task) after which they reverted back to white for 500ms (we used 1s in our task) before all objects start moving for 8 seconds (like our task). Their recall phase was slightly different as their participants had to click on the targets with a mouse and then received the visual feedback with the correct answers for 1.5s. In our task the participant gave their answer verbally before entering their confidence level with the keyboard and then received visual feedback with the correct answers for 2s. They recorded EEG activity with a 64 channels BrainCap (BrainProducts) device. On Figure 36, we see the ERSP grand average across all conditions, all channels and all participants of our study on the right (as presented on Figure 32 before) and Bland et al., 2020 on the left. We used all the trials (N=192) for all the participants (N=41). The spatial distribution on the midline was also similar to ours, showing a stronger inhibition of delta and theta in the frontal regions during tracking. Their inhibition of lower frequencies (compared to baseline) was stronger than our results so we slightly adjusted the color scale to keep a smooth range.

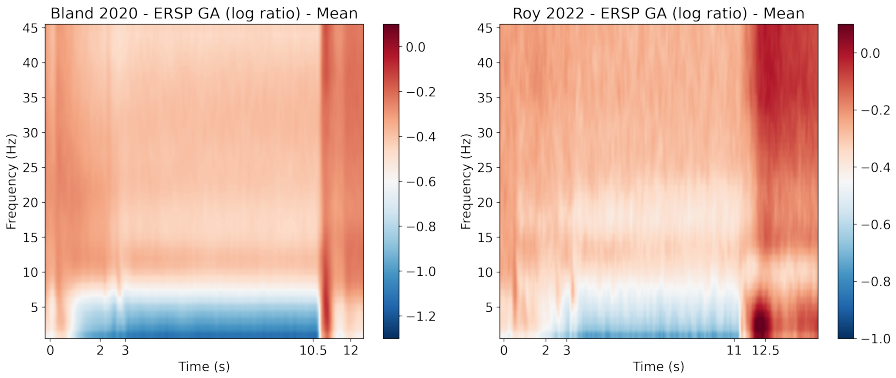


Fig. 36. ERSP: Bland, 2022 vs Roy, 2022.

One obvious limitation of our study for addressing the CDA memory load, is the number of trials per condition. According to a recent study by William X. Q. Ngiam and colleagues on the statistical power to detect set-size effects in contralateral delay activity, it requires between 30 to 50 clean trials with a sample of 25 subjects to achieve approximately 80% statistical power on detecting the presence of the CDA ([**30**]). In our study, we kept 20 subjects with an average of 45 clean (lateralized) trials after removing bad trials during preprocessing.

In order to keep the participants engaged and energized (knowing they had another task after), we asked them to give the answers orally as opposed to entering them with a keyboard. Therefore, the instructor, despite entering the answers really fast on the numpad, induced an external timing factor during the recall phase. Finally, given the interesting results we observed during the recall phase, having more granular information during that phase such as the time of entry of each answer for each participant would have allowed us to investigate deeper the neural activity during that phase.

In conclusion, we believe that we offered a more holistic view of the neural substrates during a 3D-MOT task looking at the activity across the three different phases (Identification/Indexing, Tracking, Recall/Answer) both in the time and

frequency domains. We also analysed the raw EEG data of another MOT study to compare our findings and observed similar overall trends, solidifying our findings. This study sheds more light at what seems to be some sort of a handoff between focused attention and working memory processes during tracking and recall and how the delta and theta bands in the frontal regions play a key role in the 3D-MOT task as they are being toggled like an on/off switch across phases.

Both the data and the code is available online: https://github.com/royyannick/3DMOT_EEG

References

[1] George A Alvarez and Steven L Franconeri. How many objects can you track?: Evidence for a resource-limited attentive tracking mechanism. *Journal of vision*, 7(13):14–14, 2007.

[2] Erol Başar, Canan Başar-Eroglu, Sirel Karakaş, and Martin Schürmann. Gamma, alpha, delta, and theta oscillations govern cognitive processes. *International journal of psychophysiology*, 39(2-3):241–248, 2001.

[3] Julie Justine Benoit, Eugenie Roudaia, Taylor Johnson, Trevor Love, and Jocelyn Faubert. The neuropsychological profile of professional action video game players. *PeerJ*, 8:e10211, 2020.

[4] Nicholas S Bland, Jason B Mattingley, and Martin V Sale. Gamma coherence mediates interhemispheric integration during multiple object tracking. *Journal of Neurophysiology*, 123(5):1630–1644, 2020.

[5] Patrick Cavanagh. Visual cognition. *Vision research*, 51(13):1538–1551, 2011.

[6] Patrick Cavanagh and George A Alvarez. Tracking multiple targets with multifocal attention. *Trends in cognitive sciences*, 9(7):349–354, 2005.

[7] Wei-Ying Chen, Piers D Howe, and Alex O Holcombe. Resource demands of object tracking and differential allocation of the resource. *Attention, Perception, & Psychophysics*, 75(4):710–725, 2013.

[8] Nelson Cowan. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and brain sciences*, 24(1):87–114, 2001.

[9] Patrick Dendorfer, Hamid Rezatofighi, Anton Milan, Javen Shi, Daniel Cremers, Ian Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixé. Mot20: A benchmark for multi object tracking in crowded scenes. *arXiv preprint arXiv:2003.09003*, 2020.

[10] Trafton Drew, Todd S Horowitz, Jeremy M Wolfe, and Edward K Vogel. Delineating the neural signatures of tracking spatial position and working memory during attentive tracking. *Journal of Neuroscience*, 31(2):659–668, 2011.

[11] Trafton Drew and Edward K Vogel. Neural measures of individual differences in selecting and tracking multiple moving objects. *Journal of Neuroscience*, 28(16):4183–4191, 2008.

[12] Jarrod Eisma, Eric Rawls, Stephanie Long, Russell Mach, and Connie Lamm. Frontal midline theta differentiates separate cognitive control strategies while still generalizing the need for cognitive control. *Scientific Reports*, 11(1):1–14, 2021.

[13] Jocelyn Faubert. Professional athletes have extraordinary skills for rapidly learning complex and neutral dynamic visual scenes. *Scientific reports*, 3(1):1–3, 2013.

[14] Jocelyn Faubert and Lee Sidebottom. Perceptual-cognitive training of athletes. *Journal of Clinical Sport Psychology*, 6(1):85–102, 2012.

[15] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, et al. Meg and eeg data analysis with mne-python. *Frontiers in neuroscience*, 7:267, 2013.

[16] Bahar Güntekin and Erol Başar. Review of evoked and event-related delta responses in the human brain. *International Journal of Psychophysiology*, 103:43–52, 2016.

[17] Saskia Haegens, Helena Cousijn, George Wallis, Paul J Harrison, and Anna C Nobre. Inter-and intra-individual variability in alpha peak frequency. *Neuroimage*, 92:46–55, 2014.

[18] Jaclyn Hoke, Christopher Reuter, Thomas Romeas, Maxime Montariol, Thomas Schnell, and Jocelyn Faubert. Perceptual-cognitive & physiological assessment of training effectiveness. In *Interservice/Industry Training, Simulation, and Education Conference, Orlando, FL*, 2017.

[19] Lucica Iordanescu, Marcia Grabowecky, and Satoru Suzuki. Demand-based dynamic distribution of attention and monitoring of velocities during multiple-object tracking. *Journal of vision*, 9(4):1–1, 2009.

[20] Mainak Jas, Denis A Engemann, Yousra Bekhti, Federico Raimondo, and Alexandre Gramfort. Autoreject: Automated artifact rejection for meg and eeg data. *NeuroImage*, 159:417–429, 2017.

[21] Sirel Karakaş. A review of theta oscillation and its functional correlates. *International Journal of Psychophysiology*, 157:82–99, 2020.

[22] Moritz Köster, Uwe Friese, Benjamin Schöne, Nelson Trujillo-Barreto, and Thomas Gruber. Theta–gamma coupling during episodic retrieval in the human eeg. *Brain research*, 1577:57–68, 2014.

[23] Mikael Lundqvist, Jonas Rose, Pawel Herman, Scott L Brincat, Timothy J Buschman, and Earl K Miller. Gamma and beta bursts underlie working memory. *Neuron*, 90(1):152–164, 2016.

[24] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

[25] Roy Luria, Halely Balaban, Edward Awh, and Edward K Vogel. The contralateral delay activity as a neural measure of visual working memory. *Neuroscience & Biobehavioral Reviews*, 62:100–108, 2016.

[26] Gerald T Mangine, Jay R Hoffman, Adam J Wells, Adam M Gonzalez, Joseph P Rogowski, Jeremy R Townsend, Adam R Jajtner, Kyle S Beyer, Jonathan D Bohner, Gabriel J Pruna, et al. Visual tracking speed is related to basketball-specific measures of performance in nba players. *The Journal of Strength & Conditioning Research*, 28(9):2406–2414, 2014.

[27] Christian Merkel, J-M Hopf, H-J Heinze, and Mircea Ariel Schoenfeld. Neural correlates of multiple object tracking strategies. *Neuroimage*, 118:63–73, 2015.

[28] Christian Merkel, Jens-Max Hopf, and Mircea Ariel Schoenfeld. Electrophysiological hallmarks of location-based and object-based visual multiple objects tracking. *European Journal of Neuroscience*, 55(5):1200–1214, 2022.

[29] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016.

[30] William XQ Ngiam, Kirsten CS Adam, Colin Quirk, Edward K Vogel, and Edward Awh. Estimating the statistical power to detect set-size effects in contralateral delay activity. *Psychophysiology*, 58(5):e13791, 2021.

[31] Lauri Oksama and Jukka Hyönä. Position tracking and identity tracking are separate systems: Evidence from eye movements. *Cognition*, 146:393–409, 2016.

[32] Yesul Park, L Minh Dang, Sujin Lee, Dongil Han, and Hyeonjoon Moon. Multiple object tracking in deep learning approaches: A survey. *Electronics*, 10(19):2406, 2021.

[33] Brendan Parsons and Jocelyn Faubert. Enhancing learning in a perceptual-cognitive training paradigm using eeg-neurofeedback. *Scientific Reports*, 11(1):1–10, 2021.

[34] Brendan Parsons, Tara Magill, Alexandra Boucher, Monica Zhang, Katrine Zogbo, Sarah Bérubé, Olivier Scheffer, Mario Beauregard, and Jocelyn Faubert. Enhancing cognitive function using perceptual-cognitive training. *Clinical EEG and neuroscience*, 47(1):37–47, 2016.

[35] Zenon Pylyshyn. The role of location indexes in spatial perception: A sketch of the finst spatial-index model. *Cognition*, 32(1):65–97, 1989.

[36] Zenon W Pylyshyn and Ron W Storm. Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial vision*, 3(3):179–197, 1988.

[37] Rishi Rajalingham, Elias B Issa, Pouya Bashivan, Kohitij Kar, Kailyn Schmidt, and James J DiCarlo. Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *Journal of Neuroscience*, 38(33):7255–7269, 2018.

[38] Thomas Romeas, Antoine Guldner, and Jocelyn Faubert. 3d-multiple object tracking training task improves passing decision-making accuracy in soccer players. *Psychology of Sport and Exercise*, 22:1–9, 2016.

[39] Yannick Roy and Jocelyn Faubert. Is the contralateral delay activity (cda) a robust neural correlate for visual working memory (vwm) tasks? a reproducibility study. *arXiv preprint*, 2022.

[40] Paul Sauseng, Wolfgang Klimesch, Kirstin F Heise, Walter R Gruber, Elisa Holz, Ahmed A Karim, Mark Glennon, Christian Gerloff, Niels Birbaumer, and Friedhelm C Hummel. Brain oscillatory substrates of visual short-term memory capacity. *Current biology*, 19(21):1846–1852, 2009.

[41] Won Mok Shim, George A Alvarez, and Yuhong V Jiang. Spatial separation between targets constrains maintenance of attention on multiple objects. *Psychonomic bulletin & review*, 15(2):390–397, 2008.

[42] Lana M Trick and Zenon W Pylyshyn. What enumeration studies can show us about spatial attention: evidence for limited capacity preattentive processing. *Journal of Experimental Psychology: Human Perception and Performance*, 19(2):331, 1993.

[43] Domenico Tullo, Jocelyn Faubert, and Armando Bertone. Examining the benefits of training attention with multiple object-tracking for individuals diagnosed with a neurodevelopmental condition: A cross-over, cognitive training study. *Journal of Vision*, 18(10):1021–1021, 2018.

[44] Nash Unsworth, Keisuke Fukuda, Edward Awh, and Edward K Vogel. Working memory delay activity predicts individual differences in cognitive abilities. *Journal of Cognitive Neuroscience*, 27(5):853–865, 2015.

[45] Oshin Vartanian, Lori Coady, and Kristen Blackler. 3d multiple object tracking boosts working memory span: Implications for cognitive training in military populations. *Military Psychology*, 28(5):353–360, 2016.

[46] Edward K Vogel and Maro G Machizawa. Neural activity predicts individual differences in visual working memory capacity. *Nature*, 428(6984):748–751, 2004.

[47] Andreas Wutz, Agnese Zazio, and Nathan Weisz. Oscillatory bursts in parietal cortex reflect dynamic attention between multiple objects and ensembles. *Journal of Neuroscience*, 40(36):6927–6937, 2020.

[48] Steven Yantis. Multielement visual tracking: Attention and perceptual organization. *Cognitive psychology*, 24(3):295–340, 1992.

5. Supplementary Material

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

Table 4. Electrode Clusters

Third Article.

Deep learning-based electroencephalography analysis: a systematic review

by

Yannick Roy[1], Hubert Banville[2], Isabela Albuquerque[3], Alexandre Gramfort[4], Tiago H. Falk[5], and Jocelyn Faubert[6]

([1]) Faubert Lab, Université de Montréal, Canada
([2]) Inria, Université Paris-Saclay, France
([3]) MuSAE Lab, INRS, Canada
([4]) Inria, Université Paris-Saclay, France
([5]) MuSAE Lab, INRS, Canada
([6]) Faubert Lab, Université de Montréal, Canada

This article was submitted in Journal of Neural Engineering (JNE).

Contributions. The first author did most of the breakdown and analysis of the DL-EEG papers reviewed, the second author did most of the writing and the graphs. The other authors helped with the writing and reviewing.

RÉSUMÉ. L'électroencéphalographamme (EEG) est un signal complexe et peut nécessiter plusieurs années d'expérience, ainsi que des techniques avancées de traitement du signal et d'extraction des caractéristiques afin de l'interpréter. L'apprentissage profond a un potentiel intéressant pour la classification des signaux EEG. Dans cette revue de littérature, nous avons analysé 154 papiers publiés entre janvier 2010 et juillet 2018 à travers différents domaines tel que l'épilepsie, le sommeil, les interfaces cerveau-machine et la mesure des états affectifs et cognitifs. Notre analyse révèle que la quantité de donnée EEG utilisé varie grandement à travers les articles allant de quelques minutes à plusieurs milliers d'heures. Environ 40% des études ont utilisé un réseau neuronal convolutif, pour 13% avec un réseau récurrent, généralement avec un total allant de 3 à 10 couches. Finalement, le gain médian avec une approche profonde comparée à une approche traditionnelle est de 5.4% à travers toutes les études.

Mots clés : EEG, apprentissage profond, réseaux de neurones, review, survey

Abstract.

Context. Electroencephalography (EEG) is a complex signal and can require several years of training, as well as advanced signal processing and feature extraction methodologies to be correctly interpreted. Recently, deep learning (DL) has shown great promise in helping make sense of EEG signals due to its capacity to learn good feature representations from raw data. Whether DL truly presents advantages as compared to more traditional EEG processing approaches, however, remains an open question.

Objective. In this work, we review 154 papers that apply DL to EEG, published between January 2010 and July 2018, and spanning different application domains such as epilepsy, sleep, brain-computer interfacing, and cognitive and affective monitoring. We extract trends and highlight interesting approaches from this large body of literature in order to inform future research and formulate recommendations.

Methods. Major databases spanning the fields of science and engineering were queried to identify relevant studies published in scientific journals, conferences, and electronic preprint repositories. Various data items were extracted for each study pertaining to 1) the data, 2) the preprocessing methodology, 3) the DL design choices, 4) the results, and 5) the reproducibility of the experiments. These items were then analyzed one by one to uncover trends.

Results. Our analysis reveals that the amount of EEG data used across studies varies from less than ten minutes to thousands of hours, while the number of samples seen during training by a network varies from a few dozens to several millions, depending on how epochs are extracted. Interestingly, we saw that more than half the studies used publicly available data and that there has also been a clear shift from intra-subject to inter-subject approaches over the last few years. About 40% of the studies used convolutional neural networks (CNNs), while 13% used recurrent neural networks (RNNs), most often with a total of 3 to 10 layers. Moreover, almost one-half of the studies trained their models on raw or preprocessed EEG time series. Finally, the median gain in accuracy of DL approaches over traditional baselines was 5.4% across all relevant studies. More importantly, however, we noticed studies often suffer from poor reproducibility: a majority of papers would be hard or impossible to reproduce given the unavailability of their data and code.

Keywords: EEG and electroencephalogram and deep learning and neural networks and review and survey

1. Introduction

1.1. Measuring brain activity with EEG

EEG, the measure of the electrical fields produced by the active brain, is a brain mapping and neuroimaging technique widely used inside and outside the clinical domain [**66, 150, 19**]. Specifically, EEG picks up the electric potential differences, on the order of tens of μV, that reach the scalp when tiny excitatory post-synaptic potentials produced by pyramidal neurons in the cortical layers of the brain sum together. The potentials measured therefore reflect neuronal activity and can be used to study a wide array of brain processes.

Thanks to the great speed at which electric fields propagate, EEG has an excellent temporal resolution: events occurring at millisecond timescales can typically be captured. However, EEG suffers from low spatial resolution, as the electric fields generated by the brain are smeared by the tissues, such as the skull, situated between the sources and the sensors. As a result, EEG channels are often highly correlated spatially. The source localization problem, or inverse problem, is an active area of research in which algorithms are developed to reconstruct brain sources given EEG recordings [**70**].

There are many applications for EEG. For example, in clinical settings, EEG is often used to study sleep patterns [**1**] or epilepsy [**3**]. Various conditions have also been linked to changes in electrical brain activity, and can therefore be monitored to various extents using EEG. These include attention deficit hyperactivity disorder (ADHD) [**10**], disorders of consciousness [**46, 41**], depth of anaesthesia [**59**], etc. EEG is also widely used in neuroscience and psychology research, as it is an excellent tool for studying the brain and its functioning. Applications such as cognitive and affective monitoring are very promising as they could allow unbiased measures of, for example, an individual's level of fatigue, mental workload, [**18, 12**], mood, or emotions [**5**].

Finally, EEG is widely used in brain-computer interfaces (BCIs) - communication channels that bypass the natural output pathways of the brain - to allow brain activity to be directly translated into directives that affect the user's environment [**104**].

1.2. Current challenges in EEG processing

Although EEG has proven to be a critical tool in many domains, it still suffers from a few limitations that hinder its effective analysis or processing. First, EEG has a low signal-to-noise ratio (SNR) [**20, 77**], as the brain activity measured is often buried under multiple sources of environmental, physiological and activity-specific noise of similar or greater amplitude called "artifacts". Various filtering and noise reduction techniques have to be used therefore to minimize the impact of these noise sources and extract true brain activity from the recorded signals.

EEG is also a non-stationary signal [**30, 56**], that is its statistics vary across time. As a result, a classifier trained on a temporally-limited amount of user data might generalize poorly to data recorded at a different time on the same individual. This is an important challenge for real-life applications of EEG, which often need to work with limited amounts of data.

Finally, high inter-subject variability also limits the usefulness of EEG applications. This phenomenon arises due to physiological differences between individuals, which vary in magnitude but can severely affect the performance of models that are meant to generalize across subjects [**29**]. Since the ability to generalize from a first set of individuals to a second, unseen set is key to many practical applications of EEG, a lot of effort is being put into developing methods that can handle inter-subject variability.

To solve some of the above-mentioned problems, processing pipelines with domain-specific approaches are often used. A significant amount of research has been put into developing processing pipelines to clean, extract relevant features, and classify EEG

data. State-of-the-art techniques, such as Riemannian geometry-based classifiers and adaptive classifiers [**103**], can handle these problems with varying levels of success.

Additionally, a wide variety of tasks would benefit from a higher level of automated processing. For example, sleep scoring, the process of annotating sleep recordings by categorizing windows of a few seconds into sleep stages, currently requires a lot of time, being done manually by trained technicians. More sophisticated automated EEG processing could make this process much faster and more flexible. Similarly, real-time detection or prediction of the onset of an epileptic seizure would be very beneficial to epileptic individuals, but also requires automated EEG processing. For each of these applications, most common implementations require domain-specific processing pipelines, which further reduces the flexibility and generalization capability of current EEG-based technologies.

1.3. Improving EEG processing with deep learning

To overcome the challenges described above, new approaches are required to improve the processing of EEG towards better generalization capabilities and more flexible applications. In this context, deep learning (DL) [**88**] could significantly simplify processing pipelines by allowing automatic end-to-end learning of preprocessing, feature extraction and classification modules, while also reaching competitive performance on the target task. Indeed, in the last few years, DL architectures have been very successful in processing complex data such as images, text and audio signals [**88**], leading to state-of-the-art performance on multiple public benchmarks - such as the Large Scale Visual Recognition challenge [**35**] - and an ever-increasing role in industrial applications.

DL, a subfield of machine learning, studies computational models that learn hierarchical representations of input data through successive non-linear transformations [**88**]. Deep neural networks (DNNs), inspired by earlier models such as the perceptron

[**136**], are models where: 1) stacked layers of artificial "neurons" each apply a linear transformation to the data they receive and 2) the result of each layer's linear transformation is fed through a non-linear activation function. Importantly, the parameters of these transformations are learned by directly minimizing a cost function. Although the term "deep" implies the inclusion of many layers, there is no consensus on how to measure depth in a neural network and therefore on what really constitutes a deep network and what does not [**53**].

Fig. 38 presents an overview of how EEG data (and similar multivariate time series) can be formatted to be fed into a DL model, along with some important terminology (see Section 1.4), as well as an illustration of a generic neural network architecture. Usually, when c channels are available and a window has length l samples, the input of a neural network for EEG processing consists of an array $X_i \in \mathbb{R}^{c \times l}$ containing the l samples corresponding to a window for all channels. This two-dimensional array can be used directly as an example for training a neural network, or could first be unrolled into a n-dimensional array (where $n = c \times l$) as shown in Fig. 38. As for the m-dimensional output, it could represent the number of classes in a multi-class classification problem. Variations of this end-to-end formulation can be imagined where the window X_i is first passed through a preprocessing and feature extraction pipeline (e.g., time-frequency transform), yielding an example X_i' which is then used as input to the neural network instead.

Different types of layers are used as building blocks in neural networks. Most commonly, those are fully-connected (FC), convolutional or recurrent layers. We refer to models using these types of layers as FC networks, convolutional neural networks (CNNs) [**89**] and recurrent neural networks (RNNs) [**140**], respectively. Here, we provide a quick overview of the main architectures and types of models. The interested reader is referred to the relevant literature for more in-depth descriptions of DL methodology [**88, 53, 149**].

FC layers are composed of fully-connected neurons, i.e., where each neuron receives as input the activations of every single neuron of the preceding layer. Convolutional layers, on the other hand, impose a particular structure where neurons in a given layer only see a subset of the activations of the preceding one. This structure, akin to convolutions in signal or image processing from which it gets its name, encourages the model to learn invariant representations of the data. This property stems from another fundamental characteristic of convolutional layers, which is that parameters are shared across different neurons - this can be interpreted as if there were filters looking for the same information across patches of the input. In addition, pooling layers can be introduced, such that the representations learned by the model become invariant to slight translations of the input. This is often a desirable property: for instance, in an object recognition task, translating the content of an image should not affect the prediction of the model. Imposing these kinds of priors thus works exceptionally well on data with spatial structure. In contrast to convolutional layers, recurrent layers impose a structure by which, in its most basic form, a layer receives as input both the preceding layer's current activations and its own activations from a previous time step. Models composed of recurrent layers are thus encouraged to make use of the temporal structure of data and have shown high performance in natural language processing (NLP) tasks [**225, 206**].

Additionally, outside of purely supervised tasks, other architectures and learning strategies can be built to train models when no labels are available. For example, autoencoders (AEs) learn a representation of the input data by trying to reproduce their input given some constraints, such as sparsity or the introduction of artificial noise [**53**]. Generative adversarial networks (GANs) [**54**] are trained by opposing a generator (G), that tries to generate fake examples from an unknown distribution of interest, to a discriminator (D), that tries to identify whether the input it receives has been artificially generated by G or is an example from the unknown distribution of interest. This dynamic can be compared to the one between a thief (G) making

fake money and the police (D) trying to distinguish fake money from real money. Both agents push one another to get better, up to a point where the fake money looks exactly like real money. The training of G and D can thus be interpreted as a two-player zero-sum minimax game. When equilibrium is reached, the probability distribution approximated by G converges to the real data distribution [**54**].

Overall, there are multiple ways in which DL improve and extend existing EEG processing methods. First, the hierarchical nature of DNNs means features could potentially be learned on raw or minimally preprocessed data, reducing the need for domain-specific processing and feature extraction pipelines. Features learned through a DNN might also be more effective or expressive than the ones engineered by humans. Second, as is the case in the multiple domains where DL has surpassed the previous state-of-the-art, it has the potential to produce higher levels of performance on different analysis tasks. Third, DL facilitates the development of tasks that are less often attempted on EEG data such as generative modelling [**52**] and domain adaptation [**15**]. The use of deep learning-based methods allowed the synthesis of high-dimensional structured data such as images [**25**] and speech [**120**]. Generative models can be leveraged to learn intermediate representations or for data augmentation [**52**]. In the case of domain adaptation, the use deep neural networks along with techniques such as correlation alignment [**164**] allows the end-to-end learning of domain-invariant representations, while preserving task-dependent information. Similar strategies can also be applied to EEG data in order to learn better representations and thus improve the performance of EEG-based models across different subjects and tasks.

On the other hand, there are various reasons why DL might not be optimal for EEG processing and that may justify the skepticism of some of the EEG community. First and foremost, the datasets typically available in EEG research contain far fewer examples than what has led to the current state-of-the-art in DL-heavy domains such as computer vision (CV) and NLP. Data collection being relatively expensive and data accessibility often being hindered by privacy concerns - especially with clinical data -

openly available datasets of similar sizes are not common. Some initiatives have tried to tackle this problem though [**64**]. Second, the peculiarities of EEG, such as its low SNR, make EEG data different from other types of data (e.g, images, text and speech) for which DL has been most successful. Therefore, the architectures and practices that are currently used in DL might not be readily applicable to EEG processing.

