Enabling real-world EEG applications with deep learning

Hubert Banville

université PARIS-SACLAY

Enabling real-world EEG applications with deep learning

Apprentissage profond pour la mise en application de l'EEG en conditions réelles

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 580, Sciences et technologies de l'information et de la communication (STIC)
Spécialité de doctorat : Mathématiques et informatique
Graduate School : Informatique et sciences du numérique, Référent : Faculté des sciences d'Orsay

Thèse préparée dans l'unité de recherche Inria Saclay-Île-de-France (Université Paris-Saclay, Inria), sous la direction d'**Alexandre GRAMFORT**, directeur de recherche et le co-encadrement de **Denis-Alexander ENGEMANN**, chercheur

Thèse soutenue à Paris-Saclay, le 24 janvier 2022, par

Hubert BANVILLE

Composition du jury

Slim ESSID Professeur, Télécom Paris, Institut Polytechnique de Paris	Président
Maarten DE VOS Professeur, KU Leuven	Rapporteur & Examinateur
Guillaume DUMAS Professeur adjoint, Université de Montréal	Rapporteur & Examinateur
Suzanna BECKER Professeure, McMaster University	Examinatrice
Sebastian STOBER Professeur, Otto-von-Guericke-Universität Magdeburg	Examinateur
Alexandre GRAMFORT Directeur de recherche, Inria, Université Paris-Saclay	Directeur de thèse

THESE DE DOCTORAT
NNT : 2022UPASG005

Acknowledgments

My first coding sprint was a pretty fruitful week: not only did I get an inspiring hands-on introduction to the world of open source science, unbeknownst to me at the time, I also got to meet my two future PhD advisors, Alex and Denis. To Alex I extend my deepest gratitude for welcoming me to his lab, and supporting me through the challenges of this transatlantic thesis project. I admire his pragmatism and surgical effectiveness, and of course his dedication to the open source software community, which I can only aspire to emulate. I am also very grateful to Denis, whose scientific rigour, attention to detail and unfaltering enthusiasm made this thesis an enriching learning experience.

I am particularly grateful to my colleagues in the Parietal team at Inria. Despite spending less time than I initially hoped on that side of the ocean, their camaraderie and support made my time in France both pleasant and gratifying. Thanks to Sik, Hicham, Quentin, Antonia, David, Valentin, Omar, Maëliss, Kshitij, Guillaume F., Juan, Patricio, Mathurin, Pierre A., Olivier, Thomas M., Marine, Jérôme, Gaston and everyone else who I might have forgotten. Gratitude to Aapo, for his patience and his sense of humor.

This thesis would not have been possible without the continued support of the extraordinary people at InteraXon. I would like to thank the many mentors who I had the chance to learn from and who ultimately brought me to start a PhD. Thanks to Lou, who welcomed me to the team and supported me through the end of my MSc thesis. I am greatly indebted to Graeme, who on top of preparing me for a new chapter in France, provided incredible support in the inception of this thesis project, and played a crucial role in its realization. My deepest gratitude to Chris, whose unwavering passion, creative energy and leadership had a profound impact on me over the years and who was a constant source of motivation. Many thanks to Derek, who believed in this project and audaciously supported InteraXon's first ever sponsored thesis. I must also extend my gratitude to the research team members I had the chance to work with over the years, many of whom have provided great support and help during this thesis: Sean, Nicole, Subash, Aravind, Maurice, Oishe, Ben, Javier, Jonathan, Kry, George and Matt. Thanks to Tracy and Naseem, who were also of great help in setting up this project and ensuring its progressed smoothly.

I extend my gratitude to Sue for so kindly welcoming me to her lab as a visiting student at McMaster University, and to her students Saurabh, Yarden, Isaac and Lauren. Thanks to Yannick and Isabela, with whom it was a real pleasure to work on the literature review project and the following portal.

Finalement, je tiens à remercier ma famille pour leur soutien indéfectible à travers mes aventures torontoises et parisiennes. Merci à mes parents, Line et Guy, d'avoir toujours su m'encourager et d'entre autres n'avoir jamais hésité à venir me rendre visite en terre étrangère. Probablement le plus gros de tous ces mercis va à Colleen, ma compagne de tous les moments, qui m'a soutenu depuis les tout débuts de ce projet de thèse entre deux continents avec amour, patience, intelligence, générosité et humour.

Abstract

Our understanding of the brain has improved considerably in the last decades, thanks to groundbreaking advances in the field of neuroimaging. Now, with the invention and wider availability of personal wearable neuroimaging devices, such as low-cost mobile EEG, we have entered an era in which neuroimaging is no longer constrained to traditional research labs or clinics. "Real-world" EEG comes with its own set of challenges, though, ranging from a scarcity of labelled data to unpredictable signal quality and limited spatial resolution. In this thesis, we draw on the field of deep learning to help transform this century-old brain imaging modality from a purely clinical- and research-focused tool, to a practical technology that can benefit individuals in their day-to-day life.

First, we study how unlabelled EEG data can be utilized to gain insights and improve performance on common clinical learning tasks using self-supervised learning. We present three such self-supervised approaches that rely on the temporal structure of the data itself, rather than onerously collected labels, to learn clinically-relevant representations. Through experiments on large-scale datasets of sleep and neurological screening recordings, we demonstrate the significance of the learned representations, and show how unlabelled data can help boost performance in a semi-supervised scenario.

Next, we explore ways to ensure neural networks are robust to the strong sources of noise often found in out-of-the-lab EEG recordings. Specifically, we present Dynamic Spatial Filtering, an attention mechanism module that allows a network to dynamically focus its processing on the most informative EEG channels while de-emphasizing any corrupted ones. Experiments on large-scale datasets and real-world data demonstrate that, on sparse EEG, the proposed attention block handles strong corruption better than an automated noise handling approach, and that the predicted attention maps can be interpreted to inspect the functioning of the neural network.

Finally, we investigate how weak labels can be used to develop a biomarker of neurophysiological health from real-world EEG. We translate the brain age framework, originally developed using lab and clinic-based magnetic resonance imaging, to real-world EEG data. Using recordings from more than a thousand individuals performing a focused attention exercise or sleeping overnight, we show not only that age can be predicted from wearable EEG, but also that age predictions encode information contained in well-known brain health biomarkers, but not in chronological age.

Overall, this thesis brings us a step closer to harnessing EEG for neurophysiological monitoring outside of traditional research and clinical contexts, and opens the door to new and more flexible applications of this technology.

Keywords: Deep learning, representation learning, self-supervised learning, electroencephalography, neuroimaging, wearable neurotechnology

Résumé

Au cours des dernières décennies, les avancées révolutionnaires en neuroimagerie ont permis de considérablement améliorer notre compréhension du cerveau. Aujourd'hui, avec la disponibilité croissante des dispositifs personnels de neuroimagerie portables, tels que l'EEG mobile « à bas prix », une nouvelle ère s'annonce où cette technologie n'est plus limitée aux laboratoires de recherche ou aux contextes cliniques. Les applications de l'EEG en « conditions réelles » présentent cependant leur lot de défis, de la rareté des données étiquetées à la qualité imprévisible des signaux et leur résolution spatiale limitée. Dans cette thèse, nous nous appuyons sur le domaine de l'apprentissage profond afin de transformer cette modalité d'imagerie cérébrale centenaire, purement clinique et axée sur la recherche, en une technologie pratique qui peut bénéficier à l'individu au quotidien.

Tout d'abord, nous étudions comment les données d'EEG non étiquetées peuvent être mises à profit via l'apprentissage auto-supervisé pour améliorer la performance d'algorithmes d'apprentissage entraînés sur des tâches cliniques courantes. Nous présentons trois approches auto-supervisées qui s'appuient sur la structure temporelle des données elles-mêmes, plutôt que sur des étiquettes souvent difficiles à obtenir, pour apprendre des représentations pertinentes aux tâches cliniques étudiées. Par le biais d'expériences sur des ensembles de données à grande échelle d'enregistrements de sommeil et d'examens neurologiques, nous démontrons l'importance des représentations apprises, et révélons comment les données non étiquetées peuvent améliorer la performance d'algorithmes dans un scénario semi-supervisé.

Ensuite, nous explorons des techniques pouvant assurer la robustesse des réseaux de neurones aux fortes sources de bruit souvent présentes dans l'EEG hors laboratoire. Nous présentons le Filtrage Spatial Dynamique, un mécanisme attentionnel qui permet à un réseau de dynamiquement concentrer son traitement sur les canaux EEG les plus instructifs tout en minimisant l'apport des canaux corrompus. Des expériences sur des ensembles de données à grande échelle, ainsi que des données du monde réel, démontrent qu'avec l'EEG à peu de canaux, notre module attentionnel gère mieux la corruption qu'une approche automatisée de traitement du bruit, et que les cartes d'attention prédites reflètent le fonctionnement du réseau de neurones.

Enfin, nous explorons l'utilisation d'étiquettes faibles afin de développer un biomarqueur de la santé neurophysiologique à partir d'EEG collecté dans le monde réel. Pour ce faire, nous transposons à ces données d'EEG le principe d'âge cérébral, originellement développé avec l'imagerie par résonance magnétique en laboratoire et en clinique. À travers l'EEG de plus d'un millier d'individus enregistré pendant un exercice d'attention focalisée ou le sommeil nocturne, nous démontrons non seulement que l'âge peut être prédit à partir de l'EEG portable, mais aussi que ces prédictions encodent des informations contenues dans des biomarqueurs de santé cérébrale, mais absentes dans l'âge chronologique.

Dans l'ensemble, cette thèse franchit un pas de plus vers l'utilisation de l'EEG pour le suivi neurophysiologique en dehors des contextes de recherche et cliniques traditionnels, et ouvre la porte à de nouvelles applications plus flexibles de cette technologie.

Mots-clés : Apprentissage profond, apprentissage de représentations, apprentissage auto-supervisé, électroencéphalographie, neuroimagerie, neurotechnologie portable

List of Figures

List of Tables

List of Acronyms

AE autoencoder. 35, 36, 39, 40, 57

BCI brain-computer interface. 6, 7, 9

BOLD blood oxygenation level-dependent. 3

CBAM convolutional block attention module. 16

CNN convolutional neural network. 11, 13, 14, 16, 55, 57, 62, 66, 88

CPC contrastive predictive coding. 27, 29, 31–34, 36–40, 42, 44–47, 50, 51

CSP common spatial patterns. 60

CV computer vision. 15, 16, 19

DL deep learning. 2, 8, 9, 11, 19, 20

DNN deep neural network. 8, 19, 20

DSF dynamic spatial filtering. 21, 54, 57, 60, 62, 63, 67, 68, 70–80, 82–84, 86, 97, 99, 100, 109, 112

ECG electrocardiography. 28, 88, 112

EEG electroencephalography. xvi, 2–11, 13, 15, 16, 18–21, 23–26, 28–30, 32–57, 59–63, 65–68, 70, 73, 74, 76–80, 82–93, 95–97, 99–101, 103–109, 111, 112

EMG electromyography. 37, 88

EOG electroculography. 37, 38

ERP event-related potential. 88

FBCSP filter bank common spatial patterns. 33, 62, 97

FC fully-connected. 11, 12

fMRI functional magnetic resonance imaging. 3, 87

fNIRS functional near-infrared spectroscopy. 3, 112

GAN generative adversarial network. 56

GAP global average pooling. 14

GPU graphical processing unit. 17, 36, 62, 99

GRU gated recurrent unit. 31, 33

ICA independent component analysis. 27, 28, 50, 51, 56

iEEG intracranial EEG. 2

LSTM long short-term memory. 57

MAE mean absolute error. 11, 87, 89, 98, 99

MEG magnetoencephalography. 2, 3, 55, 87

MLP multilayer perceptron. 11, 59

MRI magnetic resonance imaging. 86–88

MSE mean squared error. 11, 16, 35, 40

NLP natural language processing. 14–16, 19, 27, 82

PCA principal component analysis. 56

PPG photoplethysmography. 112

PSP post-synaptic potential. 4

ReLU rectified linear unit. 12

RF random forest. 62, 64

RNN recurrent neural network. 11

RP relative positioning. 29–32, 34, 36–40, 42, 44–47, 50, 51

SE Squeeze-and-Excitation. 16

seq2seq sequence-to-sequence. 15

SFA slow feature analysis. 28

SNR signal-to-noise ratio. 19

SPD symmetric positive definite. 61

SSL self-supervised learning. 18, 20, 24–29, 32, 35, 36, 39–43, 45–49, 51, 52, 54, 62, 111, 112

SVM support vector machine. 61

TS temporal shuffling. 29–32, 34, 36–40, 42, 43, 45–47, 50, 51

UMAP Uniform Manifold Approximation and Projection. 41, 42

Notation

General

a	A scalar
$\mathbf{a}$	A vector
$\mathbf{A}$	A matrix
A	A tensor
a_i	Element of vector $\mathbf{x}$ at position i (with the first index being 1)
$A_{i,j}$	Element of matrix $\mathbf{X}$ at row i and column j
$\mathsf{A}_{i,j,k}$	Element (i, j, k) of tensor X
$\mathbf{I}$	Identity matrix
$\text{diag}(\mathbf{a})$	Square matrix filled with zeros whose diagonal is the vector $\mathbf{a}$
$\mathbf{1}_{\text{condition}}$	Indicator function returning 1 if condition is true, and 0 otherwise
$\mathbb{X}$	A set of N examples $\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}\}$, where $^{(i)}$ indexes the example in the dataset
$[\![q]\!]$	Set $\{1, \dots, q\}$ for any integers $p, q \in \mathbb{N}$
$[\![p, q]\!]$	Set $\{p, \dots, q\}$for any integers $p, q \in \mathbb{N}$
$\lvert\cdot\rvert$	Absolute value, applied element-wise
$\mathcal{N}(\mu, \sigma^2)$	Gaussian distribution with mean μ and standard deviation σ
$\mathcal{U}(a, b)$	Uniform distribution in the closed interval $[a, b]$

EEG time series

$\mathbf{S}$	Multivariate time series $\mathbf{S} \in \mathbb{R}^{C \times M}$, where M is the number of time samples and C is the dimension of samples (*e.g.,* channels)
$\mathbf{X}$	Non-overlapping window $\mathbf{X} \in \mathbb{R}^{C \times T}$ extracted from $\mathbf{S}$, with T the number of time samples in a window.

Contents

1 Motivation and contributions

Contents

Understanding the brain and how it functions has been a core scientific endeavour ever since its role in sensory and cognitive functions was first identified. Groundbreaking work in the nineteenth and twentieth centuries, such as the first recording of electrical brain activity (Caton, 1875), the discovery that neurons are the discrete units of the nervous system (Ramón y Cajal, 1894), and the functional mapping of the cortex (Penfield and Rasmussen, 1950), has laid the foundation for investigating how human health relates to brain anatomy and function.

Since then, the neurophysiological underpinnings of various phenomena such as sleep, consciousness and cognition have been studied. Pathologies that originate in the brain, such as epilepsy, dementia, and sleep disorders, are also the subject of ongoing research. Ultimately, a better understanding of how the brain works could help prevent, treat or manage diseases, and even identify ways to optimize its overall performance.

As the tools we use to record brain activity become more precise, affordable and even portable, the data needing to be analyzed increases in complexity and volume. Relying on human expertise to interpret that data becomes increasingly difficult. As a result, signal processing and machine learning have become critical tools for processing and

analyzing brain data, in both research and clinical settings. Deep learning (DL), in particular, has recently shown great promise in automating pattern recognition tasks performed on raw data (LeCun et al., 2015). Together, neuroimaging and machine learning have tremendous potential to enable revolutionary applications outside of traditional controlled environments.

In this thesis, methodological advances in the use of deep learning for the processing of real-world EEG data are presented. To set the stage, this introduction chapter covers the core principles and applications of EEG and of DL. The main contributions of this thesis, along with the resulting publications, concludes the chapter.

1.1 Electroencephalography: a window into human brain function

In this section, we give an overview of functional neuroimaging, with a particular focus on EEG, one of the most common techniques for studying brain function and the one that is at the core of this thesis.

1.1.1 Functional neuroimaging

Over the last decades, powerful tools have been developed to image the brain and measure its functioning. Functional neuroimaging modalities leverage different physical properties of brain tissues to capture brain activity in a direct or indirect manner. For instance, *electrophysiological* modalities pick up the electromagnetic fields that are generated as brain cells communicate with each other. *Hemodynamic* modalities, in contrast, measure how blood flow-related metrics vary when energy demands increase in different regions of the brain. We describe these two categories of modalities next.

Electrophysiological modalities As neurons communicate, electrochemical activity in organized brain structures gives rise to weak electromagnetic fields. As a result, these fields reflect instantaneous brain activity (see Section 1.1.2). The electrical component of these fields can be measured with *electrodes*, *i.e.,* sensors made out of conductive material and placed directly on or close to the area to be recorded. The voltage difference between two electrodes is then indicative of the activity occurring in the brain volume between these two locations. This can be done invasively by placing electrodes directly on the cortical surface or inserting them inside brain structures (intracranial EEG (iEEG) (Jasper and Penfield, 1949; Parvizi and Kastner, 2018)). Non-invasive measures can also be taken by placing electrodes on the scalp (surface or scalp EEG, often simply referred to as EEG (Berger, 1929; Hari and Puce, 2017)). The magnetic component of the field, in contrast, can be measured non-invasively using sensors called *magnetometers* and *gradiometers*, as is done in magnetoencephalography (MEG) (Cohen, 1968; Baillet, 2017; Hari and Puce, 2017). Both scalp EEG and MEG are non-invasive and therefore pose little to no health risk (as opposed to iEEG which requires opening the skull), making them excellent choices for studying or monitoring brain activity. Moreover, thanks to the direct relationship between neuronal activity and the electromagnetic fields that are captured, these modalities enable high temporal resolution monitoring on the order of milliseconds. While the spatial resolution of EEG is limited by the smearing of electrical fields by tissues, which does not occur in MEG, EEG is found more commonly in clinical and research settings, as it is portable and more affordable (by orders of magnitude) than MEG. Recent developments in more

portable MEG sensor technology might however improve this in the foreseeable future (Tierney et al., 2019).

Hemodynamics-based modalities As a brain region becomes active, its neurons require more glucose and oxygen to support widespread or sustained neuronal activity. As a result, more blood needs to be delivered to this brain region. This coupling between neural activity and cerebral blood flow is called the *hemodynamic response* (Iadecola, 2017). Typically, there is a delay of about four to six seconds between neuronal activation and the peak in localized blood flow (Miezin et al., 2000). The resulting increase in blood flow can therefore be used as a (delayed) proxy of neural activity. This hemodynamic information can be measured by optical and magnetic means. Optical neuroimaging modalities such as functional near-infrared spectroscopy (fNIRS) rely on the fact that hemoglobin, the molecule that carries oxygen in the bloodstream, has different light absorption characteristics in the near-infrared spectrum, depending on whether it has oxygen molecules attached to it or not (Jöbsis, 1977; Ferrari and Quaresima, 2012). Changes in the concentration of oxygenated hemoglobin (HbO) and deoxygenated hemoglobin (HbR) can therefore be measured using near-infrared light sources placed on the scalp and nearby photodetectors that pick up the scattered photons that have travelled through the skull all the way to the cortex and back. Functional magnetic resonance imaging (fMRI) similarly monitors hemoglobin levels as an indirect measure of neural activity, though leveraging the magnetic, rather than optical, properties of HbR molecules (Ogawa et al., 1990), giving rise to a signal known as the blood oxygenation level-dependent (BOLD) signal (Logothetis et al., 2001). Unlike in fNIRS where measurements are limited to the cortical surface, the BOLD signal can be recorded in the entire brain. This has made fMRI a prevailing modality for functional neuroimaging since the 1990s. However, contrary to fNIRS, fMRI is a highly expensive modality and is not portable.

Because it has excellent temporal resolution, is non-invasive, relatively inexpensive and also increasingly portable, EEG is not only a common choice for studying brain activity, it also opens the door to neurophysiological monitoring in entirely new environments and contexts. Next, we delve deeper into EEG: we describe its generation mechanism, some typical applications, and discuss how its portability enables new scientific and technological developments.

1.1.2 Electroencephalography

Reports of electrical brain activity measurements date back to the second half of the nineteenth century, when invasive recordings in animal models suggested sensory activation could be measured with a galvanometer (Caton, 1875). Almost half a century later, the first human scalp EEG recordings were reported by Berger (1929). Berger's main finding was what he called the "alpha" rhythm, a component found at the back of the head, oscillating at roughly 10 Hz. This same component is, still today, one of the most studied EEG components, and is of critical importance in the discussion of some of the results reported in this thesis (Chapter 4).

Physiological generators of EEG signals The human brain comprises approximately 86 billion cells called neurons (Azevedo et al., 2009) which transfer information by propagating electrochemical currents along their main axis (Hari and Puce, 2017). About 19% of these neurons are found in the outer layer of the brain, called the *cortex*

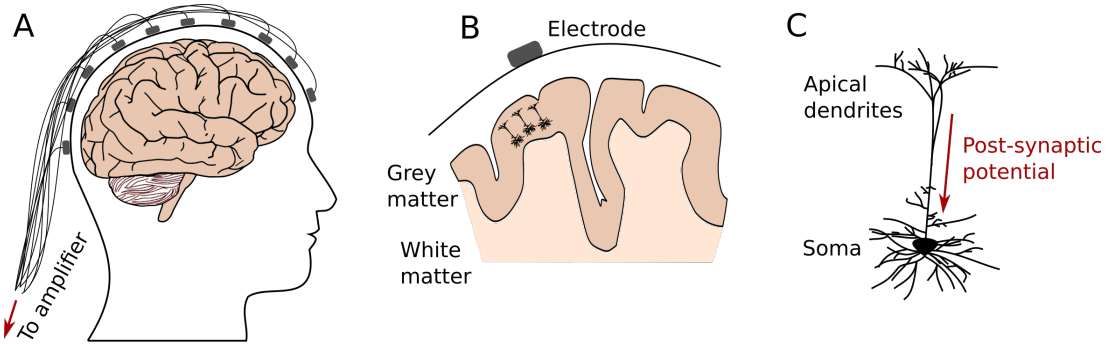


Figure 1.1 – EEG signal generation mechanism. (A) Electrodes are placed on the scalp and connected to an amplifier. (B) Under the skull, the outer layer of the brain, called *cortex*, contains pyramidal neurons that are aligned perpendicularly to its surface. (C) Post-synaptic potentials propagate as intracellular current flow along the main axis of pyramidal neurons. When these potentials occur in synchrony for a large enough population of nearby pyramidal neurons, the resulting potential can be detected at the surface of skull using electrodes.

(Figure 1.1). Only 3-4 mm thick, the cortex is a highly convoluted structure, of which about two-thirds of the surface is folded over itself. Despite accounting for only a portion of total brain volume, the cortex is where most of the electrical activity picked up by EEG originates. This is because the primary cortical neurons, called *pyramidal* neurons due to their characteristic triangular cell bodies, are highly organized: they are aligned perpendicularly to the cortical surface. As a result, when a group of thousands of such neurons are simultaneously activated, tiny electrical potentials called post-synaptic potentials (PSPs) add up and give rise to electrical potentials large enough to be measured on the scalp, *e.g.,* on the order of tens of μV. However, because the medium through which the electrical potential propagate (*i.e.,* the cerebrospinal fluid, skull and scalp) is not homogeneous the resulting EEG signals are smeared spatially. Consequently, the information picked up at a specific location is actually the summation of multiple electric fields originating from different regions of the brain. This gives rise to the localization problem, or inverse problem, an active area of research in which algorithms are developed to reconstruct brain sources given EEG recordings (Baillet et al., 2001; He et al., 2018). Moreover, thanks to the direct relationship between the signals measured on the scalp and neuronal activity, and due to the near instantaneous propagation of electric fields in the brain tissues, EEG has an excellent temporal resolution allowing the recording of brain activity at a millisecond timescale.