1.4. Terminology used in this review

Some terms are sometimes used in the fields of machine learning, deep learning, statistics, EEG and signal processing with different meanings. For example, in machine learning, "sample" usually refers to one example of the input received by a model, whereas in statistics, it can be used to refer to a group of examples taken from a population. It can also refer to the measure of a single time point in signal processing and EEG. Similarly, in deep learning, the term "epoch" refers to one pass through the whole training set during training; in EEG, an epoch is instead a grouping of consecutive EEG time points extracted around a specific marker. To avoid the confusion, we include in Table 5 definitions for a few terms as used in this review. Fig. 37 gives a visual example of what these terms refer to.

1.5. Objectives of the review

This systematic review covers the current state-of-the-art in DL-based EEG processing by analyzing a large number of recent publications. It provides an overview of the field for researchers familiar with traditional EEG processing techniques and who are interested in applying DL to their data. At the same time, it aims to introduce the field applying DL to EEG to DL researchers interested in expanding the types of data they benchmark their algorithms with, or who want to contribute to EEG research. For readers in any of these scenarios, this review also provides detailed methodological information on the various components of a DL-EEG pipeline to

	Definition used in this review
Point or sample	A measure of the instantaneous electric potential picked up by the EEG sensors, typically in μV.
Example	An instantiation of the data received by a model as input, typically denoted by $\mathbf{x}_i$ in the machine learning literature.
Trial	A realization of the task under study, e.g., the presentation of one image in a visual ERP paradigm.
Window or segment	A group of consecutive EEG samples extracted for further analysis, typically between 0.5 and 30 seconds.
Epoch	A window extracted around a specific trial.

Table 5. Disambiguation of common terms used in this review.

inform their own implementation[1]. In addition to reporting trends and highlighting interesting approaches, we distill our analysis into a few recommendations in the hope of fostering reproducible and efficient research in the field.

1.6. Organization of the review

The review is organized as follows: Section 1 briefly introduces key concepts in EEG and DL, and details the aims of the review; Section 2 describes how the systematic review was conducted, and how the studies were selected, assessed and analyzed; Section 3 focuses on the most important characteristics of the studies selected and

[1]Additional information with more fine-grained data can be found in our data items table available at `http://dl-eeg.com`.

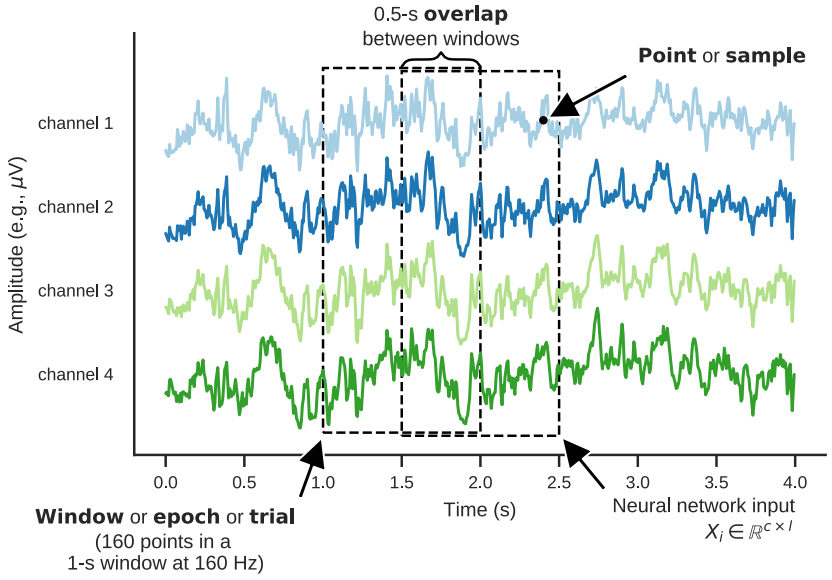


Fig. 37. Overlapping windows (which may correspond to trials or epochs in some cases) are extracted from multichannel EEG recordings.

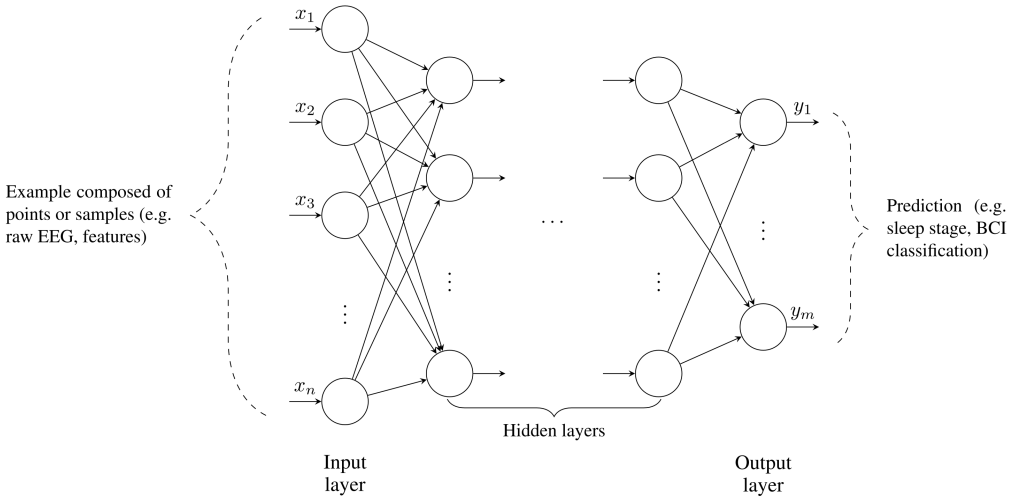


Fig. 38. Illustration of a general neural network architecture.

describes trends and promising approaches; Section 4 discusses critical topics and challenges in DL-EEG, and provides recommendations for future studies; and Section 5

concludes by suggesting future avenues of research in DL-EEG. Finally, supplementary material containing our full data collection table, as well as the code used to produce the graphs, tables and results reported in this review, are made available online.

How to use this review

We advise readers interested in a specific application domain of EEG to use the review in the following way:

(1) Read or glance over the **main result sections and corresponding discussion sections** covering general data items. This should give the reader a broad overview of the current practices and design choices used in the field of DL-EEG.

(2) If the reader is interested in a specific application (e.g., brain-computer interfacing) or in a specific type of architecture (e.g., CNNs), **identify relevant references**.

(3) **Consult the detailed summary of the relevant references** - which includes the data items introduced in Table 39 as well as many more (e.g., detailed preprocessing and feature extraction methodology, software implementation, values of specific hyperparameters, etc.) - contained in the data items spreadsheet available online at `http://dl-eeg.com`.

(4) Check the data items spreadsheet for **regular updates** (including additional studies) and the online repository for an updated version of the figures included in this review. We aim to provide readers with **evolving and up-to-date information** on various domains and architectures; therefore the table will remain open for external contributions from the community and authors of DL-EEG studies.

(5) **Use our checklist provided in Appendix 5** to ensure that you include all the relevant information in your future DL-EEG publications.

2. Methods

English journal and conference papers, as well as electronic preprints, published between January 2010 and July 2018, were chosen as the target of this review. PubMed, Google Scholar and arXiv were queried[2] to collect an initial list of papers containing specific search terms in their title or abstract.[3] Additional papers were identified by scanning the reference sections of these papers. The databases were queried for the last time on July 2, 2018.

The following search terms were used to query the databases: (1) EEG, (2) electroencephalogra*, (3) deep learning, (4) representation learning, (5) neural network*, (6) convolutional neural network*, (7) ConvNet, (8) CNN, (9) recurrent neural network*, (10) RNN, (11) long short-term memory, (12) LSTM, (13) generative adversarial network*, (14) GAN, (15) autoencoder, (16) restricted boltzmann machine*, (17) deep belief network* and (18) DBN. The search terms were further combined with logical operators in the following way: (1 OR 2) AND (3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 OR 11 OR 12 OR 13 OR 14 OR 15 OR 16 OR 17 OR 18). The papers were then included or excluded based on the criteria listed in Table 6.

To assess the eligibility of the selected papers, the titles were read first. If the title did not clearly indicate whether the inclusion and exclusion criteria were met, the abstract was read as well. Finally, when reading the full text during the data collection process, papers that were found to be misaligned with the criteria were rejected.

Non-peer reviewed papers, such as arXiv electronic preprints[4], are a valuable source of state-of-the-art information as their release cycle is typically shorter than that of

[2]The queries used for each database are available at `http://dl-eeg.com`.

[3]Since the Google Scholar search engine only allows searching full text or titles, and not titles and abstracts, the query was performed using the flag *allintitle* to search titles only. On arXiv and PubMed, however, both abstracts and titles were queried.

[4]`https://arxiv.org/`

Table 6. Inclusion and exclusion criteria.

Inclusion criteria	Exclusion criteria
• Training of one or multiple deep learning architecture(s) to process non-invasive EEG data.	• Studies focusing solely on invasive EEG (e.g., electrocorticography (ECoG) and intracortical EEG) or magnetoencephalography (MEG). • Papers focusing solely on software tools. • Review articles.

peer-reviewed publications. Moreover, unconventional research ideas are more likely to be shared in such repositories, which improves the diversity of the reviewed work and reduces the bias possibly introduced by the peer-review process [**124**]. Therefore, non-peer reviewed preprints were also included in our review. However, whenever a peer-reviewed publication followed a preprint submission, the peer-reviewed version was used instead.

A data extraction table was designed containing different data items relevant to our research questions, based on previous reviews with similar scopes and the authors' prior knowledge of the field. Following a first inspection of the papers with the data extraction sheet, data items were added, removed and refined. Each paper was initially reviewed by a single author, and then reviewed by a second if needed. For each article selected, around 70 data items were extracted covering five categories: origin of the article, rationale, data used, EEG processing methodology, DL methodology and reported results. Table 39 lists and defines the different items included in each of these categories. We make this data extraction table openly available for interested readers to reproduce our results and dive deeper into the data collected. We also

invite authors of published work in the field of DL and EEG to contribute to the table by verifying its content or by adding their articles to it.

The first category covers the origin of the article, that is whether it comes from a journal, a conference publication or a preprint repository, as well as the country of the first author's affiliation. This gives a quick overview of the types of publication included in this review and of the main actors in the field. Second, the rationale category focuses on the domains of application of the selected studies. This is valuable information to understand the extent of the research in the field, and also enables us to identify trends across and within domains in our analysis. Third, the data category includes all relevant information on the data used by the selected papers. This comprises both the origin of the data and the data collection parameters, in addition to the amount of data that was available in each study. Through this section, we aim to clarify the data requirements for using DL on EEG. Fourth, the EEG processing parameters category highlights the typical transformations required to apply DL to EEG, and covers preprocessing steps, artifact handling methodology, as well as feature extraction. Fifth, details of the DL methodology, including architecture design, training procedures and inspection methods, are reported to guide the interested reader through state-of-the-art techniques. Sixth, the reported results category reviews the results of the selected articles, as well as how they were reported, and aims to clarify how DL fares against traditional processing pipelines performance-wise. Finally, the reproducibility of the selected articles is quantified by looking at the availability of the data and code. The results of this section support the critical component of our discussion.

3. Results

The database queries yielded 553 different results that matched the search terms (see Fig. 40). 49 additional papers were then identified using the reference sections

Category	Data item	Description
Origin of article	Type of publication	Whether the study was published as a journal article, a conference paper or in an electronic preprint repository
	Venue	Publishing venue, such as the name of a journal or conference
	Country of first author affiliation	Location of the affiliated university, institute or research body of the first author
Study rationale	Domain of application	Primary area of application of the selected study. In the case of multiple domains of application, the domain that was the focus of the study was retained
Data	Quantity of data	Quantity of data used in the analysis, reported both in total number of samples and total minutes of recording
	Hardware	Vendor and model of the EEG recording device used
	Number of channels	Number of EEG channels used in the analysis. May differ from the number of recorded channels
	Sampling rate	Sampling rate (reported in Hertz) used during the EEG acquisition
	Subjects	Number of subjects used in the analysis. May differ from the number of recorded subjects
	Data split and cross-validation	Percentage of data used for training, validation, and test, along with the cross-validation technique used, if any
	Data augmentation	Data augmentation technique used, if any, to generate new examples
EEG processing	Preprocessing	Set of manipulation steps applied to the raw data to prepare it for use by the architecture or for feature extraction
	Artifact handling	Whether a method for cleaning artifacts was applied
	Features	Output of the feature extraction procedure, which aims to better represent the information of interest contained in the preprocessed data
Deep learning methodology	Architecture	Structure of the neural network in terms of types of layers (e.g. fully-connected, convolutional)
	Number of layers	Measure of architecture depth
	EEG-specific design choices	Particular architecture choices made with the aim of processing EEG data specifically
	Training procedure	Method applied to train the neural network (e.g. standard optimization, unsupervised pre-training followed by supervised fine-tuning, etc)
	Regularization	Constraint on the hypothesis class intended to improve a learning algorithm generalization performance (e.g. weight decay, dropout)
	Optimization	Parameter update rule
	Hyperparameter search	Whether a specific method was employed in order to tune the hyperparameter set
	Subject handling	Intra- versus inter-subject analysis
	Inspection of trained models	Method used to inspect a trained DL model
Results	Type of baseline	Whether the study included baseline models that used traditional processing pipelines, DL baseline models, or a combination of the two
	Performance metrics	Metrics used by the study to report performance (e.g. accuracy, f1-score, etc)
	Validation procedure	Methodology used to validate the performance of the trained models, including cross-validation and data split
	Statistical testing	Types of statistical tests used to assess the performance of the trained models
	Comparison of results	Reported results of the study, both for the trained DL models and for the baseline models
Reproducibility	Dataset	Whether the data used for the experiment comes from private recordings or from a publicly available dataset
	Code	Whether the code used for the experiment is available online or not, and if so, where

Fig. 39. Data items extracted for each article selected.

of the initial papers. Based on our inclusion and exclusion criteria, 448 papers were

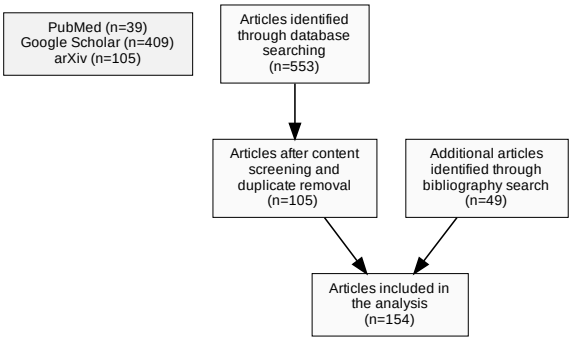


Fig. 40. Selection process for the papers.

excluded. One additional paper was excluded since it had been retracted. Therefore, 154 papers were selected for inclusion in the analysis.

3.1. Origin of the selected studies

Our search methodology returned 51 journal papers, 61 conference and workshop papers and 42 preprints that met our criteria. A total of 28 journal and conference papers had initially been made available as preprints on arXiv or bioRxiv. Popular journals included *Neurocomputing*, *Journal of Neural Engineering* and *Biomedical Signal Processing and Control*, each with three publications contained in our selected studies. We also looked at the location of the first author's affiliation to get a sense of the geographical distribution of research on DL-EEG. We found that most contributions came from the USA, China and Australia (see Fig. 41).

3.2. Domains

The selected studies applied DL to EEG in various ways (see Fig. 42). Most studies (86%) focused on using DL for the classification of EEG data, most notably for sleep staging, seizure detection and prediction, brain-computer interfaces (BCIs), as well as for cognitive and affective monitoring. Around 9% of the studies focused instead on the improvement of processing tools, such as learning features from EEG, handling

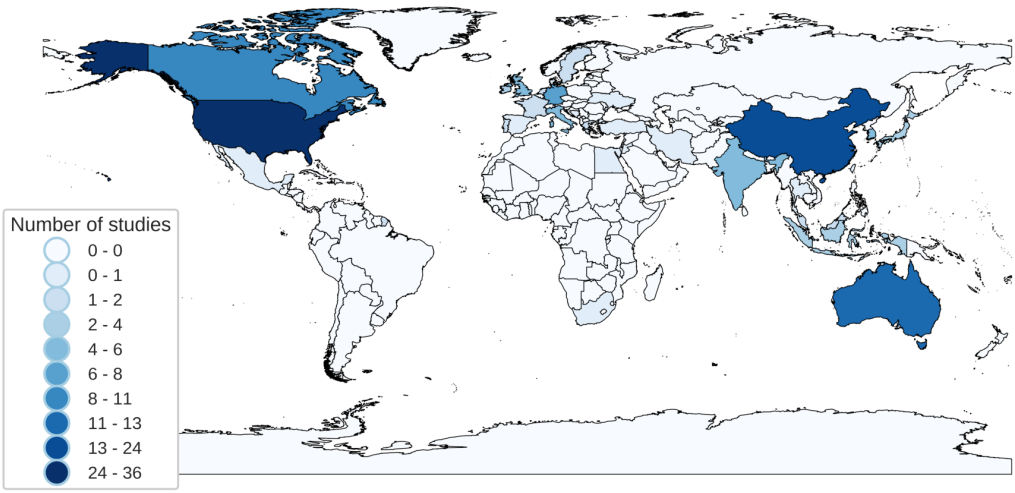


Fig. 41. Countries of first author affiliations.

artifacts, or visualizing trained models. The remaining papers (5%) explored ways of generating data from EEG, e.g. augmenting data, or generating images conditioned on EEG.

Despite the absolute number of DL-EEG publications being relatively small as compared to other DL applications such as computer vision [**88**], there is clearly a growing interest in the field. Fig. 43 shows the growth of the DL-EEG literature since 2010. The first seven months of 2018 alone count more publications than 2010 to 2016 combined, hence the relevance of this review. It is, however, still too early to conclude on trends concerning the application domains, given the relatively small number of publications to date.

3.3. Data

The availability of large datasets containing unprecedented numbers of examples is often mentioned as one of the main enablers of deep learning research in the early 2010s [**53**]. It is thus crucial to understand what the equivalent is in EEG research,

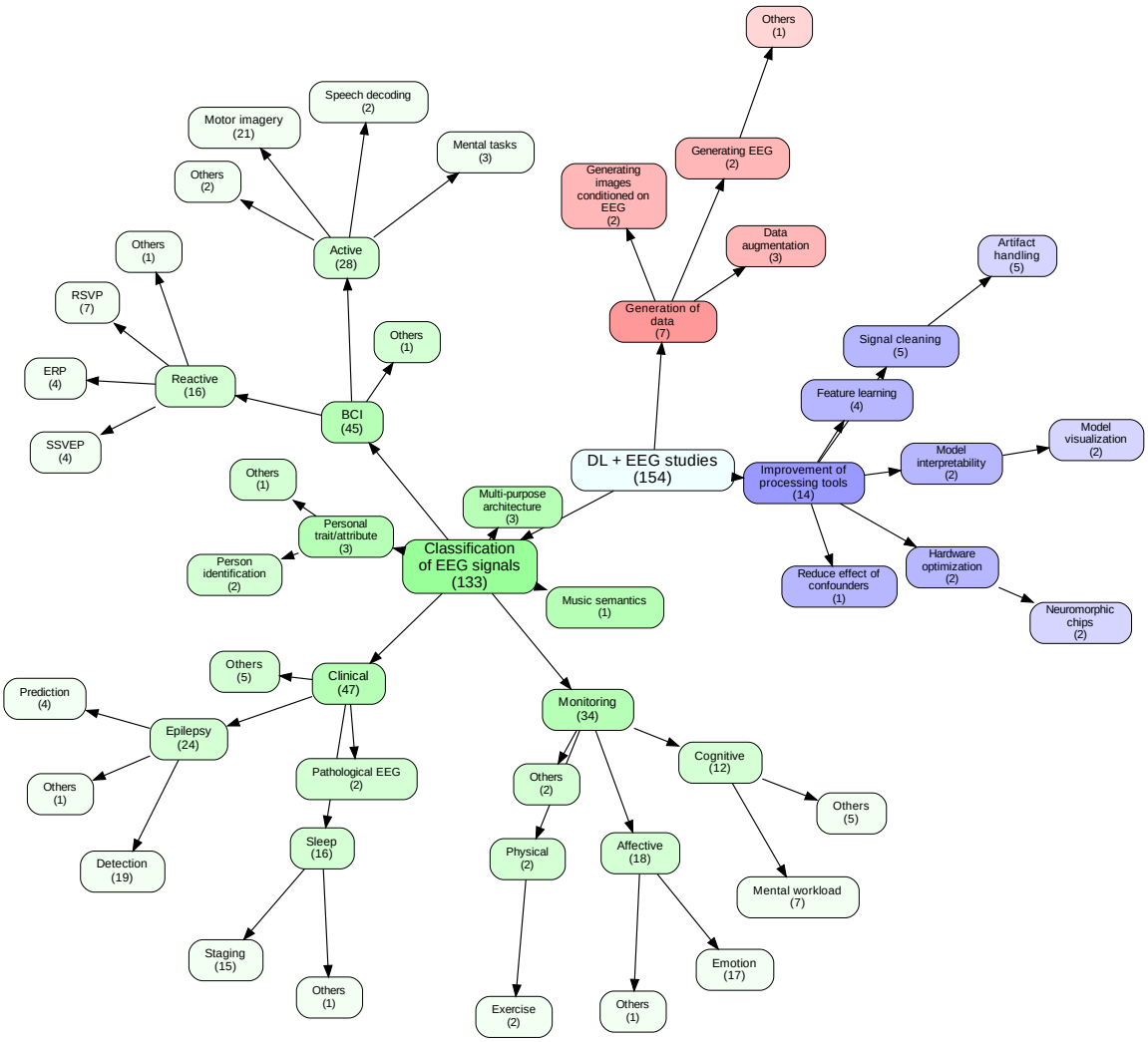


Fig. 42. Focus of the studies. The number of papers that fit in a category is showed in brackets for each category. Studies that covered more than one topic were categorized based on their main focus.

given the relatively high cost of collecting EEG data. Given the high dimensionality of EEG signals [**103**], one would assume that a considerable amount of data is required. Although our analysis cannot answer that question fully, we seek to cover as many dimensions of the answer as possible to give the reader a complete view of what has been done so far.

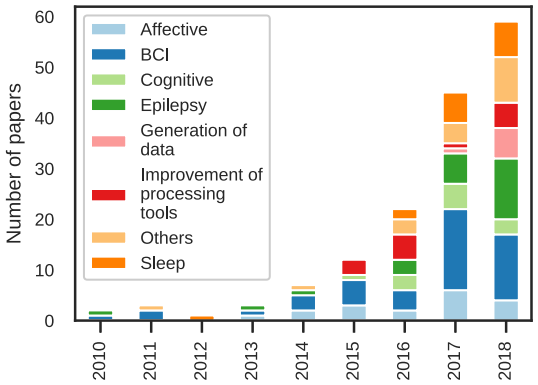


Fig. 43. Number of publications per domain per year. To simplify the figure, some of the categories defined in Fig. 42 have been grouped together.

3.3.1. Quantity of data. We make use of two different measures to report the amount of data used in the reviewed studies: 1) the number of examples available to the deep learning network and 2) the total duration of the EEG recordings used in the study, in minutes. Both measures include the EEG data used across training, validation and test phases. For an in-depth analysis of the amount of data, please see the data items table which contains more detailed information.

The left column of Fig. 44 shows the amount of EEG data, in minutes, used in the analysis of each study, including training, validation and/or testing. Therefore, the time reported here does not necessarily correspond to the total recording time of the experiment(s). For example, many studies recorded a baseline at the beginning and/or at the end but did not use it in their analysis. Moreover, some studies recorded more classes than they used in their analysis. Also, some studies used sub-windows of recorded epochs (e.g. in a motor imagery BCI, using 3 s of a 7 s epoch). The amount of data in minutes used across the studies ranges from 2 up to 4,800,000 (mean = 62,602; median = 360).

The center column of Fig. 44 shows the amount of examples available to the models, either for training, validation or test. This number presents a relevant variability

as some studies used a sliding window with a significant overlap generating many examples (e.g., 250 ms windows with 234 ms overlap, therefore generating 4,050,000 examples from 1080 minutes of EEG data [**153**]), while some other studies used very long windows generating very few examples (e.g., 15-min windows with no overlap, therefore generating 62 examples from 930 minutes of EEG data [**48**]). The wide range of windowing approaches (see Section 3.3.4) indicates that a better understanding of its impact is still required. The number of examples used ranged from 62 up to 9,750,000 (mean = 251,532; median = 14,000).

The right column of Fig. 44 shows the ratio between the amount of data in minutes and the number of examples. This ratio was never mentioned specifically in the papers reviewed but we nonetheless wanted to see if there were any trends or standards across domains and we found that in sleep studies for example, this ratio tends to be of two as most people are using 30 s non-overlapping windows. Brain-computer interfacing is seeing the most sparsity perhaps indicating a lack of best practices for sliding windows. It is important to note that the BCI field is also the one in which the exact relevant time measures were hardest to obtain since most of the recorded data isn't used (e.g. baseline, in-between epochs). Therefore, some of the sparsity on the graph could come from us trying our best to understand and calculate the amount of data used (i.e., seen by the model). Obviously, in the following categories: generation of data, improvement of processing tools and others, this ratio has little to no value as the trends would be difficult to interpret.

The amount of data across different domains varies significantly. In domains like sleep and epilepsy, EEG recordings last many hours (e.g., a full night), but in domains like affective and cognitive monitoring, the data usually comes from lab experiments on the scale of a few hours or even a few minutes.