Instrumentation A typical clinical- or research-grade EEG device has around 16 to 256 sensors, or electrodes, that are arranged on the scalp following a predetermined grid-like pattern called a *montage* (*e.g.,* the International 10-20 system, see Figure 1.2D). Signals are typically digitized at a sampling frequency f_s of 128 to 1024 Hz. The measurements are made between one electrode and another *reference* electrode, which is commonly set to the top of the head or to linked mastoids, as these locations are far away from the most common artifact sources (see Section 1.1.3). Typically, conductive gel, paste, or a saline solution is used to ensure a good electrical connection between the electrodes and the skin. Dry electrodes have more recently been introduced and, because they can work without these kinds of solutions, are usually faster to set up and more convenient to wear, though they can have a higher sensitivity to noise and

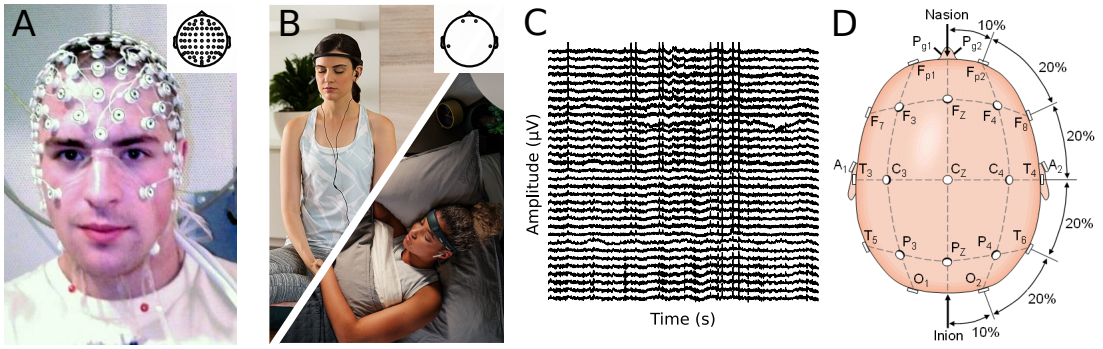


Figure 1.2 – Illustration of common EEG instrumentation and related concepts. (A) A typical research/clinical-grade EEG device.[1](B) Examples of wearable mobile EEG used for neurofeedback and sleep tracking applications (Muse headband, InteraXon Inc., Toronto, Canada). Images shared with permission from Interaxon Inc. (C) Example EEG time series collected with a 32-channel EEG montage. (D) The International 10-20 standard positioning system, as viewed from the top of the head. Letters refer to the different brain regions (F: frontal, C: central, T: temporal, P: parietal and O: occipital) while numbers indicate how far from the sagittal line an electrode is found (taken from Malmivuo et al. (1995)). The electrodes are positioned such that the distance between them is 10 or 20% of the distance between the front and back, or the right and left sides, of the head. Insets in (A) and (B) indicate where on the head the different EEG electrodes are placed.

artifacts (Lopez-Gordo et al., 2014).

Applications EEG can be used to study a wide array of brain processes (Hari and Puce, 2017; Schomer and Da Silva, 2012; Biasiucci et al., 2019). For instance, as one of the main tools used in clinical sleep studies, EEG is key to diagnosing and studying sleep disorders such as apnea and narcolepsy (Aboalayon et al., 2016; Ghassemi et al., 2018; Bathgate and Edinger, 2019). Critically, EEG is the gold standard for *sleep staging*, *i.e.,* the process of identifying in which stage of sleep an individual is during the night (Berry et al., 2012). Due to the importance of sleep in EEG research, multiple experiments presented in this thesis make use of sleep data, including on the sleep staging task (Chapters 2-4). EEG is also routinely used in clinical contexts to screen individuals for neurological conditions such as epilepsy (Smith, 2005; Acharya et al., 2013), dementia (Micanovic and Pal, 2014), attention deficit hyperactivity disorder (Arns et al., 2013), disorders of consciousness (Giacino et al., 2014; Engemann et al., 2018a) and depth of anaesthesia (Hagihira, 2015). The *pathology detection* task, which we study alongside sleep staging in Chapters 2 and 3, can be seen as a higher level classification task where one must whether an individual's EEG is indicative of a pathology or not. Recently, EEG has also been proposed to build biomarkers of neurophysiological health through the *brain age* framework (Franke et al., 2010, 2012; Al Zoubi et al., 2018; Sabbagh et al., 2020). By comparing an individual's chronological age to the age predicted by a neuroimaging-based model trained on a normative healthy population, we can identify individuals whose brain look "older" than other people in the same age group, suggesting premature or pathological aging (Cole and Franke, 2017; Cole et al., 2018). We

[1]Image from https://en.wikipedia.org/wiki/Electroencephalography#/media/File:EEG_cap.jpg (public domain).

present an extension of the brain age framework to real-world EEG data in Chapter 4. Neuroscience and psychology research also make frequent use of EEG with applications such as cognitive and affective monitoring (Berka et al., 2007; Thorsten and Christian, 2011; Al-Nafjan et al., 2017). Finally, EEG is a popular modality for 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 a user's environment (Lotte et al., 2015; McFarland and Wolpaw, 2017).

1.1.3 Unsolved challenges in EEG processing

Although EEG has proven to be a critical tool in many domains, it still suffers from a few limitations, which we briefly discuss here.

Artifacts and noise Though they contain information about brain activity, the signals picked up by EEG instrumentation often contain noise originating from various other sources and which can be orders of magnitude larger than actual brain signals (Hari and Puce, 2017). For this reason, proper care must be taken to avoid these sources of noise if possible, and to handle them otherwise. Common noise categories include physiological artifacts, movement artifacts, instrumentation noise, and environmental noise. Physiological artifacts are large signals that are generated by electrical current sources outside the brain, such as heart activity, eye or tongue movement, muscle contraction, etc. Depending on the EEG electrode montage and the setting of the recording (*e.g.,* eyes-open or eyes-closed), these artifacts are more or less likely to disrupt measurements of the brain activity of interest. Movement artifacts, on the other hand, are caused by the relative displacement of EEG electrodes with respect to the scalp, and can introduce noise of varying spectral content in the affected electrodes during the movement itself. If an electrode cannot properly connect with the skin (*e.g.,* after a movement artifact or because it was not correctly set up initially), its reading will likely contain little or no physiological information and instead pick up instrumentation and environmental noise. Electrodes suffering from this problem are commonly referred to as *bad* or *missing channels* in the literature. In the context of this thesis, we refer to them as *corrupted channels* to explicitly reflect the fact that such channels may still contain usable information, in addition to the noise. This is investigated in Chapter 3. Importantly, because these sources of noise are characterized by widely different morphologies and spatial distributions, noise handling techniques typically need to be designed with a specific type of noise in mind. This challenge, along with existing solutions, will be extensively discussed in Chapter 3 when presenting corruption-robust architectures for mobile EEG.

Non-stationarity EEG is also a non-stationary signal (Gramfort et al., 2013; Cole and Voytek, 2018), *i.e.,* 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, as the statistics of the signals might have changed in the meantime. This is an important challenge for real-life applications of EEG, which often need to work with limited amounts of data.

Inter-subject variability Inter-subject variability arises due to physiological differences between individuals, which vary in magnitude but can severely affect the performance of models that need to generalize across subjects (Clerc et al., 2016). 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, such as transfer learning, to handle inter-subject variability (see Saha and Baumert (2020) and Wan et al. (2021) for reviews).

Domain-specific processing pipelines To solve some of the above-mentioned problems, processing pipelines are often developed with a specific application domain in mind. Indeed, a significant amount of research has been dedicated to developing domain-specific processing pipelines to clean, extract relevant features, and classify EEG data. As a result, there are no "one-size-fits-all" methods to processing EEG. However, more recent state-of-the-art techniques, such as Riemannian geometry-based classifiers and adaptive classifiers might be more re-usable across domains (Lotte et al., 2018).

Need for automated processing Finally, a wide variety of tasks would benefit from a higher level of automated processing. For example, sleep staging, the process of annotating sleep recordings by categorizing windows of a few seconds into sleep stages, currently requires trained technicians to visually inspect and manually label the data. The development of 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. Finally, the ability to fully automate processing has the potential to greatly improve reproducibility in EEG and neuroimaging research in general, as the exact same processing pipelines could be re-applied to a dataset, or re-used over multiple datasets (Jas et al., 2018).

1.1.4 Toward ubiquitous neurophysiological monitoring with low-cost mobile EEG

In recent years, the emergence of low-cost mobile EEG devices has made it possible to monitor and record EEG in entirely new environments, and to dramatically improve access to the technology (Mihajlović et al., 2014; Casson, 2019). In the context of this thesis, we refer to this new paradigm as "real-world EEG", to emphasize the fact that the technology is no longer confined to the controlled environment of research labs and clinical settings.

Consumer-focused EEG devices typically have a few channels only, are wireless, use dry electrodes, and are available below a USD 1,000 price point. Thanks to their affordability and ease of use, they enable recording brain activity in at-home settings or anywhere medical or research infrastructure is not available, with applications such as sleep monitoring, pathology screening, neurofeedback, brain-computer interfacing and anaesthesia monitoring (Mihajlović et al., 2014; Dhindsa, 2017; Kreuzer, 2017; Krigolson et al., 2017; Johnson and Picard, 2020; Hohmann et al., 2020; Krigolson et al., 2021; Mikkelsen et al., 2021). This in turn enables the collection of unprecedented amounts of EEG data from diverse populations across the world and opens the door to performing neurophysiological health assessments on a day-to-day basis.

However, low-cost mobile EEG technology comes with its own set of challenges. First, spatial information is often limited, *i.e.,* only a few channels are available, as compared to typical clinical- and research-grade devices. As a result, localized activity in specific parts of the brain might be impossible to monitor if the available electrodes do not cover all the regions of interest. For instance, BCIs that rely on detecting the activation

of the motor cortex to distinguish imagined movements require coverage of the central (*i.e.,* the top) of the head to function (Pfurtscheller and Neuper, 1997). Therefore, a sparse EEG montage without electrodes in this region might be blind to the actual patterns of interest.

Second, while channel corruption affects EEG recordings in all contexts, it is more likely to occur in real-world mobile EEG recordings than in controlled laboratory settings. Indeed, trained experimenters can monitor and remedy bad electrodes during laboratory recordings, which is not typically the case in real-world EEG recordings. Low-cost mobile EEG also often makes use of dry electrodes, which, despite being less complicated to set up, are more sensitive to environmental noise than regular wet or gel electrodes. This further increases the likelihood of noise being injected into the recordings. Moreover, as opposed to controlled experiments where dense electrode montages allow interpolating missing channels offline, the limited spatial information provided by mobile EEG devices makes this approach much more challenging. Therefore, special care must be given to the problem of channel corruption in sparse mobile EEG settings. This topic will be further addressed in Chapter 3.

Thirdly, with the increasing availability of low-cost mobile EEG devices, the volume of data generated exceeds the capacity of human experts (*e.g.,* neurologists, sleep technicians, etc.) to analyze and manually annotate every single recording, as is traditionally done in research and clinical settings. Novel methods facilitating clinical and research applications in real-world settings, especially with sparse EEG montages, are therefore needed. In response to this challenge, new unsupervised paradigms can be designed to allow the use of this significant amount of data (see Chapter 2). Similarly, new ways of leveraging the available metadata (or "weak labels") can be designed to develop proxy measures of brain health (see Chapter 4).

Now that we have introduced EEG, discussed common applications and motivated its use "outside of the lab", we turn our attention to deep learning, on which we rely in this thesis to further enable the use of EEG in the real world.

1.2 Deep learning on EEG time series

To overcome the challenges that arise from the various applications of EEG (Section 1.1.3), new approaches are required that should be more robust to noise and non-stationarity, and lead to better generalization. In this context, a promising approach is deep learning (DL), a subfield of machine learning which studies computational models that learn hierarchical representations of data through successive non-linear transformations (LeCun et al., 2015). DL has the potential to significantly simplify processing pipelines by allowing automatic end-to-end learning of preprocessing, feature extraction and classification modules, while also reaching competitive performances on target learning tasks. Over the last few years, DL architectures have already proven to be very successful in processing complex data such as images, text and audio signals (LeCun et al., 2015), leading to state-of-the-art performances on multiple public benchmarks - such as the Large Scale Visual Recognition challenge (Deng et al., 2012) - and an ever-increasing role in industrial applications.

Deep neural networks (DNNs), inspired by earlier models such as the perceptron (Rosenblatt, 1958) (itself a simplified model of biological neurons), are models where: 1) stacked layers of artificial "neurons" each apply a linear transformation to the data they

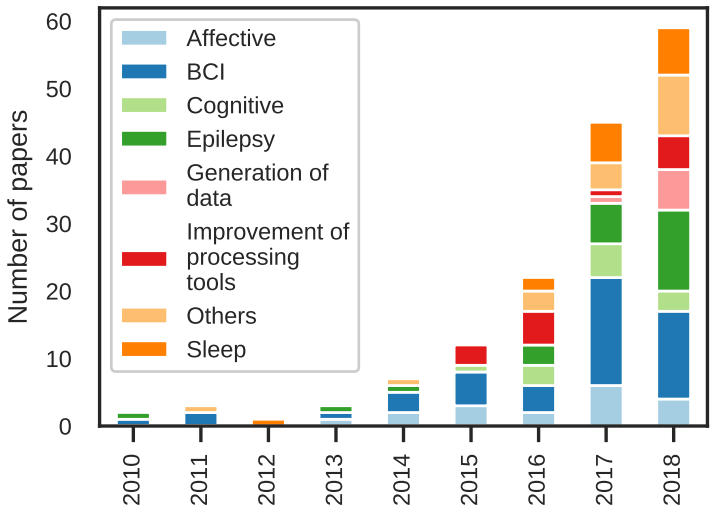


Figure 1.3 – Number of DL-EEG publications per domain per year, as reported in Roy et al. (2019a).

receive as input and 2) the output of each layer's linear transformation is fed through a non-linear activation function.

While neural networks with a single hidden layer are universal function approximators (Hornik et al., 1989), composing multiple successive hidden layers has been shown empirically to help learning by enabling hierarchical transformations of the input data LeCun et al. (2015). Importantly, the parameters for 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 the point at which a neural network becomes a deep one (Goodfellow et al., 2016).

Although it is still a relatively new approach to processing EEG, DL has already been the subject of a significant amount of research looking at applying it to EEG data. For instance, the number of DL-EEG publications in different application categories (*e.g.,* BCIs, epilepsy and sleep) has been steadily growing over the last few years, as shown by a recent comprehensive review of the literature (Figure 1.3, from Roy et al. (2019a)). Despite these early results, multiple challenges and questions remain to be tackled.

In this section, we introduce the core concepts behind using deep learning to learn from EEG data. We simultaneously present relevant results taken from Roy et al. (2019a) to highlight important trends in DL-EEG research.

1.2.1 Learning problem

A parametric machine learning model $f_{\boldsymbol{\theta}}$ with parameters $\boldsymbol{\theta}$ is trained to learn the mapping between $\mathbf{x} \in \mathcal{X}$, an *example*, and $y \in \mathcal{Y}$, the *target*, *i.e.,* a variable that represents class membership or a value $\mathbf{x}$ is associated to. In this thesis, we often consider $f_{\boldsymbol{\theta}}$ to be a deep neural network. Both $\mathcal{X}$ and $\mathcal{Y}$ depend on the dataset and the task at hand. Looking first at the input, in EEG-related tasks, $\mathcal{X}$ is often $\mathbb{R}^{C \times T}$, *i.e.,* a matrix $\mathbf{X}$ containing the amplitude values (*e.g.,* in μV) of C channels across T

time points. Sometimes, such as in classical machine learning settings, $\mathbf{x}$ might also be a transformed version of the EEG signals, *i.e.,* processed through a *feature extraction* step before being passed to $f_{\boldsymbol{\theta}}$. The goal of the feature extraction step is to represent the data more efficiently and/or to highlight the relevant information that it contains in order to facilitate the learning task. For instance, $\mathbf{X}$ can be represented by its covariance matrices in specific frequency bands (Ang et al., 2008)[2].

When it comes to the target, two common machine learning tasks focused on learning the mapping between $\mathbf{x}$ and y are *classification* and *regression*. Because the target y is required for the algorithm to learn such mappings, these two tasks are said to be *supervised*. In classification problems, $y^{(i)}$ represents the category an example $\mathbf{x}^{(i)}$ belongs to, *i.e.,* $\mathcal{Y}$ is $[\![L]\!]$ for a classification problem with L classes. For instance, the sleep staging task introduced in Section 1.1.2 is typically cast as a 5-class classification problem where a 30-s EEG window (*i.e.,* $T = 30 \times f_s$ where f_s is the sampling frequency) must be mapped to the sleep stage (W, N1, N2, N3, R) in which it occurred. In regression problems, the target $y^{(i)}$ is usually a continuous value, *i.e.,* $\mathcal{Y}$ is $\mathbb{R}$ or a specific portion of the real line. For instance, the brain age prediction task (Chapter 4) focuses on predicting the age, in years, of an individual given one or multiple windows drawn from their EEG.

In order to know how well a model $f_{\boldsymbol{\theta}}$ performs on a given learning task and how to improve its performance, we use a *loss function* $\mathcal{L}$. Loss functions can be derived using *e.g.,* the *maximum likelihood estimation* framework. In this framework, our model is taken to represent a parametric probability distribution $p_{\text{model}}(\mathbf{x}; \boldsymbol{\theta})$ which estimates the true distribution of the data. We then seek to find the parameters $\boldsymbol{\theta}^*$ that maximize the conditional log-likelihood on pairs $(\mathbf{x}^{(i)}, y^{(i)})$ sampled from the true underlying distribution p_{data}:

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{x},y\sim p_{\text{data}}} \log p_{\text{model}}(y|\mathbf{x}; \boldsymbol{\theta}) \ . \tag{1.1}$$

Importantly, p_{data} is not accessible in practice, and so we have to resort to using a finite collection of examples sampled from this distribution, *i.e.,* a dataset $\mathbb{X} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}\}$ along with the associated targets $\{y^{(1)}, \ldots, y^{(N)}\}$. However, these N examples form a different distribution $\hat{p}_{\text{data}}(\mathbf{x}, y)$, called the *empirical distribution*, which likely differs from $p_{\text{data}}(\mathbf{x}, y)$. Despite this, if the training examples are independently and identically distributed (*i.i.d.*)[3], $\hat{p}_{\text{data}}$ should be representative of the true distribution and we can rely on the empirical risk to compute and optimize the conditional likelihood of $f_{\boldsymbol{\theta}}$. This is equivalent to minimizing the cross-entropy of the two distributions p_{model} and $\hat{p}_{\text{data}}$ (Goodfellow et al., 2016). For instance, in classification problems, the *categorical cross-entropy* loss function, which corresponds to p_{model} following a Multinouilli distribution, can be implemented as:

$$\mathcal{L}_{\text{categorical}}(\boldsymbol{\theta}) = -\sum_{i=1}^{N}\sum_{k=1}^{L} \mathbf{1}_{y^{(i)}=k} \log p_{\text{model}}(y = k|\mathbf{x}^{(i)}; \boldsymbol{\theta}) \ . \tag{1.2}$$

[2]This follows from the assumption that a short time window $\mathbf{X}$, once bandpass filtered, is centered around zero and mostly Gaussian, and therefore that the covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{C\times C}$ is a sufficient statistic of the distribution of multivariate EEG samples.

[3]Of note, for EEG windows to be independently distributed, they must not overlap.

On the other hand, for regression problems, it is common to use the mean squared error (MSE) loss function (*i.e.,* a Gaussian prior) which directly uses the prediction $\hat{y}^{(i)} = f_{\boldsymbol{\theta}}(\mathbf{x}^{(i)})$:

$$\mathcal{L}_{MSE}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N}\left(y^{(i)} - f_{\boldsymbol{\theta}}(\mathbf{x}^{(i)})\right)^2 \quad , \tag{1.3}$$

or the mean absolute error (MAE), which we will use in Chapter 4 for brain age prediction experiments (*i.e.,* a Laplacian prior):

$$\mathcal{L}_{MAE}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N}\left|y^{(i)} - f_{\boldsymbol{\theta}}(\mathbf{x}^{(i)})\right| \quad . \tag{1.4}$$

For instance, MSE might be preferred over MAE if predictions that are further away from the true target should be penalized more strongly.

Importantly, while the loss function is evaluated on the training data, we are actually interested in the ability of the model $f_{\boldsymbol{\theta}}$ to *generalize*, *i.e.,* to correctly predict the target of an example that it has never seen. It is therefore common practice to split the available data into *training* and *testing* sets. The *training* set contains examples on which $f_{\boldsymbol{\theta}}$ will be trained, while the *testing* set is used to obtain an unbiased estimate of the performance of the model. Additionally, if comparing multiple variations of a model to pick the best one, an additional set is needed to avoid optimistically biasing the final results. For this purpose, it is common to further divide the training examples into a smaller training set and a *validation* set. When working with multi-subject EEG datasets, we often divide the data such that no examples from the same subject or recording is in more than one set at a time. This allows a more faithful evaluation of the ability of a trained model to generalize to new unseen individuals or recordings.

1.2.2 Architectures

As stated above, deep neural networks perform successive non-linear transformations on their input. To do so, different types of *layers* are used as building blocks. Most commonly, these are fully-connected (FC), convolutional or recurrent layers. The types and numbers of layers define a neural network's *architecture*. For instance, we refer to models using these types of layers as FC networks, convolutional neural networks (CNNs) (LeCun et al., 1989) and recurrent neural networks (RNNs) (Rumelhart et al., 1986), respectively. Recently, attention mechanisms have also led to major improvements in the performance and interpretability of neural networks (Niu et al., 2021). Here, we provide a quick overview of the main architectures found in the DL literature. In-depth descriptions of DL methodology can be found in Schmidhuber (2015); LeCun et al. (2015); Goodfellow et al. (2016).

Fully-connected layers

FC layers are composed of fully-connected neurons, *i.e.,* each neuron receives as input the activations of every single neuron of the preceding layer (Figure 1.4). Neural networks composed of only FC layers are called fully-connected neural networks or multilayer perceptrons (MLPs) (Goodfellow et al., 2016). At its core, an FC layer h with d' hidden units applies an affine transformation to the input vector $\mathbf{x} \in \mathbb{R}^d$ using a matrix $\mathbf{W} \in \mathbb{R}^{d' \times d}$ and a bias vector $\mathbf{b} \in \mathbb{R}^{d'}$, followed by a non-linear function g (called

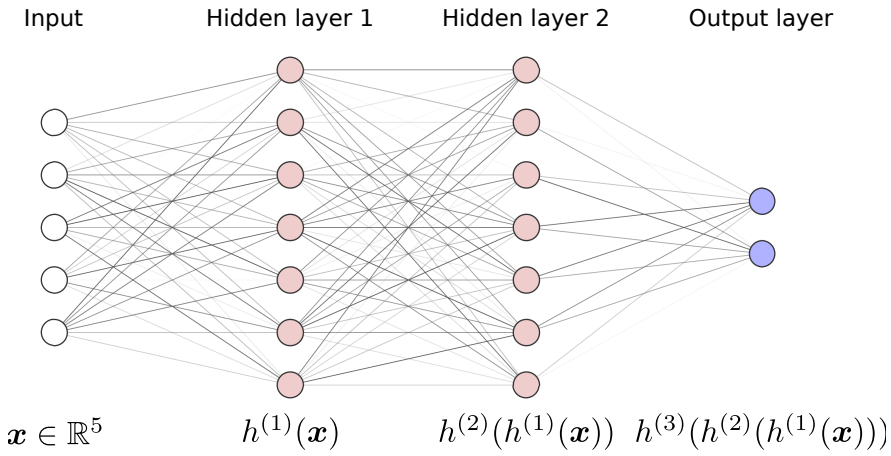


Figure 1.4 – A fully-connected neural network with two hidden layers. Edges represent weights between the output of a layer and units of the next layer. The opacity of the edges illustrate the relative weights of randomly initialized parameters. Adapted from LeNail (2019).

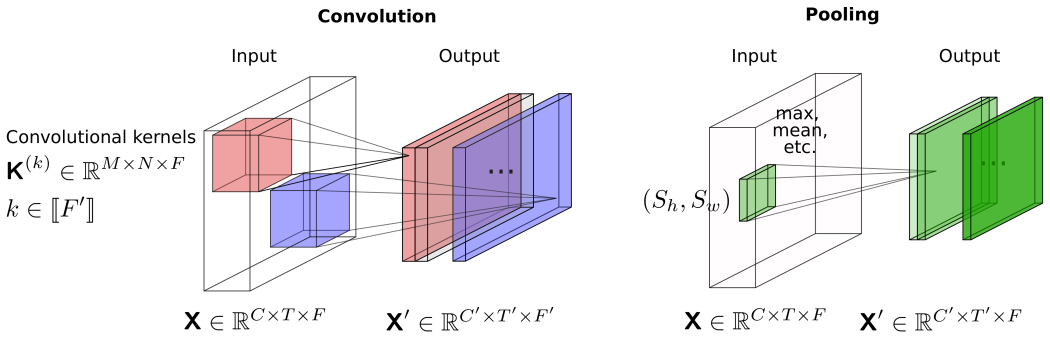


Figure 1.5 – The two core operations in convolutional neural networks: convolution and pooling.

activation function), which makes it possible to learn a non-linear mapping between $\mathbf{x}$ and the output of the network:

$$h(\mathbf{x}; \mathbf{W}, \mathbf{b}) = g(\mathbf{W}\mathbf{x} + \mathbf{b}) \ . \tag{1.5}$$

A common choice for the activation function is the element-wise rectified linear unit (ReLU) function:

$$g(\mathbf{z}) = \max(\mathbf{z}, 0) \ . \tag{1.6}$$

A k-layer FC neural network is then composed by cascading layers:

$$f_{\boldsymbol{\theta}_{\text{FC}}}(\mathbf{x}) = h^{(k)}(\quad \ldots \quad h^{(2)}(h^{(1)}(\mathbf{x}))) \ , \tag{1.7}$$

where $\boldsymbol{\theta}_{\text{FC}} = \{\mathbf{W}^{(1)}, \mathbf{W}^{(2)}, ..., \mathbf{W}^{(k)}, \mathbf{b}^{(1)}, \mathbf{b}^{(2)}, ..., \mathbf{b}^{(k)}\}$.

Convolutional layers

Convolutional layers, as opposed to fully-connected ones, impose a particular structure where neurons in a given layer are only connected to a subset of the neurons in the preceding layer (Figure 1.5, *left*) (LeCun et al., 1989). This structure is akin to the convolution operation in signal or image processing from which it gets its name, and which is defined as:

$$s(t) = (x * w)(t) = \int_{-\infty}^{\infty} x(a)w(t-a)da \ , \tag{1.8}$$

i.e., a weighting function w is shifted around t and multiplies the input x.

In convolutional layers, discrete convolution[4] is often used to process a 3D input $\mathbf{X} \in \mathbb{R}^{C \times T \times F}$ (where F is the number of *convolutional channels*, *e.g.,* they could be thought of as the number of frequency bands in a filterbank representation[5]) with a collection of P *kernels* $\mathbf{K} \in \mathbb{R}^{M \times N \times F}$ (where M and N are the dimensions of the kernel in the first and second dimensions of $\mathbf{X}$). The parameters of each kernel $\mathbf{K}$, also called *convolutional filter*, are learned during training and transform the input by picking out *e.g.,* different spatial, temporal and/or spectral patterns. We can write a convolutional layer as:

$$\mathbf{X'}_{i,j,k} = g\left((\mathbf{X} * \mathbf{K}^{(k)})_{i,j,k}\right) = g\left(\sum_{c=1}^{C}\sum_{t=1}^{T}\sum_{f=1}^{F}\mathbf{X}_{c,t,f}\mathbf{K}^{(k)}_{i-c,j-t,f}\right) \ , \tag{1.9}$$

where $k \in [\![P]\!]$ and g is an element-wise non-linear function, as defined above. The slices of $\mathbf{X'}$ along its third dimension are called *convolutional channels*[6] or *feature maps*.

Equation 1.9 illustrates the *parameter sharing* property of convolution layers: each neuron of a layer k only sees a portion of the layer's input, and shares its weights (*i.e.,* the kernel $\mathbf{K}^{(k)}$) with all other neurons of its feature map. From this perspective, a convolutional layer learns filters that "look" for the same information across patches of the input. This also helps make CNNs more efficient, as fewer parameters need to be tuned.

Importantly, the local structure of a convolutional layer encourages the model to learn translation-equivariant representations of the data, *i.e.,* where the representations for an input $\mathbf{X'}$ that is a translated version of $\mathbf{X}$ are the translated representations themselves.

In deep learning architectures designed for EEG, it is common to encounter 1D convolutions, *i.e.,* convolutional layers in which convolutions are applied to either the temporal or spatial dimension ($M = 1$ or $N = 1$). This is analogous to the temporal and spatial filtering steps commonly used in traditional signal processing of EEG signals (Makeig et al., 1996; McFarland et al., 1997; Parra et al., 2005; Blankertz et al., 2007; de Cheveigné and Simon, 2008; Lotte and Guan, 2010; Nikulin et al., 2011).

[4] In practice, deep learning libraries implement a *cross-correlation* operation, which is identical except for its sign. Since the parameters are learned, this difference has no effect on the underlying capacity of the model.

[5] Typically, when working with EEG, the input to a CNN has $F = 1$, however after the first layer the number of convolutional channels increases.

[6] Not to be confused with *EEG channels* as introduced in Section 1.1.2.

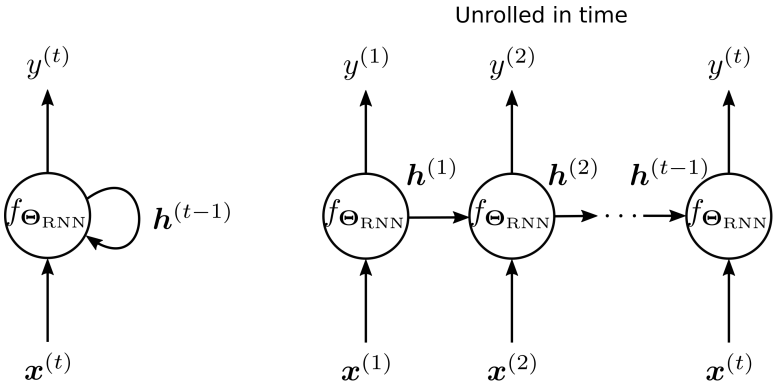


Figure 1.6 – A recurrent neural network $f_{\boldsymbol{\theta}_{\text{RNN}}}$, in a *folded* (left) and *unfolded* (right) configuration. Adapted from Goodfellow et al. (2016).

A second key component of CNNs is *pooling*, where activations are locally aggregated (Figure 1.5, *right*). In CNNs, pooling operations typically follow blocks of convolutions. One example is the *max-pooling* operation, which is defined as:

$$\mathbf{X}'_{i,j,k} = \max_{m=[\![M]\!]} \max_{n=[\![N]\!]} \mathbf{X}_{S_h(i-1)+m, S_w(j-1)+n, k} \ , \tag{1.10}$$

where S_h, S_w are the strides of the pooling operation in the first and second dimensions of $\mathbf{X}$ (*i.e.,* the height and width dimensions).

It is specifically the combination of convolutions and pooling that enables models to learn representations that are approximately invariant to 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 usually affect the prediction of the model. Moreover, pooling can improve the computational efficiency of a layer by reducing the dimensionality of its output, when input regions are pooled with a stride larger than one. Pooling is also sometimes applied to the convolutional channel dimension (*i.e.,* the third dimension of $\mathbf{X}$), to instead provide invariance to the transformations applied by the different kernels of convolutional layers, in which case it is referred to as *global pooling*, *e.g.,* global average pooling (GAP) (Lin et al., 2013).

Recurrent layers

In contrast to convolutional layers, recurrent layers impose a structure whereby a layer receives both the preceding layer's activations and its own activations from the previous time step as input (Figure 1.6). Models composed of recurrent layers are thus encouraged to make use of the sequential structure of data, allowing them to achieve high performance in tasks such as natural language processing (NLP) (Zhou and Xu, 2015; Yogatama et al., 2017). The operations carried out by a vanilla recurrent layer can be written as:

$$\mathbf{h}^{(t)} = g_h(\mathbf{W}\mathbf{h}^{(t-1)} + \mathbf{U}\mathbf{x}^{(t)} + \mathbf{b}) \ , \tag{1.11}$$

$$\mathbf{o}^{(t)} = g_o(\mathbf{V}\mathbf{h}^{(t)}) + \mathbf{c} \ , \tag{1.12}$$

where $\mathbf{W} \in \mathbb{R}^{d' \times d'}$, $\mathbf{U} \in \mathbb{R}^{d' \times d}$, $\mathbf{V} \in \mathbb{R}^{d' \times d''}$ are weight matrices, $\mathbf{b} \in \mathbb{R}^{d'}$ and $\mathbf{c} \in \mathbb{R}^{d''}$, and g_h and g_o are non-linear functions applied on the state vector (or *hidden state*) $\mathbf{h}$ or the output $\mathbf{o}$, respectively. This structure provides a different kind of parameter

sharing where the same parameters are reused from one time step to another (instead of from one patch to another, as is the case for convolutional layers).

To train such networks, outputs must be "unfolded" in time (Figure 1.6, *right*). This leads to a major challenge when using gradient descent (see Section 1.2.3), as gradients run the risk of vanishing or exploding (Goodfellow et al., 2016). A widely adopted solution to this problem is to use *gated* recurrent units layers instead of the vanilla recurrent layers (Hochreiter and Schmidhuber, 1997; Cho et al., 2014). Gated units have a more complicated structure that allows them to accumulate information over multiple time steps, while also forgetting information when needed.

Attention mechanisms

Attention mechanisms are (generally differentiable) modules that can be inserted between the layers of a neural network to help the network focus its processing on the most relevant parts of the data (Bahdanau et al., 2014; Luong et al., 2015; Vaswani et al., 2017; Niu et al., 2021). Even setting aside the empirical results that have shown significant performance gains thanks to its use, attention is intuitively a powerful framework as 1) it helps focus the computational power of a model on the information that matters given the task at hand, and 2) it allows the inner workings of a neural network to be introspected through *attention maps*, *i.e.,* by observing which parts of the input the network focused on. These two properties of attention mechanisms will be leveraged in Chapter 3 to design noise-robust and interpretable deep learning EEG models.

Attention over sequences Attention mechanisms were initially designed in the context of NLP where they were used to improve performance on sequence-to-sequence (seq2seq) tasks. For instance, tasks such as machine translation require the processing of sequences of arbitrary length (*e.g.,* translating a sentence from French to English). Early work using an encoder-decoder architecture where the input sequence was summarized into a single vector (Sutskever et al., 2014) struggled with long-range dependencies: the longer the sequence, the harder it was for a neural network to reliably encode the information into a unique vector. To address this, *additive attention* (Bahdanau et al., 2014; Luong et al., 2015) uses a dedicated scoring function to estimate the relative importance of each input word for predicting each output word. The internal representation of the input words, along with information on the sequence being generated, can then all be aggregated (with a weighted sum), with the result then being used to predict the next word in the sequence. The introduction of additive attention greatly improved modelling of longer sequences. Recently, a major improvement to attention-based architectures came with the self-attention *Transformer* architecture (Vaswani et al., 2017). Self-attention refers to attention mechanisms that evaluate the importance of each item of a sequence relative to the others, which is key in tasks such as language understanding. Much of the recent NLP literature relies on Transformers to achieve high-performance in different tasks (Devlin et al., 2018; Radford et al., 2019; Liu et al., 2019).

Attention over images While attention mechanisms have proved critical in recent years for sequence learning tasks, similar strategies can also be very useful for other, non-sequential data. In the computer vision (CV) literature, *scaling attention* mechanisms are designed to help networks focus on specific portions of an input (*i.e.,* spatial information) or specific feature maps (*i.e.,* specific transformations of an input). For in-

stance, the Squeeze-and-Excitation (SE) block (Hu et al., 2018) does so by transforming the feature map outputted by a convolutional layer using the information contained in the feature map itself. This makes it easier for a model to boost relevant features while minimizing the contribution of less relevant ones. Similarly, the convolutional block attention module (CBAM) is an attention module that sequentially applies channel attention and spatial attention (Woo et al., 2018). First, as in an SE block, channels are remapped based on an aggregated version of their input. Second, a spatial attention map is obtained by transforming the aggregated feature map across channels. This spatial attention map is then used to re-weight the different parts of the feature map before passing it on to the next layer. Finally, the Transformer networks used in NLP have also been adapted to work on CV tasks. This requires extracting patches from the input images before feeding them to the network as to mimic the sequential structure expected by the Transformer architecture. This approach achieved state-of-the-art performance on image classification tasks even rivalling well-established CNNs (Dosovitskiy et al., 2020).

DL architectures for EEG

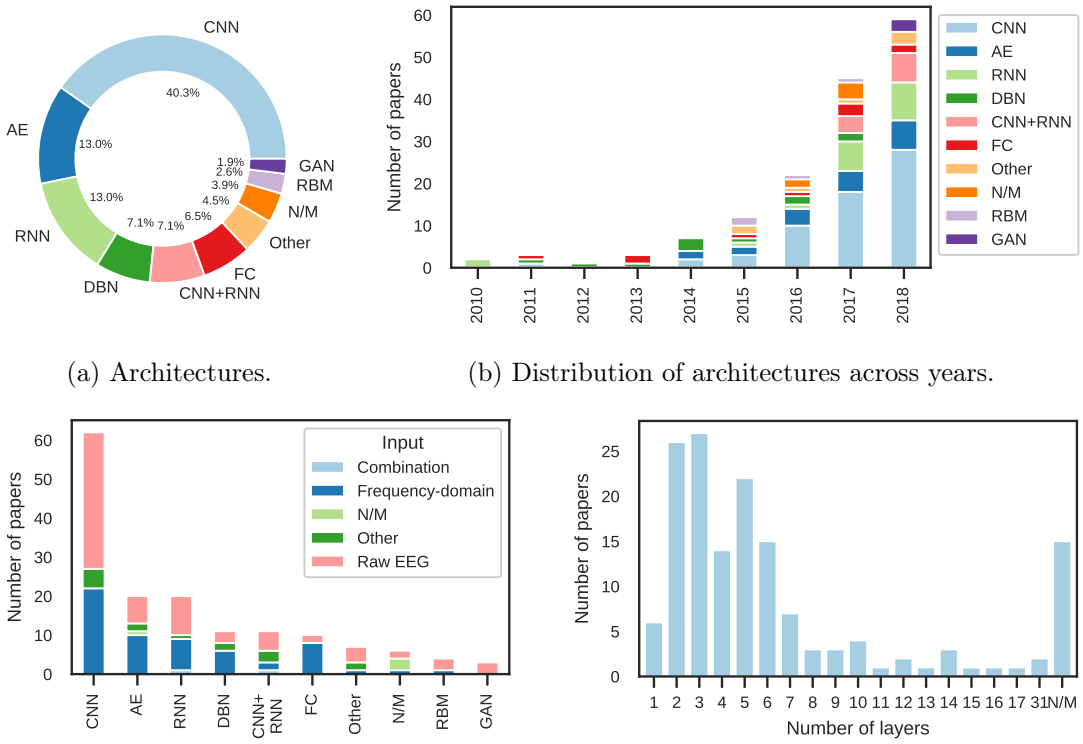
As shown in Figure 1.7, CNNs were found to be the most popular type of architecture used with EEG in Roy et al. (2019a) (40% of all papers reviewed). A large proportion of studies using CNNs further used raw EEG as input. Moreover, the great majority of studies surveyed used architectures with at most 10 layers, suggesting relatively shallow models can already yield satisfactory performance on EEG data.

EEG-focused architectures are commonly designed to replicate the usual steps of a signal processing pipeline. For instance, many architectures have been designed to process temporal and spatial information separately in the earlier stages of the network, replicating the temporal and spatial filtering steps typically employed when analyzing EEG (Manor and Geva, 2015; Kwak et al., 2017; Zafar et al., 2017; Behncke et al., 2017; Schirrmeister et al., 2017; Chambon et al., 2018). Two of these models were used extensively in this thesis and will be further described in Chapter 2: ShallowNet (also called FBCSPNet) (Schirrmeister et al., 2017) and StagerNet (Chambon et al., 2018).

1.2.3 Training neural networks

Neural networks are commonly trained using *gradient descent* and the *backpropagation* algorithm. We briefly describe these two foundational algorithms next.

Gradient descent is an optimization technique used to find a value $\mathbf{w}^*$ which minimizes a function $y = f(\mathbf{w})$ (Cauchy et al., 1847). Starting from an initial (*e.g.,* random) guess $\mathbf{w}_0$, a small step is taken in the opposite direction of the gradient $\nabla_{\mathbf{w}} y$, which indicates the direction in which f increases the most rapidly. The size of the step is controlled by the *learning rate* λ: the larger it is, the more rapidly the function might decrease, however a large λ also increases the probability of overshooting and getting worse results. With a well calibrated λ though, we are likely to find a value $\mathbf{w}_1$ that returns a lower value of y, *i.e.,* $f(\mathbf{w}_1) < f(\mathbf{w}_0)$. This procedure can be repeated multiple times to find the *minimum* of f. This minimum might be *local* (the smallest value of all neighboring values) or *global* (the smallest possible value of f). When training neural networks, this technique is used to find the parameters $\boldsymbol{\theta}^*$ (*i.e.,* the weights and biases of the different layers) that minimize a cost function, such as a categorical cross-entropy or MSE (Section 1.2.1).



(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.

Figure 1.7 – Deep learning architectures used in the studies included in Roy et al. (2019a). "N/M" stands for "Not mentioned" and includes the papers in which no information was provided about the deep learning methodology component under analysis. CNN: convolution neural network; AE: autoencoder; RNN: recurrent neural network; DBN: deep belief network; FC: fully-connected network; RBM: restricted Boltzmann machine; GAN: generative adversarial network.

To apply gradient descent to deep learning, we must therefore compute the gradient $\nabla_{\boldsymbol{\theta}}\mathcal{L}$ of a neural network's cost function with respect to all of its parameters. Backpropagation is an efficient procedure for computing this gradient (Rumelhart et al., 1986). First, an input $\mathbf{x}$ is *forward propagated* through the neural network $f_{\boldsymbol{\theta}}$ up to the output layer, yielding a prediction $\hat{y}$. Second, $\hat{y}$ is used to estimate the value of the loss function $\mathcal{L}$. Finally, the chain rule of calculus is used to compute the gradient of the loss function with respect to the different parameters in the network, starting from the output layer and moving backwards all the way to the input layer. Once these gradients are calculated, a gradient descent step is taken to update the parameters such that the model now produces a lower value for the cost function.

In practice, neural networks are often trained using *minibatch* gradient descent, *i.e.,* the gradient is computed over a set of multiple examples. This both speeds up computation thanks to parallelization on graphical processing units (GPUs) and regulates the level of noise present in the estimate of the gradient. Finally, modern deep learning optimizers often use additional tricks to make optimization less sensitive to *hyperparameters*, *i.e.,* learning algorithm settings that are fixed externally, and not tuned through the learning procedure (Goodfellow et al., 2016). For instance, the Adam algorithm

(Kingma and Ba, 2014) makes the optimization process more robust to the choice of initial learning rate λ_0 by adaptively rescaling the gradient of each parameter.

1.2.4 Learning tasks

When introducing the learning problem (Section 1.2.1) and describing applications of EEG (Section 1.1.2), we focused on *supervised* learning problems. In supervised learning, *e.g.,* classification or regression tasks, the learning algorithm has access to targets or *labels* $y \in \mathbb{Y}$ which are used to learn the mapping $f_{\boldsymbol{\theta}} : \mathcal{X} \to \mathcal{Y}$. This is akin to having access to a teacher who corrects us after we guess the answer to a question, allowing us to quickly identify our mistakes and change our internal model to better answer the next time. The supervised learning setting is widely studied from both theoretical and empirical perspectives. As a result, procedures to train machine learning models in this scenario are fairly well understood. However, obtaining targets for entire datasets can be time-consuming, expensive, or even impossible. This means it is not always possible to train a model with a supervised objective, or that the amount of available labels may, in some cases, be too limited to expect good performance.

In *unsupervised* learning, in contrast, there are no labels. Therefore, it is not possible to directly learn a mapping like in supervised learning, although different tasks can be used to learn the structure of the data *e.g.,* density estimation, clustering and representation learning. Because, by definition, there is no need for resource-hungry targets in unsupervised learning, it is often the case that the amount of data available to a model is orders of magnitude larger than for supervised problems. However, devising an objective function that will lead to good results can be challenging. Recently, self-supervised learning (SSL) has been proposed as an effective unsupervised learning strategy where supervision is provided by the very structure of the data under study (Jing and Tian, 2021). By reframing an unsupervised learning problem as a supervised one, SSL allows the use of standard, better understood optimization procedures. Specifically, in SSL, a model is trained on a *pretext task*, which must be sufficiently related to the *downstream task* (*i.e.,* the task we are actually interested in, but for which there are limited or no annotations) such that similar representations are likely to be learned to carry it out. We will discuss SSL in more detail in Chapter 2, when presenting self-supervised algorithms for learning on clinical EEG time series.

Finally, in some settings, only part of the data is labelled. This gives rise to a *semi-supervised* learning problem. Different methods have been proposed to improve performance in this scenario, *e.g.,* self-training (Yarowsky, 1995) and label propagation (Zhu and Ghahramani, 2002). For instance, in self-training, a classifier trained with a limited number of labelled examples is used to predict labels on the unlabelled examples. Examples for which the classifier produces predictions with a high enough level of confidence are then added to the training set, and the model is trained again. This process can be repeated a specific number of times, or until no new examples are added during the prediction phase. An alternative approach, popular in the world of deep learning, is to leverage representation learning methods to first learn a good representation of the input data in an unsupervised manner, and then to reuse this representation to train shallow classifiers with the labelled data or use the pre-trained neural network as an initialization for the supervised network. This semi-supervised setting will be studied in Chapter 2 as well when investigating SSL approaches for EEG.

1.2.5 Opportunities for deep learning and EEG

There are multiple ways in which DL may improve and extend existing EEG processing methods. First, the hierarchical nature of DNNs means features can potentially be learned on raw or minimally preprocessed data, reducing the need for domain-specific processing and feature extraction pipelines. Because they are developed in a purely data-driven way, features learned through a DNN may also be more effective or expressive than the ones engineered by humans. Importantly, DL makes it possible to learn highly informative features in a completely unsupervised manner with *e.g.,* self-supervision.

Second, as is the case in the multiple domains where DL has surpassed the previous state of the art, it has the potential to substantially improve model performance on a variety of tasks. Such improvements have already been hinted at by previous reports (Roy et al., 2019a), although more work will be necessary to confirm these findings.

Third, DL facilitates the learning of tasks that are less often attempted on EEG data, such as generative modelling (Goodfellow, 2016) and domain adaptation (Ben-David et al., 2007). Finally, when used jointly with attention mechanisms, DL also makes it possible to explain decisions through the visualization of attention maps.

1.2.6 Open questions and challenges for deep learning and EEG

There remain multiple open questions and challenges concerning the use of deep learning for EEG tasks. First and foremost, the amount of data that is required to train deep learning models on EEG is still to be determined. In many cases, the datasets typically available in EEG research contain a limited number of examples (as compared *e.g.,* to the quantities of data that have enabled the current state-of-the-art in DL-heavy domains such as CV and NLP). With data collection being relatively expensive (when compared to collecting images or text, for instance) and data accessibility often being hindered by privacy concerns - especially for clinical data - openly available datasets of comparable sizes are not common. Some initiatives are actively tackling this problem though, *e.g.,* the TUH EEG corpus, a publicly available dataset of more than 30,000 clinical EEG recordings (Harati et al., 2014) or the National Sleep Research Resource, which contains more than 31,000 sleep EEG recordings (Zhang et al., 2018). Generally speaking, opportunities for collecting larger-scale datasets do exist. Low-cost mobile EEG, in particular, makes the collection of such large unlabelled datasets much easier. However, in many cases the data will be unlabelled, as the labelling process can become prohibitively expensive or time-consuming when it requires clinical expertise or carefully crafted experimental protocols. We tackle this problem in Chapter 2 and suggest a further alternative when only some weak labels are available in Chapter 4.

Second, the peculiarities of EEG, such as its low signal-to-noise ratio (SNR), make EEG data different from other types of data for which DL has been most successful, *e.g.,* images, text and speech. Therefore, the architectures and practices that are currently used in DL might not be directly applicable to EEG processing. Specifically, the strong noise characteristics that can be expected in real-world EEG scenarios might affect deep learning models in a different way from "traditional" deep learning tasks. We study this scenario and propose a solution to mitigating this problem in Chapter 3.

Third, DNNs are notoriously seen as black boxes, when compared to more traditional "shallow" methods. Indeed, straightforward model inspection techniques such as visualizing the weights of a linear classifier are not applicable to deep neural networks and

as a result their decisions are much harder to interpret. This is problematic in clinical settings for instance, where understanding and explaining the choices 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. In response to this problem, multiple model inspection techniques have been proposed (Roy et al., 2019a), with some work specifically tailored to EEG (Schirrmeister et al., 2017; Hartmann et al., 2018; Ghosh et al., 2018). In addition to these strategies, attention mechanisms can also be used to improve the interpretability of DNNs (Niu et al., 2021): through the visualization of the resulting attention maps, it is possible to infer the importance of the different dimensions and elements of a neural network's input. Overall, 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.