3.3.2. Subjects. Often correlated with the amount of data, the number of subjects also varies significantly across studies (see Fig. 45). Half of the datasets used in the

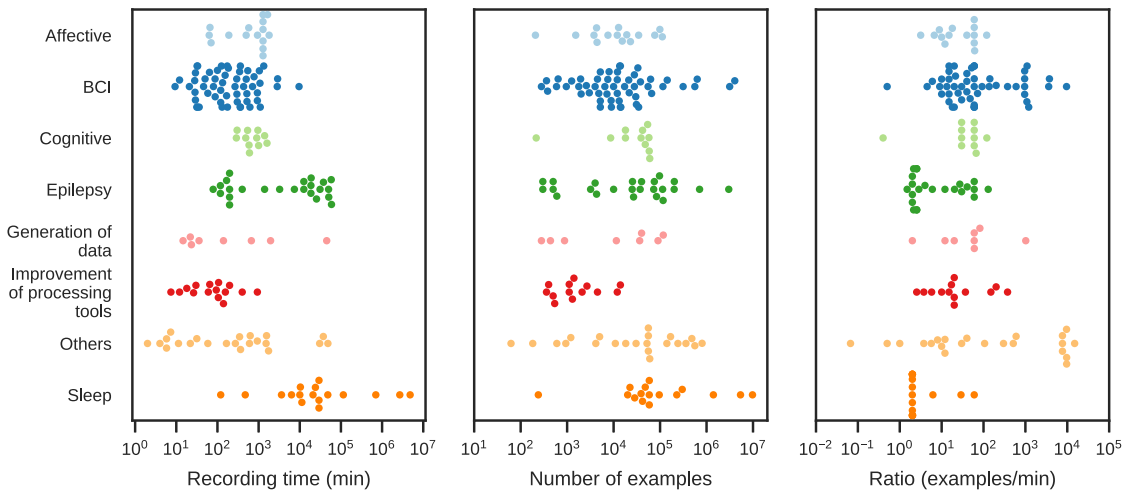


Fig. 44. Amount of data used by the selected studies. Each dot represents one dataset. The left column shows the datasets according to the total length of the EEG recordings used, in minutes. The center column shows the number of examples that were extracted from the available EEG recordings. The right column presents the ratio of number of examples to minutes of EEG recording.

selected studies contained fewer than 13 subjects. Six studies, in particular, used datasets with a much greater number of subjects: [**127, 159, 185, 147**] all used datasets with at least 250 subjects, while [**22**] and [**49**] used datasets with 10,000 and 16,000 subjects, respectively. As explained in Section 3.7.4, the untapped potential of DL-EEG might reside in combining data coming from many different subjects and/or datasets to train a model that captures common underlying features and generalizes better. In [**198**], for example, the authors trained their model using an existing public dataset and also recorded their own EEG data to test the generalization on new subjects. In [**189**], an increase in performance was observed when using more subjects during training before testing on new subjects. The authors tested using from 1 to 30 subjects with a leave-one-subject-out cross-validation scheme, and reported an increase in performance with noticeable diminishing returns above 15 subjects.

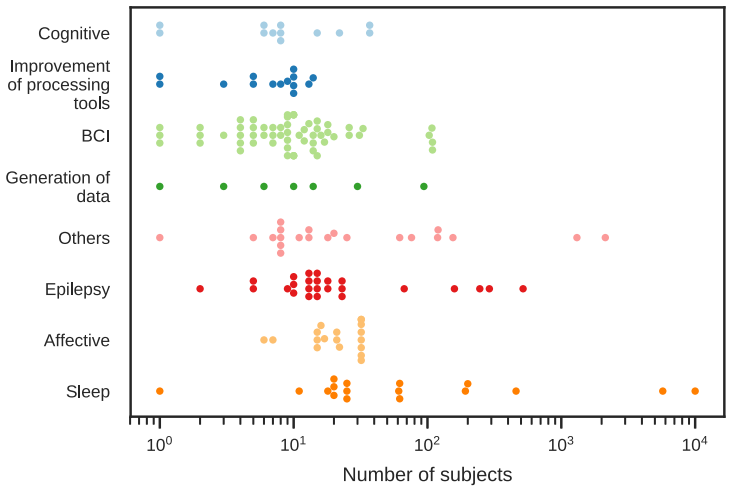


Fig. 45. Number of subjects per domain in datasets. Each point represents one dataset used by one of the selected studies.

3.3.3. Recording parameters. As shown later in Section 3.8, 42% of reported results came from private recordings. We look at the type of EEG device that was used by the selected studies to collect their data, and additionally highlight low-cost, often called "consumer" EEG devices, as compared to traditional "research" or "medical" EEG devices (see Fig. 46). We loosely defined low-cost EEG devices as devices under the USD 1,000 threshold (excluding software, licenses and accessories). Among these devices, the Emotiv EPOC was used the most, followed by the OpenBCI, Muse and Neurosky devices. As for the research grade EEG devices, the BioSemi ActiveTwo was used the most, followed by BrainProducts devices.

The EEG data used in the selected studies was recorded with 1 to 256 electrodes, with half of the studies using between 8 and 62 electrodes (see Fig. 47). The number of electrodes required for a specific task or analysis is usually arbitrarily defined as no fundamental rules have been established. In most cases, adding electrodes will improve possible analyses by increasing spatial resolution. However, adding an electrode close to other electrodes might not provide significantly different information,

while increasing the preparation time and the participant's discomfort and requiring a more costly device. Higher density EEG devices are popular in research but hardly ecological. In [**152**], the authors explored the impact of the number of channels on the specificity and sensitivity for seizure detection. They showed that increasing the number of channels from 4 up to 22 (including two referential channels) resulted in an increase in sensitivity from 31% to 39% and from 40% to 90% in specificity. They concluded, however, that the position of the referential channels is very important as well, making it difficult to compare across datasets coming from different neurologists and recording sites using different locations for the reference(s) channel(s).

Similarly, in [**28**], the impact of different electrode configurations was assessed on a sleep staging task. The authors found that increasing the number of electrodes from two to six produced the highest increase in performance, while adding additional sensors, up to 22 in total, also improved the performance but not as much. The placement of the electrodes in a 2-channel montage also impacted the performance, with central and frontal montages leading to better performance than posterior ones on the sleep staging task.

Furthermore, the recording sampling rates varied mostly between 100 and 1000 Hz in the selected studies. As described in Section 3.4, however, it is common to decrease the EEG sampling rate before further processing - a process called downsampling, by which a signal is resampled to reduce its dimensionality, often by keeping every other N points. Around 50% of studies used sampling rates of 250 Hz or less and the highest sampling rate used was 5000 Hz [**68**].

3.3.4. Data augmentation. Data augmentation is a technique by which new data examples are artificially generated from the existing training data. Data augmentation has proven efficient in other fields such as computer vision, where data manipulations including rotations, translations, cropping and flipping can be applied to generate more training examples [**129**]. Adding more training examples allows the use of

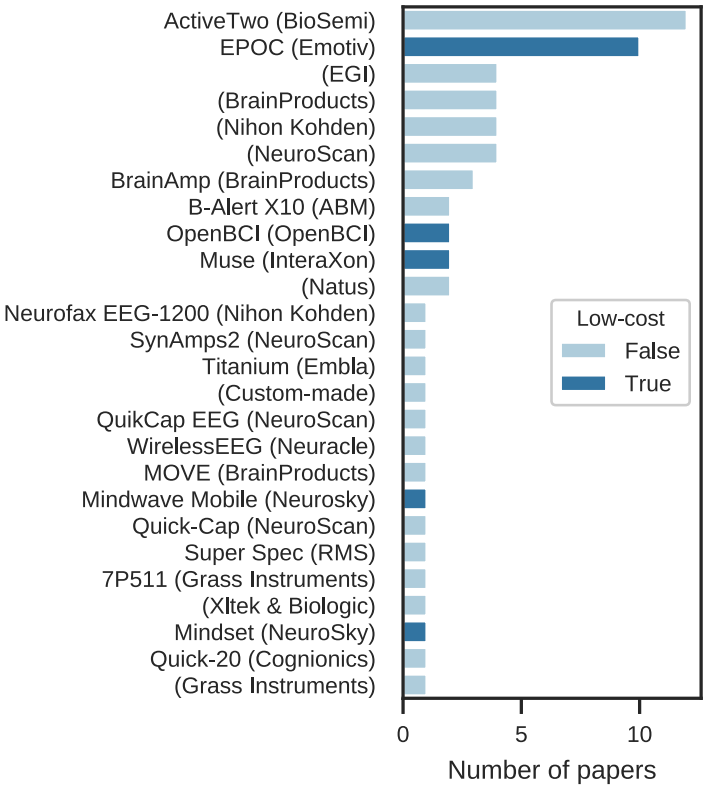


Fig. 46. EEG hardware used in the studies. The device name is followed by the manufacturer's name in parentheses. Low-cost devices (defined as devices below $1,000 excluding software, licenses and accessories) are indicated by a different color.

more complex models comprising more parameters while reducing overfitting. When done properly, data augmentation increases accuracy and stability, offering a better generalization on new data [**211**].

Out of the 154 papers reviewed, three papers explicitly explored the impact of data augmentation on DL-EEG ([**190, 215, 151**]). Interestingly, each one looked at it from the perspective of a different domain: sleep, affective monitoring and BCI. Also, all three are from 2018, perhaps showing an emerging interest in data augmentation. First, in [**190**], Gaussian noise was added to the training data to

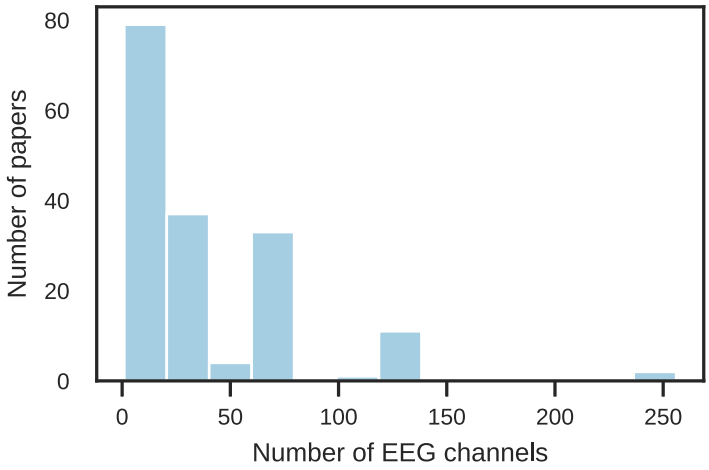


Fig. 47. Distribution of the number of EEG channels.

obtain new examples. This approach was tested on two different public datasets for emotion classification (SEED [**223**] and MAHNOB-HCI [**158**]). They improved their accuracy on the SEED dataset using LeNet ([**90**]) from 49.6% (without augmentation) to 74.3% (with augmentation), from 34.2% (without) to 75.0% (with) using ResNet ([**71**]) and from 40.8% (without) to 45.4% (with) on MAHNOB-HCI dataset using ResNet. Their best accuracy was obtained with a standard deviation of 0.2 and by augmenting the data to 30 times its original size. Despite impressive results, it is important to note that they also compared LeNet and ResNet to an SVM which had an accuracy of 74.2% (without) and 73.4% (with) on the SEED dataset. This might indicate that the initial amount of data was insufficient for LeNet or ResNet but adding data clearly helped bring the performance up to par with the SVM. Second, in [**215**], a conditional deep convolutional generative adversarial network (cDCGAN) was used to generate artificial EEG signals on one of the BCI Competition motor imagery datasets. Using a CNN, it was shown that data augmentation helped improve accuracy from 83% to around 86% to classify motor imagery. In [**151**], the authors explicitly targeted the class imbalance problem of under-represented sleep stages by

generating Fourier transform (FT) surrogates of raw EEG data on the CAPSLPDB dataset. They improved their accuracy up to 24% on some classes.

An additional 30 papers explicitly used data augmentation in one form or another but only a handful investigated the impact it has on performance. In [**83, 13**], noise was added to 2D feature images, although it did not improve results in [**13**]. In [**76**], artifacts such as eye blinks and muscle activity, as well as Gaussian white noise, were used to augment the data and improve robustness. In [**205**] and [**204**], Gaussian noise was added to the input feature vector. This approach increased the accuracy of the SDAE model from around 76.5% (without augmentation) to 85.5% (with).

Multiple studies also used overlapping windows as a way to augment their data, although many did not explicitly frame this as data augmentation. In [**183, 118**], overlapping windows were explicitly used as a data augmentation technique. In [**84**], different shift lengths between overlapping windows (from 10 ms to 60 ms out of a 2-s window) were compared, showing that by generating more training samples with smaller shifts, performance improved significantly. In [**148**], the concept of overlapping windows was pushed even further: 1) redundant computations due to EEG samples being in more than one window were simplified thanks to "cropped training", which ensured these computations were only done once, thereby speeding up training and 2) the fact that overlapping windows share information was used to design an additional term to the cost function, which further regularizes the models by penalizing decisions that are not the same while being close in time.

Other procedures used the inherent spatial and temporal characteristics of EEG to augment their data. In [**34**], the authors doubled their data by swapping the right and left side electrodes, claiming that as the task was a symmetrical problem, which side of the brain expresses the response would not affect classification. In [**16**], the authors augmented their multimodal (EEG and EMG) data by duplicating samples and keeping the values from one modality only, while setting the other modality values to 0 and vice-versa. In [**42**], the authors made use of the data that is usually

thrown away when downsampling EEG in the preprocessing stage. It is common to downsample a signal acquired at higher sampling rate to 256 Hz or less. In their case, they reused the data thrown away during that step as new samples: a downsampling by a factor of N would therefore allow an augmentation of N times.

Finally, classification of rare events where the number of available samples are orders of magnitude smaller than their counterpart classes [**151**] is another motivation for data augmentation. In EEG classification, epileptic seizures or transitional sleep stages (e.g. S1 and S3) often lead to such unbalanced classes. In [**187**], the class imbalance problem was addressed by randomly balancing all classes while sampling for each training epoch. Similarly, in [**28**], balanced accuracy was maximized by using a balanced sampling strategy. In [**181**], EEG segments from the interictal class were split into smaller subgroups of equal size to the preictal class. In [**159**], cost-sensitive learning and oversampling were used to solve the class imbalance problem for sleep staging but the overall performance using these approaches did not improve. In [**139**], the authors randomly replicated subjects from the minority class to balance classes. Similarly, in [**166, 38, 39, 107**], oversampling of the minority class was used to balance classes. Conversely, in [**173, 153**], the majority class was subsampled. In [**179**], an overlapping window with a subject-specific overlap was used to match classes. Similar work by the same group [**178**] showed that when training a GAN on individual subjects, augmenting data with an overlapping window increased accuracy from 60.91% to 74.33%. For more on imbalanced learning, we refer the interested reader to [**154**].

3.4. EEG processing

One of the oft-claimed motivation for using deep learning on EEG processing is automatic feature learning [**127, 76, 45, 69, 110, 11, 209**]. This can be explained by the fact that feature engineering is a time-consuming task [**95**]. Additionally,

preprocessing and cleaning EEG signals from artifacts is a demanding step of the usual EEG processing pipeline. Hence, in this section, we look at aspects related to data preparation, such as preprocessing, artifact handling and feature extraction. This analysis is critical to clarify what level of preprocessing EEG data requires to be successfully used with deep neural networks.

3.4.1. Preprocessing. Preprocessing EEG data usually comprises a few general steps, such as downsampling, band-pass filtering, and windowing. Throughout the process of reviewing papers, we found that a different number of preprocessing steps were employed in the studies. In [**72**], it is mentioned that "a substantial amount of preprocessing was required" for assessing cognitive workload using DL. More specifically, it was necessary to trim the EEG trials, downsample the data to 512 Hz and 64 electrodes, identify and interpolate bad channels, calculate the average reference, remove line noise, and high-pass filter the data starting at 1 Hz. On the other hand, Stober et al. [**162**] applied a single preprocessing step by removing the bad channels for each subject. In studies focusing on emotion recognition using the DEAP dataset [**82**], the same preprocessing methodology proposed by the researchers that collected the dataset was typically used, i.e., re-referencing to the common average, downsampling to 256 Hz, and high-pass filtering at 2 Hz.

We separated the papers into three categories based on whether or not they used preprocessing steps: "Yes", in cases where preprocessing was employed; "No", when the authors explicitly mentioned that no preprocessing was necessary; and not mentioned ("N/M") when no information was provided. The results are shown in Fig. 48.

A considerable proportion of the reviewed articles (72%) employed at least one preprocessing method such as downsampling or re-referencing. This result is not surprising, as applications of DNNs to other domains, such as computer vision, usually require some kind of preprocessing like cropping and normalization as well.

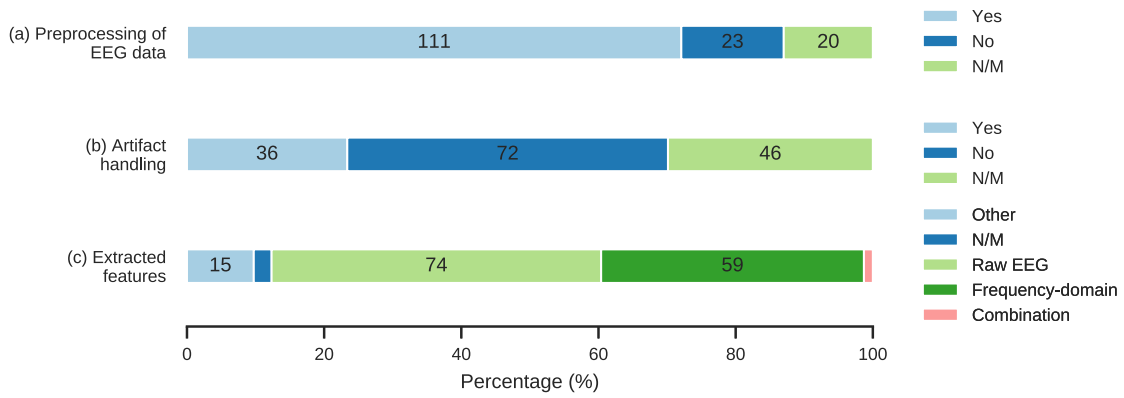


Fig. 48. EEG processing choices. (a) Number of studies that used preprocessing steps, such as filtering, (b) number of studies that included, rejected or corrected artifacts in their data and (c) types of features that were used as input to the proposed models.

3.4.2. Artifact handling. Artifact handling techniques are used to remove specific types of noise, such as ocular and muscular artifacts [**184**]. As emphasized in [**199**], removal of artifacts may be crucial for achieving good EEG decoding performance. Adding this to the fact that cleaning EEG signals might be a time-consuming process, some studies attempted to apply only minimal preprocessing such as removing bad channels and leave the burden of learning from a potentially noisy signal on the neural network [**162**]. With that in mind, we decided to look at artifact handling separately.

Artifact removal techniques usually require the intervention of a human expert [**115**]. Different techniques leverage human knowledge to different extents, and might fully rely on an expert, as in the case of visual inspection, or require prior knowledge to simply tune a hyperparameter, as in the case of wavelet-enhanced independent component analysis (wICA) [**26**]. Among the studies which handled artifacts, a myriad of techniques were applied. Some studies employed methods which rely on human knowledge such as amplitude thresholding [**110**], manual identification of high-variance

segments [**72**], and handling EEG blinking-related noise based on high-amplitude EOG segments [**107**]. Moreover, in [**165, 203, 204, 45, 126, 128**], independent component analysis (ICA) was used to separate ocular components from EEG data [**106**].

In order to investigate the necessity of removing artifacts from EEG when using deep neural networks, we split the selected papers into three categories, in a similar way to the preprocessing analysis (see Fig. 48). Almost half the papers (47%) did not use artifact handling methods, while 23% did. Additionally, 30% of the studies did not mention whether artifact handling was necessary to achieve their results. Given those results, we are encouraged to believe that using DNNs on EEG might be a way to avoid the explicit artifact removal step of the classical EEG processing pipeline without harming task performance.

3.4.3. Features. Feature engineering is one of the most demanding steps of the traditional EEG processing pipeline [**95**] and the main goal of many papers considered in this review [**127, 76, 45, 69, 110, 11, 209**] is to get rid of this step by employing deep neural networks for automatic feature learning. This aspect appears to be of interest to researchers in the field since its early stages, as indicated by the work of Wulsin et al. [**195**], which, in 2011, compared the performance of deep belief networks (DBNs) on classification and anomaly detection tasks using both raw EEG and features as inputs. More recently, studies such as [**163, 67**] achieved promising results without the need to extract features.

On the other hand, a considerable proportion of the reviewed papers used hand-engineered features as the input to their deep neural networks. In [**172**], for example, authors used a time-frequency domain representation of EEG obtained via the short-time Fourier transform (STFT) for detecting binary user-preference (*like* versus *dislike*). Similarly, Truong et al. [**179**], used the STFT as a 2-dimensional EEG representation for seizure prediction using CNNs. In [**214**], EEG frequency-domain information

was also used. Widely adopted by the EEG community, the power spectral density (PSD) of classical frequency bands from around 1 Hz to 40 Hz were used as features. Specifically, authors selected the delta (1-4 Hz), theta (5-8 Hz), alpha (9-13 Hz), lower beta (14-16 Hz), higher beta (17-30 Hz), and gamma (31-40 Hz) bands for mental workload state recognition. Moreover, other studies employed a combination of features, for instance [**48**], which used PSD features, as well as entropy, kurtosis, fractal component, among others, as input of the proposed CNN for ischemic stroke detection.

Given that the majority of EEG features are obtained in the frequency-domain, our analysis consisted in separating the reviewed articles into four categories according to the respective input type. Namely, the categories were: “Raw EEG” (which includes EEG time series that have been preprocessed, e.g., filtered or artifact-corrected), “Frequency-domain”, “Combination” (in case more than one type of feature was used), and “Other” (for papers using neither raw EEG nor frequency-domain features). Studies that did not specify the type of input were assigned to the category “N/M” (not mentioned). Notice that, here, we use “feature” and “input type” interchangeably.

Fig. 48 presents the result of our analysis. One can observe that 49% of the papers used only raw EEG data as input, whereas 49% used hand-engineered features, from which 38% corresponded to frequency domain-derived features. Finally, 2% did not specify the type of input of their model. According to these results, we find indications that DNNs can be in fact applied to raw EEG data and achieve state-of-the-art results.

3.5. Deep learning methodology

3.5.1. Architecture. A crucial choice in the DL-based EEG processing pipeline is the neural network architecture to be used. In this section, we aim at answering a few questions on this topic, namely: 1) "What are the most frequently used architectures?",

2) "How has this changed across years?", 3) "Is the choice of architecture related to input characteristics?" and 4) "How deep are the networks used in DL-EEG?".

To answer the first three questions, we divided and assigned the architectures used in the 154 papers into the following groups: CNNs, RNNs, AEs, restricted Boltzmann machines (RBMs), DBNs, GANs, FC networks, combinations of CNNs and RNNs (CNN+RNN), and "Others" for any other architecture or combination not included in the aforementioned categories. Fig. 49(a) shows the percentage of studies that used the different architectures. 40% of the papers used CNNs, whereas RNNs and AEs were both the architecture choice of about 13% of the works, respectively. Combinations of CNNs and RNNs, on the other hand, were used in 7% of the studies. RBMs and DBNs corresponded together to almost 10% of the architectures. FC neural networks were employed by 6% of the papers. GANs and other architectures appeared in 6% of the considered cases. Notice that 4% of the analyzed papers did not report their choice of architecture.

In Fig. 49, we provide a visualization of the distribution of architecture types across years. Until the end of 2014, DBNs and FC networks comprised the majority of the studies. However, since 2015, CNNs have been the architecture type of choice in most studies. This can be attributed to the their capabilities of end-to-end learning and of exploiting hierarchical structure on the data [**175**], as well as their success and subsequent popularity on computer vision tasks, such as the ILSVRC 2012 challenge [**35**]. Interestingly, we also observe that as the number of papers grows, the proportion of studies using CNNs and combinations of recurrent and convolutional layers has been growing steadily. The latter shows that RNNs are increasingly of interest for EEG analysis. On the other hand, the use of architectures such as RBMs, DBNs and AEs has been decreasing with time. Commonly, models employing these architectures utilize a two-step training procedure consisting of 1) unsupervised feature learning and 2) training a classifier on top of the learned features. However, we notice that recent studies leverage the hierarchical feature learning capabilities of CNNs to achieve

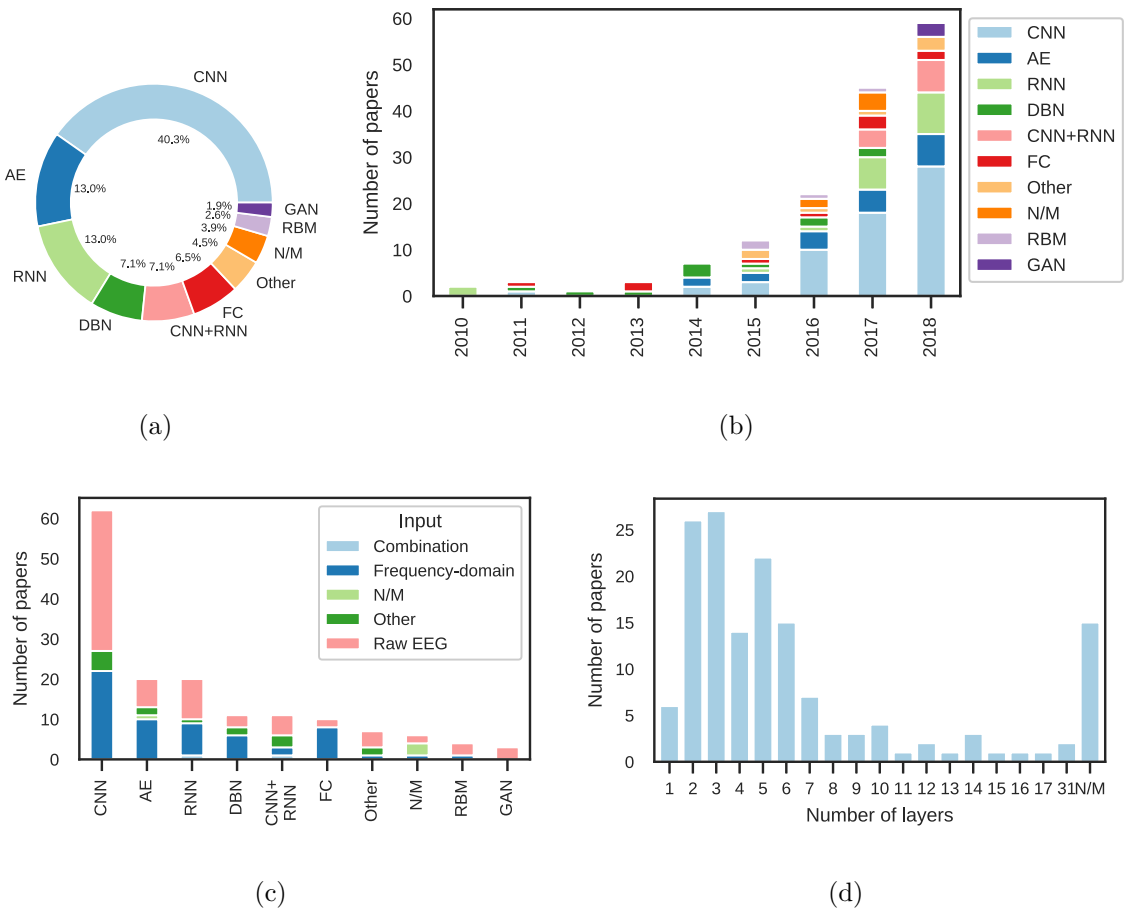


Fig. 49. Deep learning architectures used in the selected studies. 'N/M' stands for 'Not mentioned' and accounts for papers which have not reported the respective deep learning methodology aspect under analysis. (a) Architectures. (b) Distribution of architectures across years. (c) Distribution of input type according to the architecture category. (d) Distribution of number of neural network layers.

end-to-end supervised feature learning, i.e., training both a feature extractor and a classifier simultaneously.