Finally, one major challenge the field of DL-EEG has been facing is *reproducibility*. Indeed, one of the conclusions of the comprehensive literature review of Roy et al. (2019a) was that the results of most papers included in the analysis were hard or even impossible to reproduce, due to missing information, lack of code sharing, or the use of results obtained on private datasets only. This lack of reproducibility in turn makes it harder to provide a good answer as to whether deep learning is generally better suited to EEG than classical machine learning that relies on feature engineering and shallow models (preliminary or domain-specific answers can be found in Figure 14 of Roy et al. (2019a) or in the benchmarked results reported by Gemein et al. (2020)). Open source libraries such as braindecode[7] (Schirrmeister et al., 2017) and initiatives such as the DL-EEG Portal[8] (Roy et al., 2019b) will be key to fostering reproducibility in the field.

1.3 Contributions

The thesis is organized as follows. First, the introduction (this chapter) provides background information on both EEG and deep learning, which are the two common threads throughout this thesis. Additionally, some results from a comprehensive review of the literature on deep learning and EEG are presented to motivate the importance and timeliness of this topic (Roy et al., 2019a). The collected metadata and associated code developed for the literature review have been made publicly available[9], along with an online portal (in development) to help the community foster reproducibility[10].

In Chapter 2, we present experiments on using SSL to leverage unlabelled EEG data in unsupervised and semi-supervised scenarios. Motivated by the difficulty obtaining expert labels on EEG data presents, we propose three self-supervised tasks that can be used to learn representations of EEG data without any expert supervision. Experiments on two large public datasets and on two distinct EEG classification tasks (sleep staging and pathology detection) demonstrate that the clinical structure of EEG data can in fact be learned by leveraging its temporal structure alone, and that our proposed methodology can improve downstream task performance and reduce overall reliance on labelled data. The results presented in this chapter have been published in Banville

[7] https://github.com/braindecode/braindecode
[8] https://dl-eeg.com/
[9] https://github.com/hubertjb/dl-eeg-review
[10] https://dl-eeg.com/

et al. (2019, 2021a). Additionally, an implementation of the core proposed method has been made publicly available in the braindecode library[11].

Next, Chapter 3 focuses on the challenge of channel corruption in real-world EEG. We introduce dynamic spatial filtering (DSF), an attention module that is specifically designed to handle channel corruption in sparse real-world EEG. We present experiments on the sleep staging and pathology detection classification tasks, including a dataset of real-world EEG data collected by users of a low-cost mobile EEG headset. We demonstrate the usability and interpretability of the DSF approach, and show that, in sparse EEG settings, it outperforms a commonly used automated noise handling strategy. The work presented in this chapter has been shared as a preprint (Banville et al., 2021b) and is currently under revision at *NeuroImage*. Code for reproducing these experiments is publicly available online[12].

Finally, Chapter 4 presents the first application of the brain age prediction framework on real-world, low-cost mobile EEG. This framework, which uses age predictions as a proxy measure of pathological aging, can be seen as a variation of the self-supervised techniques presented in Chapter 2. Through experiments on more than 1,500 at-home EEG recordings, we demonstrate that brain age can indeed be predicted from out-of-the-lab EEG and that it is a promising biomarker of pathological aging. These experiments make use of the DSF attention module introduced in Chapter 3 and further show its usability on real-world EEG data. A manuscript based on the work presented in this chapter is currently being prepared for submission at the *Journal of Sleep Research*.

1.4 Publications

The work presented in this thesis led to the following publications:

- 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, 2019a
  *Shared first authorship.

- **Hubert Banville**, Graeme Moffat, Isabela Albuquerque, Denis-Alexander Engemann, Aapo Hyvärinen, and Alexandre Gramfort. Self-supervised representation learning from electroencephalography signals. In *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2019

- **Hubert Banville**, Omar Chehab, Aapo Hyvärinen, Denis-Alexander Engemann, and Alexandre Gramfort. Uncovering the structure of clinical EEG signals with self-supervised learning. *Journal of Neural Engineering*, 18(4):046020, 2021a

- **Hubert Banville**, Sean UN Wood, Chris Aimone, Denis-Alexander Engemann, and Alexandre Gramfort. Robust learning from corrupted EEG with dynamic spatial filtering. *arXiv preprint arXiv:2105.12916*, 2021c

[11] https://braindecode.org/auto_examples/plot_relative_positioning.html
[12] https://github.com/hubertjb/dynamic-spatial-filtering

- **Hubert Banville**, Sean UN Wood, Maurice Abou Jaoude, Chris Aimone, Alexandre Gramfort, and Denis-Alexander Engemann. Brain age as a proxy measure of neurophysiological health using low-cost mobile EEG. Manuscript in preparation, 2021b

2

Uncovering the structure of clinical EEG signals with self-supervised learning

Contents

For common applications of EEG, such as sleep staging and pathology detection, labels are needed in order to train supervised machine learning models. However, in many cases, obtaining labels is a costly and time-consuming process: for instance, it can require trained experts to visually analyze the collected EEG data, window by window. While such expert annotators may be available in some contexts, such as small-scale data collection in a lab or clinic, or in some prominent data collection projects (Goldberger et al., 2000; Obeid and Picone, 2016; Zhang et al., 2018), they are typically a bottleneck for EEG researchers. As the capacity to collect large EEG datasets increases, thanks to advances in mobile and consumer-focused technology, we need new ways to leverage unlabelled EEG data.

In this chapter, we explore self-supervised learning (SSL) as a way to learn representations of EEG signals with deep learning in a purely unsupervised fashion. We present three SSL tasks that rely on predicting the temporal context of EEG windows and evaluate them on sleep staging and pathology detection tasks. Specifically, we evaluate these methods in a semi-supervised scenario, where the objective is to improve model performance when only a small amount of labelled EEG data, but large amounts of unlabelled EEG data, are available.

Through experiments on two large public datasets with thousands of recordings, we show that SSL-learned features consistently outperform purely supervised deep neural networks in low-labelled data regimes, while remaining competitive when all labels are available. Additionally, the representations learned with each method reveal clear latent structures related to physiological and clinical phenomena, such as age. Our results suggest that SSL is a promising technique for learning representations on unlabelled EEG data, improving our ability to make use of real-world EEG in research and clinical applications.

Since the publication of the articles on which this chapter is based (Banville et al., 2019, 2021a), other groups have produced research on SSL and EEG (Cheng et al., 2020; Mohsenvand et al., 2020; Xu et al., 2020a; Kostas et al., 2021; Han et al., 2021; Ye et al., 2021), in many cases citing and building on the methods we introduced. Their results further validate the idea that self-supervision is a powerful representation learning task, with great potential for low-labelled data regimes.

This chapter is based on the following work:

- **Hubert Banville**, Graeme Moffat, Isabela Albuquerque, Denis-Alexander Engemann, Aapo Hyvärinen, and Alexandre Gramfort. Self-supervised representation learning from electroencephalography signals. In 2*019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2019

- **Hubert Banville**, Omar Chehab, Aapo Hyvärinen, Denis-Alexander Engemann, and Alexandre Gramfort. Uncovering the structure of clinical EEG signals with self-supervised learning. *Journal of Neural Engineering*, 18(4):046020, 2021a

Implementations of the proposed SSL tasks are available in the braindecode[1] Python library. (Schirrmeister et al., 2017)

2.1 Introduction

Electroencephalography (EEG) and other biosignal modalities have enabled numerous applications inside and outside of the clinical domain, *e.g.,* studying sleep patterns and their disruption (Ghassemi et al., 2018), monitoring seizures (Acharya et al., 2013) and brain-computer interfacing (Lotte et al., 2018). In the last few years, the availability and portability of these devices has increased dramatically, effectively democratizing their use and unlocking the potential for positive impact on people's lives (Mihajlović et al., 2014; Casson, 2019). For instance, applications such as at-home sleep staging and apnea detection, pathological EEG detection, mental workload monitoring, etc. are now entirely possible.

[1] https://github.com/braindecode/braindecode

In all these scenarios, monitoring modalities generates an ever-increasing amount of data which needs to be interpreted. Therefore, predictive models that can classify, detect and ultimately "understand" physiological data are required. Traditionally, this type of modelling has mostly relied on supervised approaches, where large datasets of annotated examples are required to train models with high performance.

However, obtaining accurate annotations on physiological data can prove expensive, time consuming or simply impossible. For example, annotating sleep recordings requires trained technicians to go through hours of data visually and label 30-s windows one-by-one (Malhotra and Avidan, 2013). Clinical recordings such as those used to diagnose epilepsy or brain lesions must be reviewed by neurologists, who might not always be available. More broadly, noise in the data and the complexity of brain processes of interest can make it difficult to interpret and annotate EEG signals, which can lead to high inter-rater variability, *i.e.,* label noise (Younes et al., 2016; Engemann et al., 2018b). Furthermore, in some cases, knowing exactly what the participants were thinking or doing in cognitive neuroscience experiments can be challenging, making it hard to obtain accurate labels. In imagery tasks, for instance, the subjects might not be following instructions or the process under study might be difficult to quantify objectively (*e.g.,* meditation, emotions). Therefore, a new paradigm that does not rely primarily on supervised learning is necessary for making use of large unlabelled sets of recordings such as those generated in the scenarios described above. However, traditional unsupervised learning approaches such as clustering and latent factor models do not offer fully satisfying answers as their performance is not as straightforward to quantify and interpret as supervised ones.

Self-supervised learning (SSL) is an unsupervised learning approach that learns representations from unlabelled data, exploiting the structure of the data to provide supervision (**?**) (introduced in Section 1.2.4). SSL comprises a *pretext* and a *downstream* task. The downstream task is the task we are actually interested in but for which there are limited or no annotations. The pretext task, on the other hand, must be sufficiently related to the downstream task such that similar representations should be employed to carry it out. Importantly, it must also be possible to generate the annotations for this pretext task using the unlabelled data alone. For example, in a computer vision scenario, one could use a jigsaw puzzle task where patches are extracted from an image, scrambled randomly and then fed to a neural network that is trained to recover the original spatial ordering of the patches (Noroozi and Favaro, 2016). If the network performs this task reasonably well, then it is conceivable that it has learned some of the structure of natural images, and that the trained network could be reused as a feature extractor or weight initialization on a smaller-scale supervised learning problem such as object recognition. Apart from facilitating the downstream task and/or reducing the number of annotated examples necessary, self-supervision can also uncover more general and robust features than those learned in a specialized supervised task (van den Oord et al., 2018). Therefore, given the potential benefits of SSL, can it be used to enhance the analysis of EEG?

To date, most applications of SSL have focused on domains where plentiful annotated data is already available, such as computer vision (**?**) and natural language processing (Mikolov et al., 2013; Devlin et al., 2018). Particularly in computer vision, deep networks are often trained with fully supervised tasks (*e.g.,* ImageNet pretraining). In this case, enough labelled data is available such that direct supervised learning on the downstream task is already in itself competitive (He et al., 2019b). SSL has an even

greater potential in domains where low-labelled data regimes are common and supervised learning's effectiveness is limited, *e.g.,* biosignal and EEG processing. Despite this, few studies on SSL and biosignals have been published.[2] These studies either focus on limited downstream tasks and datasets (Yuan et al., 2017), or test their approach on signals other than EEG (Sarkar and Etemad, 2020).

Therefore, it still remains to be shown whether self-supervision can truly bring improvements over standard supervised approaches on EEG, and if this is the case, what the best ways of applying it are. Specifically, can we learn generic representations of EEG with self-supervision and, in doing so, reduce the need for costly EEG annotations? Given the growing popularity of deep learning as a processing tool for EEG (Roy et al., 2019a), the answer could have a significant impact on current practices in the field. Indeed, while deep learning is notoriously data-hungry, an overwhelmingly large part of all neuroscience research happens in the low-labelled data regime, including EEG research: clinical studies with a few hundred subjects are often considered to be big data, while large-scale studies are much rarer and usually originate from research consortia (Shafto et al., 2014; Obeid and Picone, 2016; Zhang et al., 2018; Bycroft et al., 2018; Engemann et al., 2020). Therefore, it is to be expected that the performance reported by most deep learning-EEG studies - often in low-labelled data regimes - has so far remained limited and does not clearly outperform those of conventional approaches (Roy et al., 2019a). By leveraging unlabelled data, SSL can effectively create a lot more examples, which could enable more successful applications of deep learning to EEG.

In this chapter, we investigate the use of self-supervision as a general approach to learning representations from EEG data. To the best of our knowledge, we present the first detailed analysis of SSL tasks on multiple types of EEG recordings. We aim to answer the following questions:

1. What are good SSL tasks that capture relevant structure in EEG data?
2. How do SSL features compare to other unsupervised and supervised approaches in terms of downstream classification performance?
3. What are the characteristics of the features learned by SSL? Specifically, can SSL capture physiologically- and clinically-relevant structure from unlabelled EEG?

The rest of the chapter is structured as follows. Section 2.2 presents an overview of the SSL literature at the time of our publication (Banville et al., 2021a), then describes the different SSL tasks and learning problems considered in our study. We also introduce the neural architectures, baseline methods and data used in our experiments. Next, Section 2.3 reports the results of our experiments on EEG. Lastly, we discuss the results in Section 2.4.

2.2 Methods

2.2.1 State-of-the-art self-supervised learning approaches

Although it has not always been known as such, SSL has already been used in many other fields. In computer vision, multiple approaches have been proposed that rely on the spatial structure of images and the temporal structure of videos. For example,

[2] At the time of writing the two papers on which this chapter is based (Banville et al., 2019, 2021a).

a context prediction task was used to train feature extractors on unlabelled images in Doersch et al. (2015) by predicting the position of a randomly sampled image patch relative to a second patch. Using this approach to pretrain a neural network, the authors reported improved performance as compared to a purely supervised model on the Pascal VOC object detection challenge. These results were among the first showing that self-supervised pretraining could help improve performance when limited annotated data is available. Similarly, the jigsaw puzzle task mentioned above (Noroozi and Favaro, 2016) led to improved downstream performance on the same dataset. In the realm of video processing, approaches based on temporal structure have also been proposed: for instance, in Misra et al. (2016), predicting whether a sequence of video frames was ordered or shuffled was used as a pretext task and tested on a human activity recognition downstream task. The interested reader can find other applications of SSL to images in **?**.

Similarly, modern natural language processing (NLP) tasks often rely on self-supervision to learn word embeddings, which are at the core of many applications (Turian et al., 2010). For instance, the original *word2vec* model was trained to predict the words around a center word or a center word based on the words around it (Mikolov et al., 2013), and then reused on a variety of downstream tasks (Nayak et al., 2016). More recently, a dual-task self-supervised approach, *BERT*, led to state-of-the-art performance on 11 NLP tasks such as question answering and named entity recognition (Devlin et al., 2018). The high performance achieved by this approach showcases the potential of SSL for learning general-purpose representations.

Lately, more general pretext tasks inspired by early work on representation learning (Becker and Hinton, 1992; Becker, 1993) along with new improved methodology have led to strong results that have begun to rival purely supervised approaches. For instance, contrastive predictive coding (CPC), an autoregressive prediction task in latent space, was successfully used for images, text and speech (van den Oord et al., 2018). Given an encoder and an autoregressive model, the task consists of predicting the output of the encoder for future windows (or image patches or words) given a context of multiple windows. The authors presented several improved results on various downstream tasks; a follow-up further showed that higher-capacity networks could improve downstream performance even more, especially in low-labelled data regimes (Hénaff et al., 2019). Momentum contrast (MoCo), rather than proposing a new pretext task, is an improvement upon contrastive tasks, *i.e.,* tasks where a classifier must predict which of two or more inputs is the true sample (He et al., 2019a; Chen et al., 2020b). By improving the sampling of negative examples in contrastive tasks, MoCo helped boost the efficiency of SSL training as well as the quality of the representations learned. Similarly, the *SimCLR* approach of Chen et al. (2020a) showed that using the right data augmentation transforms (*e.g.,* random cropping and color distortion on images) and increasing batch size could lead to significant improvements in downstream performance.

The ability of SSL-trained features to demonstrably generalize to downstream tasks justifies a closer look at their statistical structure. A general and theoretically grounded approach was recently formalized in Hyvärinen and Morioka (2017); Hyvärinen et al. (2019) from the perspective of nonlinear independent component analysis (ICA). In this generalized framework, an observation $\mathbf{x}$ is embedded using an invertible neural network, and contrasted against an auxiliary variable $\mathbf{u}$ (*e.g.,* the time index, the index of a segment or the history of the data). A discriminator classifies the pair by learning to predict whether $\mathbf{x}$ is paired with its corresponding auxiliary variable $\mathbf{u}$ or a

perturbed (random) one $\mathbf{u}^*$. When the data exhibits certain structure (*e.g.,* autocorrelation, non-stationarity, non-gaussianity), the embedder trained on this contrastive task will perform identifiable nonlinear ICA (Hyvärinen et al., 2019). Most of the previously introduced SSL tasks can be viewed through this framework. Given the widespread use of linear ICA as a preprocessing and feature extraction tool in the EEG community (Makeig et al., 1997; Jung et al., 1997; Parra et al., 2005; Ablin et al., 2018), an extension to the nonlinear regime is a natural step forward and could help improve traditional processing pipelines.

Remarkably, very few studies have applied SSL to biosignals despite its potential to leverage large quantities of unlabelled data. In Yuan et al. (2017), a model inspired by *word2vec*, called *wave2vec*, was developed to work with EEG and electrocardiography (ECG) time series. Representations were learned by predicting the features of neighbouring windows from the concatenation of time-frequency representations of EEG signals and demographic information. This approach was however only tested on a single EEG dataset and was not benchmarked against fully supervised deep learning approaches or expert feature classification. SSL has also been applied to electrocardiography (ECG) as a way to learn features for a downstream emotion recognition task: in Sarkar and Etemad (2020), a transformation discrimination pretext task was used in which the model had to predict which transformations had been applied to the raw signal. While these results show the potential of self-supervised learning for biosignals, a more extensive analysis of SSL targeted at EEG is required to pave the way for practical applications.

2.2.2 Self-supervised learning pretext tasks for EEG

In this section, we describe the three SSL pretext tasks used in this chapter. A visual explanation of the tasks can be found in Figure 2.1. Implementations of the proposed SSL tasks are available in the braindecode[3] Python library (Schirrmeister et al., 2017).

Relative positioning

To produce labelled samples from the multivariate time series $\mathbf{S}$, we propose to sample pairs of time windows $(\mathbf{X}_t, \mathbf{X}_{t'})$ where each window $\mathbf{X}_t$, $\mathbf{X}_{t'}$ is in $\mathbb{R}^{C\times T}$ and T is the duration of each window, and where the index t indicates the time sample at which the window starts in $\mathbf{S}$. The first window $\mathbf{X}_t$ is referred to as the *anchor window*. Our assumption is that an appropriate representation of the data should evolve slowly over time (akin to the driving hypothesis behind slow feature analysis (SFA) (Földiák, 1991; Becker, 1993; Wiskott and Sejnowski, 2002)) suggesting that time windows close in time should share the same label. In the context of sleep staging, for instance, sleep stages usually last between 1 to 40 minutes (Altevogt and Colten, 2006); therefore, nearby windows likely come from the same sleep stage, whereas faraway windows likely come from different sleep stages. Given $\tau_{pos} \in \mathbb{N}$, which controls the duration of the positive context, and $\tau_{neg} \in \mathbb{N}$, which corresponds to the negative context around each window $\mathbf{X}_i$, we sample N labelled pairs:

$$\mathcal{Z}_N = \{((\mathbf{X}_{t_i}, \mathbf{X}_{t_i'}), y_i) \mid i \in [\![N]\!], (t_i, t_i') \in \mathcal{T}, y_i \in \mathcal{Y}\},$$

[3] https://github.com/braindecode/braindecode

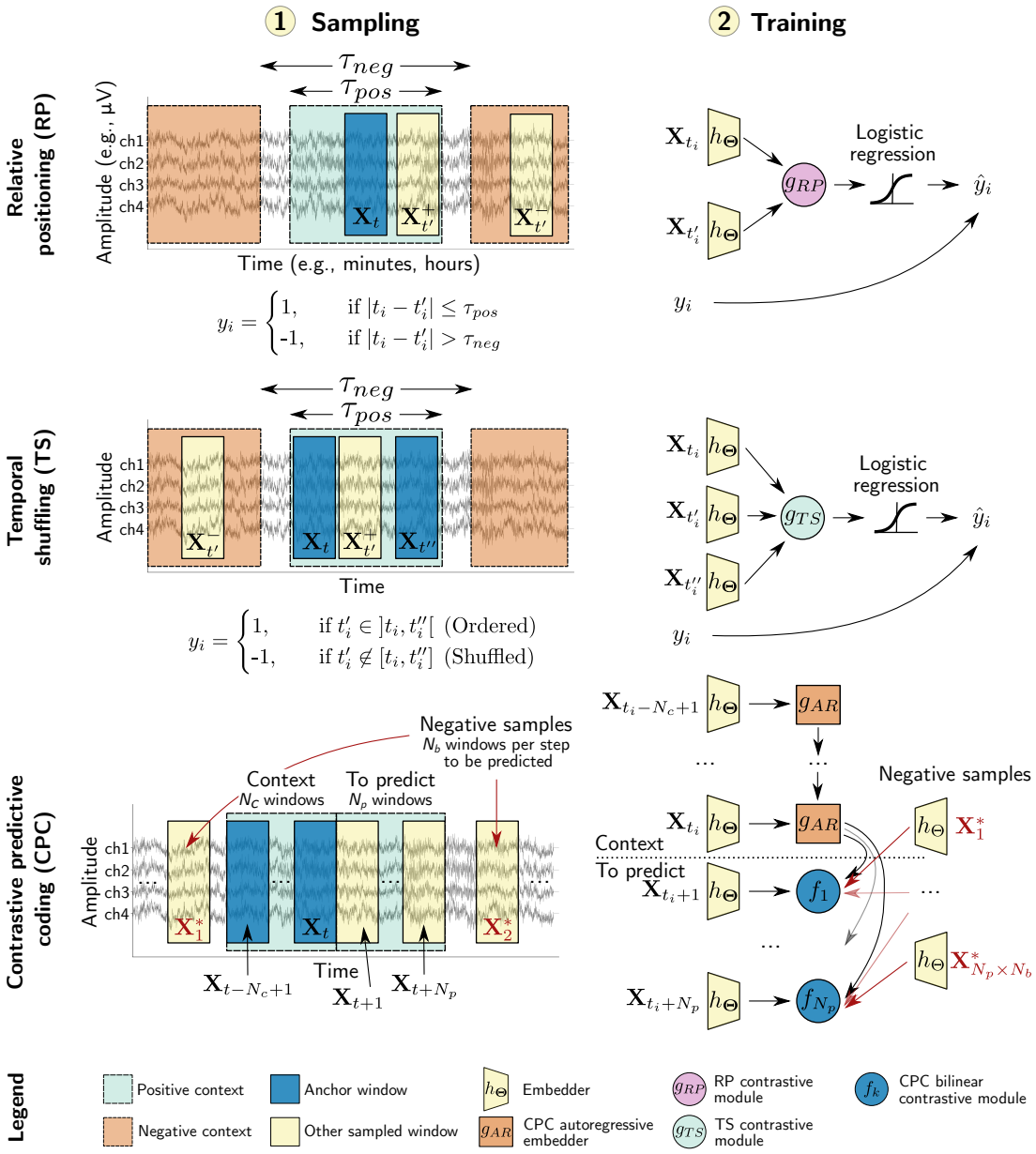


Figure 2.1 – Visual explanation of the three proposed SSL pretext tasks (relative positioning (RP), temporal shuffling (TS) and contrastive predictive coding (CPC)). The first column illustrates the sampling process by which examples are extracted from a time series $\mathbf{S}$ (EEG recording) in each pretext task. The second column describes the training process, where sampled examples are used to train a feature extractor $h_{\boldsymbol{\Theta}}$ end-to-end. **RP**: Pairs of windows are sampled from $\mathbf{S}$ such that the two windows of a pair are either close in time ("positive pairs") or farther away ("negative pairs"). $h_{\boldsymbol{\Theta}}$ is then trained to predict whether a pair is positive or negative. **TS**: Triplets of windows (rather than pairs) are sampled from $\mathbf{S}$. A triplet is given a positive label if its windows are ordered or a negative label if they are shuffled. $h_{\boldsymbol{\Theta}}$ is then trained to predict whether the windows of a triplet are ordered or shuffled. **CPC**: Sequences of $N_c + N_p$ consecutive windows are sampled from $\mathbf{S}$ along with random distractor windows ("negative samples"). Given the first N_c windows of a sequence (the "context"), a neural network is trained to identify which window out of a set of distractor windows actually follows the context.

where $\mathcal{Y} = \{-1, 1\}$ and $\mathcal{T} = \{(t, t') \in [\![M - T + 1]\!]^2 \mid \left|t - t'\right| \leq \tau_{pos}$ or $\left|t - t'\right| > \tau_{neg}\}$. Intuitively, $\mathcal{T}$ is the set of all pairs of time indices (t, t') which can be constructed from windows of size T in a time series of size M, given the duration constraints imposed by the particular choices of τ_{pos} and τ_{neg}[4]. Here $y_i \in \mathcal{Y}$ is specified by the positive or negative contexts parameters:

$$y_i = \begin{cases} 1, & \text{if } |t_i - t'_i| \leq \tau_{pos} \\ -1, & \text{if } |t_i - t'_i| > \tau_{neg} \end{cases} . \tag{2.1}$$

We ignore window pairs where $\mathbf{X}_{t'}$ falls outside of the positive and negative contexts of the anchor window $\mathbf{X}_t$. In other words, the label indicates whether two time windows are closer together than τ_{pos} or farther apart than τ_{neg} in time. Noting the connection with the task of Doersch et al. (2015), we call this pretext task relative positioning (RP).

In order to learn end-to-end how to discriminate pairs of time windows based on their relative position, we introduce two functions $h_{\boldsymbol{\Theta}}$ and g_{RP}. $h_{\boldsymbol{\Theta}} : \mathbb{R}^{C \times T} \to \mathbb{R}^D$ is a feature extractor with parameters $\boldsymbol{\Theta}$ which maps a window $\mathbf{X}$ to its representation in the feature space. Ultimately, we expect $h_{\boldsymbol{\Theta}}$ to learn an informative representation of the raw EEG input which can be reused in different downstream tasks. A contrastive module g_{RP} is then used to aggregate the feature representations of each window. For the RP task, $g_{RP} : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}^D$ combines representations from pairs of windows by computing an elementwise absolute difference, denoted by the $|\cdot|$ operator: $g_{RP}(h_{\boldsymbol{\Theta}}(\mathbf{X}), h_{\boldsymbol{\Theta}}(\mathbf{X}')) = |h_{\boldsymbol{\Theta}}(\mathbf{X}) - h_{\boldsymbol{\Theta}}(\mathbf{X}')| \in \mathbb{R}^D$. The role of g_{RP} is to aggregate the feature vectors extracted by $h_{\boldsymbol{\Theta}}$ on the two input windows and highlight their differences to simplify the contrastive task. Finally, a linear context discriminative model with coefficients $\mathbf{w} \in \mathbb{R}^D$ and bias term $w_0 \in \mathbb{R}$ is responsible for predicting the associated target y. Using the binary logistic loss on the predictions of g_{RP} we can write a joint loss function $\mathcal{L}(\boldsymbol{\Theta}, \mathbf{w}, w_0)$ as

$$\mathcal{L}(\boldsymbol{\Theta}, \mathbf{w}, w_0) = \sum_{(\mathbf{X}_t, \mathbf{X}_{t'}, y) \in \mathcal{Z}_N} \log(1 + \exp(-y[\mathbf{w}^\top g_{RP}(h_{\boldsymbol{\Theta}}(\mathbf{X}_t), h_{\boldsymbol{\Theta}}(\mathbf{X}_{t'})) + w_0])) \tag{2.2}$$

which we assume to be fully differentiable with respect to the parameters $(\boldsymbol{\Theta}, \mathbf{w}, w_0)$. Given the convention used for y, the predicted target is the sign of $\mathbf{w}^\top g(h_{\boldsymbol{\Theta}}(\mathbf{X}_t), h_{\boldsymbol{\Theta}}(\mathbf{X}_{t'})) + w_0$.

Temporal shuffling

We also introduce a variation of the RP task that we call temporal shuffling (TS), in which we instead sample two anchor windows $\mathbf{X}_t$ and $\mathbf{X}_{t''}$ from the positive context, and a third window $\mathbf{X}_{t'}$ that is either between the first two windows or in the negative context. We then construct window triplets that are either temporally ordered ($t < t' < t''$) or shuffled ($t < t'' < t'$ or $t' < t < t''$). We augment the number of possible triplets by also considering the mirror image of the previous triplets, *e.g.*, $(\mathbf{X}_t, \mathbf{X}_{t'}, \mathbf{X}_{t''})$ becomes $(\mathbf{X}_{t''}, \mathbf{X}_{t'}, \mathbf{X}_t)$. The label y_i then indicates whether the three windows are ordered or have been shuffled, similar to Misra et al. (2016).

The contrastive module for TS is defined as $g_{TS} : \mathbb{R}^D \times \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}^{2D}$ and is implemented by concatenating the absolute differences:

$$g_{TS}(h_{\boldsymbol{\Theta}}(x), h_{\boldsymbol{\Theta}}(\mathbf{X}'), h_{\boldsymbol{\Theta}}(\mathbf{X}'')) = (|h_{\boldsymbol{\Theta}}(\mathbf{X}) - h_{\boldsymbol{\Theta}}(\mathbf{X}')|, |h_{\boldsymbol{\Theta}}(\mathbf{X}') - h_{\boldsymbol{\Theta}}(\mathbf{X}'')|) \in \mathbb{R}^{2D} .$$

[4]The values of τ_{pos} and τ_{neg} can be selected based on prior knowledge of the signals and/or with a hyperparameter search.

Moreover, Eq. (2.2) is extended to TS by replacing g_{RP} by g_{TS} and introducing $\mathbf{X}_{t''}$ to obtain:

$$\mathcal{L}(\mathbf{\Theta}, \mathbf{w}, w_0) = \sum_{(\mathbf{X}_t, \mathbf{X}_{t'}, \mathbf{X}_{t''}, y) \in \mathcal{Z}_N} \log(1 + \exp(-y[\mathbf{w}^\top g_{TS}(h_{\mathbf{\Theta}}(\mathbf{X}_t), h_{\mathbf{\Theta}}(\mathbf{X}_{t'}), h_{\mathbf{\Theta}}(\mathbf{X}_{t''})) + w_0])) \ . \quad (2.3)$$

TS shares similarities with the unsupervised metric learning approach of Franceschi et al. (2019), however the sampling procedure and loss function both differ.

Contrastive predictive coding

The contrastive predictive coding (CPC) pretext task, introduced in van den Oord et al. (2018), is defined here in comparison to RP and TS, as all three tasks share key similarities. Indeed, CPC can be seen as an extension of RP, where the single anchor window $\mathbf{X}_t$ is replaced by a sequence of N_c non-overlapping windows that are summarized by an autoregressive encoder $g_{AR} : \mathbb{R}^{D \times N_c} \to \mathbb{R}^{D_{AR}}$ with parameters $\mathbf{\Theta}_{AR}$[5]. This way, the information in the context can be represented by a single vector $\mathbf{c}_t \in \mathbb{R}^{D_{AR}}$. g_{AR} can be implemented for example as a recurrent neural network with gated recurrent units (GRUs).

The context vector $\mathbf{c}_t$ is paired with not one, but N_p future windows (or *steps*) which immediately follow the context. Negative windows are then sampled in a similar way as with RP and TS when $\tau_{neg} = 0$, *i.e.,* the negative context is relaxed to include the entire time series. For each future window, N_b negative windows $\mathbf{X}^*$ are sampled inside each multivariate time series $\mathbf{S}$ ("same-recording negative sampling") or across all available $\mathbf{S}$ ("across-recording negative sampling"). For the sake of simplicity and to follow the notation of the original CPC article, we modify our notation slightly: we now denote a time window by $\mathbf{X}_t$ where t is the index of the window in the list of all non-overlapping windows of size T that can be extracted from a time series $\mathbf{S}$. Therefore, the procedure for building a dataset with N examples boils down to sampling sequences X^c, X^p and X^n in the following manner:

$$\begin{aligned}
X_i^c &= (\mathbf{X}_{t_i - N_c + 1}, \ldots, \mathbf{X}_{t_i}) && (N_c \text{ context windows}) \\
X_i^p &= (\mathbf{X}_{t_i + 1}, \ldots, \mathbf{X}_{t_i + N_p}) && (N_p \text{ future windows}) \\
X_i^n &= (\mathbf{X}_{t^*_{i_{1,1}}}, \ldots, \mathbf{X}_{t^*_{i_{1,N_b}}}, \ldots, \mathbf{X}_{t^*_{i_{N_p,1}}}, \ldots, \mathbf{X}_{t^*_{i_{N_p,N_b}}}) && (N_p N_b \text{ random negative windows})
\end{aligned}$$

where $t_i \in [\![N_c, M - N_p]\!]$. We denote with t^* time indices of windows sampled uniformly at random. The dataset then reads:

$$\mathcal{Z}_N = \{(X_i^c, X_i^p, X_i^n) \mid i \in [\![N]\!]\} \ . \quad (2.4)$$

As with RP and TS, the feature extractor $h_{\mathbf{\Theta}}$ is used to extract a representation of size D from a window $\mathbf{X}_t$. Finally, whereas the contrastive modules g_{RP} and g_{TS} explicitly relied on the absolute value of the difference between embeddings h, here for each future window $\mathbf{X}_{t+k}$ where $k \in [\![N_p]\!]$ a bilinear model f_k parametrized by $\mathbf{W}_k \in \mathbb{R}^{D \times D_{AR}}$ is used to predict whether the window chronologically follows the context $\mathbf{c}_t$ or not:

$$f_k(\mathbf{c}_t, h_{\mathbf{\Theta}}(\mathbf{X}_{t+k})) = h_{\mathbf{\Theta}}(\mathbf{X}_{t+k})^\top \mathbf{W}_k \mathbf{c}_t \quad (2.5)$$

[5]CPC's encoder g_{AR} has parameters $\mathbf{\Theta}_{AR}$, however we omit them from the notation for brevity.

The whole CPC model is trained end-to-end using the InfoNCE loss (van den Oord et al., 2018) (a categorical cross-entropy loss) defined as

$$\mathcal{L}(\boldsymbol{\Theta},\boldsymbol{\Theta}_{AR},\mathbf{W}_k,\ldots,\mathbf{W}_{k+N_p-1}) = \\ -\sum_{\substack{(X_i^c,X_i^p,X_i^n)\in\mathcal{Z}_N \\ \mathbf{c}_{t_i}=g_{AR}(X_i^c)}}\sum_{k=1}^{N_p}\log\left[\frac{\exp(f_k(\mathbf{c}_{t_i},h_{\boldsymbol{\Theta}}(\mathbf{X}_{t_i+k})))}{\exp(f_k(\mathbf{c}_{t_i},h_{\boldsymbol{\Theta}}(\mathbf{X}_{t_i+k})))+\sum_{j\in[\![N_b]\!]}\exp(f_k(\mathbf{c}_{t_i},h_{\boldsymbol{\Theta}}(\mathbf{X}_{t^*_{i_{k,j}}})))}\right] \tag{2.6}$$

While in RP and TS the model must predict *whether* a pair is positive or negative, in CPC the model must pick *which* of $N_b + 1$ windows actually follows the context. In practice, we sample batches of $N_b + 1$ sequences and for each sequence use the N_b other sequences in the batch to supply negative examples.

2.2.3 Downstream tasks

We performed empirical benchmarks of EEG-based SSL on two clinical problems that are representative of the current challenges in machine learning-based analysis of EEG: sleep monitoring and pathology screening. These two clinical problems commonly give rise to classification tasks, albeit with different numbers of classes and distinct data-generating mechanisms: sleep monitoring is concerned with biological events (*event level*) while pathology screening is concerned with single patients as compared to the population (*subject level*). These two clinical problems have generated considerable attention in the research community, which has led to the curation of large public databases. To enable fair comparisons with supervised approaches, we benchmarked SSL on the Physionet Challenge 2018 (Ghassemi et al., 2018; Goldberger et al., 2000) and the TUH Abnormal EEG (López et al., 2017) datasets.

First, we considered sleep staging, which is a critical component of a typical sleep monitoring assessment and is key to diagnosing and studying sleep disorders such as apnea and narcolepsy (Bathgate and Edinger, 2019). Sleep staging has been extensively studied in the machine (and deep) learning literature (Chambon et al., 2018; Motamedi-Fakhr et al., 2014; Roy et al., 2019a) (approximately 10% of papers reviewed in Roy et al. (2019a)), though not through the lens of SSL. Achieving fully automated sleep staging could have a substantial impact on clinical practice as 1) agreement between human raters is often limited (Younes et al., 2016) and 2) the annotation process is time-consuming and still largely manual (Malhotra and Avidan, 2013). Sleep staging typically gives rise to a 5-class classification problem where the possible predictions are W (wake), N1, N2, N3 (different levels of sleep) and R (rapid eye movement periods). Here, the task consists of predicting the sleep stages that correspond to 30-s windows of EEG.

Second, we applied SSL to pathology detection: EEG is routinely used in a clinical context to screen individuals for neurological conditions such as epilepsy and dementia (Smith, 2005; Micanovic and Pal, 2014). However, successful pathology detection requires highly specialized medical expertise and its quality depends on the expert's training and experience. Automated pathology detection could, therefore, have a major impact on clinical practice by facilitating neurological screening. This gives rise to classification tasks at the subject level where the challenge is to infer the patient's

diagnosis or health status from the EEG recording. In the TUH dataset, medical specialists have labelled recordings as either pathological or non-pathological, giving rise to a binary classification problem. Importantly, these two labels reflect highly heterogeneous situations: a pathological recording could reflect anomalies due to various medical conditions, suggesting a rather complex data-generating mechanism. Again, various supervised approaches, some of them leveraging deep architectures, have addressed this task in the literature (Lopez et al., 2015; Schirrmeister et al., 2017; Gemein et al., 2020), although none has relied on self-supervision.

These two tasks are further described in Section 4.2.2 when discussing the data used in our experiments.

2.2.4 Deep learning architectures

We used two different deep learning architectures as embedders $h_{\mathbf{\Theta}}$ in our experiments (see Figure 2.2 for a detailed description). Both architectures were convolutional neural networks composed of spatial and temporal convolution layers, which respectively learned to perform the spatial and temporal filtering operations typical of EEG processing pipelines.

The first one, which we call StagerNet, was adapted from previous work on sleep staging where it was shown to perform well for window-wise classification of sleep stages (Chambon et al., 2018). StagerNet is a 3-layer convolutional neural network optimized to process windows of 30 s of multichannel EEG. As opposed to the original architecture, 1) we used twice as many convolutional channels (16 instead of 8), 2) we added batch normalization after both temporal convolution layers[6] 3) we did not pad temporal convolutions and 4) we changed the dimensionality of the output layer to $D = 100$ instead of the number of classes (see Figure 2.2-1). This yielded a total of 62,307 trainable parameters.

The second embedder architecture, ShallowNet, was directly taken from previous literature on the TUH Abnormal dataset (Schirrmeister et al., 2017; Gemein et al., 2020). Originally designed to be a parametrized version of the filter bank common spatial patterns (FBCSP) processing pipeline common in brain-computer interfacing, ShallowNet has a single (split) convolutional layer followed by a squaring non-linearity, average pooling, a logarithm non-linearity, and a linear output layer. Batch normalization was used after the temporal convolution layer. Despite its simplicity, this architecture was shown in the benchmark of Gemein et al. (2020) to perform almost as well as the best model on the task of pathology detection on the TUH Abnormal dataset. We therefore used it as is, except for the dimensionality of the output layer which we also changed to $D = 100$ (see Figure 2.2-2). This yielded a total of 170,860 trainable parameters.

We used a GRU with a hidden layer of size $D_{AR} = 100$ for the CPC task's g_{AR}, for experiments on both datasets.

The Adam optimizer (Kingma and Ba, 2014) with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ and learning rate 5×10^{-4} was used. The batch size for all deep models was set to 256, except for CPC where it was set to 32. Training ran for at most 150 epochs or until the validation

[6] As described in van den Oord et al. (2018); He et al. (2019a), batch normalization can harm the network's ability to learn on the CPC pretext task. However, we did not see this effect on our models (likely because their capacity is relatively small) and alternatives such as no normalization or layer normalization (Ba et al., 2016) performed unfavorably. Therefore, we also used batch normalization in CPC experiments.

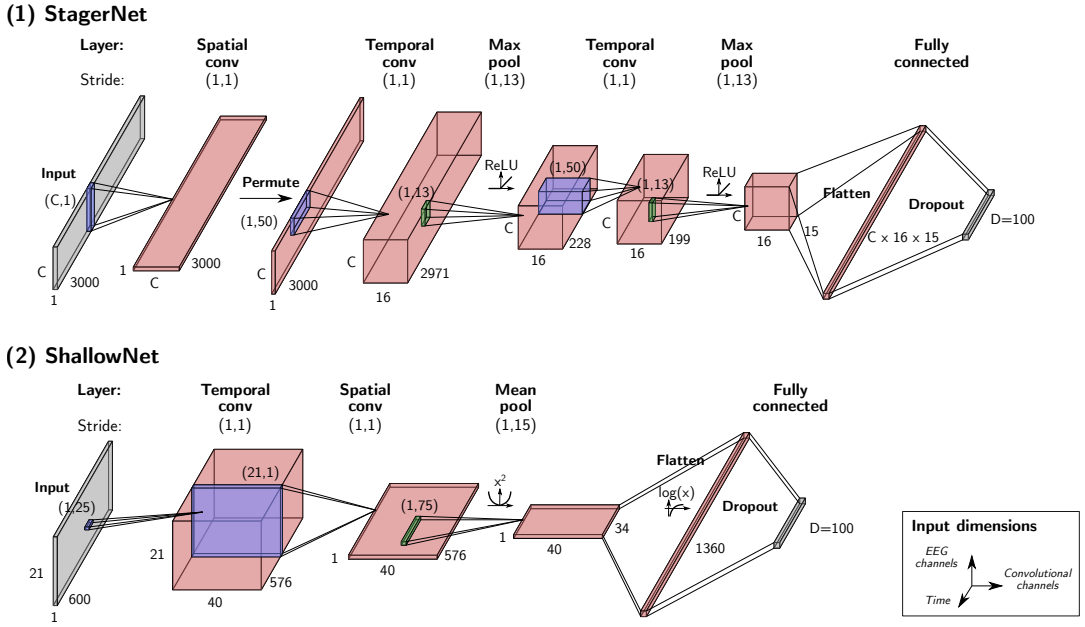


Figure 2.2 – Neural network architectures used as embedder h_Θ for 1) sleep EEG and 2) pathology detection experiments.

loss stopped decreasing for a period of a least 10 epochs (or 6 epochs for CPC). Dropout was applied to fully connected layers at a rate of 50% and a weight decay of 0.001 was applied to the trainable parameters of all layers. Finally, the parameters of all neural networks were randomly initialized using uniform He initialization (He et al., 2015).

2.2.5 Hyperparameter search procedure

The hyperparameter search in Section 2.3.3 was carried out using the following steps. First, embedders h_Θ were independently trained on the RP, TS and CPC tasks. The parameters of h_Θ were then frozen, and the different h_Θ were used as feature extractors to obtain sets of 100-dimensional feature vectors from the original input data. Finally, we trained linear logistic regression classifiers to perform the downstream tasks given the extracted features. We further varied the principal pretext task hyperparameters to understand their impact on both pretext and downstream task performance (see Table 2.1). In both cases, we compared the balanced accuracy on the validation set. For RP and TS, we focused our attention on τ_{pos} and τ_{neg}, which are used to control the size of the positive and negative contexts when sampling pairs or triplets of windows. As a first step, the values of τ_{pos} and τ_{neg} were varied jointly, *i.e.*, $\tau_{pos} = \tau_{neg}$, to avoid sampling "confusing" pairs or triplets of windows which could come from either the positive or negative classes. The best value was then used to set τ_{pos}, and a sweep over different τ_{neg} values was carried out. In a second step, we fixed τ_{neg} such that it encompassed all recordings, *i.e.*, negative windows were uniformly sampled from any recording in the dataset instead of being limited to the recording which contained the anchor window. We then varied τ_{pos} again with this second negative sampling strategy. For CPC, we studied the impact of the number of predicted windows ("# steps") (van den Oord et al., 2018) and, as for RP and TS, the type of negative sampling ("same-recording" vs. "across-recordings"). Finally, we again first varied the number of predicted windows and reused the best value to compare negative sampling strategies.

	PC18			TUAB		
	τ_{pos} (min)	τ_{neg} (min)	Negative sampling	τ_{pos} (min)	τ_{neg} (min)	Negative sampling
RP	0.5, **1**, 2, 5, 15, 30, 60, 120	0, 0.5, 1, 2, 5, **15**, 30, 60, 120	**same rec.**, across rec.	6 s, **12 s**, 0.5, 1, 2, 5, 10	0 s, 6 s, 12 s, 0.5, 1, **2**, 5, 10	same rec., **across rec.**
TS	1, 2, 5, 15, 30, 60, **120**	2, 5, 15, **30**, 60, 120	**same rec.**, across rec.	12 s, 0.5, 1, 2, 5, **10**	12 s, 0.5, 1, 2, 5, **10**	same rec., **across rec.**
	# predicted windows		Negative sampling	# predicted windows		Negative sampling
CPC	2, 4, 8, 12, **16**		**same rec.**, across rec.	2, 4, 8, **12**, 16		same rec., **across rec.**

Table 2.1 – SSL pretext task hyperparameter values considered in Section 2.3.3. Bold face indicates values that led to the highest downstream task performance.

2.2.6 Baselines

The SSL tasks were compared to four baseline approaches on the downstream tasks: 1) random weights, 2) convolutional autoencoders, 3) purely supervised learning and 4) handcrafted features.

The random weights baseline used an embedder whose weights were frozen after random initialization. The autoencoder (AE) was a more basic approach to representation learning, where a neural network made up of an encoder and a decoder learned an identity mapping between its input and its output, penalized by *e.g.,* a MSE loss (Kramer, 1991). Here, we used $h_{\mathbf{\Theta}}$ as the encoder and designed a convolutional decoder that inverts the operations of $h_{\mathbf{\Theta}}$. The purely supervised model was directly trained on the downstream classification problem, *i.e.,* it had access to the labelled data. To do so, we used the same embedder architectures as for the self-supervised tasks, but with an additional linear classification layer on top of the embedder, before training the whole model with a multi-class cross-entropy loss.

We also included traditional machine learning baselines based on handcrafted features. For sleep staging, we extracted the following features (Chambon et al., 2018): mean, variance, skewness, kurtosis, standard deviation, frequency log-power bands between (0.5, 4.5, 8.5, 11.5, 15.5, 30) Hz as well as all their possible ratios, peak-to-peak amplitude, Hurst exponent, approximate entropy and Hjorth complexity. This resulted in 37 features per EEG channel, which were concatenated into a single vector. In the event of an artefact causing missing values in the feature vector of a window, we imputed missing values feature-wise using the mean of the feature computed over the training set. For pathology detection, a pipeline based on covariance matrices, Riemannian geometry and non-linear classifiers was used, inspired by the results of Gemein et al. (2020) which

showed high accuracy on the evaluation set of the TUH Abnormal dataset. Specifically, the spatial covariance matrix of each window $\mathbf{X}$ was extracted then vectorized through a projection into its Riemannian tangent space (yielding an input dimensionality of $C(C+1)/2$ for C channels of EEG), allowing the use of a standard Euclidean-space classifier (Barachant et al., 2013b; Congedo et al., 2017; Lotte et al., 2018). Here, we did not average the covariance matrices per recording to allow a fair comparison with the other methods which work window-wise.

Finally, since we used SSL-learned features in a semi-supervised setting, *i.e.,* where a limited amount of labelled data is used in conjunction with a larger set of unlabelled examples (see Section 1.2.4), we included an additional baseline that draws on self-training (Yarowsky, 1995), for completeness.

For the downstream tasks, features learned with RP, TS, CPC and AE were classified using a linear logistic regression with L2-regularization parameter[7] $C = 1$, while hand-crafted features were classified using a random forest classifier with 300 trees, maximum depth of 15 and a maximum number of features per split of $\sqrt{F}$ (where F is the number of features)[8]. Balanced accuracy (bal acc), defined as the average per-class recall, was used to evaluate model performance on the downstream tasks. Moreover, during training, the loss was weighted to account for class imbalance. Models were trained using a combination of the `braindecode` (Schirrmeister et al., 2017), `MNE-Python` (Gramfort et al., 2014), `PyTorch` (Paszke et al., 2019), `pyRiemann` (Barachant et al., 2013b) and `scikit-learn` (Pedregosa et al., 2011) packages. Finally, deep learning models were trained on 1 or 2 Nvidia Tesla V100 GPUs for anywhere from a few minutes to 7h, depending on the amount of data, early stopping and GPU configuration.