To complement the previous result, we cross-checked the architecture and input type information provided in Fig. 48. Results are presented in Fig. 49 and clearly show

that CNNs are indeed used more often with raw EEG data as input. This corroborates the idea that researchers employ this architecture with the aim of leveraging the capabilities of deep neural networks to process EEG data in an end-to-end fashion, avoiding the time-consuming task of extracting features. From this figure, one can also notice that some architectures such as deep belief networks are typically used with frequency-domain features as inputs, while GANs, on the other hand, have been only applied to EEG processing using raw data.

Number of layers. Deep neural networks are usually composed of stacks of layers which provide hierarchical processing. Although one might think the use of *deep* neural networks implies the existence of a large number of layers in the architecture, there is no absolute consensus in the literature regarding this definition. Here we investigate this aspect and show that the number of layers is not necessarily large, i.e., larger than three, in many of the considered studies.

In Fig 49, we show the distribution of the reviewed papers according to the number of layers in the respective architecture. For studies reporting results for different architectures and number of layers, we only considered the highest value. We observed that most of the selected studies (128) utilized architectures with at most 10 layers. A total of 16 articles have not reported the architecture depth. When comparing the distribution of papers according to the architecture depth with architectures commonly used for computer vision applications, such as VGG-16 (16 layers) [**156**] and ResNet-18 (18 layers) [**71**], we observe that the current literature on DL-EEG suggests shallower models achieve better performance. The same trend is applicable to other domains such as NLP and speech processing. In unsupervised language modeling, for instance, the GPT-2 model [**134**] outperformed previous work[5] using architectures with 12 to 48 layers. Likewise, in automatic speech recognition, the state-of-the-art model[6] on

[5] `https://paperswithcode.com/sota/word-level-models-penn-treebank`

[6] `https://paperswithcode.com/sota/speech-recognition-word-error-rate-on-sw`

human-to-human communication was achieved with a 30-layer architecture containing residual blocks and recurrent layers [**145**].

Some studies specifically investigated the effect of increasing the model depth. Zhang et al. [**214**] evaluated the performance of models with depth ranging from two to 10 on a mental workload classification task. Architectures with seven layers outperformed both shallower (two and four layers) and deeper (10 layers) models in terms of accuracy, precision, F-measure and G-mean. Moreover, O'Shea et al. [**118**] compared the performance of a CNN with six and 11 layers on neonatal seizure detection. Their results show that, in this case, the deeper network presented better area under the receiver operating curve (ROC AUC) (0.971) in comparison to the shallower model, as well as a support vector machine (SVM) (0.965). In [**84**], the effect of depth on CNN performance was also studied. The authors compared results obtained by a CNN with two and three convolutional layers on the task of classifying SSVEPs under ambulatory conditions. The shallower architecture outperformed the three-layer one in all scenarios considering different amounts of training data. Canonical correlation analysis (CCA) together with a KNN classifier were also evaluated and employed as a baseline method. Interestingly, as the number of training samples increased, the shallower model outperformed the CCA-based baseline. These three examples offer a representative view of the current state of DL-EEG research, namely that it is impossible to conclude that either deeper or shallower models perform better in all contexts. Depending on factors such as the amount of data, the task to be solved, the type of architecture, the hyperparameter tuning strategy, and the available computational resources, shallower or deeper models might work best. To gain a better idea of what might be preferable to use in a specific case, we invite the reader to identify relevant studies in the data items table and explore their results.

EEG-specific design choices. Particular choices regarding the architecture might enable a model to mimic the process of extracting EEG features. An architecture can also be specifically designed to impose specific properties on the learned representations. This

is for instance the case with max-pooling, which is used to produce invariant feature maps to slight translations on the input [**53**]. In the case of EEG signals, one might be interested in forcing the model to process temporal and spatial information separately in the earlier stages of the network. In [**28, 84, 209, 14, 148, 107**], one-dimensional convolutions were used in the input layer with the aim of processing either temporal or spatial information independently at this point of the hierarchy. Other studies [**217, 166**] combined recurrent and convolutional neural networks as an alternative to the previous approach of separating temporal and spatial content. Recurrent models were also applied in cases where it was necessary to capture long-term dependencies from the EEG data [**98, 216**].

3.5.2. Training. Details regarding the training of the models proposed in the literature are of great importance as different approaches and hyperparameter choices can greatly impact the performance of neural networks. The use of pre-trained models, regularization, and hyperparameter search strategies are examples of aspects we took into account during the review process. We report our main findings in this section. Training Procedure. One of the advantages of applying deep neural networks to EEG processing is the possibility of simultaneously training a feature extractor and a model for executing a downstream task such as classification or regression. However, in some of the reviewed studies [**86, 192, 111**], these two tasks were executed separately. Usually, the feature learning was done in an unsupervised fashion, with RBMs, DBNs, or AEs. After training those models to provide an appropriate representation of the EEG input signal, the new features were then used as the input for a target task which is, in general, classification. In other cases, pre-trained models were used for a different purpose, such as object recognition, and were fine-tuned on the specific EEG task with the aim of providing a better initialization or regularization effect [**96**].

In order to investigate the training procedure of the reviewed papers, we classify each one according to the adopted training procedure. Models which have parameters

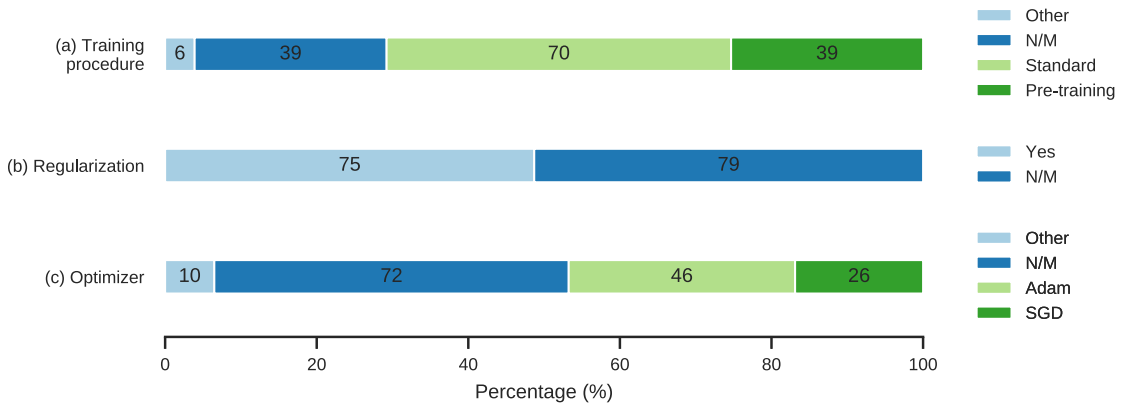


Fig. 50. Deep learning methodology choices. (a) Training methodology used in the studies, (b) number of studies that reported the use of regularization methods such as dropout, weight decay, etc. and (c) type of optimizer used in the studies.

learned without using any kind of pre-training were assigned to the "Standard" group. The remaining studies, which specified the training procedure, were included in the "Pre-training" class, in case the parameters were learned in more than one step. Finally, papers employing different methodologies for training, such as co-learning [**34**], were included in the "Other" group.

In Fig. 50a) we show how the reviewed papers are distributed according to the training procedure. "N/M" refers to studies which have not reported this aspect. Almost half the papers did not employ any pre-traning strategy, while 25% did. Even though the training strategy is crucial for achieving good performance with deep neural networks, 25% of the selected studies have not explicitly described it in their paper.

Regularization. In the context of our literature review, we define regularization as any constraint on the set of possible functions parametrized by the neural network intended to improve its performance on unseen data during training [**53**]. The main goal when regularizing a neural network is to control its complexity in order to obtain better

generalization performance [**21**], which can be verified by a decrease on test error in the case of classification problems. There are several ways of regularizing neural networks, and among the most common are weight decay (L2 and L1 regularization) [**53**], early stopping [**133**], dropout [**167**], and label smoothing [**168**]. Notice that even though the use of pre-trained models as initialization can also be interpreted as a regularizer [**96**], in this work we decided to include it in the training procedure analysis instead.

As the use of regularization might be fundamental to guarantee a good performance on unseen data during training, we analyzed how many of the reviewed studies explicitly stated that they have employed it in their models. Papers were separated in two groups, namely: "Yes" in case any kind of regularization was used, and "N/M" otherwise. In Fig. 50 we present the proportion of studies in each group.

From Fig. 50, one can notice that more than half the studies employed at least one regularization method. Furthermore, regularization methods were frequently combined in the reviewed studies. Hefron et al. [**72**] employed a combination of dropout, L1- and L2- regularization to learn temporal and frequency representations across different participants. The developed modelwas trained for recognizing mental workload states elicited by the MATB task [**31**]. Similarly, Längkvist and Loutfi [**86**], combined two types of regularization with the aim of developing a model tailored to an automatic sleep stage classification task. Besides L2-regularization, they added a penalty term to encourage weight sparsity, defined as the KL-divergence between the mean activation of each hidden unit over all training examples in a training batch and a hyperparameter ρ.

Optimization. Learning the parameters of a deep neural network is, in practice, an optimization problem. The best way to tackle it is still an open research question in the deep learning literature, as there is often a compromise between finding a good solution in terms of minimizing the cost function and the performance of a local optimum expressed by the generalization gap, i.e. the difference between the training

error and the true error estimated on the test set. In this scenario, the choice of a parameter update rule, i.e. the *learning algorithm* or *optimizer*, might be key for achieving good results.

The most commonly used optimizers are reported in Fig. 50. One surprising finding is that even though the choice of optimizer is a fundamental aspect of the DL-EEG pipeline, 47% of the considered studies did not report which parameter update rule was applied. Moreover, 30% used Adam [**81**] and 17% Stochastic Gradient Descent [**135**] (notice that we also refer to the mini-batch case as SGD). 6% of the papers utilized different optimizers, such as RMSprop [**176**], Adagrad [**40**], and Adadelta [**210**].

Another interesting finding the optimizer analysis provided is the steady increase in the use of Adam. Indeed, from 2017 to 2018, the percentage of studies using Adam increased from 28.9% to 54.2%. Adam was proposed as a gradient-based method with the capability of adaptively tuning the learning rate based on estimates of first and second order moments of the gradient. It became very popular in general deep neural networks applications (accumulating approximately 15,000 citations since 2014[7]). Interestingly, we notice a proportional decrease from 2017 to 2018 of the number of papers which did not report the optimizer utilized.

Hyperparameter search. From a practical point-of-view, tuning the hyperparameters of a learning algorithm often takes up a great part of the time spent during training. GANs, for instance, are known to be sensitive to the choices of optimizer and architecture hyperparameters [**57, 97**]. In order to minimize the amount of time spent finding an appropriate set of hyperparameters, several methods have been proposed in the literature. Examples of commonly applied methods are grid search [**17**] and Bayesian optimization [**157**]. Grid search consists in determining a range of values for each parameter to be tuned, choosing values in this range, and evaluating the model,

[7]Google scholar query run on 30/11/2018.

usually in a validation set considering all combinations. One of the advantages of grid search is that it is highly parallelizable, as each set of hyperparameter is independent of the other. Bayesian optimization, in turn, defines a posterior distribution over the hyperparameters space and iteratively updates its values according to the performance obtained by the model with a hyperparameter set corresponding to the expected posterior.

Given the importance of finding a good set of hyperparameters and the difficulty of achieving this in general, we calculate the percentage of papers that employed some search method for tuning their models and optimizers, as well as the amount of articles that have not included any information regarding this aspect. Results indicate that 80% of the reviewed papers have not mentioned the use of hyperparameters search strategies. It is important to highlight that among those articles, it is not clear how many have not done any tuning at all and how many have just not considered to include this information in the paper. From the 20% that declared to have searched for an appropriate set of hyperparameters, some have manually done this by trial and error (e.g. [**2, 38, 181, 127**]), while others employed grid search (e.g. [**203, 197, 39, 204, 99, 11, 86**]), and a few used other strategies such as Bayesian methods (e.g. [**161, 162, 151**]).

3.6. Inspection of trained models

In this section, we review if, and how, studies have inspected their proposed models. Out of the selected studies, 27% reported inspecting their models. Two studies focused more specifically on the question of model inspection in the context of DL and EEG [**68, 45**]. See Table 7 for a list of the different techniques that were used by more than one study. For a general review of DL model inspection techniques, see [**75**].

The most frequent model inspection techniques involved the analysis of the trained model's weights [**130, 207, 86, 34, 87, 197, 180, 117, 169, 224, 162, 107, 200**].

This often requires focusing on the weights of the first layer only, as their interpretation in regard to the input data is straightforward. Indeed, the absolute value of a weight represents the strength with which the corresponding input dimension is used by the model - a higher value can therefore be interpreted as a rough measure of feature importance. For deeper layers, however, the hierarchical nature of neural networks means it is much harder to understand what a weight is applied to.

The analysis of model activations was used in multiple studies [**208, 191, 87, 84, 204, 166, 153, 107**]. This kind of inspection method usually involves visualizing the activations of the trained model over multiple examples, and thus inferring how different parts of the network react to known inputs. The input-perturbation network-prediction correlation map technique, introduced in [**147**], pushes this idea further by trying to identify causal relationships between the inputs and the decisions of a model. The impact of the perturbation on the activations of the last layer's units then shines light onto which characteristics of the input are important for the classifier to make a correct prediction. To do this, the input is first perturbed, either in the time- or frequency-domain, to alter its amplitude or phase characteristics [**68**], and then fed into the network. Occlusion sensitivity techniques [**92, 28, 173**] use a similar idea, by which the decisions of the network when different parts of the input are occluded are analyzed.

Several studies used backpropagation-based techniques to generate input maps that maximize activations of specific units [**185, 139, 159, 13**]. These maps can then be used to infer the role of specific neurons, or the kind of input they are sensitive to.

Finally, some model inspection techniques were used in a single study. For instance, in [**45**], the class activation map (CAM) technique was extended to overcome its limitations on EEG data. To use CAMs in a CNN, the channel activations of the last convolutional layer must be averaged spatially before being fed into the model's penultimate layer, which is a FC layer. For a specific input image, a map can then be created to highlight parts of the image that contributed the most to the decision,

Table 7. Model inspection techniques used by more than one study.

	Citation
Analysis of weights	[**130, 207, 86, 34, 87, 197, 180, 117, 169, 224, 162, 107, 200, 85, 27**]
Analysis of activations	[**208, 191, 87, 84, 204, 166, 153, 107**]
Input-perturbation network-prediction correlation maps	[**147, 189, 68, 14, 148**]
Generating input to maximize activation	[**185, 139, 159, 13**]
Occlusion of input	[**92, 28, 173**]

by computing a weighted average of the last convolutional layer's channel activations. Other techniques include Deeplift [**87**], saliency maps [**187**], input-feature unit-output correlation maps [**148**], retrieval of closest examples [**34**], analysis of performance with transferred layers [**62**], analysis of most-activating input windows [**68**], analysis of generated outputs [**67**], and ablation of filters [**87**].

3.7. Reporting of results

The performance of DL methods on EEG is of great interest as it is still not clear whether DL can outperform traditional EEG processing pipelines [**103**]. Thus, a major question we thus aim to answer in this review is: "Does DL lead to better performance than traditional methods on EEG?" However, answering this question is not straightforward, as benchmark datasets, baseline models, performance metrics and reporting methodology all vary considerably between the studies. In contrast, other application domains of DL, such as computer vision and NLP, benefit from standardized datasets and reporting methodology [**53**].

Therefore, to provide as satisfying an answer as possible, we adopt a two-pronged approach. First, we review how the studies reported their results by focusing on

directly quantifiable items: 1) the type of baseline used as a comparison in each study, 2) the performance metrics, 3) the validation procedure, and 4) the use of statistical testing. Second, based on these points and focusing on studies that reported accuracy comparisons with baseline models, we analyze the reported performance of a majority of the reviewed studies.

3.7.1. Type of baseline. When contributing a new model, architecture or methodology to solve an already existing problem, it is necessary to compare the performance of the new model to the performance of state-of-the-art models commonly used for the problem of interest. Indeed, without a baseline comparison, it is not possible to assess whether the proposed method provides any advantage over the current state-of-the-art.

Points of comparison are typically obtained in two different ways: 1) (re)implementing standard models or 2) referring to published models. In the first case, authors will implement their own baseline models, usually using simpler models, and evaluate their performance on the same task and in the same conditions. Such comparisons are informative, but often do not reflect the actual state-of-the-art on a specific task. In the second case, authors will instead cite previous literature that reported results on the same task and/or dataset. This second option is not always possible, especially when working on private datasets or tasks that have not been explored much in the past.

In the case of typical EEG classification tasks, state-of-the-art approaches usually involve traditional processing pipelines that include feature extraction and shallow/classical machine learning models. With that in mind, 68.2% of the studies selected included at least one traditional processing pipeline as a baseline model (see Fig. 55). Some studies instead (or also) compared their performance to DL-based approaches, to highlight incremental improvements obtained by using different architectures or training methodology: 34.4% of the studies therefore included at least one DL-based model as a baseline model. Out of the studies that did not compare their

models to a baseline, six did not focus on the classification of EEG. Therefore, in total, 20.8% of the studies did not report baseline comparisons, making it impossible to assess the added value of their proposed methods in terms of performance.

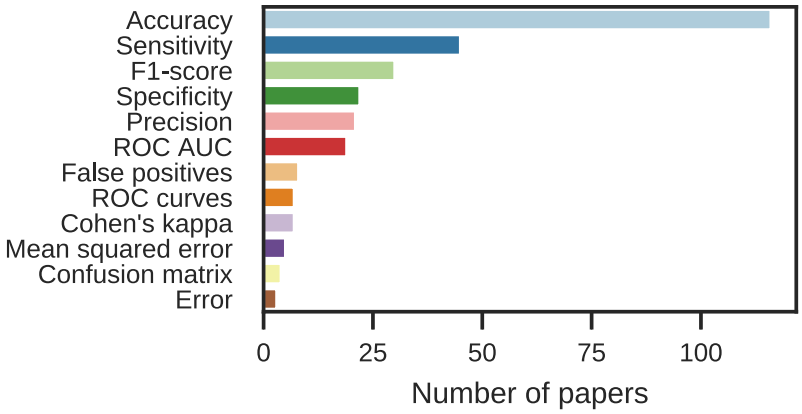


Fig. 51. Type of performance metrics used in the selected studies. Only metrics that appeared in at least three different studies are included in this figure.

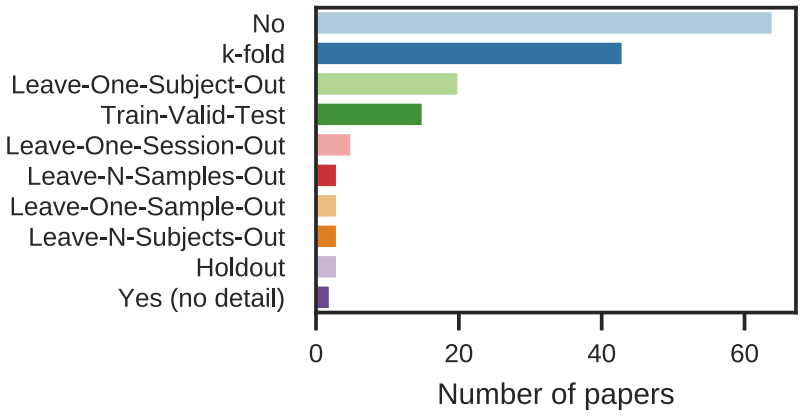


Fig. 52. Cross-validation approaches.

3.7.2. Performance metrics. The types of performance metrics used by studies focusing on EEG classification are shown in Fig. 51. Unsurprisingly, most studies used metrics derived from confusion matrices, such as accuracy, sensitivity, f1-score, ROC AUC and precision. As highlighted in [**28, 197**], it is often preferable to use metrics that are robust to class imbalance, such as balanced accuracy, f1-score, and the ROC AUC for binary problems. This is often the case in sleep or epilepsy recordings, where clinical events are rare.

Studies that did not focus on the classification of EEG signals also mainly used accuracy as a metric. Indeed, these studies generally used a classification task to evaluate model performance, although their main purpose was different (e.g., correcting artifacts). In other cases, performance metrics specific to the study's purpose, such as generating data, were used, e.g., the inception score ([**144**]), the Fréchet inception distance ([**74**]), as well as custom metrics.

3.7.3. Validation procedure. When evaluating a machine learning model, it is important to measure its generalization performance, i.e., how well it performs on unseen data. In order to do this, it is common practice to divide the available data into a training and a test sets. When hyperparameters need to be tuned, the performance on the test set cannot be used anymore as an unbiased evaluation of the generalization performance of the model. Therefore, the training set is divided to obtain a third set called a "validation set" which is used to select the best hyperparameter configuration, leaving the test set to evaluate the performance of the best model in an unbiased way. However, when the amount of data available is small, dividing the data into different sets and only using a subset for training can seriously undermine the performance of data-hungry models. A procedure known as "cross-validation" is used in these cases, where the data is broken down into different partitions, which will then successively be used as either training or validation data.

The cross-validation techniques used in the selected studies are shown in Fig. 52. Some studies mentioned using cross-validation but did not provide any details. The category 'Train-Valid-Test' includes studies doing random permutations of train/valid, train/test or train/valid/test, as well as studies that mentioned splitting their data into training, validation and test sets but did not provide any details on the validation method. The Leave-One-Out variations correspond to the special case where $N = 1$ in the Leave-N-Out versions. 42% of the studies did not use any form of cross-validation. Interestingly, in [**102**], the authors proposed a 'warm restart' technique to improve performance and/or generalization of stochastic gradient descent and to relax the need to access a validation set by providing a recommendation solution as the latest solution of the latest completed cycle/restart.

3.7.4. Subject handling. Whether a study focuses on intra- or inter-subject classification has an impact on the performance. Intra-subject models, which are trained and used on the data of a single subject, often lead to higher performance since the model has less data variability to account for. However, this means the data the model is trained on is obtained from a single subject, and thus often comprises only a few recordings. In inter-subject studies, models generally see more data, as multiple subjects are included, but must contend with greater data variability, which introduces different challenges.

In the case of inter-subject classification, the choice of the validation procedure can have a big impact on the reported performance of a model. The Leave-N-Subject-Out procedure, which uses different subjects for training and for testing, may lead to lower performance, but is applicable to real-life scenarios where a model must be used on a subject for whom no training data is available. In contrast, using k-fold cross-validation on the combined data from all the subjects often means that the same subjects are seen in both the training and testing sets. In the selected studies, 23 out

of the 108 studies using an inter-subject approach used a Leave-N-Subjects-Out or Leave-One-Subjects-Out procedure.

In the selected studies, 26% focused only on intra-subject classification, 62% focused only on inter-subject classification, 8% focused on both, and 4% did not mention it. Obviously, 'N/M' studies necessarily fall under one of the three previous categories. The 'N/M' might be due to certain domains using a specific type of experiment (i.e. intra or inter-subject) almost exclusively, thereby obviating the need to mention it explicitly.

Fig. 53 shows that there has been a clear trend over the last few years to leverage DL for inter-subject rather than intra-subject analysis. In [**34**], the authors used a large dataset and tested the performance of their model both on new (unseen) subjects and on known (seen) subjects. They obtained 38% accuracy on unseen subjects and 75% on seen subjects, showing that classifying EEG data from unseen subjects can be significantly more challenging than from seen ones.

In [**182**], the authors compared their model on both intra- and inter-subject tasks. Despite the former case providing the model with less less training data than the latter, it led to better results. In [**61**], the authors compared different DL models and showed that cross-subject (37 subjects) models always performed worse than within-subject models. In [**125**], a hybrid system trained on multiple subjects and then fine-tuned on subject-specific data led to the best performance. Finally, in [**173**], the authors compared their DNN to a state-of-the-art traditional approach and showed that deep networks generalize better, although their performance on intra-subject classification is still higher than on inter-subject classification.

3.7.5. Statistical testing. To assess whether a proposed model is actually better than a baseline model, it is useful to use statistical tests. In total, 19.5% of the selected studies used statistical tests to compare the performance of their models to

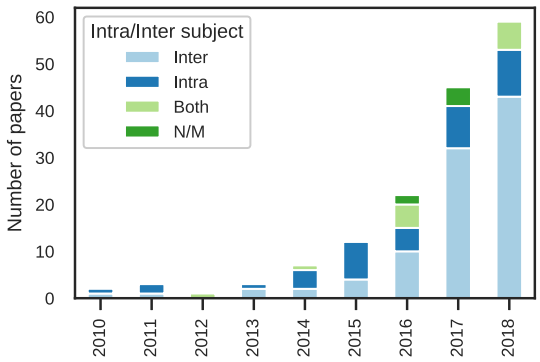


Fig. 53. Distribution of intra- vs. inter-subject studies per year.

baseline models. The tests most often used were Wilcoxon signed-rank tests, followed by ANOVAs.

3.7.6. Comparison of results. Although, as explained above, many factors make this kind of comparison imprecise, we show in this section how the proposed approaches and traditional baseline models compared, as reported by the selected studies.

We focus on a specific subset of the studies to make the comparison more meaningful. First, we focus on studies that report accuracy as a direct measure of task performance. As shown in Fig. 51, this includes the vast majority of the studies. Second, we only report studies which compared their models to a traditional baseline, as we are interested in whether DL leads to better results than non-DL approaches. This means studies which only compared their results to other DL approaches are not included in this comparison. Third, some studies evaluated their approach on more than one task or dataset. In this case, we report the results on the task that has the most associated baselines. If that is more than one, we either report all tasks, or aggregate them if they are very similar (e.g., binary classification of multiple mental tasks, where performance is reported for each possible pair of tasks). In the case of multimodal studies, we only report the performance on the EEG-only task, if it is available. Finally, when reporting accuracy differences, we focus on the difference between the best proposed

model and the best baseline model, per task. Following these constraints, a total of 102 studies/tasks were left for our analysis.

Figure 54 shows the difference in accuracy between each proposed model and corresponding baseline per domain type (as categorized in Fig. 42), as well as the corresponding distribution over all included studies and tasks.

The median gain in accuracy with DL is of 5.4%, with an interquartile range of 9.4%. Only four values were negative values, meaning the proposed DL approach led to a lower performance than the baseline. We notice a slight, although not significant, difference in the median accuracy difference of the preprint and peer-reviewed groups (4.7% and 6.00%), respectively; Mann-Whitney test $p = 0.072$). While this difference is minor and exemplifies the same trend of slightly higher performance of DL models over traditional methods, it might originate from the lower publication standards of non-peer-reviewed research.