2.2.7 Data

The experiments were conducted on two publicly available EEG datasets, which are described in Tables 2.2 and 2.3.

PC18 (train)			
	# windows		
W	158,020	# unique subjects	994
N1	136,858	# recordings	994
N2	377,426	Sampling frequency	200 Hz
N3	102,492	# EEG channels	6
R	116,872	Reference	M1 or M2
Total	891,668		

Table 2.2 – Description of the Physionet Challenge 2018 (PC18) dataset used in this study for sleep staging experiments.

[7] Varying C had little impact on downstream performance, and therefore we used a value of 1 across experiments.

[8] Random forest hyperparameters were selected using a grid search with maximum depth in $\{3, 5, 7, 9, 11, 13, 15\}$, and maximum number of features per tree in $\{\sqrt{F}, \log_2 F\}$ using the validation sets as described in Section 4.2.2.

TUAB				
	train	eval	# unique subjects	2329
	# recordings	# recordings	# recordings	2993
Normal	1371	150	Sampling frequency	250, 256, 512 Hz
Abnormal	1346	126	# EEG channels	27 to 36
Total	2717	276	Reference	Common average

Table 2.3 – Description of the TUH Abnormal (TUAB) dataset used in this study for EEG pathology detection experiments.

Physionet Challenge 2018 dataset

First, we conducted sleep staging experiments on the Physionet Challenge 2018 (PC18) dataset (Ghassemi et al., 2018; Goldberger et al., 2000). This dataset was initially released in the context of an open-source competition on the detection of arousals in sleep recordings, *i.e.,* short moments of wakefulness during the night. A total of 1,983 different individuals with (suspected) sleep apnea were monitored overnight and their EEG, electroculography (EOG), chin electromyography (EMG), respiration airflow and oxygen saturation measured. Specifically, 6 EEG channels from the international 10/20 system were recorded at 200 Hz: F3-M2, F4-M1, C3-M2, C4-M1, O1-M2 and O2-M1. The recorded data was then annotated by 7 trained scorers following the AASM manual (Berry et al., 2012) into sleep stages (W, N1, N2, N3 and R). Moreover, 9 different types of arousal and 4 types of sleep apnea events were identified in the recordings. As the sleep stage annotations are only publicly available on about half the recordings (used as the training set during the competition), we focused our analysis on these 994 recordings. In this subset of the data, mean age is 55 years old (min: 18, max: 93) and 33% of participants are female.

TUH Abnormal EEG dataset

We used the TUH Abnormal EEG dataset v2.0.0 (TUAB) to conduct experiments on pathological EEG detection (López et al., 2017). This dataset, a subset of the entire TUH EEG Corpus (Obeid and Picone, 2016), contains 2,993 recordings of 15 minutes or more from 2,329 different patients who underwent a clinical EEG in a hospital setting. Each recording was labelled as "normal" (1,385 recordings) or "abnormal" (998 recordings) based on detailed physician reports. Most recordings were sampled at 250 Hz (although some were sampled at 256 or 512 Hz) and contained between 27 and 36 electrodes. Moreover, the corpus is divided into a training and an evaluation set with 2,130 and 253 recordings each. The mean age across all recordings is 49.3 years old (min: 1, max: 96) and 53.5% of recordings are of female patients.

Data splits and sampling

We split the available recordings from PC18 and TUAB into training, validation and testing sets such that the examples from each recording were only in one of the sets (see Table 2.4).

For PC18, we used a 60-20-20% random split, meaning there were 595, 199 and 199 recordings in the training, validation and testing sets respectively. For RP and TS, 2,000 pairs or triplets of windows were sampled from each recording. For CPC, the

	PC18 # recordings	RP/TS # tuples	CPC # sequences	**TUAB** # recordings	RP/TS # tuples	CPC # sequences
Train	595	1,190,000	877,792	2,171	868,400	642,144
Valid	199	398,000	294,272	543	217,200	160,224
Test	199	398,000	292,608	276	110,400	81,184
Total	993	1,986,000	1,464,672	2,990	1,196,000	883,552

Table 2.4 – Number of recordings used in the training, validation and testing sets with PC18 and TUAB, as well as the number of examples for each pretext task.

number of batches to extract from each recording was computed as 0.05 times the number of windows in that recording; moreover, we set the batch size to 32.

For TUAB, we used the provided evaluation set as the test set. The recordings of the development set were split 80-20% into a training and a validation set. Therefore, we used 2,171, 543 and 276 recordings in the training, validation and testing sets. Since the recordings were shorter for TUAB, we randomly sampled 400 RP pairs or TS triplets instead of 2000 from each recording. We used the same CPC sampling parameters as for PC18.

Preprocessing

The preprocessing of the EEG recordings differed for the two datasets. On PC18, the raw EEG was first filtered using a 30 Hz FIR lowpass filter with a Hamming window, to reject higher frequencies that are not critical for sleep staging (Chambon et al., 2018; Aboalayon et al., 2016). The EEG channels were then downsampled to 100 Hz to reduce the dimensionality of the input data. For the same reason, we focused our analysis on two channels only. We selected F3-M2 and F4-M1 as, being closer to the eyes, they pick up more of the EOG activity critical for the classification of stage R. These channels are also close to the forehead region, whose lack of hair makes it a popular location for at-home polysomnography systems. Lastly, non-overlapping windows of 30 s of size (3000 x 2) were extracted.

On TUAB, a similar procedure to the one reported in Gemein et al. (2020) was used. The first minute of each recording was cropped to remove noisy data that occurs at the beginning of recordings. Longer files were also cropped such that a maximum of 20 minutes was used from each recording. Then, 21 channels that are common to all recordings were selected (Fp1, Fp2, F7, F8, F3, Fz, F4, A1, T3, C3, Cz, C4, T4, A2, T5, P3, Pz, P4, T6, O1 and O2). EEG channels were downsampled to 100 Hz and clipped at $\pm 800\,\mu V$ to mitigate the effect of large artifactual deflections in the raw data. Non-overlapping 6-s windows were extracted, yielding windows of size (600×21).

Finally, windows from both datasets with peak-to-peak amplitude below $1\,\mu V$ were rejected. The remaining windows were normalized channel-wise to have zero-mean and unit standard deviation.

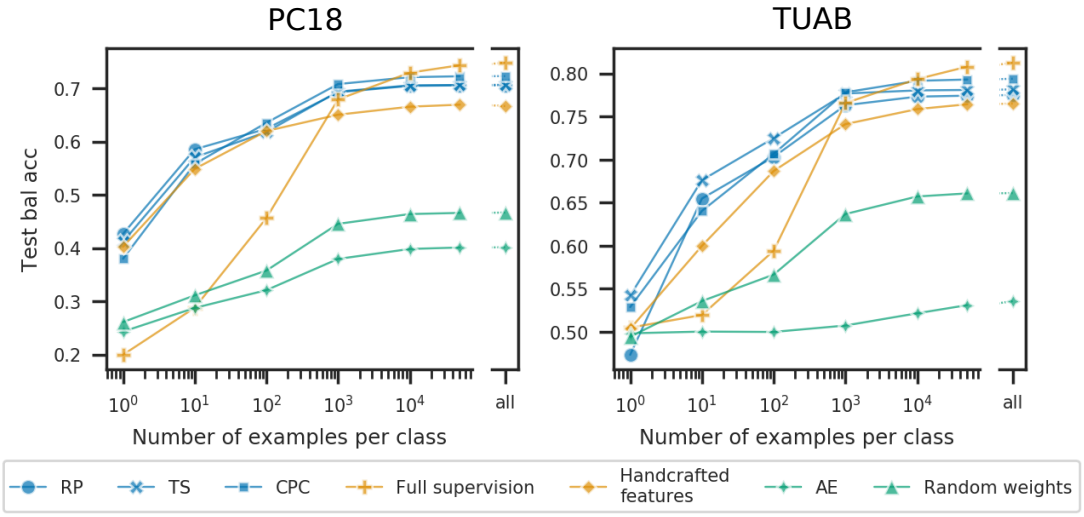


Figure 2.3 – Impact of number of labelled examples per class on downstream performance for SSL tasks (blue), supervised approaches (yellow) and unsupervised baselines (green). Feature extractors were trained with an autoencoder (AE), the relative positioning (RP), temporal shuffling (TS) and contrastive predictive coding (CPC) tasks, or left untrained ("random weights"), and then used to extract features on PC18 and TUAB. Following a hyperparameter search (Section 2.3.3), we used same-recording negative sampling on PC18 and across-recording negative sampling on TUAB. Downstream task performance was evaluated by training linear logistic regression models on the extracted features for the labelled examples, with at least one and up to all existing labelled examples in the training set ("All"). Additionally, fully supervised models were trained directly on labelled data and random forests were trained on handcrafted features. Results are the average of five runs with same initialization but different randomly selected examples (see Figure 2.4 below for a version with standard deviation). While more labelled examples led to better performance, SSL models achieved much higher performance than a fully supervised model when only few were available.

2.3 Results

2.3.1 SSL models learn representations of EEG and facilitate downstream tasks with limited annotated data

Can the suggested pretext tasks enable SSL on clinical EEG data and reduce the need for labelled EEG data in clinical tasks? To address this question, we applied the pretext tasks to two clinical datasets (PC18 and TUAB) and compared their downstream performance to that of established approaches such as fully supervised learning, while varying the number of labelled examples available.

The impact of the number of labelled samples on downstream performance is presented in Figure 2.3 (and in Figure 2.4 for a more complete view of the results with a measure of the spread around each performance estimate). First, SSL-learned features led to above-chance downstream performance across all data regimes: on PC18, maximum performance was 72.3% balanced accuracy (5-class, chance=20%) while on TUAB it was 79.4% (2-class, chance=50%). Second, SSL-learned features were competitive with other baseline approaches and could even outperform supervised approaches. On sleep data (Figure 2.3A), all three SSL models outperformed alternative approaches including

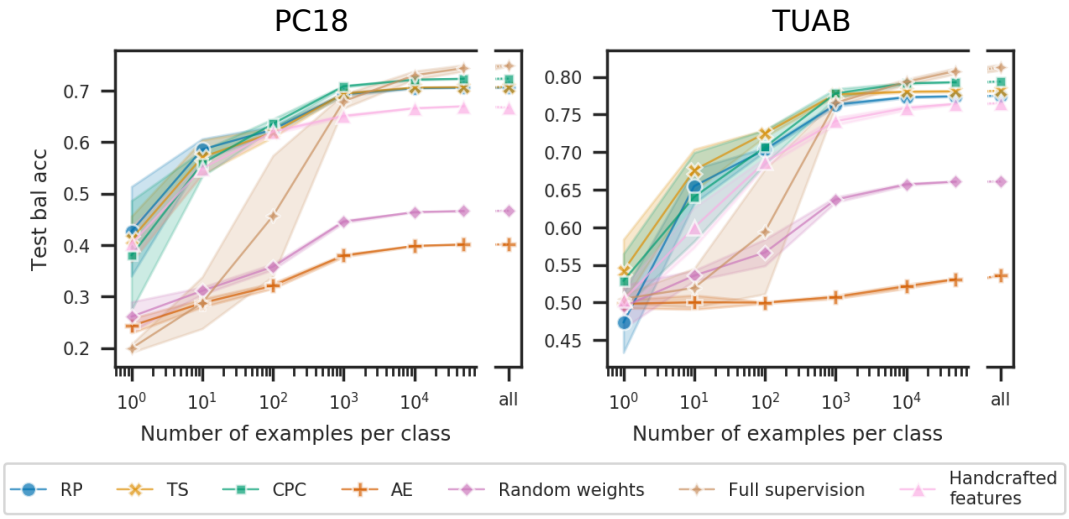


Figure 2.4 – Impact of number of labelled examples per class on downstream performance with uncertainty estimates. As compared to Figure 2.3, we present the standard deviation over five runs as the shaded area around each line. As could be expected, standard deviation was high when few labelled examples were available, but decreased as more labelled examples were provided.

full supervision and handcrafted features in most data regimes. The performance gap between SSL and full supervision reached 22.8 points when only one example per class was available. SSL remained better up until 10,000 examples per class, where full supervision began to exceed SSL performance, however by a 1.6-3.5% margin only. Moreover, SSL outperformed handcrafted features about 100 examples per class, *e.g.,* by up to 5.6 points for CPC.

Other baselines (random weights and autoencoding) achieved much lower performance, suggesting learning informative features for sleep staging is not trivial and requires more than the inductive bias of a convolutional neural network alone or a pure reconstruction task. Interestingly, the autoencoder's poor performance can be attributed to its MSE loss. This encourages the model to focus on low frequencies, which, due to 1/f power-law dynamics have the largest amplitudes in biosignals like EEG. Yet, low frequency signals only capture a small portion of the neurobiological information in EEG signals.

Next, we applied SSL to pathology detection, where classes ("normal" and "abnormal") are likely to be more heterogeneous than in sleep staging. Again, SSL-learned features outperformed baseline approaches in most data regimes: CPC outperformed full supervision under 10,000 labelled examples per class, while the performance gap between the two methods was around 1% when all examples were available. RP, TS and CPC also consistently outperformed handcrafted features, albeit by a smaller amount (*e.g.,* 3.8-4.8 point difference for CPC). Again, AE and random weights features could not compete with the other methods. Notably, AE's downstream performance never exceeded 53.0%.

Finally, for the sake of completeness, we investigated whether a classical semi-supervised approach, namely self-training, could be used to leverage unlabelled data in a similar fashion to SSL. The comparison between a purely supervised approach (handcrafted features baseline) and self-training is presented in Figure 2.5 on both PC18 and TUAB. Instead of improving performance, self-training systematically led to a decrease in per-

formance as compared to the handcrafted baseline approach, except for minimal improvements when only one or 10 labelled examples were available per class. Increasing the threshold mitigated this decrease, but did not lead to improved performance over a purely supervised model either. Given a potential limitation arising from the probability estimation accuracy of tree-based models (Tanha et al., 2017), we also tested a logistic regression classifier which should produce more reliable probability estimates. Performance was overall negatively impacted again. These results suggest that while classical semi-supervised approaches have shown potential to leverage large amounts of unlabelled data, their use is not straightforward in EEG classification problems and domain-specific efforts might have to be made in order to accommodate their use.

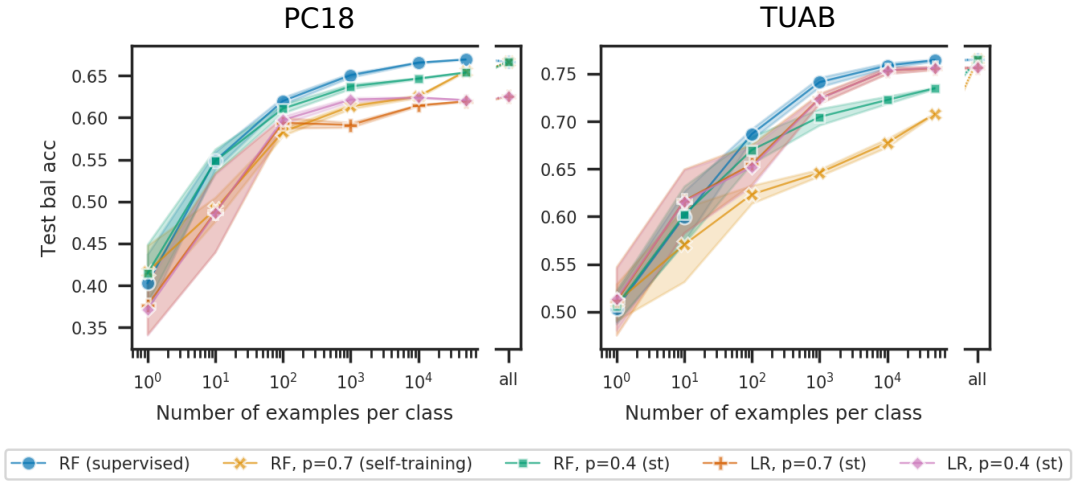


Figure 2.5 – Impact of number of labelled examples per class on downstream performance for a self-training semi-supervised baseline, as compared to the handcrafted features approach. We conducted self-training experiments with a random forest (RF) and logistic regression (LR) using the same hyperparameters as described in Section 2.2.6 and a probability threshold of 0.7 or 0.4 and maximum number of iterations of 5. Self-training overall harmed downstream performance for both datasets.

Taken together, our results demonstrate that the proposed SSL pretext tasks were general enough to handle two fundamentally different types of EEG classification problems. All SSL tasks systematically outperformed or matched other approaches in low-to-medium labelled data regimes and remained competitive in a high labelled data regime.

2.3.2 SSL models capture physiologically and clinically meaningful features

While SSL-learned features yielded competitive performance for sleep staging and pathology detection, it is unclear what structure was captured by SSL. Hence, we examined the relationship between the embeddings, annotations and metadata available in the clinical datasets. We projected the 100-dimensional embeddings obtained on PC18 and TUAB to two dimensions using UMAP (McInnes et al., 2018) and using the models with the highest downstream task performance as identified in Section 2.3.3.[9] This allows a qualitative analysis of local and global structure in SSL-learned features.

[9]Similar results were obtained with other well-performing hyperparameter configurations.

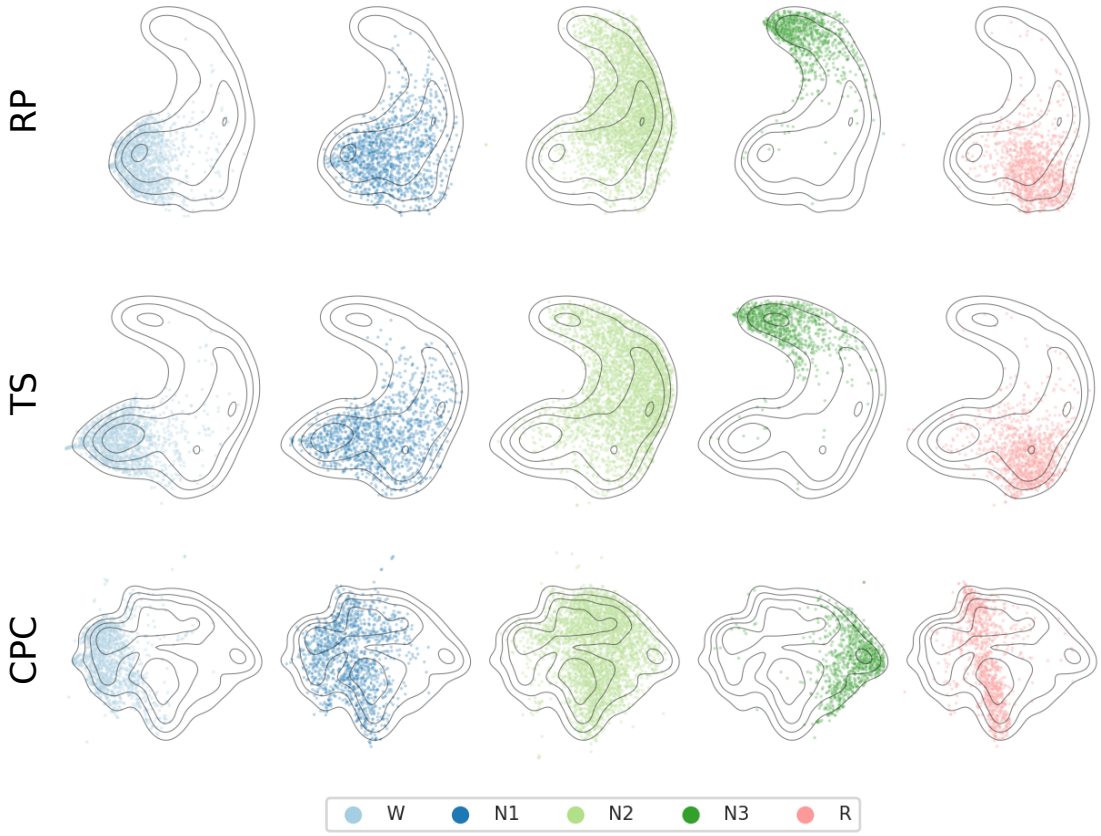


Figure 2.6 – Uniform Manifold Approximation and Projection (UMAP) visualization of SSL features on the PC18 dataset. The subplots show the distribution of the 5 sleep stages as scatterplots for RP (firt row), TS (second row) and CPC (third row) features. Contour lines correspond to the density levels of the distribution across all stages and are used as visual reference. Finally, each point corresponds to the features extracted from a 30-s window of EEG by the RP, TS and CPC embedders with the highest downstream performance as identified in Section 2.3.3 and Table 2.1. All available windows from the train, validation and test sets of PC18 were used. In all three cases, there is clear structure related to sleep stages although no labels were available during training.

Results on sleep data are shown in Figure 2.6. A structure that closely follows the different sleep stages is visible in RP, TS and CPC embeddings of PC18. Upon inspection of the distribution of examples from the different stages, clear groups emerged. These groups not only correspond to the labelled sleep stages, but are also sequentially arranged: moving from one end of the embedding to another, we could draw a line that passes through W, N1, N2 and N3. Stage R, finally, mostly overlaps with N1. These results are in line with previous observations on the structure of the sleep-wakefulness continuum (Pardey et al., 1996; Lopour et al., 2011).

Intuitively, the largest sources of variation in sleep EEG data are linked to changes in sleep stages and corresponding microstructure (slow waves, sleep spindles, etc.). To explore other sources of variations, the clinical information available in PC18 was used to color the embeddings: apnea events and subject age. The results are presented in the first row of Figure 2.7, 2.8 and 2.9 for TS, RP and CPC, respectively. Similar conclusions apply to all three methods. First, apnea-related structure can be seen in

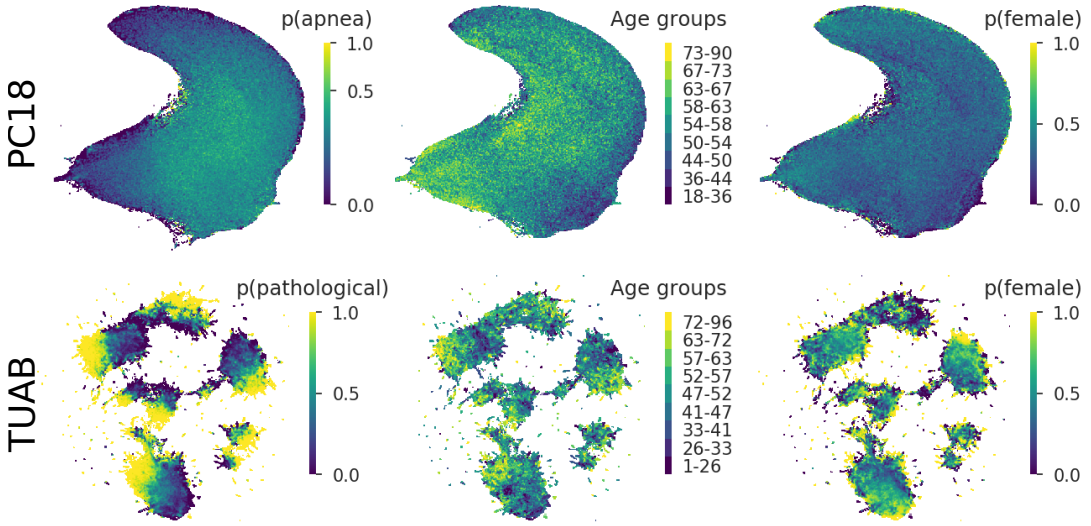


Figure 2.7 – Structure learned by the embedders trained on the TS task. The models with the highest downstream performance as identified in Section 2.3.3 and Table 2.1 were used to embed the combined train, validation and test sets of the PC18 and TUAB datasets. The embeddings were then projected to two dimensions using UMAP and discretized into 500 x 500 "pixels". For binary labels ("apnea", "pathological" and "gender"), we visualized the probability as heatmaps, *i.e.,* the color indicates the probability that the label is true (*e.g.,* that a window in that region of the embedding overlaps with an apnea annotation). For age, the subjects of each dataset were divided into 9 quantiles, and the color indicates which group was the most frequent in each bin. The features learned with SSL capture physiologically-relevant structure, such as pathology, age, apnea and gender.

the middle of the embeddings, overlapping with the area where stage N2 was prevalent (first column of Figure 2.7-2.9). At the same time, very few apnea events occurred at the extremities of the embedding, for instance over W regions, naturally, but also over N3 regions. Although this structure likely reflects the correlation between sleep stages, age and apnea-induced EEG patterns, this nonetheless shows the potential of SSL to learn features that relate to clinical phenomena. Second, age structure was revealed in two distinct ways in the embeddings (second column of Figure 2.7-2.9). The first is related to sleep macrostructure, *i.e.,* the sequence of sleep stages and their relative length. Indeed, younger subjects predominantly occupied the R region, while older subjects were more frequently found over the W region. This is in line with well-documented phenomena such as increased sleep fragmentation and sleep onset latency in older individuals, as well as a subtle reduction in REM sleep with age (Mander et al., 2017). Sleep microstructure-related information is also observed in the embeddings. For instance, looking at N2-N3 regions, older individuals are more likely in the leftmost side of the blob, while younger subjects are more likely found on its rightmost side. This suggests the characteristics of N2-N3 sleep vary across age groups, *e.g.,* sleep spindles (Purcell et al., 2017). Finally, there is also gender-related structure, with discernible low and high probability regions in the embeddings (third column of Figure 2.7-2.9).

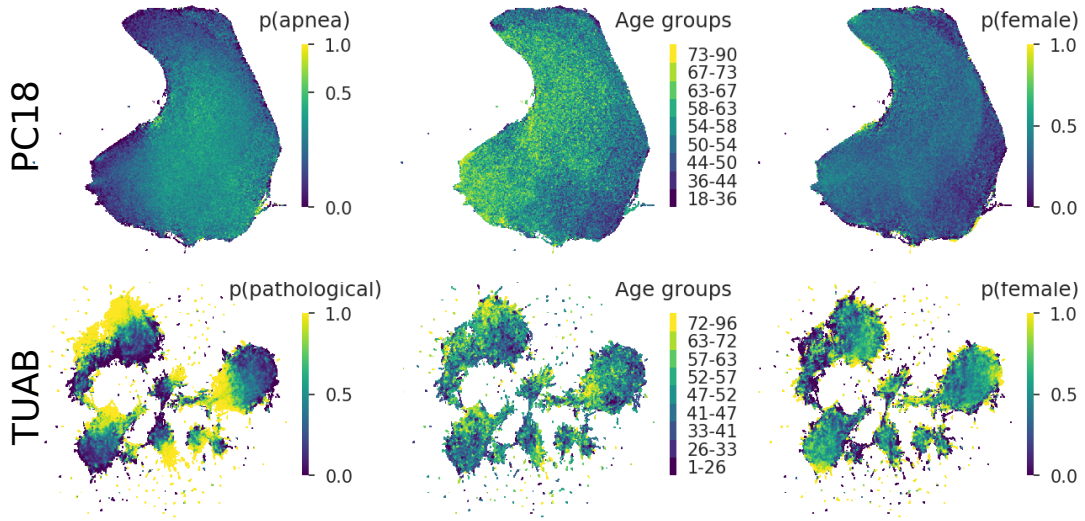


Figure 2.8 – Structure learned by the embedders trained on the RP task. See Figure 2.7 for a complete description.

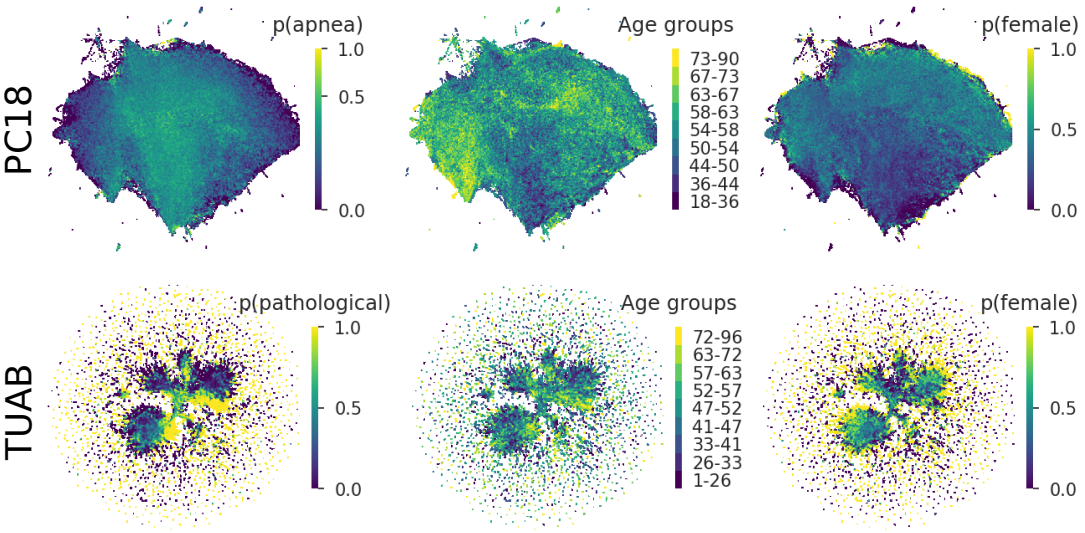


Figure 2.9 – Structure learned by the embedders trained on the CPC task. See Figure 2.7 for a complete description.

Can similar structure be learned on a different type of EEG recording? We conducted the same analysis for TUAB, this time focusing on pathology, age and gender. Results are shown in the second row of Figure 2.7, 2.8 and 2.9. The embeddings exhibited a primary multi-cluster structure, with similar gradient-like structure inside each cluster. For instance, pathology-related structure is apparent in the two embeddings (column 1), with an increasing probability of EEG being abnormal when moving from one end of the different clusters to the other. Likewise, an age-related gradient emerged inside each cluster (column 2), in a similar direction as the pathology gradient, while a gender-associated gradient appears orthogonal to the first two (last column). To understand

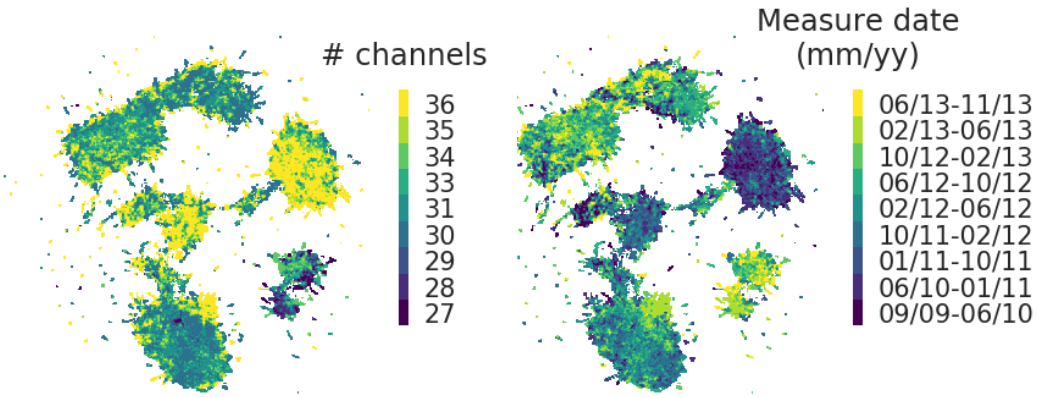


Figure 2.10 – Structure related to the original recording's number of EEG channels and measurement date in TS-learned features on the entire TUAB dataset. The overall different number of EEG channels and measurement date in each cluster reveals that the cluster-like structure reflects differences in experimental setups. See Figure 2.7 for a description of how the density plots are computed.

what these different clusters actually represent, we plotted experimental setup-related labels (original number of EEG channels and recording date) in Figure 2.10. Each cluster was predominantly composed of examples with a particular number of channels and a specific range of measurement dates. This suggests that SSL models have partially learned data collection-related noise. Indeed, the TUAB dataset was collected over several years across different sections of the Temple University Hospital, by different EEG technicians and with various EEG devices and montages (Ferrell et al., 2019). Most likely, the impact of this noise on the embedding could be mitigated by stronger preprocessing (*e.g.,* bandpass filtering) or by only sampling negative examples within recordings from the same cohort.

In conclusion, this experiment showed that SSL can encode clinically-relevant structure such as sleep stages, pathology, age, apnea and gender from EEG data, while revealing interactions (such as younger age and REM sleep), without any access to labels.

2.3.3 SSL pretext task hyperparameters strongly influence downstream task performance

How should SSL pretext task hyperparameters be tuned to make full use of self-supervision in clinical EEG tasks? In this section, we describe how hyperparameters were tuned in the experiments above and study the impact of key hyperparameters on downstream performance.

To benchmark pretext tasks across datasets, we tracked pretext and downstream performance across different choices of hyperparameters (see Section 2.2.5 for a complete description of the search procedure). The comparison is depicted in Figure 2.11. Pretext tasks performed significantly above chance on all datasets: RP and TS reached a maximum performance of 98.0% (2-class, chance=50%) while CPC yielded performances as high as 95.4% (32-class, chance= 3.1%). On the downstream tasks, SSL always performed above chance as reported in Section 2.3.1. Interestingly though, configurations with high pretext performance did not necessarily yield high downstream performance,

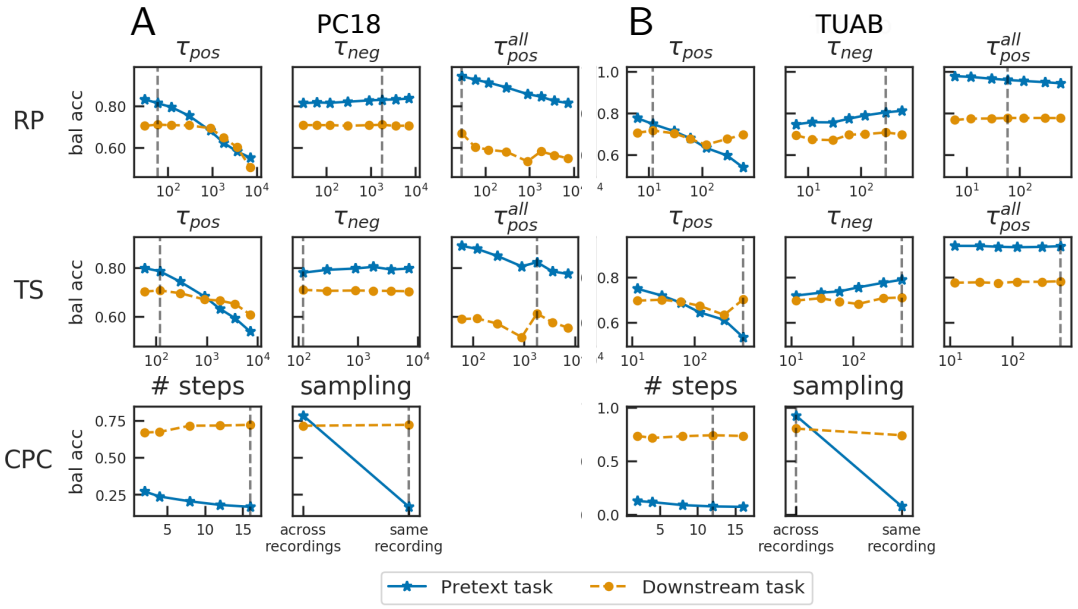


Figure 2.11 – Impact of principal hyperparameters on pretext (blue) and downstream (yellow) task performance, measured with balanced accuracy on the validation set of (A) PC18 and (B) TUAB. Each row corresponds to a different SSL pretext task. For both RP and TS, we varied the hyperparameters that control the length of the positive and negative contexts (τ_{pos}, τ_{neg}, in seconds); the exponent "all" indicates that negative windows were sampled across all recordings instead of within the same recording. For CPC, we varied the number of predicted windows and the type of negative sampling. Finally, the best hyperparameter values in terms of downstream task performance are emphasized using vertical dashed lines. See text for more details on the hyperparameter search procedure.

which highlights the necessity of appropriate hyperparameter selection.

Next, we examined the influence of hyperparameters on each pretext task (rows of Figure 2.11) to identify optimal configurations. First, we focused on same-recording negative sampling (anchor window(s) and negative examples are sampled from the same recording). With RP, increasing τ_{pos} always made the pretext task harder. Indeed, the larger the positive context, the more probable it is to get positive example pairs that are composed of distant (and thus potentially dissimilar) windows. On sleep data, we noticed a plateau effect: downstream performance was roughly constant below $\tau_{pos} = 20\,\text{min}$, suggesting EEG autocorrelation properties might change at this temporal scale. On TUAB, downstream performance decreased above $\tau_{pos} = 30\,\text{s}$ and increased again after $\tau_{pos} = 2\,\text{min}$. On the other hand, varying τ_{neg} with τ_{pos} fixed did not affect downstream performance consistently or significantly, although larger τ_{neg} generally led to easier pretext tasks.

Do these results hold when negative windows are sampled across all recordings? Interestingly, negative sampling has a considerable effect on downstream performance (columns 3 and 6 of Figure 2.11). On sleep staging, downstream performance dropped significantly and degraded faster as τ_{pos} was increased, while the opposite effect could be seen on the pathology detection task (higher, more stable performance). This effect might be explained by the nature of the downstream task: in sleep staging, major

changes in the EEG occur within a given recording, therefore distinguishing between windows of a same recording is key to identifying sleep stages. On the other hand, in pathology detection, each recording is given a single label ("normal" or "pathological") and so distinguishing between a window from the same recording (necessarily having the same label) or from another one (possibly having the opposite label) intuitively appears more useful. In other words, the distribution that is chosen for sampling negative examples determines the kind of invariance that the network is forced to learn. Overall, similar results hold for TS.

A similar analysis on CPC shows that while increasing the number of windows to predict made the pretext task harder, predicting further ahead in the future yielded better representations for sleep staging (bottom row of Figure 2.11). Pretext performances of around 20% might seem low, however they are in fact significantly higher than chance level (3.1%) on this 32-class problem. Remarkably, negative sampling had a minor effect on downstream performance on sleep data (71.6 vs. 72.2% bal acc), but had a considerable effect on pathology detection (74.1 vs. 80.4%), akin to RP and TS above. This result echoes the report of van den Oord et al. (2018) where subject-specific negative sampling led to the highest downstream performance on phoneme classification.

In this last experiment, we confirmed that our SSL pretext tasks are not trivial, and that certain pretext task hyperparameters have a measurable impact on downstream performance.

2.4 Discussion

In this chapter, we introduced self-supervised learning (SSL) as a way to learn representations on EEG data. Specifically, we proposed two SSL tasks designed to capture structure in EEG data, relative positioning (RP) and temporal shuffling (TS) and adapted a third approach, contrastive predictive coding (CPC), to work on EEG data. As compared to previous work on representation learning for EEG (Yuan et al., 2017) and ECG (Sarkar and Etemad, 2020), these methods do not require manual feature extraction, additional demographic information or a pretraining step, and do not rely on data augmentation transforms that have yet to be defined on EEG. We showed that these tasks can be used to learn generic features which capture clinically relevant structure from unlabelled EEG, such as sleep micro- and macrostructure, and pathology. Moreover, we performed a rigorous comparison of SSL methods to traditional unsupervised and supervised methods on EEG, and showed that downstream classification performance can be significantly improved by using SSL, particularly in low-labelled data regimes. These results hold for two large-scale EEG datasets comprising sleep and pathological EEG data, both with thousands of recordings.

2.4.1 Using SSL to improve performance in semi-supervised scenarios

We showed that SSL can be used to improve downstream performance when plentiful unlabelled data is available but labelled data is scarce, *i.e.,* in a semi-supervised learning scenario (Section 1.2.4). For instance, CPC-learned features outperformed fully supervised learning on sleep data by about 20% when only one labelled example per class was available. Similarly, on pathology detection a ∼15% improvement was obtained with SSL when only 10 labelled examples per class were available. In practice, SSL

has the potential to boost classification performance when annotations are expensive, a common scenario when working with biosignals like EEG.

While the SSL pretext tasks included in this work are applicable to multivariate time series in general, their successful application does require adequate recording length and dataset size. EEG recordings must be sufficiently long so that a reasonable number of windows can be sampled given the positive and negative contexts and the window length. This is typically not an issue for clinical recordings (*e.g.,* sleep monitoring and pathology screening produce recordings of tens of minutes to many hours). Other stimulus-presentation based protocols might also be used when entire recordings are available (rather than event-related windows only). Second, the current results may suggest large datasets are necessary to enable SSL on EEG as analyses were based on two of the largest publicly available EEG datasets with thousands of recordings each. However, similar results hold on much smaller datasets containing fewer than 100 recordings (Banville et al., 2019). Intuitively, as long as the unlabelled dataset is representative of the variability of the test data, representations that are useful for the pretext task should be transferable to a related downstream task.

One might argue that the performance gain of SSL over supervised approaches is minor in the moderate-to-large data regime. Nevertheless, our results open the door to further developments which may lead to substantial performance improvements. In this chapter, we limited our experiments to the linear evaluation protocol, where the downstream task is carried out by a linear classifier trained on SSL-learned features, in order to focus on the properties of the learned representations. Finetuning the parameters of the embedders on the downstream task (van den Oord et al., 2018; Tian et al., 2020) could likely improve downstream performance. Preliminary experiments (not shown) suggested that a 3-4-point improvement can be obtained in some data regimes with finetuning and that performance with all data points equals that of a purely supervised model. However, using nonlinear classifiers (here, random forests) as suggested in van den Oord et al. (2018) on SSL-learned features did not improve results, suggesting our pretext and downstream tasks might be sufficiently close that relevant information is already linearly accessible.

Generally speaking, self-supervision also presents interesting opportunities to improve model performance as compared to purely supervised approaches. First, due to their sampling methodology, most SSL tasks can "create" a combinatorially large number of distinct examples going far beyond the number of labelled examples typically available in a supervised task. For instance, on PC18, our training set contained close to 5×10^5 labelled examples while our SSL embedders were trained on more than twice that number of examples (to fit computational time requirements; more could have been easily sampled). The much higher number of available examples in SSL opens the door to larger deep neural networks (Hénaff et al., 2019). Given the relatively shallow architectures currently used in deep learning-EEG research (León et al., 2020) and the limited number of examples typically available as compared to other deep learning application domains (Roy et al., 2019a), SSL could be key to training deeper models and improving the state of the art on various EEG tasks.

Second, in most applications, it is desirable that a trained model generalizes across individuals, including individuals not in the training set. Therefore, many classification approaches rely on subject-dependent training to reach optimal performance (Roy et al., 2019a; Sabbagh et al., 2020). This however comes at the cost of requiring subject-specific labelled data. Along with work on neural network architectures (Zhang et al., 2020)

and transfer learning strategies (Xu et al., 2020b), among others, SSL is likely to help overcome this challenge. Indeed, given larger (unlabelled) datasets, SSL pretraining can improve the diversity of examples captured by the learned feature space and, in doing so, act as a strong regularizer against overfitting on the individuals of the training set.

Finally, although an increasing number of deep learning-EEG studies choose raw EEG as input to their neural networks, handcrafted feature representations are still used in a large portion of recent papers (Roy et al., 2019a). This raises the question of what optimal handcrafted features are for a specific task (Maiorana, 2020). SSL brings an interesting solution to this problem, as the features it learns can capture important physiological parameters from raw data, and their reuse in a deep neural network is straightforward (*e.g.,* as weight initialization). Although handcrafted features are inherently more interpretable, recent work on model inspection techniques has shown learned features can be meaningfully interpreted as well (Hartmann et al., 2018). Therefore, given the potential of SSL-learned features to capture true statistical sources (under certain conditions; see Section 2.4.3), SSL might close the gap between raw EEG- and handcrafted features-based approaches.

2.4.2 Sleep-wakefulness continuum and inter-rater reliability

We have demonstrated that the embeddings learned with SSL capture clinically-relevant information. Sleep, pathology, age and gender structure appeared in the learned feature space (Figure 2.6-2.7). The variety of metadata that is visible in the embeddings highlights the capacity of the proposed SSL tasks to uncover important factors of variation in noisy biosignals in a purely unsupervised manner. Critically though, this structure is not discrete, but continuous. Indeed, sleep stages are not cleanly separated into five clusters (or two for normal and abnormal EEG), but instead the embeddings display a smooth continuum of sleep-wakefulness (or of normal-abnormal EEG). Is this gradient-like structure meaningful, or is it a mere artefact of our experimental setup? We speculate that the continuous nature of SSL-learned features is inherent to the neurophysiological phenomena under study. Conveniently, this offers interesting opportunities to improve the analysis of physiological data. For instance, the concept of sleep stages and their taxonomy is the product of incremental standardization efforts in the sleep research community (Loomis et al., 1937; Kales et al., 1968; Moser et al., 2009; Berry et al., 2012). Although this categorization is convenient, critics still stress the limitations of stage-based systems under the evidence of sub-stages of sleep and the interplay between micro- and macrostructure (Schulz, 2008). Moreover, even trained experts using identical rules do not perfectly agree on their predictions, showing that the definition of stages remains ambiguous: in Younes et al. (2016), an overall agreement of 82.6% was obtained between the predictions of more than 2,500 trained sleep scorers (63.0% for N1). Consequently, could a data-driven representation of sleep EEG, such as the one learned with self-supervision, alleviate some of these challenges? While previous research suggests sleep might indeed be measured using a continuous metric derived from a supervised machine learning model (Pardey et al., 1996) or a computational mean-field model (Lopour et al., 2011), we additionally demonstrated that the rich feature space learned with SSL can simultaneously capture sleep-related structure and variability caused by age and apnea. Importantly, the data-driven nature of the SSL representation alleviates the subjectivity of manual sleep staging. This suggests SSL-learned representations could provide more fine-grained information about the multiple factors at play during sleep and, in doing so, enable a more precise study of sleep.

Similarly, many EEG pathologies are described by a clinical classification system which defines discrete subtypes of diseases or disorders, *e.g.,* epilepsy (Pack, 2019; England et al., 2012) and dementia (Walters, 2010). As for sleep EEG, inter-rater agreement is limited (Gemein et al., 2020). This suggests that there is an opportunity for these pathologies to be interpreted on a continuum as well.

2.4.3 Finding the right pretext task for EEG

With the large number of possible self-supervised pretext tasks, and the even larger number of possible EEG downstream tasks, how can we choose a combination of pretext task and hyperparameters for a given setting? To answer this question, many more empirical experiments will have to be conducted. However, our results give some insight as to what may be important to consider. In this work, we developed pretext tasks that proved effective on two different classification problems by combining 1) prior knowledge about EEG signals, 2) assumptions about the statistical structure of the features to be learned, 3) thorough hyperparameter search and 4) computational considerations.

First, we designed the RP and TS tasks by relying on prior knowledge about EEG. Specifically, sleep EEG signals have a clear temporal structure originating from the succession of sleep stages during the night. Therefore, two windows that are close in time are likely sharing the same sleep stage annotation and statistical structure. Learning to differentiate close-by from faraway windows should intuitively be related to learning to differentiate sleep stages. Similar approaches in the computer vision literature (Doersch et al., 2015; Misra et al., 2016) rely on properties of natural images that generally do not hold for EEG. For instance, whereas two EEG windows $\mathbf{X}_t$ and $\mathbf{X}_{t'}$ that are close in time likely look alike, there is typically no physiological information in these windows that allows one to determine whether $t < t'$ or $t > t'$.[10] Therefore, tasks that rely on proximity rather than absolute positioning appear to be a better match for EEG. We included CPC in our experiments as it naturally extends RP and TS and yielded promising results on other kinds of data (van den Oord et al., 2018).

Second, assumptions about the statistical structure of the latent factors to recover was used to support our choice of tasks. Given its similarity with permutation contrastive learning (PCL, a self-supervised method for nonlinear ICA (Hyvärinen and Morioka, 2017)), RP likely relies on general temporal dependencies (including autocorrelations) in EEG signals to recover informative features.[11] Since TS and CPC can both be seen as extensions of RP with more elaborate sampling strategies and contrastive procedures (Section 2.2.2), all three tasks might rely on similar structure to discover features.

Third, the careful selection of pretext task hyperparameters was essential to selecting the right pretext task configuration. For instance, RP, TS and CPC yielded similar downstream performance once optimal hyperparameters were selected. Particularly, the negative sampling strategy proved to be critical (Section 2.3.3). Indeed, sleep staging benefited from same-recording negative sampling whereas pathology detection worked best with across-recording negative sampling. This simple change allowed RP, TS and CPC to compete with purely supervised approaches on pathology detection, although

[10] Exceptions would include transitions between sleep stages that are more likely than others, such as from lighter to deeper sleep stages; however these transitions occur rarely during the night; more often a back-and-forth between sleep stages is observed.

[11] PCL can be obtained by setting RP's τ_{pos} to the length of a single window and τ_{neg} to 0. Incidentally, we found that the optimal value of τ_{pos} and τ_{neg} were relatively small on the datasets considered, suggesting hyperparameters close to those of PCL are optimal.