The highest improvement in accuracy (76.7%), obtained in [**160**], was shown to be caused by flawed experimental design and preprocessing strategy in a replication study [**94**]. Therefore, the improvement obtained in [**205**] (35.3% on a mental workload level classification task) was the highest achieved in the articles reviewed. In that study, a naive Bayes classifier trained on various features (including spectral and information theoretic features) preceded by a principal component analysis (PCA), was used as baseline.

3.8. Reproducibility

Reproducibility is a cornerstone of science [**109**]: having reproducible results is fundamental to moving a field forward, especially in a field like machine learning where new ideas spread very quickly. Here, we evaluate ease with which the results of the selected papers can be reproduced by the community using two key criteria: the availability of their data and the availability of their code.

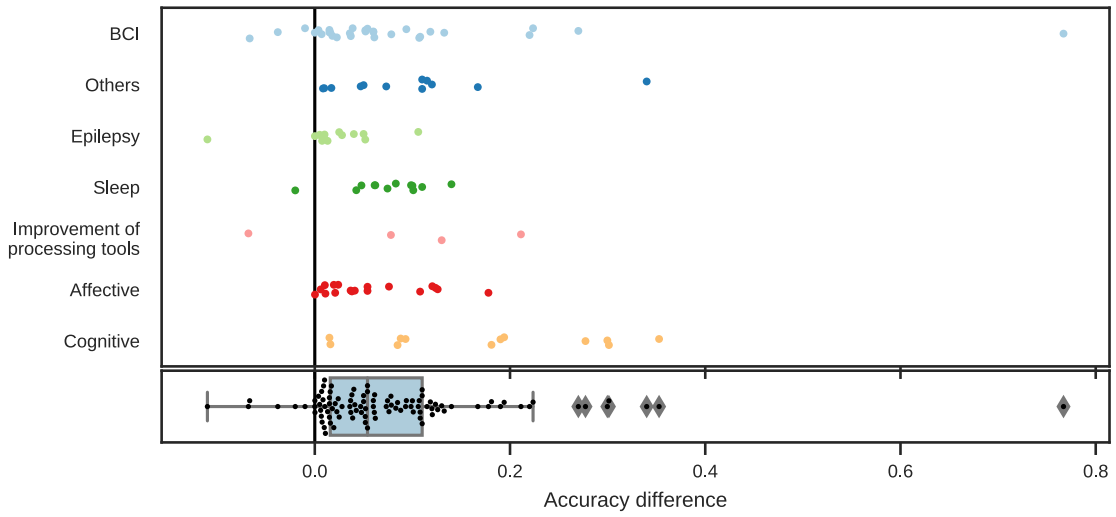


Fig. 54. Difference in accuracy between each proposed DL model and corresponding baseline model for studies reporting accuracy (see Section 3.7.6 for a description of the inclusion criteria). The difference in accuracy is defined as the difference between the best DL model and the best corresponding baseline. In the top figure, each study/task is represented by a single point, and studies are grouped according to their respective domains. The bottom figure is a box plot representing the overall distribution. *The result which achieved an accuracy difference of nearly* 77% [**160**] *was found to be caused by a flawed design in* [**94**] *and should therefore be considered as an outlier.*

From the 154 studies reviewed, 53% used public data, 42% used private data[8], and 4% used both public and private data. In particular, studies focusing on BCI, epilepsy, sleep and affective monitoring made use of openly available datasets the most (see Table 8). Interestingly, in cognitive monitoring, no publicly available datasets were used, and papers in that field all relied on internal recordings.

[8]Data that is not freely available online was considered private regardless of when and where it was recorded. Moreover, three of the reviewed studies mentioned that their data was available upon request but were included in the "private" category.

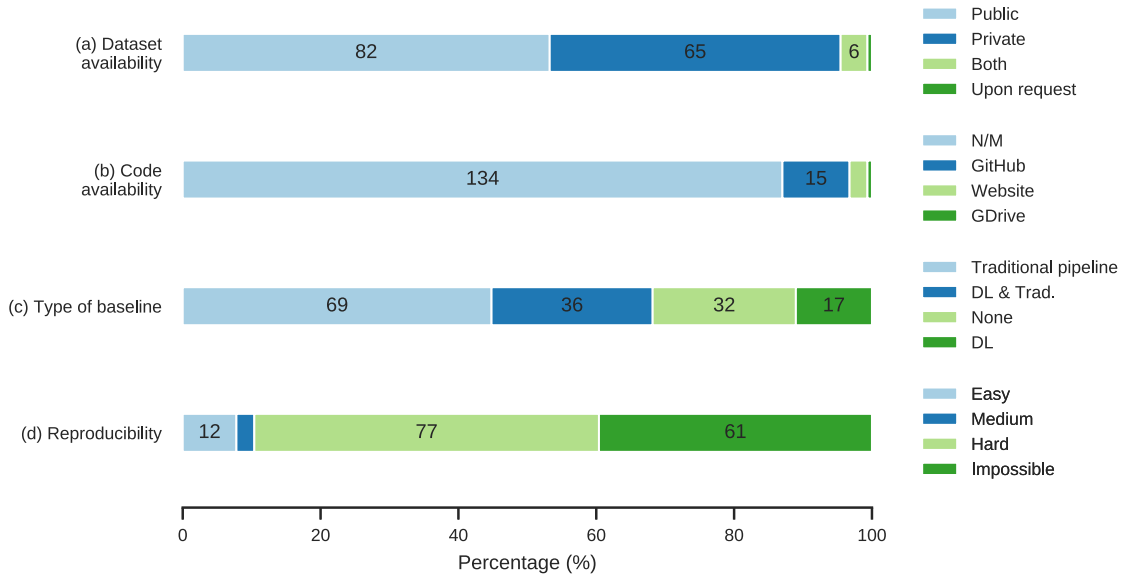


Fig. 55. Reproductibility of the selected studies. (a) Availability of the datasets used in the studies, (b) availability of the code, shown by where the code is hosted, (c) type of baseline used to evaluate the performance of the trained models and (d) estimated reproducibility level of the studies (Easy: both the data and the code are available, Medium: the code is available but some data is not publicly available, Hard: either the code or the data is available but not both, Impossible: neither the data nor the code are available).

Fittingly, a total of 33 papers (21%) explicitly mentioned that more publicly available data is required to support research on DL-EEG. In clinical settings, the lack of labeled data, rather than the quantity of data, was specifically pointed out as an obstacle.

As for the source code, only 20 papers (13%) chose to make it available online [**83, 147, 191, 159, 221, 194, 151, 87, 148, 217, 219, 166, 220, 218, 102, 160, 13, 162, 161, 85**] and as illustrated in Fig 55, GitHub is by far the preferred code sharing platform. Needless to say, having access to the source code behind published

results can drastically reduce time and increase incentive to reproduce a paper's results.

Therefore, taking both data and code availability into account, only 12 out of 154 studies (8%) could easily be reproduced using both the same data and code [**147, 191, 159, 151, 217, 219, 166, 218, 102, 160, 162, 85**]. 4 out of 154 studies (3%) shared their code but tested on both private and public data making their studies only partially reproducible [**221, 87, 148, 220**], see Fig. 55. As follows, a significant number of studies (61) did not have publicly available data or code, making them almost impossible to reproduce.

It is important to note, moreover, that for the results of a study to be perfectly reproduced, the authors would also need to share the weights (i.e. parameters) of the network. Sharing the code and the architecture of the network might not be sufficient since retraining the network could converge to a different minimum. On the other hand, retraining the network could also end up producing better results if a better performing model is obtained. For recommendations on how to best share the results, the code, the data and relevant information to make a study easy to reproduce, please see the discussion section and the checklist provided in Appendix 5.

4. Discussion

In this section, we review the most important findings from our results section, and discuss the significance and impact of various trends highlighted above. We also provide recommendations for DL-EEG studies and present a checklist to ensure reproducibility in the field.

4.1. Rationale

It was expected that most papers selected for the review would focus on the classification of EEG data, as DL has historically led to important improvements

Table 8. Most often used datasets by domain. Datasets that were only used by one study are grouped under "Other" for each category.

Main domain	Dataset	# articles	References
Affective	DEAP [**82**]	9	[**98, 6, 16, 100, 197, 101, 42, 79, 93**]
	SEED [**224**]	3	[**216, 101, 224**]
BCI	BCI Competition [**23, 24, 141**]	13	[**43, 87, 142, 148, 148, 169, 169, 107, 143, 200, 37, 37, 27**]
	Other	8	[**69, 87, 148, 62, 62, 62, 165, 8**]
	eegmmidb [**146**]	8	[**212, 217, 105, 219, 222, 36, 116, 58**]
	Keirn & Aunon (1989)[1]	2	[**123, 128**]
Cognitive	Other	4	[**83, 60, 60, 60**]
	EEG Eye State[2]	1	[**113**]
Epilepsy	Bonn University [**9**]	7	[**76, 183, 4, 170, 2, 119, 112**]
	CHB-MIT [**155**]	7	[**181, 208, 179, 178, 125, 173, 182**]
	TUH [**65**]	5	[**51, 152, 50, 49, 201**]
	Other	3	[**179, 50, 171**]
	Freiburg Hospital[3]	2	[**179, 178**]
Generation of data	BCI Competition [**23, 24, 141**]	2	[**33, 215**]
	MAHNOB [**158**]	1	[**190**]
	Other	1	[**151**]
	SEED [**224**]	1	[**190**]
Improvement of processing tools	BCI Competition [**23, 24, 141**]	3	[**198, 163, 199**]
	Other	3	[**202, 117, 162**]
	Bonn University [**9**]	1	[**192**]
	CHB-MIT [**155**]	1	[**192**]
	MAHNOB [**158**]	1	[**39**]
Others	TUH [**65**]	3	[**147, 138, 221**]
	eegmmidb [**146**]	3	[**221, 220, 218**]
	Other	2	[**185, 221**]
	EEG Eye State[2]	1	[**91**]
Sleep	MASS [**121**]	4	[**132, 28, 166, 38**]
	Sleep EDF [**80**]	4	[**187, 166, 196, 180**]
	Other	3	[**159, 177, 47**]
	UCDDB[4]	3	[**86, 108, 85**]

[1] `http://www.cs.colostate.edu/eeg/main/data/1989_Keirn_and_Aunon`

[2] `https://archive.ics.uci.edu/ml/datasets/EEG+Eye+State`

[3] `http://epilepsy.uni-freiburg.de/freiburg-seizure-prediction-project/eeg-database`

[4] `https://physionet.org/pn3/ucddb/`

on supervised classification problems [**88**]. Interestingly though, several papers also focused on new applications that were made possible or facilitated by DL: for instance, generating images conditioned on EEG, generating EEG, transfer learning between subjects, or feature learning. One of the main motivations for using DL cited by the papers reviewed was the ability to use raw EEG with no manual feature extraction steps. We expect these kinds of applications that go beyond using DL as a replacement for traditional processing pipelines to gain in popularity.

4.2. Data

A critical question concerning the use of DL with EEG data remains "How much data is enough data?". In Section 3.3, we explored this question by looking at various descriptive dimensions: the number of subjects, the amount of EEG recorded, the number of training/test/validation examples, the sampling rate and data augmentation schemes used.

Although a definitive answer cannot be reached, the results of our meta-analysis show that the amount of data necessary to at least match the performance of traditional approaches is already available. Out of the 154 papers reviewed, only six reported lower performance for DL methods over traditional benchmarks. To achieve these results with limited amounts of data, shallower architectures were often preferred. Data augmentation techniques were also used successfully to improve performance when only limited data was available. However, more work is required to clearly assess their advantages and disadvantages. Indeed, although many studies used overlapping sliding windows, there seems to be no consensus on the best overlapping percentage to use, e.g., the impact of using a sliding window with 1% overlap versus 95% overlap is still not clear. BCI studies had the highest variability for this hyperparameter, while clinical applications such as sleep staging already appeared more standardized with most studies using 30 s non-overlapping windows.

Many authors concluded their paper suggesting that having access to more data would most likely improve the performance of their models. With large datasets becoming public, such as the TUH Dataset [**64**] and the National Sleep Research Resource [**213**], deeper architectures similar to the ones used in computer vision might become increasingly usable. However, it is important to note that the availability of data is quite different across domains. In clinical fields such as sleep and epilepsy, data usually comes from hospital databases containing years of recordings from several patients, while other fields usually rely on data coming from lab experiments with a limited number of subjects.

The potential of DL in EEG also lies in its ability (at least in theory) to generalize across subjects and to enable transfer learning across tasks and domains. Although intra-subject models still work best when only limited data is available, given the inherent subject variability of EEG data, transfer learning might be the key to moving past this limitation. Indeed, Page and colleagues [**125**] showed that with hybrid models, one can train a neural network on a pool of subjects and then fine-tune it on a specific subject, achieving good performances without needing as much data from a specific subject.

While the amount of data is critical in achieving high performance on machine learning tasks (and particularly for deep learning), the *quality* of the data is also very important. In many fields of application of DL, input data usually has a high SNR: in both CV and NLP, for instance, virtually noise-free images and natural language excerpts are easy to obtain. EEG data, on the other hand, can accumulate noise at many different levels, which makes learning from it much harder. Most often, once the data is recorded, the noise is impossible or very difficult to mitigate. With that in mind, high quality and well-maintained hardware is crucial to collecting clean EEG data, however the capacity of the experimenter to prepare and use the equipment properly will ultimately determine signal quality. Prepping participants to ensure their compliance with the recording protocol is also fundamental to obtaining

meaningful data. Similarly, reliable recording requires well planned out experimental design, including stimulus presentation when applicable. Furthermore, while naturally modulated by its end purpose, the quality of the data is influenced by its diversity, e.g., how many different individuals and how different they are. A balanced number of examples in each class can also drastically improve the usefulness of a large dataset. In brief, we believe both the quantity and the quality of the data must be taken into account when assessing the usefulness of a dataset, which is particularly true with electrophysiological data.

While we did report the sampling rate, we did not investigate its effect on performance because no relationship stood out particularly in any of the reviewed papers. The impact of the number of channels though, was specifically studied. For example, in [**28**], the authors showed that they could achieve comparable results with a lower number of channels. As shown in Fig. 46, a few studies used low-cost EEG devices, typically limited to a lower number of channels. These more accessible devices might therefore benefit from DL methods, but could also enable faster data collection on a larger-scale, thus facilitating DL in return.

As DL-EEG is highly data-driven, it is important when publishing results to clearly specify the amount of data used and to clarify terminology (see Table 5 for an example). We noticed that many studies reviewed did not clearly describe the EEG data that they used (e.g., the number of subjects, number of sessions, window length to segment the EEG data, etc.) and therefore made it hard or impossible for the reader to evaluate the work and compare it to others. Moreover, reporting learning curves (i.e. performance as a function of the number of examples) would give the reader valuable insights on the bias and variance of the model.

4.3. EEG processing

According to our findings, the great majority of the reviewed papers preprocessed the EEG data before feeding it to the deep neural network or extracting features. Despite observing this trend, we also noticed that recent studies outperformed their respective baseline(s) using completely raw EEG data. Almogbel et al. [**7**] used raw EEG data to classify cognitive workload in vehicle drivers, and their best model achieved a classification accuracy approximately 4% better than their benchmarks which employed preprocessing on the EEG data. Similarly, Aznan et al. [**11**] outperformed the baselines by a 4% margin on SSVEP decoding using no preprocessing. Thus, the answer to whether it is necessary to preprocess EEG data when using DNNs remains elusive.

As most of the works considered did not use, or explicitly mention using, artifact removal methods, it appears that this EEG processing pipeline step is in general not required. However, one should observe that in specific cases such as tasks that inherently elicit quick eye movements (MATB-II [**31**]), artifact handling might still be crucial to obtaining desired performance.

One important aspect we focused on is whether it is necessary to use EEG features as inputs to DNNs. After analyzing the type of input used by each paper, we observed that there was no clear preference for using features or raw EEG time-series as input. We noticed though that most of the papers using CNNs used raw EEG as input. With CNNs becoming increasingly popular, one can conclude that there is a trend towards using raw EEG instead of hand-engineered features. This is not surprising, as we observed that one of the main motivations mentioned for using DNNs on EEG processing is to automatically learn features. Furthermore, frequency-based features, which are widely used as hand-crafted features in EEG [**103**], are very similar to the temporal filters learned by a CNN. Indeed, these features are often extracted using Fourier filters which apply a convolutive operation. This is also the case for

the temporal filters learned by a CNN although in the case of CNNs the filters are learned.

From our analysis, we also aimed to identify which input type should be used when trying to solve a problem from scratch. While the answer depends on many factors such as the domain of application, we observed that in some cases raw EEG as input consistently outperformed baselines based using classically extracted features. For example, for seizure classification, recently proposed models using raw EEG data as input [**63, 183, 122**] achieved better performances than classical baseline methods, such as SVMs with frequency-domain features. For this particular task, we believe following the current trend of using raw EEG data is the best way to start exploring a new approach.

4.4. Deep learning methodology

Another major topic this review aimed at covering is the DL methodology itself. Our analysis focused on architecture trends and training decisions, as well as on model selection techniques.

4.4.1. Architecture. Given the inherent temporal structure of EEG, we expected RNNs would be more widely employed than models that do not explicitly take time dependencies into account. However, almost half of the selected papers used CNNs. This observation is in line with recent discussions and findings regarding the effectiveness of CNNs for processing time series [**12**]. We also noticed that the use of energy-based models such as RBMs has been decreasing, whereas on the other hand, popular architectures in the computer vision community such as GANs have started to be applied to EEG data as well. As suggested by a Kruskal-Wallis test ($p = 0.043$), the choice of architecture seems to have had an impact on the reported accuracy improvement over traditional baselines: in the reviewed papers, CNNs and DBNs generally led to higher improvements, while AE-based models led to the lowest

improvements. Although this might reflect an actual advantage of convolutional architectures or of DBN-based unsupervised pretraining over vanilla recurrent or fully connected architectures, the considerable variability in the experiments reported in the reviewed papers makes it impossible to draw any conclusion yet. Instead, we believe focused studies will be necessary to evaluate the impact of architectural choices on performance on a domain-by-domain basis.

Moreover, regarding architecture depth, most of the papers used fewer than five layers. When comparing this number with popular object recognition models such as VGG and ResNet for the ImageNet challenge comprising 19 and 34 layers respectively, we conclude that for EEG data, shallower networks are currently necessary. Schirrmeister et al. [**175**] specifically focused on this aspect, comparing the performance of architectures with different depths and structures, such as fully convolutional layers and residual blocks, on different tasks. Their results showed that in most cases, shallower fully convolutional models outperformed their deeper counterpart and architectures with residual connections. However, the authors later found the weight initialization to be critical in successfully training deeper architectures such as ResNet on an intracranial task [**188**], suggesting hyperparameter tuning might be key to using deeper architectures on neurophysiological data (personal communication, April 17, 2019).

4.4.2. Training and optimization. Although crucial to achieving good results when using neural networks, only 20% of the papers employed some hyperparameter search strategy. Even fewer studies provided detailed information about the method used. Amongst these, Stober et al. [**162**] described their hyperparameter selection method and cited its corresponding implementation; in addition, the available budget in number of iterations per searching trial as well as the cross-validation split were mentioned in the paper.

4.4.3. Model inspection. Inspecting trained DL models is important, as DNNs are notoriously seen as black boxes, when compared to more traditional methods. Indeed, straightforward model inspection techniques such as visualizing the weights of a linear classifier are not applicable to deep neural networks; their decisions are thus much harder to understand. This is problematic in clinical settings for instance, where understanding and explaining the choice made by a classification model might be critical to making informed clinical choices. Neuroscientists might also be interested by what drives a model's decisions and use that information to shape hypotheses about brain function.

Although it can manifest with any machine model based on elaborate EEG features, the problem of identifying whether or not informative patterns stem from brain or artifactual activity is exacerbated by DL. Especially when considering end-to-end models trained on raw data (which is the case of almost half of the studies included in this review), any pattern correlated with the target of the learning task might end up being used by a model to drive decisions. When no artifact handling is done (at least 46% of the studies), it then becomes likely that artifactual components, which are typically much stronger in amplitude than actual EEG sources, are being used somehow by a DL model. In many applications where the unique concern is classification performance (e.g., BCI, sleep staging, seizure detection) and for subjects for whom artifacts are robust covariates of the measured condition, this might not be problematic. However, if the end goal requires the system to solely rely on brain activity (e.g., BCIs for locked-in individuals who can't rely on residual muscle activity or in neuroscience-specific investigations), it is necessary to implement artifact handling procedures and, as far as possible, inspect the models trained. Since artifactual signatures are usually well-characterized, it should be possible to use methods like those mentioned in Table 7 to assess whether brain activity or artifacts drive decisions in a DL model.

About 27% of the reviewed papers looked at interpreting their models. Interesting work on the topic, specifically tailored to EEG, was reviewed in [**148, 68, 45**]. Sustained efforts aimed at inspecting models and understanding the patterns they rely on to reach decisions are necessary to broaden the use of DL for EEG processing.

4.5. Reported results

Our meta-analysis focused on how studies compared classification accuracy between their models and traditional EEG processing pipelines on the same data. Although a great majority of studies reported improvements over traditional pipelines, this result has to be taken with a grain of salt. First, the difference in accuracy does not tell the whole story, as an improvement of 10%, for example, is typically more difficult to achieve from 80 to 90% than from 40 to 50%. More importantly though, very few articles reported negative improvements, which could be explained by a publication bias towards positive results.

The reported baseline comparisons were highly variable: some used simple models (e.g., combining straightforward spectral features and linear classifiers), others used more sophisticated pipelines (including multiple features and non-linear approaches), while a few reimplemented or cited state-of-the-art models that were published on the same dataset and/or task. Often, the description of baseline models is also too succinct to effectively assess whether the baselines are optimal for a given task: for instance, the performance on the training set can be used to assess whether the baseline models are in the overfitting or underfitting regime. Since the observed improvement will likely be higher when comparing to simple baselines than to state-of-the-art results, the values that we report might be biased positively. For instance, only two studies used Riemannian geometry-based processing pipelines as baseline models [**11, 87**], although these methods have set a new state-of-the-art in multiple EEG classification tasks [**103**].

Moreover, many different tasks and thus datasets were used. These datasets are often private, meaning there is very limited or no previous literature reporting results on them. On top of this, the lack of reproducibility standards can lead to low accountability: study results are not expected to be replicated and can be inflated by non-standard practices such as omitting cross-validation.

Notwithstanding the limits of the improvement in accuracy as a performance metric (as described above), we ran a series of non-parametric statistical tests to assess whether any of the collected data items seem to covary with accuracy improvement. We used Mann-Whitney rank-sum tests for binary data items, Kruskal-Wallis analysis of variance for data items with more than two possible values, and Spearman's rank correlation for numerical data items, and considered their p-value. All p-values were found to be above a significance level of 0.05, except for the data item "Architecture". This result was discussed in Section 4.4.1 above. As for the other data items, the inconclusiveness of the statistical tests most likely stem from highly variable and imprecise baseline comparisons across studies. Therefore, the impact of each of these data items remains better described by well-controlled domain-specific (and even dataset-specific) studies which might not be generalizable across domains. We tried to highlight studies that reported such interesting comparisons in both the Results and Discussion sections of this review.

Different approaches have been taken to solve the problem of heterogeneity of result reporting and benchmarking in the field of machine learning. For instance, OpenML [**186**] is an online platform that facilitates the sharing and running of experiments, as well as the benchmarking of models. As of November 2018, the platform already contained one EEG dataset and multiple submissions. The MOABB [**78**], a solution tailored to the field of brain-computer interfacing, is a software framework for ensuring the reproducibility of BCI experiments and providing public benchmarks for many BCI datasets. In [**73**], a similar approach, but for DL specifically, is proposed.

Additionally, a few EEG/MEG/ECoG classification online competitions have been organized in the last years, for instance the Physionet challenge [**44**] or various competitions on the Kaggle platform (see Table 1 of [**32**]). These competitions informally act as benchmarks: they provide a standardized dataset with training and test splits, as well as a leaderboard listing the performance achieved by every competitor. These platforms can then be used to evaluate the state-of-the-art as they provide a publicly available comparison point for new proposed architectures. For instance, the IEEE NER 2015 Conference competition on error potential decoding could have been used as a benchmark for the studies reviewed that focused on this topic. Generally speaking, rigorous studies of the impact of different methodologies on specific datasets will be necessary to set up clear benchmarks that can be built upon (e.g., [**137**] for the TUH Seizure corpus).

Making use of these tools, or extending them to other EEG-specific tasks, appears to be one of the greatest challenges for the field of DL-EEG at the moment, and might be the key to more efficient and productive development of practical EEG applications. Whenever possible, authors should make sure to provide as much information as possible on the baseline models they have used, and explain how to replicate their results (see Section 4.6).

4.6. Reproducibility

The significant use of public EEG datasets across the reviewed studies suggests that open data has greatly contributed to recent developments in DL-EEG. On the other hand, 42% of studies used data not publicly available - notably in domains such as cognitive monitoring. To move the field forward, it is thus important to create new benchmark datasets and share internal recordings. Moreover, the great majority of papers did not make their code available. Many papers reviewed are thus more difficult to reproduce: the data is not available, the code has not been shared, and the

baseline models that were used to compare the performances of the models are either non-existent or not available.

Recent initiatives to promote best practices in data and code sharing would benefit the field of DL-EEG. FAIR neuroscience [**193**] and the Brain Imaging Data Structure (BIDS) [**55**] both provide guidelines and standards on how to acquire, organize and share data and code. BIDS extensions specific to EEG [**131**] and MEG [**114**] were also recently proposed. Moreover, open source software toolboxes are available to perform DL experiments on EEG. For example, the recent toolbox developed by Schirrmeister and colleagues, called BrainDecode [**148**], enables faster and easier development cycles by providing the basic functionality required for DL-EEG analysis while offering high level and easy to use functions to the user. The use of common software tools could facilitate reproducibility in the community. Beyond reproducibility, we believe simplifying access to data, making domain knowledge accessible and sharing code will enable more people to jump into the field of DL-EEG and contribute, transforming what has traditionally been a domain-specific problem into a more general problem that can be tackled with machine learning and DL methods.

To move forward in that direction, we are planning a follow-up to this literature review in the form of an online public portal listing in greater detail results of published research on multiple openly available datasets. We believe such a portal will be critical in the advancement of the DL-EEG literature.

4.7. Recommendations

To improve the quality and reproducibility of the work in the field of DL-EEG, we propose six guidelines in Table 9. Moreover, Appendix 5 presents a checklist of items that are critical to ensuring reproducibility and should be included in future studies.

A similar checklist, but specifically targeting machine learning publications, has also recently been proposed.[9]

Table 9. Recommendations for future DL-EEG studies. See Appendix 5 for a detailed list of items to include.

	Recommendation	Description
1	**Clearly describe the architecture.**	Provide a table or figure clearly describing your model (e.g., see [**28, 51, 148**]).
2	**Clearly describe the data used.**	Make sure the number of subjects, the number of examples, the data augmentation scheme, etc. are clearly described. Use unambiguous terminology or define the terms used (for an example, see Table 5).
3	**Use existing datasets.**	Whenever possible, compare model performance on public datasets.
4	**Include state-of-the-art baselines.**	If focusing on a research question that has already been studied with traditional machine learning, clarify the improvements brought by using DL.
5	**Share internal recordings.**	Whenever possible.
6	**Share reproducible code.**	Share code (including hyperparameter choices and model weights) that can easily be run on another computer, and potentially reused on new data.

4.7.1. Supplementary material. Along with the current paper, we make our data items table and related code available online at `http://dl-eeg.com`. We encourage interested readers to consult it in order to dive deeper into data items that are of specific interest to them - it should be straightforward to reproduce and extend the results and figures presented in this review using the code provided. The data item

[9] `https://www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist.pdf`

table is intended to be updated frequently with new articles, therefore results will be brought up to date periodically.

Authors of DL-EEG papers not included in the review are invited to submit a summary of their article following the format of our data items table to our online code repository. We also invite authors whose papers are already included in the review to verify the accuracy of our summary. Eventually, we would like to indicate which studies have been submitted or verified by the original authors.

By updating the data items table regularly and inviting researchers in the community to contribute, we hope to keep the supplementary material of the review relevant and up-to-date as long as possible.

4.8. Limitations

In this section, we quickly highlight some limitations of the present work. First, although the search methodology used to identify relevant studies is well-founded, it undeniably did not capture all of the existing literature on the topic. Therefore, we have abstained from drawing absolute conclusions on the different data items, and instead focused on highlighting trends. As described in Section 2, our search terms were not biased toward any type of architecture, and so we are confident the results we present in this review are sound.

Second, our decision to include preprints from arXiv and bioRxiv in the database search requires some justification. It is important to note that preprints are not peer-reviewed. Therefore, some of the studies we selected might not be of the same quality and scientific rigor as the ones coming from peer-reviewed journals or conferences. For this reason, whenever a preprint was followed by a publication in a peer-reviewed venue, we focused our analysis on the peer-reviewed version. Nonetheless, we did not find significant differences between the preprints and the peer-reviewed studies in terms of reported improvement in accuracy. ArXiv has been largely adopted by the

DL community as a means to quickly disseminate results and encourage fast research iteration cycles. Since the field of DL-EEG is still young and a limited number of publications was available at the time of writing, we decided to include all the papers we could find, knowing that some of the newer trends would be mostly visible in repositories such as arXiv. Our goal with this review was to provide a transparent and objective analysis of the trends in DL-EEG. By including preprints, we feel we provided a better view of the current state-of-the-art, and are also in a better position to give recommendations on how to share results of DL-EEG studies moving forward.

Third, in order to keep this review reasonable in length, we decided to focus our analysis on the points that we judged most interesting and valuable. As a result, various factors that impact the performance of DL-EEG were not covered in the review. For example, we did not cover weight initialization: in [**51**], the authors compared 10 different initialization methods and showed an impact on the specificity metric, with ranged from 85.1% to 96.9%. Similarly, multiple data items were collected during the review process, but were not included in the analysis. These items, which include data normalization procedures, software toolboxes, hyperparameter values, loss functions, training hardware, training time, etc., remain available online for the interested reader. We are confident other reviews or research articles will be able to focus on more specific elements.

Finally, as any literature review in a field that is quickly evolving, the relevance of our analysis decays with time as new articles are being published and new trends are established. Since our last database search, we have already identified other articles that should eventually be added to the analysis. Again, making this work a living review by providing the data and code online will hopefully ensure the review will be of value and remain relevant for years to come.

5. Conclusion

The usefulness of EEG as a functional neuroimaging tool is unequivocal: clinical diagnosis of sleep disorders and epilepsy, monitoring of cognitive and affective states, as well as brain-computer interfacing all rely heavily on the analysis of EEG. However, various challenges remain to be solved. For instance, time-consuming tasks currently carried out by human experts, such as sleep staging, could be automated to increase the availability and flexibility of EEG-based diagnosis. Additionally, better generalization performance between subjects will be necessary to truly make BCIs useful. DL has been proposed as a potential candidate to tackle these challenges. Consequently, the number of publications applying DL to EEG processing has seen an exponential increase over the last few years, clearly reflecting a growing interest in the community in these kinds of techniques.

In this review, we highlighted current trends in the field of DL-EEG by analyzing 154 studies published between January 2010 and July 2018 applying DL to EEG data. We focused on several key aspects of the studies, including their origin, rationale, the data they used, their EEG processing methodology, DL methodology, reported results and level of reproducibility.

Among the major trends that emerged from our analysis, we found that 1) DL was mainly used for classifying EEG in domains such as brain-computer interfacing, sleep, epilepsy, cognitive and affective monitoring, 2) the quantity of data used varied a lot, with datasets ranging from 1 to over 16,000 subjects (mean = 223; median = 13), producing 62 up to 9,750,000 examples (mean = 251,532; median = 14,000) and from two to 4,800,000 minutes of EEG recording (mean = 62,602; median = 360), 3) various architectures have been used successfully on EEG data, with CNNs, followed by RNNs and AEs, being most often used, 4) there is a clear growing interest towards using raw EEG as input as opposed to handcrafted features, 5) almost all studies reported a small improvement from using DL when compared to other baselines and

benchmarks (median = 5.4%), and 6) while several studies used publicly available data, only a handful shared their code - the great majority of studies reviewed thus cannot easily be reproduced.

This review also shows that more targeted work needs to be done around the amount of data required to fully exploit the potential advantages of DL in EEG processing. Such work could explore the relationship between performance and the amount of data, the relationship between performance and data augmentation and the relationship between performance, the amount of data and the depth of the network.

Moreover, given the high variability in how results were reported, we made six recommendations to ensure reproducibility and fair comparison of results: 1) clearly describe the architecture, 2) clearly describe the data used, 3) use existing datasets, whenever possible, 4) include state-of-the-art baselines, ideally using the original authors' code, 5) share internal recordings, whenever possible, and 6) share code, as it is the best way to allow others to pick up where your work leaves off. We also provided a checklist (see Appendix 5) to help authors of DL-EEG studies make sure all the relevant information is available in their publications to allow straightforward reproduction.

Finally, to help the DL-EEG community maintain an up-to-date list of published work, we made our data items table open and available online. The code to reproduce the statistics and figures of this review as well as the full summaries of the papers are also available at `http://dl-eeg.com`.

The current general interest in artificial intelligence and DL has greatly benefited various fields of science and technology. Advancements in other field of application will most likely benefit the neuroscience and neuroimaging communities in the near future, and enable more pervasive and powerful applications based on EEG processing. We hope this review will constitute a good entry point for EEG researchers interested in applying DL to their data, as well as a good summary of the current state of the field for DL researchers looking to apply their knowledge to new types of data.

A planned follow-up to this review will be an online portal providing clear and reproducible benchmarks for deep learning-based analysis of EEG data, accessible at `http://dl-eeg.com`.

Acknowledgments

We thank Raymundo Cassani, Colleen Gillon, João Monteiro and William Thong for comments that greatly improved the manuscript.

Funding

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC-RDC) for JF and YR (reference number: RDPJ 514052-17), NSERC research funds for JF, HB, IA and THF, the Fonds québécois de la recherche sur la nature et les technologies (FRQNT) for YR and InteraXon Inc. (graduate funding support) for HB.

References

[1] Khald Ali I. Aboalayon, Miad Faezipour, Wafaa S. Almuhammadi, and Saeid Moslehpour. Sleep stage classification using EEG signal analysis: A comprehensive survey and new investigation. *Entropy*, 18(9):272, 2016.

[2] U Rajendra Acharya, Shu Lih Oh, Yuki Hagiwara, Jen Hong Tan, and Hojjat Adeli. Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals. *Computers in Biology and Medicine*, (August):1–9, 2017.

[3] U Rajendra Acharya, S Vinitha Sree, G Swapna, Roshan Joy Martis, and Jasjit S Suri. Automated EEG analysis of epilepsy: a review. *Knowledge-Based Systems*, 45:147–165, 2013.

[4] David Ahmedt-Aristizabal, Clinton Fookes, Kien Nguyen, and Sridha Sridharan. Deep Classification of Epileptic Signals. *arXiv preprint*, pages 1–4, 2018.

[5] Abeer Al-Nafjan, Manar Hosny, Yousef Al-Ohali, and Areej Al-Wabil. Review and Classification of Emotion Recognition Based on EEG Brain-Computer Interface System Research: A Systematic Review. *Applied Sciences*, 7(12):1239, 2017.

[6] Salma Alhagry, Aly Aly Fahmy, and Reda A El-Khoribi. Emotion Recognition based on EEG using LSTM Recurrent Neural Network. *IJACSA) International Journal of Advanced Computer Science and Applications*, 8(10):8–11, 2017.

[7] Mohammad A Almogbel, Anh H Dang, and Wataru Kameyama. EEG-Signals Based Cognitive Workload Detection of Vehicle Driver using Deep Learning. *20th International Conference on Advanced Communication Technology*, 7:256–259, 2018.

[8] Jinwon An and Sungzoon Cho. Hand motion identification of grasp-and-lift task from electroencephalography recordings using recurrent neural networks. *2016 International Conference on Big Data and Smart Computing, BigComp 2016*, pages 427–429, 2016.

[9] Ralph G Andrzejak, Klaus Lehnertz, Florian Mormann, Christoph Rieke, Peter David, and Christian E Elger. Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state. *Physical Review E*, 64(6):061907, 2001.

[10] Martijn Arns, C. Keith Conners, and Helena C. Kraemer. A Decade of EEG Theta/Beta Ratio Research in ADHD: A Meta-Analysis. *Journal of Attention Disorders*, 17(5):374–383, 2013.

[11] Nik Khadijah Nik Aznan, Stephen Bonner, Jason D Connolly, Noura Al Moubayed, and Toby P Breckon. On the Classification of SSVEP-Based Dry-EEG Signals via Convolutional Neural Networks. *arXiv preprint*, 2018.

[12] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. *arXiv preprint*, 2018.

[13] Pouya Bashivan, Irina Rish, Mohammed Yeasin, and Noel Codella. Learning Representations from EEG with Deep Recurrent-Convolutional Neural Networks. *arXiv preprint*, pages 1–15, 2016.

[14] Joos Behncke, Robin Tibor Schirrmeister, Wolfram Burgard, and Tonio Ball. The signature of robot action success in EEG signals of a human observer: Decoding and visualization using deep convolutional neural networks. *arXiv*, 2017.

[15] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *Advances in neural information processing systems*, pages 137–144, 2007.

[16] Ahmed Ben Said, Amr Mohamed, Tarek Elfouly, Khaled Harras, and Z Jane Wang. Multimodal deep learning approach for Joint EEG-EMG Data compression and classification. *IEEE Wireless Communications and Networking Conference, WCNC*, 2017.

[17] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *The Journal of Machine Learning Research*, 13(1):281–305, 2012.

[18] Chris Berka, Daniel J Levendowski, Michelle N Lumicao, Alan Yau, Gene Davis, Vladimir T Zivkovic, Richard E Olmstead, Patrice D Tremoulet, and Patrick L Craven. {EEG} correlates of task engagement and mental workload in vigilance, learning, and memory tasks. *Aviation, space, and environmental medicine*, 78(5):B231—-B244, 2007.

[19] Andrea Biasiucci, Benedetta Franceschiello, and Micah M Murray. Electroencephalography. *Current Biology*, 29(3):R80–R85, 2019.

[20] Nima Bigdely-Shamlo, Tim Mullen, Christian Kothe, Kyung-Min Su, and Kay A. Robbins. The PREP pipeline: standardized preprocessing for large-scale EEG analysis. *Frontiers in Neuroinformatics*, 9:16, 2015.

[21] Christopher M. Bishop. *Neural Networks for Pattern Recognition*, volume 92. Oxford university press, 1995.

[22] Siddharth Biswal, Joshua Kulas, Haoqi Sun, Balaji Goparaju, M Brandon Westover, Matt T Bianchi, and Jimeng Sun. SLEEPNET: Automated Sleep Staging System via Deep Learning. *arXiv preprint*, pages 1–17, 2017.

[23] Benjamin Blankertz, K-R Muller, Dean J Krusienski, Gerwin Schalk, Jonathan R Wolpaw, Alois Schlogl, Gert Pfurtscheller, Jd R Millan, Michael Schroder, and Niels Birbaumer. The bci

competition iii: Validating alternative approaches to actual bci problems. *IEEE transactions on neural systems and rehabilitation engineering*, 14(2):153–159, 2006.

[24] Benjamin Blankertz, Klaus-Robert Müller, Gabriel Curio, Theresa M Vaughan, Gerwin Schalk, Jonathan R Wolpaw, Alois Schlögl, Christa Neuper, Gert Pfurtscheller, Thilo Hinterberger, et al. The bci competition 2003. *IEEE Trans. Biomed. Eng*, 51(6):1044–51, 2004.

[25] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[26] Nazareth P Castellanos and Valeri A Makarov. Recovering eeg brain signals: artifact suppression with wavelet enhanced independent component analysis. *Journal of neuroscience methods*, 158(2):300–312, 2006.

[27] Hubert Cecotti and Axel Gräser. Convolutional neural networks for P300 detection with application to brain-computer interfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(3):433–445, 2011.

[28] Stanislas Chambon, Mathieu N Galtier, Pierrick J Arnal, Gilles Wainrib, Alexandre Gramfort, Telecom Paristech, and M L Nov. A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, pages 1–12, 2017.

[29] Maureen Clerc, Laurent Bougrain, and Fabien Lotte. *Brain-Computer Interfaces 1: Foundations and Methods*. Wiley, 2016.

[30] Scott R Cole and Bradley Voytek. Cycle-by-cycle analysis of neural oscillations. *bioRxiv*, 2018.

[31] J. R. Comstock. Mat - Multi-Attribute Task Battery for Human Operator Workload and Strategic Behavior Research. (January), 1994.

[32] Marco Congedo, Alexandre Barachant, and Rajendra Bhatia. Riemannian geometry for EEG-based brain-computer interfaces; a primer and a review. *Brain-Computer Interfaces*, 4(3):155–174, 2017.

[33] Isaac A Corley and Yufei Huang. Deep EEG Super-resolution: Upsampling EEG Spatial Resolution with Generative Adversarial Networks. In *IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, number March, pages 4–7, 2018.

[34] Olivier Deiss, Siddharth Biswal, Jing Jin, Haoqi Sun, M Brandon Westover, and Jimeng Sun. HAMLET: Interpretable Human And Machine co-LEarning Technique. *arXiv preprint*, 2018.

[35] J Deng, A Berg, S Satheesh, H Su, A Khosla, and L Fei-Fei. ILSVRC-2012, 2012. *URL http://www. image-net. org/challenges/LSVRC*, 2012.

[36] Tejas Dharamsi, Payel Das, Tejaswini Pedapati, Gregory Bramble, Vinod Muthusamy, Horst Samulowitz, Kush R Varshney, Yuvaraj Rajamanickam, John Thomas, and Justin Dauwels. Neurology-as-a-Service for the Developing World. *arXiv preprint*, (Nips):1–5, 2017.

[37] Shifei Ding, Nan Zhang, Xinzheng Xu, Lili Guo, and Jian Zhang. Deep Extreme Learning Machine and Its Application in EEG Classification. *Mathematical Problems in Engineering*, 2015, 2015.

[38] Hao Dong, Akara Supratak, Wei Pan, Chao Wu, Paul M. Matthews, and Yike Guo. Mixed Neural Network Approach for Temporal Sleep Stage Classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(2):324–333, 2018.

[39] Alexandre Drouin-Picaro and Tiago H. Falk. Using deep neural networks for natural saccade classification from electroencephalograms. In *2016 IEEE EMBS International Student Conference: Expanding the Boundaries of Biomedical Engineering and Healthcare, ISC 2016 - Proceedings*, pages 1–4. IEEE, 2016.

[40] J Duchi, E Hazan, and Y Singer. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.

[41] Denis A Engemann, Federico Raimondo, Jean-Remi King, Benjamin Rohaut, Gilles Louppe, Frédéric Faugeras, Jitka Annen, Helena Cassol, Olivia Gosseries, Diego Fernandez-Slezak, et al. Robust EEG-based cross-site and cross-protocol classification of states of consciousness. *Brain*, 141(11):3179–3192, 2018.

[42] Arvid Frydenlund and Frank Rudzicz. Emotional Affect Estimation Using Video and EEG Data in Deep Neural Networks. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9091, pages 273–280, 2015.

[43] Guangchun Gao, Lina Shang, Kai Xiong, Jian Fang, Cui Zhang, and Xuejun Gu. EEG classification based on sparse representation and deep learning. *NeuroQuantology*, 16(6):789–795, 2018.

[44] Mohammad M Ghassemi, Benjamin E Moody, LH Lehman, Christopher Song, Qiao Li, Haoqi Sun, Roger G Mark, M Brandon Westover, and Gari D Clifford. You snooze, you win: the physionet/computing in cardiology challenge 2018. *Computing in Cardiology*, 45:1–4, 2018.

[45] Arna Ghosh, Fabien Dal Maso, Marc Roig, Georgios D Mitsis, and Marie-Hélène Boudrias. Deep Semantic Architecture with discriminative feature visualization for neuroimage analysis. *arXiv preprint*, 2018.

[46] Joseph T Giacino, Joseph J Fins, Steven Laureys, and Nicholas D Schiff. Disorders of consciousness after acquired brain injury: the state of the science. *Nature Reviews Neurology*, 10(2):99, 2014.

[47] Endang Purnama Giri, Mohamad Ivan Fanany, and Aniati Murni Arymurthy. Combining Generative and Discriminative Neural Networks for Sleep Stages Classification. *arXiv preprint*, pages 1–13, 2016.

[48] Endang Purnama Giri, Mohamad Ivan Fanany, and Aniati Murni Arymurthy. Ischemic Stroke Identification Based on EEG and EOG using 1D Convolutional Neural Network and Batch Normalization. *arXiv preprint*, pages 484–491, 2016.

[49] Meysam Golmohammadi, Amir Hossein Harati Nejad Torbati, Silvia Lopez de Diego, Iyad Obeid, and Joseph Picone. Automatic Analysis of EEGs Using Big Data and Hybrid Deep Learning Architectures. *arXiv preprint*, 2017.

[50] Meysam Golmohammadi, Saeedeh Ziyabari, Vinit Shah, Silvia Lopez de Diego, Iyad Obeid, and Joseph Picone. Deep Architectures for Automated Seizure Detection in Scalp EEGs. *arXiv preprint*, 2017.

[51] Meysam Golmohammadi, Saeedeh Ziyabari, Vinit Shah, E Von Weltin, C Campbell, Iyad Obeid, and Joseph Picone. Gated recurrent networks for seizure detection. *2017 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pages 1–5, 2017.

[52] Ian Goodfellow. Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016.

[53] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*, volume 1. MIT press Cambridge, 2016.

[54] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[55] Krzysztof J. Gorgolewski, Tibor Auer, Vince D. Calhoun, R. Cameron Craddock, Samir Das, Eugene P. Duff, Guillaume Flandin, Satrajit S. Ghosh, Tristan Glatard, Yaroslav O. Halchenko, Daniel A. Handwerker, Michael Hanke, David Keator, Xiangrui Li, Zachary Michael, Camille Maumet, B. Nolan Nichols, Thomas E. Nichols, John Pellman, Jean Baptiste Poline, Ariel Rokem, Gunnar Schaefer, Vanessa Sochat, William Triplett, Jessica A. Turner, Gaël Varoquaux, and Russell A. Poldrack. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific Data*, 3:160044, 2016.

[56] Alexandre Gramfort, Daniel Strohmeier, Jens Haueisen, Matti S Hämäläinen, and Matthieu Kowalski. Time-frequency mixed-norm estimates: Sparse M/EEG imaging with non-stationary source activations. *NeuroImage*, 70:410–422, 2013.

[57] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved Training of Wasserstein GANs. In *Advances in Neural Information Processing Systems*, pages 5767–5777, 2017.

[58] Mohammad H., Aya Samaha, and Khaled AlKamha. Automated Classification of L/R Hand Movement EEG Signals using Advanced Feature Extraction and Machine Learning. *International Journal of Advanced Computer Science and Applications*, 4(6):6, 2013.

[59] S. Hagihira. Changes in the electroencephalogram during anaesthesia and their physiological basis. *British Journal of Anaesthesia*, 115(suppl_1):i27–i31, 2015.

[60] Mehdi Hajinoroozi, Zijing Mao, and Yufei Huang. Prediction of driver's drowsy and alert states from EEG signals with deep learning. *2015 IEEE 6th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing, CAMSAP 2015*, pages 493–496, 2015.

[61] Mehdi Hajinoroozi, Zijing Mao, Tzyy Ping Jung, Chin Teng Lin, and Yufei Huang. EEG-based prediction of driver's cognitive performance by deep convolutional neural network. *Signal Processing: Image Communication*, 47:549–555, 2016.

[62] Mehdi Hajinoroozi, Zijing Mao, and Yuan-pin Lin. Deep Transfer Learning for Cross-subject and Cross-experiment Prediction of Image Rapid Serial Visual Presentation Events from EEG Data. In *International Conference on Augmented Cognition*, volume 10284, pages 45–55, 2017.

[63] Yongfu Hao, Hui Ming Khoo, Nicolas von Ellenrieder, Natalja Zazubovits, and Jean Gotman. DeepIED: An epileptic discharge detector for EEG-fMRI based on deep learning. *NeuroImage: Clinical*, 17(November 2017):962–975, 2018.

[64] A Harati, S López, I Obeid, and J Picone. THE TUH EEG CORPUS : A Big Data Resource for Automated EEG Interpretation. In *Signal Processing in Medicine and Biology Symposium (SPMB), 2014 IEEE*, pages 1–5. IEEE, 2014.

[65] A Harati, S López, I Obeid, and J Picone. THE TUH EEG CORPUS : A Big Data Resource for Automated EEG Interpretation. In *Signal Processing in Medicine and Biology Symposium (SPMB), 2014 IEEE*, pages 1–5. IEEE, 2014.

[66] Riitta Hari and Aina Puce. *MEG-EEG Primer*. Oxford University Press, 2017.

[67] Kay Gregor Hartmann, Robin Tibor Schirrmeister, and Tonio Ball. EEG-GAN: Generative adversarial networks for electroencephalograhic (EEG) brain signals. *arXiv preprint*, 2018.

[68] Kay Gregor Hartmann, Robin Tibor Schirrmeister, and Tonio Ball. Hierarchical internal representation of spectral features in deep convolutional networks trained for EEG decoding. In *2018 6th International Conference on Brain-Computer Interface, BCI 2018*, volume 2018-Janua, pages 1–6. IEEE, 2018.

[69] Md Musaddaqul Hasib, Tapsya Nayak, and Yufei Huang. A hierarchical LSTM model with attention for modeling EEG non-stationarity for human decision prediction. *2018 IEEE EMBS International Conference on Biomedical and Health Informatics, BHI 2018*, 2018-Janua(March):104–107, 2018.

[70] Bin He, Abbas Sohrabpour, Emery Brown, and Zhongming Liu. Electrophysiological source imaging: A noninvasive window to brain dynamics. *Annual Review of Biomedical Engineering*, 20(1):171–196, 2018. PMID: 29494213.

[71] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2015.

[72] Ryan Hefron, Brett Borghetti, Christine Schubert Kabban, James Christensen, and Justin Estepp. Cross-Participant EEG-Based Assessment of Cognitive Workload Using Multi-Path Convolutional Recurrent Neural Networks. *Sensors*, 18(5):1339, apr 2018.

[73] Felix A. Heilmeyer, Robin T. Schirrmeister, Lukas D. J. Fiederer, Martin Völker, Joos Behncke, and Tonio Ball. A large-scale evaluation framework for EEG deep learning architectures. *ArXiv e-prints*, jun 2018.

[74] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017.

[75] Fred Matthew Hohman, Minsuk Kahng, Robert Pienta, and Duen Horng Chau. Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers. *IEEE Transactions on Visualization and Computer Graphics*, 2018.

[76] Ramy Hussein, Hamid Palangi, Rabab Ward, and Z. Jane Wang. Epileptic Seizure Detection: A Deep Learning Approach. *Arxiv*, pages 1–12, 2018.

[77] Mainak Jas, Denis A. Engemann, Yousra Bekhti, Federico Raimondo, and Alexandre Gramfort. Autoreject: Automated artifact rejection for MEG and EEG data. *NeuroImage*, 159:417–429, 2017.

[78] Vinay Jayaram and Alexandre Barachant. MOABB: trustworthy algorithm benchmarking for BCIs. *Journal of neural engineering*, 15(6):066011, 2018.

[79] Suwicha Jirayucharoensak, Setha Pan-Ngum, and Pasin Israsena. EEG-Based Emotion Recognition Using Deep Learning Network with Principal Component Based Covariate Shift Adaptation. *Scientific World Journal*, 2014, 2014.

[80] Bob Kemp, Aeilko H Zwinderman, Bert Tuk, Hilbert AC Kamphuisen, and Josefien JL Oberye. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the eeg. *IEEE Transactions on Biomedical Engineering*, 47(9):1185–1194, 2000.

[81] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv preprint*, 2014.

[82] Sander Koelstra, Christian Mühl, Mohammad Soleymani, Jong Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. DEAP: A database for emotion analysis; Using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, 2012.

[83] Shiba Kuanar, Vassilis Athitsos, Nityananda Pradhan, Arabinda Mishra, and K R Rao. Cognitive Analysis of Working Memory Load from EEG, by a Deep Recurrent Neural Network. In *IEEE Signal Processing Society*, 2018.

[84] No Sang Kwak, Klaus Robert Müller, and Seong Whan Lee. A convolutional neural network for steady state visual evoked potential classification under ambulatory environment. *PLoS ONE*, 12(2):1–20, 2017.

[85] Martin Längkvist, Lars Karlsson, and Amy Loutfi. Sleep Stage Classification Using Unsupervised Feature Learning. *Advances in Artificial Neural Systems*, 2012:1–9, 2012.

[86] Martin Längkvist and Amy Loutfi. A Deep Learning Approach with an Attention Mechanism for Automatic Sleep Stage Classification. *Arxiv*, pages 1–18, 2018.

[87] Vernon J Lawhern, Amelia J Solon, and Nicholas R Waytowich. EEGNet : a compact convolutional neural network for EEG-based brain – computer interfaces. *Journal of neural engineering*, 2018.

[88] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436, 2015.

[89] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[90] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[91] Wei-Han Lee, Jorge Ortiz, Bongjun Ko, and Ruby Lee. Time Series Segmentation through Automatic Feature Learning. *arXiv preprint*, 2018.

[92] Yonggun Lee and Yufei Huang. Generating Target / non-Target Images of an RSVP Experiment from Brain Signals in by Conditional Generative Adversarial Network. *arXiv preprint*, (March):4–7, 2018.

[93] Kang Li, Xiaoyi Li, Yuan Zhang, and Aidong Zhang. Affective state recognition from EEG with deep belief networks. *Proceedings - 2013 IEEE International Conference on Bioinformatics and Biomedicine, IEEE BIBM 2013*, pages 305–310, 2013.

[94] Ren Li, Jared S Johansen, Hamad Ahmed, Thomas V Ilyevsky, Ronnie B Wilbur, Hari M Bharadwaj, and Jeffrey Mark Siskind. Training on the test set? an analysis of spampinato et al.[arxiv: 1609.00344]. *arXiv preprint*, 2018.

[95] Xiang Li, Peng Zhang, Dawei Song, Guangliang Yu, Yuexian Hou, and Bin Hu. EEG Based Emotion Identification Using Unsupervised Deep Feature Learning. *SIGIR2015 Workshop on Neuro- Physiological Methods in IR Research*, pages 2–4, 2015.

[96] Xuhong Li, Yves Grandvalet, and Franck Davoine. Explicit Inductive Bias for Transfer Learning with Convolutional Networks. *arXiv preprint*, 2018.

[97] Yujia Li, Alexander Schwing, Kuan-Chieh Wang, and Richard Zemel. Dualing GANs. In *Advances in Neural Information Processing Systems*, pages 5606–5616, 2017.

[98] Zhenqi Li, Xiang Tian, Lin Shu, Xiangmin Xu, and Bin Hu. Emotion Recognition from EEG Using RASM and LSTM. In *Internet Multimedia Computing and Service*, volume 819, pages 310–318. Springer, 2018.

[99] Chung-yen Liao, Rung-ching Chen, and Shao-kuo Tai. Emotion stress detection using EEG signal and deep learning technologies. *2018 IEEE International Conference on Applied System Invention (ICASI)*, (2):90–93, 2018.

[100] Wenqian Lin, Chao Li, and Shouqian Sun. Deep Convolutional Neural Network for Emotion Recognition Using EEG and Peripheral Physiological Signal. In *International Conference on Image and Graphics*, pages 385–394, 2017.

[101] Wei Liu, Wei-Long Zheng, and Bao-Liang Lu. Multimodal emotion recognition using multimodal deep learning. *arXiv preprint*, 2016.

[102] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic Gradient Descent with Warm Restarts. *arXiv preprint*, 2016.

[103] Fabien Lotte, Laurent Bougrain, Andrzej Cichocki, Maureen Clerc, Marco Congedo, Alain Rakotomamonjy, and Florian Yger. A Review of Classification Algorithms for EEG-based Brain-Computer Interfaces: A 10-year Update. *Journal of Neural Engineering*, 15(3):0–20, 2018.

[104] Fabien Lotte, Laurent Bougrain, and Maureen Clerc. *Electroencephalography (EEG)-Based Brain-Computer Interfaces*, pages 1–20. American Cancer Society, 2015.

[105] Tyler C Major and James M Conrad. The effects of pre-filtering and individualizing components for electroencephalography neural network classification. *Conference Proceedings - IEEE SOUTHEASTCON*, 2017.

[106] S Makeig, Anthony J Bell, Tzyy-Ping Jung, and Terrence J Sejnowski. Independent Component Analysis of Electroencephalographic Data. In *Advances in Neural Information Processing Systems*, volume 8, pages 145–151, 1996.

[107] Ran Manor and Amir B Geva. Convolutional Neural Network for Multi-Category Rapid Serial Visual Presentation BCI. *Frontiers in Computational Neuroscience*, 9(December):1–12, 2015.

[108] Martí Manzano, Alberto Guillén, Ignacio Rojas, and Luis Javier Herrera. Deep learning using EEG data in time and frequency domains for sleep stage classification. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 10305 LNCS, pages 132–141, 2017.

[109] Marcus R. Munafò, Brian A. Nosek, Dorothy V. M. Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J. Ware, and John P. A. Ioannidis. A manifesto for reproducible science. *Nature Human Behaviour*, 1(1):0021, 2017.

[110] Marc-Antoine Moinnereau, Thomas Brienne, Simon Brodeur, Jean Rouat, Kevin Whittingstall, and Eric Plourde. Classification of Auditory Stimuli from EEG Signals with a Regulated Recurrent Neural Network Reservoir. *Arxiv*, 2018.

[111] Francesco Carlo Morabito, Maurizio Campolo, Nadia Mammone, Mario Versaci, Silvana Franceschetti, Fabrizio Tagliavini, Vito Sofia, Daniela Fatuzzo, Antonio Gambardella, Angelo Labate, Laura Mumoli, Giovanbattista Gaspare Tripodi, Sara Gasparini, Vittoria Cianci, Chiara

Sueri, Edoardo Ferlazzo, and Umberto Aguglia. Deep Learning Representation from Electroencephalography of Early-Stage Creutzfeldt-Jakob Disease and Features for Differentiation from Rapidly Progressive Dementia. *International Journal of Neural Systems*, 27(02):1650039, 2017.

[112] Mohammad Ali Naderi and Homayoun Mahdavi-Nasab. Analysis and classification of EEG signals using spectral analysis and recurrent neural networks. In *2010 17th Iranian Conference of Biomedical Engineering (ICBME)*, number November, pages 1–4. IEEE, nov 2010.

[113] Sanam Narejo, Eros Pasero, and Farzana Kulsoom. EEG based eye state classification using deep belief network and stacked autoencoder. *International Journal of Electrical and Computer Engineering*, 6(6):3131–3141, 2016.

[114] Guiomar Niso, Krzysztof J Gorgolewski, Elizabeth Bock, Teon L Brooks, Guillaume Flandin, Alexandre Gramfort, Richard N Henson, Mainak Jas, Vladimir Litvak, Jeremy T Moreau, et al. Meg-bids, the brain imaging data structure extended to magnetoencephalography. *Scientific data*, 5:180110, 2018.

[115] Hugh Nolan, Robert Whelan, and R B Reilly. FASTER: fully automated statistical thresholding for EEG artifact rejection. *Journal of neuroscience methods*, 192(1):152–162, 2010.

[116] Rogerio Normand and Hugo Alexandre Ferreira. Superchords: the atoms of thought. *arXiv preprint*, pages 1–5, may 2015.

[117] Ewan Nurse, Benjamin S Mashford, Antonio Jimeno Yepes, Isabell Kiral-Kornek, Stefan Harrer, and Dean R Freestone. Decoding EEG and LFP signals using deep learning: heading TrueNorth. *Proceedings of the ACM International Conference on Computing Frontiers - CF '16*, pages 259–266, 2016.

[118] Alison O 'shea, Gordon Lightbody, Geraldine Boylan, and Andriy Temko. Neonatal Seizure Detection Using Convolutional Neural Networks. *arXiv*, 2017.

[119] Ibrahim Omerhodzic, Samir Avdakovic, Amir Nuhanovic, and Kemal Dizdarevic. Energy Distribution of EEG Signals: EEG Signal Wavelet-Neural Network Classifier. *arXiv preprint*, 2, 2013.

[120] Aaron van den Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George van den Driessche, Edward Lockhart, Luis C Cobo, Florian Stimberg, et al. Parallel wavenet: Fast high-fidelity speech synthesis. *arXiv preprint arXiv:1711.10433*, 2017.

[121] Christian O'reilly, Nadia Gosselin, Julie Carrier, and Tore Nielsen. Montreal archive of sleep studies: an open-access resource for instrument benchmarking and exploratory research. *Journal of sleep research*, 23(6):628–635, 2014.

[122] Alison O'Shea, Gordon Lightbody, Geraldine Boylan, and Andriy Temko. Investigating the Impact of CNN Depth on Neonatal Seizure Detection Performance. *Arxiv*, pages 15–18, 2018.

[123] Lanke Padmanabh, Rajveer Shastri, and Shashank Biradar. Mental Tasks Classification using EEG signal, Discrete Wavelet Transform and Neural Network. *Discovery*, 48(December 2015):38–41, 2017.

[124] Arsenio Paez. Gray literature: An important resource in systematic reviews. *Journal of Evidence-Based Medicine*, 10(3):233–240, 2017.

[125] Adam Page, Colin Shea, and Tinoosh Mohsenin. Wearable seizure detection using convolutional neural networks with transfer learning. *2016 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1086–1089, 2016.

[126] Viral Parekh, Ramanathan Subramanian, Dipanjan Roy, and C V Jawahar. An EEG-based image annotation system. *Communications in Computer and Information Science*, 841:303–313, 2018.

[127] Amiya Patanaik, Ju Lynn Ong, Joshua J Gooley, Sonia Ancoli-Israel, and Michael W L Chee. An end-to-end framework for real-time automatic sleep stage classification. *Sleep*, 41(5):1–11, 2018.

[128] Suprava Patnaik, Lalita Moharkar, and Amogh Chaudhari. Deep RNN Learning for EEG based Functional Brain State Inference. In *International Conference on Advances in Computing, Communication and Control (ICAC3)*, 2017.

[129] Luis Perez and Jason Wang. The Effectiveness of Data Augmentation in Image Classification using Deep Learning. *arXiv preprint*, 2017.

[130] J L Perez-Benitez, J A Perez-Benitez, and J H Espina-Hernandez. Development of a Brain Computer Interface Interface using multi-frequency visual stimulation and deep neural networks . *2018 International Conference on Electronics, Communications and Computers (CONIELECOMP)*, pages 18–24, 2018.

[131] Cyril R Pernet, Stefan Appelhoff, Guillaume Flandin, Christophe Phillips, Arnaud Delorme, and Robert Oostenveld. Bids-eeg: an extension to the brain imaging data structure (bids) specification for electroencephalography, Dec 2018.

[132] Huy Phan, Fernando Andreotti, Navin Cooray, Oliver Y. Chen, and Maarten De Vos. Joint Classification and Prediction CNN Framework for Automatic Sleep Stage Classification. *IEEE Transactions on Biomedical Engineering*, pages 1–11, 2018.

[133] Lutz Prechelt. Automatic early stopping using cross validation: quantifying the criteria. *Neural Networks*, 11(4):761–767, 1998.

[134] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, Technical report, OpenAi, 2018.

[135] S Robbins, H Monro. A stochastic approximation method. In *Statistics*, pages 102–109. Springer, 1951.

[136] F Rosenblatt. The perceptron : a probabilistic model for information storage and organization. *Psychological Review*, 65(6):386–408, 1958.

[137] Subhrajit Roy, Umar Asif, Jianbin Tang, and Stefan Harrer. Machine learning for seizure type classification: Setting the benchmark. *arXiv preprint*, 2019.

[138] Subhrajit Roy, Isabell Kiral-Kornek, and Stefan Harrer. ChronoNet: A Deep Recurrent Neural Network for Abnormal EEG Identification. *arXiv preprint*, pages 1–10, 2018.

[139] Giulio Ruffini, David Ibanez, Marta Castellano, Laura Dubreuil, Jean-Francois Gagnon, Jacques Montplaisir, and Aureli Soria-Frisch. Deep learning with EEG spectrograms in rapid eye movement behavior disorder. *bioRxiv*, pages 1–14, 2018.

[140] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533, 1986.

[141] Paul Sajda, Adam Gerson, K-R Muller, Benjamin Blankertz, and Lucas Parra. A data analysis competition to evaluate machine learning algorithms for use in brain-computer interfaces. *IEEE Transactions on neural systems and rehabilitation engineering*, 11(2):184–185, 2003.

[142] Siavash Sakhavi and Cuntai Guan. Convolutional neural network-based transfer learning and knowledge distillation using multi-subject data in motor imagery BCI. *International IEEE/EMBS Conference on Neural Engineering, NER*, pages 588–591, 2017.

[143] Siavash Sakhavi, Cuntai Guan, and Shuicheng Yan. Parallel convolutional-linear neural network for motor imagery classification. *2015 23rd European Signal Processing Conference, EUSIPCO 2015*, pages 2736–2740, 2015.

[144] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved Techniques for Training GANs. In *Advances in Neural Information Processing Systems*, pages 2234–2242, 2016.

[145] George Saon, Gakuto Kurata, Tom Sercu, Kartik Audhkhasi, Samuel Thomas, Dimitrios Dimitriadis, Xiaodong Cui, Bhuvana Ramabhadran, Michael Picheny, Lynn-Li Lim, et al. English conversational telephone speech recognition by humans and machines. *arXiv preprint arXiv:1703.02136*, 2017.

[146] Gerwin Schalk, Dennis J McFarland, Thilo Hinterberger, Niels Birbaumer, and Jonathan R Wolpaw. Bci2000: a general-purpose brain-computer interface (bci) system. *IEEE Transactions on biomedical engineering*, 51(6):1034–1043, 2004.

[147] Robin Tibor Schirrmeister, Lukas Gemein, Katharina Eggensperger, Frank Hutter, and Tonio Ball. Deep learning with convolutional neural networks for decoding and visualization of EEG pathology. *arXiv preprint*, 2017.

[148] Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for EEG decoding and visualization. *Human Brain Mapping*, 38(11):5391–5420, 2017.

[149] Jürgen Schmidhuber. Deep Learning in neural networks: An overview: read section 6.6. *Neural Networks*, 61:85–117, 2015.

[150] Donald L Schomer and Fernando Lopes Da Silva. *Niedermeyer's electroencephalography: basic principles, clinical applications, and related fields*. Lippincott Williams & Wilkins, 2012.

[151] Justus T. C. Schwabedal, John C. Snyder, Ayse Cakmak, Shamim Nemati, and Gari D. Clifford. Addressing Class Imbalance in Classification Problems of Noisy Signals by using Fourier Transform Surrogates. *arXiv preprint*, pages 1–7, 2018.

[152] Vinit Shah, Meysam Golmohammadi, Saeedeh Ziyabari, Eva Von Weltin, Iyad Obeid, and Joseph Picone. Optimizing channel selection for seizure detection. In *Signal Processing in Medicine and Biology Symposium (SPMB), 2017 IEEE*, pages 1–5. IEEE, 2017.

[153] Jared Shamwell, Hyungtae Lee, Heesung Kwon, Amar R Marathe, Vernon Lawhern, and William Nothwang. Single-trial EEG RSVP classification using convolutional neural networks. In *Micro-and Nanotechnology Sensors, Systems, and Applications VIII*, volume 9836, page 983622, 2016.

[154] Jennifer Shang, Huang Yuanyue, Guo Haixiang, Li Yijing, Gu Mingyun, and Bing Gong. Learning from class-imbalanced data: Review of methods and applications. *Expert Systems With Applications*, 73:220–239, 2017.

[155] Ali H Shoeb and John V Guttag. Application of machine learning to epileptic seizure detection. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 975–982, 2010.

[156] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv preprint*, 2014.

[157] Jasper Snoek, Hugo Larochelle, and Ryan P. Adams. Practical Bayesian Optimization of Machine Learning Algorithms. In *Advances in neural information processing systems*, pages 2951–2959, 2012.

[158] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3(1):42–55, 2012.

[159] Arnaud Sors, Stéphane Bonnet, Sébastien Mirek, Laurent Vercueil, and Jean François Payen. A convolutional neural network for sleep stage scoring from raw single-channel EEG. *Biomedical Signal Processing and Control*, 42(April 2018):107–114, 2018.

[160] C Spampinato, S Palazzo, I Kavasidis, D Giordano, N Souly, and M Shah. Deep learning human mind for automated visual classification. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-Janua:4503–4511, 2017.

[161] Sebastian Stober, Daniel J Cameron, and Jessica a Grahn. Using Convolutional Neural Networks to Recognize Rhythm Stimuli from Electroencephalography Recordings. In *Neural Information Processing Systems (NIPS) 2014*, pages 1–9, 2014.

[162] Sebastian Stober, Avital Sternin, Adrian M Owen, and Jessica A Grahn. Deep Feature Learning for EEG Recordings. *arXiv preprint*, 2015.

[163] Irene Sturm, Sebastian Bach, Wojciech Samek, and Klaus-Robert Müller. Interpretable Deep Neural Networks for Single-Trial EEG Classification. *arXiv preprint*, 33518:1–5, 2016.

[164] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision*, pages 443–450. Springer, 2016.

[165] Pengfei Sun and Jun Qin. Neural networks based EEG-Speech Models. *arXiv preprint*, pages 1–10, 2016.

[166] Akara Supratak, Hao Dong, Chao Wu, and Yike Guo. DeepSleepNet: a Model for Automatic Sleep Stage Scoring based on Raw Single-Channel EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(11):1998–2008, 2017.

[167] Ilya Sutskever, Geoffrey Hinton, Alex Krizhevsky, and Ruslan R Salakhutdinov. Dropout : A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[168] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[169] Yousef Rezaei Tabar and Ugur Halici. A novel deep learning approach for classification of EEG motor imagery signals. *Journal of Neural Engineering*, 14(1):16003, 2016.

[170] Sachin S Talathi. Deep Recurrent Neural Networks for seizure detection and early seizure detection systems. *arXiv preprint*, 2017.

[171] Arwa M Taqi, Fadwa Al-Azzo, M Mariofanna, and Jassim M Al-Saadi. Classification and discrimination of focal and non-focal EEG signals based on deep neural network. *2017 International Conference on Current Research in Computer Science and Information Technology (ICCIT)*, pages 86–92, 2017.

[172] Jason Teo, Chew Lin Hou, and James Mountstephens. Preference Classification Using Electroencephalography (EEG) and Deep Learning. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, 10(1):87–91, 2018.

[173] Pierre Thodoroff, Joelle Pineau, and Andrew Lim. Learning Robust Features using Deep Learning for Automatic Seizure Detection. *arXiv preprint*, pages 1–12, 2016.

[174] O Zander Thorsten and Kothe Christian. Towards passive brain–computer interfaces: applying brain–computer interface technology to human–machine systems in general. *Journal of Neural Engineering*, 8(2):25005, 2011.

[175] Schirrmeister Robin Tibor, Springenberg Jost Tobias, Fiederer Lukas Dominique Josef, Glasstetter Martin, Eggensperger Katharina, Tangermann Michael, Hutter Frank, Burgard Wolfram, and Ball Tonio. Deep learning with convolutional neural networks for EEG decoding and visualization. *Human Brain Mapping*, 38(11):5391–5420, 2017.

[176] Tijmen Tieleman, Geoffrey E. Hinton, Nitish Srivastava, and Kevin Swersky. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning*, 4(2):26—-31, 2012.

[177] R K Tripathy and U Rajendra Acharya. Use of features from RR-time series and EEG signals for automated classification of sleep stages in deep neural network framework. *Biocybernetics and Biomedical Engineering*, pages 1–13, 2018.

[178] Nhan Duy Truong, Levin Kuhlmann, Mohammad Reza Bonyadi, and Omid Kavehei. Semi-supervised Seizure Prediction with Generative Adversarial Networks. *arXiv preprint*, pages 1–6, 2018.

[179] Nhan Duy Truong, Anh Duy Nguyen, Levin Kuhlmann, Mohammad Reza Bonyadi, Jiawei Yang, Samuel Ippolito, and Omid Kavehei. Convolutional neural networks for seizure prediction using intracranial and scalp electroencephalogram. *Neural Networks*, 105:104–111, 2018.

[180] Orestis Tsinalis, Paul M. Matthews, Yike Guo, and Stefanos Zafeiriou. Automatic Sleep Stage Scoring with Single-Channel EEG Using Convolutional Neural Networks. *arXiv preprint*, 2016.

[181] Kostas M Tsiouris, Vasileios C Pezoulas, Michalis Zervakis, Spiros Konitsiotis, Dimitrios D Koutsouris, and Dimitrios I Fotiadis. A Long Short-Term Memory deep learning network for the prediction of epileptic seizures using EEG signals. *Computers in Biology and Medicine*, 99:24–37, 2018.

[182] J T Turner, Adam Page, Tinoosh Mohsenin, and Tim Oates. Deep Belief Networks used on High Resolution Multichannel Electroencephalography Data for Seizure Detection. *AAAI Spring Symposium Series*, pages 75–81, 2014.

[183] Ihsan Ullah, Muhammad Hussain, Emad-Ul-Haq Qazi, and Hatim Aboalsamh. An Automated System for Epilepsy Detection using EEG Brain Signals based on Deep Learning Approach. *Arxiv*, 2018.

[184] Jose Antonio Urigen and Bego{\˜{n}}a Garcia-Zapirain. EEG artifact removal state-of-the-art and guidelines. *Journal of Neural Engineering*, 12(3):031001, 2015.

[185] Michel J A M Van Putten, Sebastian Olbrich, and Martijn Arns. Predicting sex from brain rhythms with deep learning. *Scientific Reports*, 8(1):1–7, 2018.

[186] Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. OpenML: networked science in machine learning. *SIGKDD Explorations*, 15(2):49–60, 2014.

[187] Albert Vilamala, Kristoffer H Madsen, and Lars K Hansen. Deep Convolutional Neural Networks for Interpretable Analysis of EEG Sleep Stage Scoring. *arXiv preprint*, (659860), 2017.

[188] Martin Volker, Jiri Hammer, Robin T Schirrmeister, Joos Behncke, Lukas DJ Fiederer, Andreas Schulze-Bonhage, Petr Marusic, Wolfram Burgard, and Tonio Ball. Intracranial error detection

via deep learning. In *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 568–575. IEEE, 2018.

[189] Martin Völker, Robin T Schirrmeister, Lukas D J Fiederer, Wolfram Burgard, and Tonio Ball. Deep Transfer Learning for Error Decoding from Non-Invasive EEG. In *Brain-Computer Interface (BCI), 2018 6th International Conference on*, pages 1–6. IEEE, 2017.

[190] Fang Wang, Sheng Hua Zhong, Jianfeng Peng, Jianmin Jiang, and Yan Liu. Data augmentation for eeg-based emotion recognition with deep convolutional neural networks. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 10705 LNCS, pages 82–93, 2018.

[191] Nicholas R Waytowich, Vernon Lawhern, Javier O Garcia, Jennifer Cummings, Josef Faller, Paul Sajda, and Jean M Vettel. Compact Convolutional Neural Networks for Classification of Asynchronous Steady-state Visual Evoked Potentials. *arXiv preprint*, pages 1–21, 2018.

[192] Tingxi Wen and Zhongnan Zhang. Deep Convolution Neural Network and Autoencoders-Based Unsupervised Feature Learning of EEG Signals. *IEEE Access*, 6:25399–25410, 2018.

[193] Mark D Wilkinson. Comment: The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3:1–9, 2016.

[194] Zhenglin Wu, Haohan Wang, Mingze Cao, Yin Chen, and Eric P Xing. Fair Deep Learning Prediction for Healthcare Applications with Confounder Filtering. *arXiv preprint*, pages 1–17, 2018.

[195] D. F. Wulsin, J. R. Gupta, R. Mani, J. A. Blanco, and B. Litt. Modeling electroencephalography waveforms with semi-supervised deep belief nets: Fast classification and anomaly measurement. *Journal of Neural Engineering*, 8(3), 2011.

[196] Songyun Xie, Yabing Li, Xinzhou Xie, Wei Wang, and Xu Duan. The Analysis and Classify of Sleep Stage Using Deep Learning Network from Single-Channel EEG Signal. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10637 LNCS:752–758, 2017.

[197] Haiyan Xu and Konstantinos N. Plataniotis. Affective states classification using EEG and semi-supervised deep learning approaches. *2016 IEEE 18th International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6, 2016.

[198] Banghua Yang, Kaiwen Duan, Chengcheng Fan, Chenxiao Hu, and Jinlong Wang. Automatic ocular artifacts removal in EEG using deep learning. *Biomedical Signal Processing and Control*, 43:148–158, 2018.

[199] Banghua Yang, Kaiwen Duan, and Tao Zhang. Removal of EOG artifacts from EEG using a cascade of sparse autoencoder and recursive least squares adaptive filter. *Neurocomputing*, 214:1053–1060, 2016.

[200] Huijuan Yang, Siavash Sakhavi, Kai Keng Ang, and Cuntai Guan. On the use of convolutional neural networks and augmented CSP features for multi-class motor imagery of EEG signals classification. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2015-Novem:2620–2623, 2015.

[201] S. Yang, Meysam Golmohammadi, Iyad Obeid, and Joseph Picone. Semi-automated annotation of signal events in clinical EEG data, Engineering Data Consortium , Temple University , Philadelphia , Pennsylvania , USA. *Signal Processing in Medicine and Biology Symposium*, pages 1–5, 2016.

[202] Antonio Jimeno Yepes, Jianbin Tang, and Benjamin Scott Mashford. Improving classification accuracy of feedforward neural networks for spiking neuromorphic chips. *IJCAI International Joint Conference on Artificial Intelligence*, pages 1973–1979, 2017.

[203] Zhong Yin and Jianhua Zhang. Recognition of Cognitive Task Load levels using single channel EEG and Stacked Denoising Autoencoder. In *Chinese Control Conference, CCC*, volume 2016-Augus, pages 3907–3912. IEEE, jul 2016.

[204] Zhong Yin and Jianhua Zhang. Cross-session classification of mental workload levels using EEG and an adaptive deep learning model. *Biomedical Signal Processing and Control*, 33:30–47, 2017.

[205] Zhong Yin and Jianhua Zhang. Cross-subject recognition of operator functional states via EEG and switching deep belief networks with adaptive weights. *Neurocomputing*, 260:349–366, 2017.

[206] Dani Yogatama, Chris Dyer, Wang Ling, and Phil Blunsom. Generative and discriminative text classification with recurrent neural networks. *arXiv preprint arXiv:1703.01898*, 2017.

[207] Jaehong Yoon, Jungnyun Lee, and Mincheol Whang. Spatial and Time Domain Feature of ERP Speller System Extracted via Convolutional Neural Network. *Computational Intelligence and Neuroscience*, 2018, 2018.

[208] Ye Yuan, Guangxu Xun, Fenglong Ma, Qiuling Suo, Hongfei Xue, Kebin Jia, and Aidong Zhang. A Novel Channel-aware Attention Framework for Multi-channel EEG Seizure Detection via Multi-view Deep Learning. *IEEE EMBS International Conference on Biomedical & Health Informatics*, (March):4–7, 2018.

[209] Raheel Zafar, Sarat C Dass, and Aamir Saeed Malik. Electroencephalogram-based decoding cognitive states using convolutional neural network and likelihood ratio based score fusion. *arXiv preprint*, pages 1–23, 2017.

[210] Matthew D. Zeiler. ADADELTA: An Adaptive Learning Rate Method. *arXiv preprint*, 2012.

[211] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint*, 2016.

[212] Dalin Zhang, Lina Yao, Xiang Zhang, Sen Wang, Weitong Chen, and Robert Boots. Cascade and Parallel Convolutional Recurrent Neural Networks on EEG-Based Intention Recognition for Brain Computer Interface. In *Thirty-Second AAAI Conference on Artificial Intelligence*, pages 1703–1710, 2018.

[213] Guo-Qiang Zhang, Licong Cui, Remo Mueller, Shiqiang Tao, Matthew Kim, Michael Rueschman, Sara Mariani, Daniel Mobley, and Susan Redline. The national sleep research resource: towards a sleep data commons. *Journal of the American Medical Informatics Association*, 25(10):1351–1358, 2018.

[214] Jianhua Zhang, Sunan Li, and Rubin Wang. Pattern recognition of momentary mental workload based on multi-channel electrophysiological data and ensemble convolutional neural networks. *Frontiers in Neuroscience*, 11(MAY):1–16, 2017.

[215] Qiqi Zhang and Ying Liu. Improving brain computer interface performance by data augmentation with conditional Deep Convolutional Generative Adversarial Networks. *arXiv preprint*, 2018.

[216] Tong Zhang, Wenming Zheng, Zhen Cui, Yuan Zong, and Yang Li. Spatial-Temporal Recurrent Neural Network for Emotion Recognition. *IEEE Transactions on Cybernetics*, 1:1–9, 2018.

[217] X. Zhang, L. Yao, Q. Z. Sheng, S. S. Kanhere, T. Gu, and D. Zhang. Converting your thoughts to texts: Enabling brain typing via deep feature learning of eeg signals. In *2018 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 1–10, March 2018.

[218] Xiang Zhang, Lina Yao, Kaixuan Chen, Xianzhi Wang, Quanz. Sheng, and Tao Gu. DeepKey: An EEG and Gait Based Dual-Authentication System. *arXiv preprint*, 9(4):1–20, 2017.

[219] Xiang Zhang, Lina Yao, Chaoran Huang, Quan Z Sheng, and Xianzhi Wang. Intent Recognition in Smart Living Through Deep Recurrent Neural Networks. *arXiv preprint*, pages 1–11, 2017.

[220] Xiang Zhang, Lina Yao, Salil S. Kanhere, Yunhao Liu, Tao Gu, and Kaixuan Chen. Mindid: Person identification from brain waves through attention-based recurrent neural network. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2(3):149:1–149:23, September 2018.

[221] Xiang Zhang, Lina Yao, Xianzhi Wang, Wenjie Zhang, Shuai Zhang, and Yunhao Liu. Know Your Mind: Adaptive Brain Signal Classification with Reinforced Attentive Convolutional Neural Networks. *arXiv preprint*, 2018.

[222] Xiang Zhang, Lina Yao, Dalin Zhang, Xianzhi Wang, Quan Z. Sheng, and Tao Gu. Multi-person brain activity recognition via comprehensive eeg signal analysis. In *Proceedings of the 14th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, MobiQuitous 2017, pages 28–37, New York, NY, USA, 2017. ACM.

[223] Wei-Long Zheng, Wei Liu, Yifei Lu, Bao-Liang Lu, and Andrzej Cichocki. Emotionmeter: A multimodal framework for recognizing human emotions. *IEEE Transactions on Cybernetics*, (99):1–13, 2018.

[224] Wei Long Zheng and Bao Liang Lu. Investigating Critical Frequency Bands and Channels for EEG-Based Emotion Recognition with Deep Neural Networks. *IEEE Transactions on Autonomous Mental Development*, 7(3):162–175, 2015.

[225] Jie Zhou and Wei Xu. End-to-end learning of semantic role labeling using recurrent neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1127–1137, 2015.

List of acronyms

AE	autoencoder.
BCI	brain-computer interface.
CCA	canonical correlation analysis.
CNN	convolutional neural network.
CV	computer vision.
DBN	deep belief network.
DL	deep learning.
DNN	deep neural network.
ECoG	electrocorticography.
EEG	electroencephalography.
EMG	electromyography.
EOG	electroculography.
ERP	event-related potential.
FC	fully-connected.
GAN	generative adversarial network.
ICA	independent component analysis.
LSTM	long short-term memory.
MEG	magnetoencephalography.
NLP	natural language processing.
PCA	principal component analysis.
PSD	power spectral density.
RBM	restricted Boltzmann machine.
RNN	recurrent neural network.
ROC	AUC area under the receiver operating curve.
RSVP	rapid serial visual presentation.
SDAE	stacked denoising autoencoder.
SGD	stochastic gradient descent.
SNR	signal-to-noise ratio.
STFT	short-time Fourier transform.
SVM	support vector machine.
wICA	wavelet-enhanced independent component analysis.

Checklist of items to include in a DL-EEG study

This section contains a checklist of items we believe DL-EEG papers should mention to ensure their published results are readily reproducible. The following information should all be clearly stated at one point or another in the text or supplementary materials of future DL-EEG studies:
Data.

- ☐ Number of subjects (and relevant demographic data)
- ☐ Electrode montage including reference(s) (number of channels and their locations)
- ☐ Shape of one example (e.g., "256 samples × 16 channels")
- ☐ Data augmentation technique (e.g., percentage of overlap for sliding windows)
- ☐ Number of examples in training, validation and test sets

EEG processing.

- ☐ Temporal filtering, if any
- ☐ Spatial filtering, if any
- ☐ Artifact handling techniques, if any
- ☐ Resampling, if any

Neural network architecture.

- ☐ Architecture type
- ☐ Number of layers (consider including a diagram or table to represent the architecture)
- ☐ Number of learnable parameters

Training hyperparameters.

- ☐ Parameter initialization
- ☐ Loss function
- ☐ Batch size
- ☐ Number of epochs
- ☐ Stopping criterion
- ☐ Regularization (e.g., dropout, weight decay, etc.)
- ☐ Optimization algorithm (e.g., stochastic gradient descent, Adam, RMSProp, etc.)
- ☐ Learning rate schedule and optimizer parameters
- ☐ Values of **all** hyperparameters (including random seed) for the results that are presented in the paper

- ☐ Hyperparameter search method

Performance and model comparison.

- ☐ Performance metrics (e.g., f1-score, accuracy, etc.)
- ☐ Type of validation scheme (intra- vs. inter-subject, leave-one-subject-out, k-fold cross-validation, etc.)
- ☐ Description of baseline models (thorough description or reference to published work)

Fourth Article.

Passive EEG Brain-Computer Interface (BCI) for a 3D Multiple Object Tracking (3D-MOT) task.

by

Yannick Roy[1], and Jocelyn Faubert[1]

([1]) Université de Montréal

This article was submitted in Computational Intelligence and Neuroscience.

Résumé. Les tâches de suivi d'objets multiples (MOT en anglais) ont été utilisé abondamment en psychophysique et plus récemment pour l'entrainement cognitif. Ici, nous explorons le potentiel de fermer la boucle en créant un interface cerveau-machine de type passif. Nous avons utilisé les données EEG d'une étude précédente pour tenter de classifier si l'activité cérébrale a lieu durant la phase de suivi ou de rappel. Nous avons aussi tenté de classifier les essais latéralisés où les cibles sont présentées soit à gauche ou à droite. Pour la classification de la phase, nous avons obtenus 80% avec un entraînement à travers tous les sujets confondus en utilisant les électrodes frontales et les bandes de fréquence delta et thêta. Pour la classification du côté, nous avons obtenu une moyenne de 68% en entraînant un model différent par sujet, en utilisant la hauteur moyenne du signal.
Mots clés : Suivi d'objets multiples, MOT, EEG, ICM

Abstract. Multiple Object Tracking (MOT) tasks have been used extensively in psychophysics and more recently in the context of cognitive training. Here, we explore the potential for closing the loop and creating a passive BCI with a 3D-MOT task. We used the EEG data from our previous study on a 3D-MOT Task to classify EEG activity to predict if such activity was happening during the tracking or the recall phase of the 3D-MOT task. We also trained a classifier for lateralized trials to predict if the targets were presented on the left or right hemifield using EEG brain activity. For the phase classification between tracking and recall, we obtained 80% accuracy when training a SVM across subjects using the theta and delta frequency band power from the frontal electrodes and 83% accuracy when training within subjects. For the side classification we obtained an average accuracy of 68% when training within subjects in the time domain using the mean amplitude.
Keywords: MOT, EEG, BCI

1. Introduction

Multiple Object Tracking (MOT) is a paradigm that has been heavily studied both in human cognition and computer vision. In psychophysics, researchers often use a variant of the multiple-object tracking (MOT) task developed by Pylyshyn &

Storm in 1988 ([**6**]). The task is simple and usually consists of a few targets the participant has the track among distractors. The shape, the color, the number of targets and distractors, the pathing of the moving objects, the speed and the length of the trial are all parameters researchers modulate to generate a MOT task that allows them to address their research question. While the task itself is rather simple, it engages complex neural mechanisms and systems, which also makes the task a great candidate for cognitive training ([**1**]). Our ability to track multiple objects in a dynamic environment enables us to perform everyday tasks such as driving, playing team sports, and walking in a shopping mall. While MOT tasks have been used extensively in research and even commercially as a cognitive training tool, our understanding of the underlying roles and relationship between attention and working memory remains vague. Here, we hypothesize that it is possible to use machine learning to classify brain activity via electroencephalography (EEG) to distinguish at least two different phases of a 3D-MOT task as well as the hemifield in which the targets are presented. Our previous research ([**8**]) showed very different brain activity patterns during tracking than during the recall phase of the task. The tracking phase seems to rely way more on attention mechanisms while the recall phase relies more on working memory processes. Given these previous results, we believe that a passive brain-computer interface could be developed to enhance the experience of a 3D-MOT task.

The field of brain-computer interface (BCI) has come a long way since it's inception in the 1970s ([**9, 10**]). While the promise of brain-computer interfaces is still to give back some sort of communication and control to those suffering from severe neuromuscular disorders, such as amyotrophic lateral sclerosis, brainstem stroke, and spinal cord injury, as described in Jonathan Wolpaw's famous paper in the early 2000s ([**11**]), the field has now extended to various other use cases such as gaming, neuromarketing, smart appliances, and robotics, just to name a few ([**4**]). In his 2011 paper, Thorsten Zander ([**12**]) claimed that passive BCIs will be the first ones

to become pervasive. Being able to passively monitor brain activity and feed such information back into the task to offer a seamless experience between our brain and the digital world will open a new world of opportunities. In domains such as education and cognitive training, this closed-loop system could lead to new paradigms allowing to modulate the task in real time to maximize both the engagement and the learning experience.

2. Methods

Here we used the data from our previous study on a 3D-MOT Task ([**8**]). The full description of the task, the participants and the recording protocol is available in the original paper, here we will only provide a high-level overview of the dataset.

2.1. Task and Participants

NeuroTracker^{TM} is a commercially available 3D-MOT task currently used by a multitude of users in many countries around the world as a perceptual-cognitive training and assessment tool. Here, we used a modified laboratory version of the NeuroTracker^{TM}. Figure 56 shows the five different phases of the 3D-MOT task developed with the Unity engine. (A) *presentation phase* where 8 yellow spheres are shown in a 3D volume space for 2 seconds, (B) *indexing phase* where one, two or three spheres (targets) change colour (to red) and are highlighted (hallo) for 2 seconds, (C) *tracking (or movement) phase* where the targets indexed in phase 2 return to their original colour (yellow) and 1 second later start moving for 8 seconds crisscrossing and bouncing off of each other and the virtual 3D volume cube walls, (D) *recall phase* where the spheres stop moving and the observer is prompted to identify the spheres originally indexed in phase 2. Each sphere is labelled with a number between 1 to 8. After identifying the targets, the observer is asked to provide a confidence level for each answer (either 0%, 25%, 75% or 100% confident). And finally, (E) *feedback phase*

where the correct targets are clearly identified on the screen. The whole trial takes around 15s (2s + 2s + 9s + [1-4]s) depending on how long the participant takes to provide the answers.

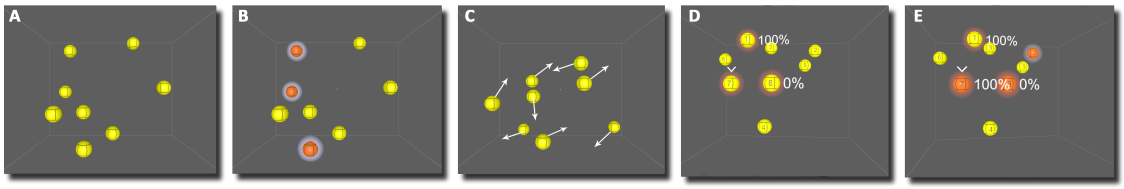


Fig. 56. 3D-MOT Task Sequence. (A) All spheres appear on screen. (B) Targets are highlighted in red for 2 seconds. (3) All the spheres are moving for 8 seconds. (D) Participant must identify the targets and provide a confidence level. (E) Feedback is provided to the participant showing the correct answers.

Twenty-four participants (thirteen females) aged between 21 and 41 years of age (M=29.3, SD=4.9) took part in this study. The experiment consisted of 4 blocks of 21 trials with 2 conditions: side and set size. The speed was kept constant. In total, 30 trials were presented in the left hemifield, 10 for each set size (1,2,3), 30 trials were presented in the right hemifield, 10 for each set size (1,2,3), and 24 trials were not lateralized and the targets could freely cross from left to right and vice-versa, 8 for each set size (1,2,3). Once the targets stopped moving, a number between 1 and 8 appeared on each of the spheres and the participants had to provide their answer by saying the number of the target(s) out loud for the instructor to enter the answers.

2.2. EEG Acquisition

The electroencephalogram (EEG), electrocardiogram (ECG), and electrooculogram (EOG) were recorded using the Biosemi ActiveTwo system (Biosemi, Amsterdam, Netherlands) with 71 Ag–Ag/Cl electrodes positioned at 64 standard International 10/20 System sites (EEG), left and right mastoids for offline EEG re-reference, 1cm lateral to the external canthi for horizontal EOG (HEOG), left and right ribs plus

right collarbone for ECG. The HEOG was used for eye movements to confirm that the participant was tracking the targets with covert and not overt attention (i.e. not moving their eyes). Electrophysiological signals were digitized at 2048Hz.

2.3. EEG Analysis

Two different pipelines were made based on results from our previous study. For the phase classification, features from a time-frequency decomposition were obtained and for the side classification, features from the time domain using a bipolar electrode configuration were used. All the analysis was performed in Python using MNE-Python ([**2**]) and scikit-learn ([**5**]) toolboxes. Like in the original publication, four participants (out of the twenty-four) were removed from the analysis because of too much eye movements during the trials or too much noise in the EEG data.

Phase classification. For the phase classification (i.e. ID vs Tracking vs Recall), the EEG channels were first re-referenced to the left and right mastoids. Second, independent component analysis (ICA) was used to remove eye blinks and eye movement artifacts. Third, the EEG data was epoched in [-1, 15]s windows where t=0s represents the stimuli/trigger of the spheres being highlighted in red. Fourth, AutoReject ([**3**]) was used to automatically remove bad trials and correct bad channels. The time-frequency decomposition was computed using Morlet wavelets for frequencies between 1 and 50Hz with varying cycles of half the frequency. The frequency features were then obtained by getting the log ratio of the power relative to the baseline power. The baseline was selected as -1s to 0s prior to the targets being colored in red (t=0.15s). Instead of using raw power, the log ratio has the advantage of normalizing the power across participants. 1s EEG segments were used to extract relevant features sent to the machine learning classifier. Given the low amount of trials available, we opted for traditional machine learning approach and not novel

deep learning models ([**7**]). We compared linear discriminant analysis (LDA) and support vector machine (SVM) models.

Side classification. For the side classification (i.e. left vs right), the EEG channels were used with a bipolar configuration, matching their opposite channels (e.g. F3 with F4, O1 with O2, etc.) the midline channels weren't used. No extra preprocessing step was done. Segments of EEG were extracted from different pairs of electrodes and the mean amplitude of the segments was added to the feature matrix.

3. Results

3.1. Phase classification

Across subjects. Using 1s EEG segment from each phase we obtained an average accuracy of 79.84% on a two-classes classification problem (*Tracking* vs *Recall*) with a SVM model using a 10-fold cross-validation across subjects. Chance level is 50%. Time-frequency features in the delta and theta bands provided the best results. The 1s window from tracking was taken arbitrarily between 5 to 6s and the 1s segment from recall was taken from 11.5s to 12.5s, when recall activity is the strongest. The time-frequency features of the two 1s segments were extracted for each trial of each subject and concatenated into a feature matrix. Different channels and combinations of channels were tested and the best results were obtained with AFz and Fpz. The resulting feature matrix was of dimension 4047 (trials) by 2 (channels) by 7 (frequencies) by 256 (time points). The classes were balanced as there was 1 sample of each phase for each trial. The classification (*ID* vs *Tracking* and *ID* vs *Recall*) yielded results too close to chance level (50%) to be included here. The three-classes classification (i.e. *ID* vs *Tracking* vs *Recall*) had an accuracy of 56.33% where the chance level is 33%.

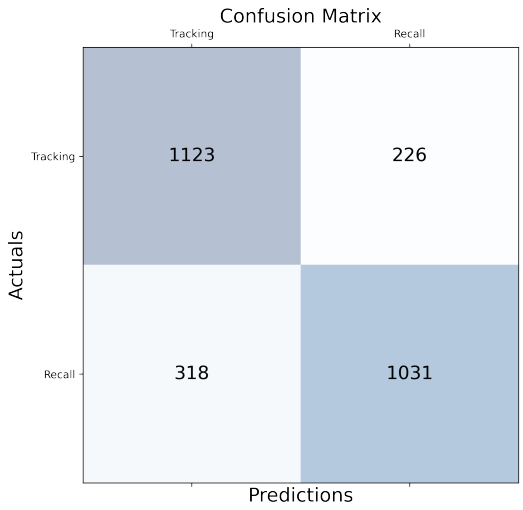


Fig. 57. Confusion Matrix. Tracking vs Recall classification.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = 79.84\%$$

$$Sensitivity = \frac{TP}{TP + FN} = 77.93\%$$

$$Specificity = \frac{TN}{FP + TN} = 82.02\%$$

Within subjects. The same approach was used for within subjects, training a different model for each participants. The average accuracy on two-classes (tracking vs recall) with a SVM model and a 5-fold cross-validation for each subject is 83.31% (std: 7.64%, min: 61.57%, max: 94.62%).

3.2. Side classification

The best results were obtained using a 200ms segment from the ID phase when the targets are identified in red ([0.4-0.6]s) as well as a 1s EEG segment from the recall phase ([11.5-12.5]s) and concatenating them in the feature space. We obtained an grand average accuracy of 68.1% on a two-classes classification problem (left vs right) with a SVM model using a 5-fold cross-validation on each subject. The best combination of channels was: O1/2, PO3/4, FT7/8, and F5/6 in a bipolar configuration.

4. Discussion

Our results show that the implementation of a passive BCI for a 3D-MOT task would be possible and that distinct brain activity is happening during tracking vs during recall. In this proof of concept work, specific 1s segment of EEG data were selected (between 5 and 6s for tracking and 11.5s to 12.5s for recall), however this is not fully representative of what an online BCI implementation would be, since new 1s windows would be continually sampled and classified. The classes would also be unbalanced as most of the time is spent in the tracking phase. The trained model would have to account for such bias.

Moreover, it is worth noting that we also used the mean and median from the time-frequency features in order to reduce the dimensionality of the feature matrix, however we obtained better results with the raw time-frequency decomposition.

For the side classification, we can see in Table 10 that we could predict the side with high-level of accuracy for half the subjects, however the other half obtained a poor classification accuracy. We did not correlate the classification performance with the actual behavioural performance.

Here we used basic machine learning models to obtain our results, yet more advance techniques might yield better results. Single trial EEG classification is challenging

Participant ID	Accuracy (%)
F26F2	97
T25F1	83
M22F2	83
T18M1	82
F12F2	85
M8M2	78
T18F2	78
F19M1	77
W10F2	77
S20M1	73
T9F2	67
T23F1	65
T23F2	65
F12M1	65
T11F1	53
M8M1	53
S21F1	52
W17M2	52
T16F2	47
F5M1	32
Avg	68.17
Min	32.00
Max	96.67
std	15.66

Table 10. Side classification accuracy of the SVM model for each subject using 5-fold cross-validation.

given the inherent poor signal-to-noise ratio (SNR) of the EEG signal. However, achieving 80% accuracy across subjects shows good potential for further investigation.

Finally, this work shows that it is possible to differentiate between the tracking and recall phases of a 3D-MOT task with a high level of accuracy from brain activity alone and to differentiate between lateralized trials where targets are presented either in the left hemifield or the right hemifield, and that across different subjects, even with new subjects the system hasn't trained on.

Both the code and the EEG data are available on github:
github.com/royyannick/3DMOT_BCI

5. Funding

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC-RDC) (reference number: RDPJ 514052-17) and an NSERC Discovery fund.

References

[1] Jocelyn Faubert and Lee Sidebottom. Perceptual-cognitive training of athletes. *Journal of Clinical Sport Psychology*, 6(1):85–102, 2012.

[2] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A. Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, and Matti S. Hämäläinen. MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, 7(267):1–13, 2013.

[3] Mainak Jas, Denis A Engemann, Yousra Bekhti, Federico Raimondo, and Alexandre Gramfort. Autoreject: Automated artifact rejection for meg and eeg data. *NeuroImage*, 159:417–429, 2017.

[4] Aleksandra Kawala-Sterniuk, Natalia Browarska, Amir Al-Bakri, Mariusz Pelc, Jaroslaw Zygarlicki, Michaela Sidikova, Radek Martinek, and Edward Jacek Gorzelanczyk. Summary of over fifty years with brain-computer interfaces—a review. *Brain Sciences*, 11(1):43, 2021.

[5] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher,

M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[6] Zenon W Pylyshyn and Ron W Storm. Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial vision*, 3(3):179–197, 1988.

[7] Yannick Roy, Hubert Banville, Isabela Albuquerque, Alexandre Gramfort, Tiago H Falk, and Jocelyn Faubert. Deep learning-based electroencephalography analysis: a systematic review. *Journal of neural engineering*, 16(5):051001, 2019.

[8] Yannick Roy and Jocelyn Faubert. Significant changes in eeg neural oscillations during different phases of three-dimensional multiple object tracking task (3d-mot). *arXiv preprint*, 2022.

[9] Jacques J Vidal. Toward direct brain-computer communication. *Annual review of Biophysics and Bioengineering*, 2(1):157–180, 1973.

[10] Jacques J Vidal. Real-time detection of brain events in eeg. *Proceedings of the IEEE*, 65(5):633–641, 1977.

[11] Jonathan R Wolpaw, Niels Birbaumer, Dennis J McFarland, Gert Pfurtscheller, and Theresa M Vaughan. Brain–computer interfaces for communication and control. *Clinical neurophysiology*, 113(6):767–791, 2002.

[12] Thorsten O Zander and Christian Kothe. Towards passive brain–computer interfaces: applying brain–computer interface technology to human–machine systems in general. *Journal of neural engineering*, 8(2):025005, 2011.

Conclusion

Over the last four chapters, we have demonstrated that (1) the CDA is a reliable neural correlate for working memory. (2) A significant hand-off between attention and working memory processes happens in the brain when switching from tracking to recall during a 3D-MOT task. (3) Deep learning has a lot of potential for the future of the field but has yet to show a real superior added value compared to simpler and well understood traditional approaches. (4) It is possible to distinguish between EEG brain activity occurring during tracking vs during recall with high-level accuracy using a basic machine learning model and low frequency spectral features.

When taken together, it shows that closing the loop in a cognitive training task is possible and that the task could eventually be modulated in real-time based on brain activity. Unfortunately, this work hasn't closed the loop in a real-time BCI context, however we will continue to work in this direction given the encouraging results we have obtained in the different studies published in this thesis.

While single trial EEG classification is the ultimate objective, it remains a very difficult challenge due to EEG inherent poor signal-to-noise ratio. This is one of the reason why the most popular BCI paradigms like motor imagery and P300 speller relies on averaging a few trials or gathering a few seconds before making a decision. Unfortunately, not all applications and paradigms are suitable for averaging multiple trials. In our case, to mitigate our misclassification while trying to classify in real-time the phases of the task, the laterality of the presented targets and ultimately the

number of targets, we could for example, cumulate the decisions from the classifier using a sliding window of EEG during the trial to become more and more confident. Or to reverse the initial decision with accumulating evidences. Moreover, a confidence level should be provided so that the system can use the classifier output when the confidence level is high, and not rely on the BCI and brain input when the confidence is low. This hybrid approach might help alleviate false positive while still providing added value when the confidence level is high. The exact amount of trials and the time it takes the BCI system to make a decision is not a one-size-fits-all straightforward answer. It depends on the context and the severity of the task, for example the error rate deemed acceptable when operating a robotic arm bringing hot coffee near your mouth isn't the same as for a casual BCI controlling a smart TV. While the field keeps evolving, BCI applications need to find a good balance between type I error, type II error and the time it takes for the BCI to make a decision.

Given the inherent low SNR of EEG, it might never be possible to obtain a very high-level (85%+) on all subjects and all trials. As we have demonstrated here, some people had brain activity easier to classify, while some others had very poor classifier performances. Perhaps having a quick profiling test to see the subject responds well or not to the classifier prior to doing the task. For example, in a real world application, perhaps the BCI closed-loop system could be tuned down or disable for people that scored poorly on the profiling test and activate the BCI closed-loop feature for people who scored higher. The previous chapters have shown that indeed the signal is there and when averaged across several trials can be easily differentiated so perhaps having more trials would allow to leverage deep learning to perform better, however, 24 participants with 82 trials each wasn't enough to provide good performances.

While we are still unable to conclude how working memory and attention work together and interlace in a task such as 3D-MOT, it is clear the the different phases of the task activate different brain mechanisms. More work in that direction would be required to isolate both mechanisms. The phase classification across subjects with

80% accuracy is encouraging and shows that indeed there is a strong change in lower frequencies, almost like an on/off switch when transitioning from tracking to recall.

Given the interest (and citations) on our deep learning review paper and the number of new DL-EEG papers published since 2019, it is clear that the trend to leverage AI and deep learning for EEG isn't fading away but only growing. In order to continue contributing to the DL-EEG field we did put together, as we'd initially mention in our review paper, an online portal as a resource for researchers in the field. The DL-EEG portal is a community-driven platform that keeps track of the scientific literature on deep learning and electroencephalography. It is both a database of published results in DL-EEG and a tool that simplifies entering metadata from DL-EEG studies. The main mission of the portal is to foster reproducibility in DL-EEG research. The project was unfortunately temporarily put on hold given the lack of funding but we certainly intend to resume working on the project shortly given the very positive response we received for our proof of concept (available at dl-eeg.com).