RP and TS were initially designed for capturing intra-recording sleep structure. This shows that negative sampling hyperparameters can be used to develop invariances to particular structure that is not desirable, *e.g.,* intra-recording changes or measurement-site effects (Figure 2.10). Ultimately, the fact that all three pretext tasks could reach similar downstream performance suggests self-supervision was able to uncover fundamental information likely related to physiology.

Finally, computational requirements and architecture-specific constraints are important to consider when choosing a pretext task. Given the similar downstream performance yielded by each pretext task after hyperparameter selection, RP might be preferred as it is the most memory-efficient and simplest. However, although CPC has more hyperparameters and requires additional forward and backward passes, its autoregressive encoder g_{AR} could yield better features for some tasks with larger-scale dependencies (van den Oord et al., 2018), *e.g.,* sleep staging. Indeed, recent studies on automated polysomnography have reported improved performance using larger-scale temporal information (Chambon et al., 2018; Supratak et al., 2017). Preliminary results (not shown) suggest CPC's autoregressive features can substantially improve both sleep staging and pathology detection performance.

Other pretext tasks could have been explored, for instance a transformation discrimination task similar to the ECG-focused work of Sarkar and Etemad (2020), or a nonlinear ICA-derived framework such as generalized contrastive learning (Hyvärinen et al., 2019) to explicitly leverage other structure present in EEG signals.

2.4.4 Limitations

We identify three principal limitations to this work: fixed hyperparameters across data regimes, restricted architecture search, and difference between reported results and state of the art.

Given the computational requirements of training neural networks on large EEG datasets, we fixed the training hyperparameters of fully supervised models (learning rate, batch size, dropout, weight decay) and reused these values across data regimes. As a result, fully supervised models typically stopped learning after only a few epochs. We tested the impact of various training hyperparameters on a subset of the models and saw that although training can be slightly improved, this effect is not strong enough to change any of our conclusions (results not shown).

Similarly, hyperparameter search was limited to pretext task hyperparameters in our experiments. However, architecture hyperparameters (*e.g.,* number of convolutional channels, embedding size, number of layers, etc.) can also play a critical role in achieving high performance using SSL (Kolesnikov et al., 2019). Sticking to a single fixed architecture for all models and data regimes means that these improvements - which could help bridge (or widen) the gap between SSL methods and the various baselines - were not taken into account in this work.

Finally, the goal of this work being to introduce self-supervision as a representation learning paradigm for EEG, we did not focus on matching state-of-the-art performance. Nonetheless, downstream performance would most likely improve significantly by aggregating temporal windows (Supratak et al., 2017; Chambon et al., 2018). On TUAB, we reused the simpler approaches from Gemein et al. (2020) instead of the best performing models. Moreover, 1) we did not use cropped decoding (Schirrmeister et al., 2017), 2) we used z-score instead of exponential moving average normalization

and 3) our train/validation data split was different. Together, these differences explain the small drop in performance between these state-of-the-art methods and the results reported here.

2.5 Conclusion

In this chapter, we introduced SSL approaches to learn representations on EEG data and showed that they could compete with and sometimes even outperform traditional supervised approaches on two large clinical datasets. Importantly, the features learned through SSL displayed a clear structure in which different physiological quantities were jointly encoded. This validates the potential of self-supervision to capture important physiological information, even in the absence of labelled data. In particular, this shows that SSL is a promising tool for making use of large-scale, unlabelled EEG data recorded in real-world conditions with mobile EEG.

Future work will have to demonstrate whether SSL can also be used successfully with other kinds of EEG recording settings and tasks. Ultimately, developing a better understanding of how best to design pretext tasks in order to target specific types of EEG structure will be critical to establishing self-supervision as a valuable component of any EEG analysis pipeline.

3

Robust learning from corrupted EEG with dynamic spatial filtering

Contents

In the last chapter, we focused on the challenges brought about by large unlabelled EEG datasets, which will become increasingly common as out-of-the-lab EEG technology matures. Now that we have demonstrated the effectiveness of SSL in extracting information from unlabelled EEG, we turn our attention to another important challenge real-world EEG applications are faced with, namely noise. Indeed, low-cost mobile EEG devices typically 1) trade-off signal quality for faster setup time and practicality (*e.g.,* by favoring sparse montages and dry electrodes over denser montages and the use of conductive gel or paste) and 2) are used in uncontrolled environments (*e.g.,* at home) without expert assistance or supervision. As a result, both the prevalence and strength of noise tend to increase, when compared to traditional research or clinical settings.

Building machine learning models for real-world EEG processing therefore requires methods that are robust to noisy data and randomly corrupted channels. In this chapter, we propose dynamic spatial filtering (DSF), an attention module that can be plugged in before the first layer of a neural network to handle corrupted EEG channels by learning to focus on good channels and to reduce the emphasis on bad ones. Specifically, DSF reweights EEG channels on a window-by-window basis according to their relevance given a predictive task. Moreover, DSF provides an interesting insight into the functioning of the neural network it is used with, through the visualization of its attention maps.

Our experiments on public and private EEG datasets, comprising more than 4,000 recordings and including real-world EEG recordings, show that when significant channel loss occurs, DSF systematically outperforms traditional noise-handling strategies. We also show how DSF's outputs can be interpreted, making it possible to monitor the usefulness of each channel in real-time. Overall, this new approach enables effective EEG analysis in challenging settings where channel corruption is likely to hamper the reading of brain signals.

This chapter is based on the following work (pending revisions in *NeuroImage*):

- **Hubert Banville**, Sean UN Wood, Chris Aimone, Denis-Alexander Engemann, and Alexandre Gramfort. Robust learning from corrupted EEG with dynamic spatial filtering. *arXiv preprint arXiv:2105.12916*, 2021c

Code used for the data analysis presented in this chapter is openly available at https://github.com/hubertjb/dynamic-spatial-filtering.

3.1 Introduction

The use of machine learning for automating EEG analysis has been the subject of much research in recent decades (Lotte et al., 2007, 2018; Roy et al., 2019a). However, state-of-the-art EEG prediction pipelines are generally benchmarked on datasets recorded in well-controlled conditions that are relatively clean when compared to data from mobile EEG. As a result, it is unclear how models designed for laboratory data will cope with signals encountered in real-world contexts. This is especially critical for mobile EEG recordings that may contain a varying number of usable channels as well as overall noisier signals, in contrast to most research- and clinical-grade recordings. In addition, the difference in number of channels between research and mobile settings also means that interpolating bad channels offline (as is commonly done in recordings with dense electrode montages) is likely to fail on mobile EEG devices given their limited spatial information. It is an additional challenge that the quality of EEG data is not static but can vary significantly within a given recording. This suggests that predictive models

should handle noise dynamically. Ideally, not only should machine learning pipelines produce predictions that are robust to (changing) sources of noise in EEG, but they should also do so in a way that is interpretable. For instance, if noise is easily identifiable, corrective action can be quickly taken by users or experimenters during a recording. Finally, practical real-world EEG devices are likely to have access to a limited amount of computational resources for processing the raw EEG (*e.g.,* mobile phone), meaning processing must be as efficient as possible, especially in applications where real-time feedback is required (*e.g.,* brain-computer interfacing and neurofeedback). With this in mind, existing noise-handling strategies from the EEG and MEG literature may not be optimal for handling the problem of channel corruption in sparse montages.

In this chapter, we propose and benchmark an attention mechanism module designed to handle corrupted channel data, based on the concept of *scaling attention* (Hu et al., 2018; Woo et al., 2018). This module can be inserted before the first layer of any convolutional neural network architecture in which activations have a spatial dimension (Schirrmeister et al., 2017; Lawhern et al., 2018; Chambon et al., 2018), and then be trained end-to-end for the prediction task at hand.

The rest of the chapter is structured as follows. Section 3.2 presents an overview of the EEG noise handling literature, then describes the attention module and denoising procedure proposed in this study. The neural architectures, baseline methods and data used in our experiments are introduced in Section 3.3. Next, Section 3.4 reports the results of our experiments on sleep and pathology EEG datasets. Lastly, we examine related work and discuss the results in Section 3.5.

3.2 Methods

3.2.1 State-of-the-art approaches to noise-robust EEG processing

Existing strategies for dealing with noisy data can be divided into three categories (Table 3.1): 1) ignoring or rejecting noisy segments, 2) implicit denoising, *i.e.,* methods that allow models to work despite noise, and 3) explicit denoising, *i.e.,* methods that rely on a separate preprocessing step to handle noise or missing channels before prediction. We now discuss existing methods employing these strategies in more detail.

The simplest way to deal with noise in EEG is to assume that it is negligible or to simply discard bad segments (Roy et al., 2019a). For instance, a manually selected amplitude or variance threshold (Manor and Geva, 2015; Hefron et al., 2018; Wang et al., 2018) or a classifier trained to recognize artifacts (Dhindsa, 2017) can be used to identify segments to be ignored. This approach, though commonplace, is ill-suited to mobile EEG settings where noise cannot be assumed to be negligible, but also to online applications where model predictions need to be continuously available. Moreover, this approach is likely to discard windows due to a small fraction of bad electrodes, potentially losing usable information from other channels.

Implicit denoising approaches can be used to design noise-robust processing pipelines that do not contain a specific noise handling step. First, implicit denoising approaches can use representations of EEG data that are robust to missing channels. For instance, multichannel EEG can be transformed into topographical maps ("topomaps") that are less sensitive to the absence of a few channels. This representation is then typically fed into a standard CNN architecture. While this approach can gracefully handle missing channels in dense montages (*e.g.,* 16 to 64 channels in Bashivan et al. (2016); Thodoroff

et al. (2016); Hagad et al. (2019)), it is likely to perform poorly on sparse montages (*e.g.,* 4 channels) as spatial interpolation might fail if channels are missing. Moreover, this approach requires computationally demanding preprocessing and feature extraction steps, undesirable in online and low-computational resources contexts. In the traditional machine learning setting, Sabbagh et al. (2020) showed that representing input windows as covariance matrices and using Riemannian geometry-aware models did not require common noise correction steps to reach high performance on a brain age prediction task. However, the robustness of this approach has not been evaluated on sparse montages. Also, its integration into neural network architectures is not straightforward with geometry-aware deep learning remaining an active field of research (Bronstein et al., 2017). Signal processing techniques can also be used to promote invariance to certain types of noise. For instance, the Lomb-Scargle periodogram can be used to extract spectral representations that are robust to missing samples (Li et al., 2015; Chu et al., 2018). However, this approach fails when channels are completely missing. Finally, implicit denoising can be achieved with traditional machine learning models that are inherently robust to noise. For instance, random forests trained on handcrafted EEG features were shown to be notably more robust to low SNR inputs than univariate models on a state-of-consciousness prediction task (Engemann et al., 2018a). Although promising, this approach is limited by its feature engineering step, as features 1) rely heavily on domain knowledge, 2) might not be optimal to the task, and 3) require an additional processing step which can be prohibitive in limited resource contexts.

Multiple studies have explicitly handled noise by correcting corrupted signals or predicting missing or additional channels from available ones. Spatial projection approaches aim at projecting the input signals to a noise-free subspace before projecting the signals back into channel-space, *e.g.,* ICA (Jung et al., 1997; Mammone et al., 2011; Winkler et al., 2011) or principal component analysis (PCA) (Uusitalo and Ilmoniemi, 1997; Kothe and Jung, 2016). While approaches such as ICA are powerful tools to mitigate artifact and noise components in a semi-automated way, their efficacy can diminish when only few channels are available. For instance, in addition to introducing an additional preprocessing step, these approaches are likely to discard important discriminative information during preprocessing because they are decoupled from the prediction task. Also, the fact that preprocessing is done independently from the supervised learning task, or the statistical testing procedure, actually makes the selection of preprocessing parameters (*e.g.,* number of good components) challenging. Motivated by the challenge of parameter selection, fully automated denoising pipelines have been proposed. FASTER (Nolan et al., 2010) and PREP (Bigdely-Shamlo et al., 2015) both combine artifact correction, noise removal and bad channel interpolation into a single automated pipeline. Autoreject (Jas et al., 2017) is another recently developed pipeline that uses cross-validation to automatically select amplitude thresholds to use for rejecting windows or flagging bad channels. These approaches are well-suited to offline analyses where the morphology of the signals is of interest, however they are typically computationally demanding and are also decoupled from the statistical modeling. Additionally, it is unclear how interpolation can be applied when using bipolar montages (*i.e.,* that do not share a single reference), as is often the case in *e.g.,* polysomnography (Berry et al., 2012) and epilepsy monitoring (Rosenzweig et al., 2014).

Finally, generic machine learning models have been proposed to recover bad channels. For instance, generative adversarial networks (GANs) have been trained to recover dense EEG montages from a few electrodes (Corley and Huang, 2018; Svantesson et al., 2020).

Other similar methods have been proposed, *e.g.,* long short-term memory (LSTM) neural networks (Paul, 2020), AEs (El-Fiqi et al., 2019), or tensor decomposition and compressed sensing (Ramakrishnan and Satyanarayana, 2016; Sole-Casals et al., 2018). However, these methods postulate that the identity of bad channels is known ahead of time, which is a non-trivial assumption in practice.

In contrast to the existing literature on channel corruption handling in EEG, we introduce an interpretable end-to-end denoising approach that can learn implicitly to work with corrupted sparse EEG data, and that does not require additional preprocessing steps.

3.2.2 Dynamic spatial filtering: Second-order attention for learning on noisy EEG signals

The key goal behind DSF is to help neural networks focus on the most important channels, at each time instant, given a specific machine learning task on EEG. To do so, we introduce a spatial attention mechanism that dynamically re-weights channels according to their predictive power. This idea is inspired by recent developments in attention mechanisms, most specifically the scaling attention approach proposed in computer vision (Hu et al., 2018; Woo et al., 2018). Notably, DSF leverages second-order information, *i.e.,* spatial covariance, to capture dependencies between EEG channels. In this section, we detail the learning problem under study, the proposed attention architecture and a data augmentation transform designed to help train noise-robust models.

We perform experiments in the supervised classification setting. A model $f_{\mathbf{\Theta}} : \mathcal{X} \to \mathcal{Y}$ with parameters $\mathbf{\Theta}$ (*e.g.,* , a CNN) is trained to predict the class y of EEG windows $\mathbf{X}$. For this, we train $f_{\mathbf{\Theta}}$ to minimize the loss $\mathcal{L}$, *e.g.,* the categorical cross-entropy loss, over the example-label pairs $(\mathbf{X}^{(i)}, y^{(i)})$:

$$\hat{f_{\mathbf{\Theta}}} = \arg\min_{\mathbf{\Theta}} \mathbb{E}_{\mathbf{X}, y \in \mathcal{X} \times \mathcal{Y}} [\mathcal{L}(f_{\mathbf{\Theta}}(\mathbf{X}), y)] \ . \tag{3.1}$$

In particular, we are interested in the performance of $f_{\mathbf{\Theta}}$ when random channels are corrupted and more specifically when channel corruption occurs at test time (*i.e.,* when training data is mostly clean). Toward this goal, we insert an attention-based module $m_{\text{DSF}} : \mathbb{R}^{C \times T} \to \mathbb{R}^{C' \times T}$ into $f_{\mathbf{\Theta}}$ which performs a (fixed) transformation $\Phi(\mathbf{X})$ to extract relevant spatial information from $\mathbf{X}$, followed by a parametrized re-weighting mechanism for the input signals.

In order to implicitly handle noise in neural network architectures, we design an attention module where second-order information is extracted from the input and used to predict weights of a linear transformation of the input EEG channels, that are optimized for the learning task (Figure 3.1). Applying such linear transforms to multivariate EEG signals is commonly referred to as *spatial filtering*, a technique that has been widely used in the field of EEG (Makeig et al., 1996; McFarland et al., 1997; Parra et al., 2005; Blankertz et al., 2007; de Cheveigné and Simon, 2008; Lotte and Guan, 2010; Nikulin et al., 2011). This enables the model to learn to ignore noisy outputs and/or to re-weight them, while still leveraging any remaining spatial information. We now show how this module can be applied to the raw input $\mathbf{X}$.

We define the dynamic spatial filter (DSF) module m_{DSF} as:

$$m_{\text{DSF}}(\mathbf{X}) = \mathbf{W}_{\text{DSF}}(\mathbf{X})\mathbf{X} + \mathbf{b}_{\text{DSF}}(\mathbf{X}) \ , \tag{3.2}$$

Table 3.1 – Existing methods for dealing with noisy EEG data.

	Approach	Examples	Notes
Ignore or reject noise	No denoising	(Schirrmeister et al., 2017; Lawhern et al., 2018; Li et al., 2019; Schirrmeister et al., 2017; Gemein et al., 2020; Supratak et al., 2017; Guillot et al., 2020; Phan et al., 2019, 2020)	Might not work in real-life applications (out of the lab/clinic)
	Removing bad epochs	(Manor and Geva, 2015; Dhindsa, 2017; Hefron et al., 2018; Wang et al., 2018)	Doesn't allow online predictions; Might discard useful information
Implicit denoising	Robust input representations	Covariance matrices in Riemannian tangent space (Sabbagh et al., 2020)	Might not work if too few channels available
		Topomaps (Bashivan et al., 2016; Thodoroff et al., 2016; Hagad et al., 2019)	Expensive preprocessing step; Might not work if too few channels available
	Robust signal processing techniques	Lomb-Scargle periodogram (Li et al., 2015; Chu et al., 2018)	Only useful for missing samples, not missing channels
	Robust machine learning classifiers	Handcrafted features and random forest (Engemann et al., 2018a)	Requires feature engineering step
Explicit denoising	Spatial projection-based approaches	Signal Space Separation (SSS) for MEG (Taulu et al., 2004)	Might not work if too few channels available; Additional preprocessing step; Preprocessing might discard important information for learning task
		ICA-based denoising (Jung et al., 1997; Mammone et al., 2011; Winkler et al., 2011)	
	Automated correction	Autoreject (Jas et al., 2017), FASTER (Nolan et al., 2010), PREP (Bigdely-Shamlo et al., 2015)	Expensive preprocessing step
	Model-based interpolation/ reconstruction	Deep learning-based super-resolution (GAN, LSTM, AE, etc.) (Han et al., 2018; Kwon et al., 2019; Corley and Huang, 2018; Svantesson et al., 2020; El-Fiqi et al., 2019)	Separate training step; Additional inference step to reconstruct at test time; Requires separate procedure to detect corrupted channels
		Tensor decomposition, compressed sensing (Sole-Casals et al., 2018; Ramakrishnan and Satyanarayana, 2016)	
Interpretable denoising	**Channel corruption-invariant architecture**	**Dynamic Spatial Filtering (this work)**	**Trained end-to-end, no additional preprocessing, interpretable, works with sparse montages**

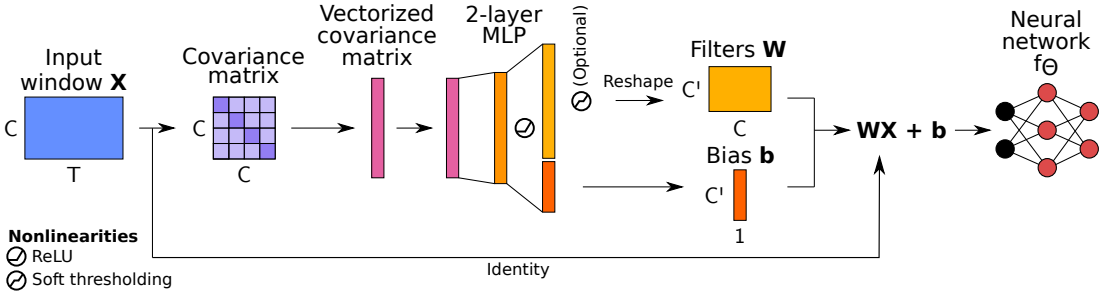


Figure 3.1 – Visual description of the Dynamic Spatial Filtering (DSF) attention module. An input window $\mathbf{X}$ with C spatial channels is processed by a 2-layer MLP to produce a set of C' spatial filters $\mathbf{W}$ and biases $\mathbf{b}$ that dynamically transform the input $\mathbf{X}$. This allows the subsequent layers of a neural network to ignore bad channels and focus on the most informative ones.

where $\mathbf{W}_{\text{DSF}} \in \mathbb{R}^{C' \times C}$ and $\mathbf{b}_{\text{DSF}} \in \mathbb{R}^{C'}$ are obtained by reshaping the output of a neural network, *e.g.,* a MLP, $h_{\mathbf{\Theta}_{DSF}}(\Phi(\mathbf{X})) \in \mathbb{R}^{C' \times (C+1)}$ (see Figure 3.1). Under this formulation, each row in $\mathbf{W}_{\text{DSF}}$ corresponds to a spatial filter that linearly transforms the input signals into another virtual channel. Here, C' can be set to the number of input spatial channels C or considered a hyperparameter of the attention module[1]. When $C' = C$, if the diagonal of $\mathbf{W}_{\text{DSF}}$ is 0, $\mathbf{W}_{\text{DSF}}$ corresponds to a linear interpolation of each channel based on the $C - 1$ others, as is commonly done in the classical EEG literature (Perrin et al., 1989) (see Section 3.5.6 for an in-depth discussion). Heavily corrupted channels can be ignored by giving them a weight of 0 in $\mathbf{W}_{\text{DSF}}$. To facilitate this behavior, we can further apply a soft-thresholding element-wise nonlinearity to $\mathbf{W}_{\text{DSF}}$:

$$\mathbf{W}'_{\text{DSF}} = \text{sign}(\mathbf{W}_{\text{DSF}}) \max(\left|\mathbf{W}_{\text{DSF}}\right| - \tau, 0) \ , \tag{3.3}$$

where τ is a threshold empirically set to 0.1, $|\cdot|$ is the element-wise absolute value and both the sign and max operators are applied element-wise.

In our experiments, the spatial information extracted by the transforms $\Phi(\mathbf{X})$ was either 1) the log-variance of each input channel or 2) the flattened upper triangular part of the matrix logarithm of the covariance matrix of $\mathbf{X}$ (see Section 3.2.3)[2]. When reporting results, we denote models as *DSFd* and *DSFm* when DSF takes the log-variance or the matrix logarithm of the covariance matrix as input, respectively. We further add the suffix "-st" to indicate the use of the soft-thresholding nonlinearity, *e.g., DSFm-st.*

Interestingly, the DSF module can be seen as a multi-head attention mechanism (Vaswani et al., 2017) with real-valued attention weights and where each head is tasked with producing a linear combination of the input spatial signals.

[1] In which case it can be used to increase the diversity of input channels in models trained on sparse montages ($C' > C$) or perform dimensionality reduction to reduce computational complexity ($C' < C$).

[2] In practice, if a channel is "flat-lining" (has only 0s) inside a window and therefore has a variance of 0, its log-variance is replaced by 0. Similarly, if a covariance matrix eigenvalue is 0 when computing the matrix logarithm (see Equation 3.8), its logarithm is replaced by 0.

Finally, we can inspect the attention given by m_{DSF} to each input channel by computing the "effective channel importance" metric[3] $\boldsymbol{\phi} \in \mathbb{R}^C$ where

$$\phi_j = \sqrt{\sum_{i=1}^{C'} {W_{ij}}^2} \ . \tag{3.4}$$

Intuitively, $\boldsymbol{\phi}$ measures how much each input channel is used by m_{DSF} to produce the output virtual channels. A normalized version

$$\hat{\boldsymbol{\phi}} = \frac{\boldsymbol{\phi}}{\max_i \boldsymbol{\phi}_i} \tag{3.5}$$

can also be used to obtain a value between 0 and 1. This straightforward way of inspecting the functioning of the DSF module facilitates the identification of important or noisy channels.

To further help our models learn to be robust to noise, we design a data augmentation procedure that randomly corrupts channels. Specifically, channel corruption is simulated by performing a masked channel-wise convex combination of input channels and Gaussian white noise $\mathbf{Z} \in \mathbb{R}^{C \times T}$:

$$\tilde{\mathbf{X}} = (1 - \eta)\,\text{diag}(\boldsymbol{\nu})\mathbf{X} + \eta\,\text{diag}(\boldsymbol{\nu})\mathbf{Z} + \text{diag}(1 - \boldsymbol{\nu})\mathbf{X} \ , \tag{3.6}$$

where $Z_{i,j} \sim \mathcal{N}(0, \sigma_n^2)$ for $i \in [\![T]\!]$ and $j \in [\![C]\!]$, $\eta \in [0, 1]$ controls the relative strength of the noise, and $\boldsymbol{\nu} \in \{0, 1\}^C$ is a masking vector that controls which channels are corrupted. The operator $\text{diag}(\mathbf{x})$ creates a square matrix filled with zeros whose diagonal is the vector $\mathbf{x}$. Here, $\boldsymbol{\nu}$ is sampled from a Multinouilli distribution with parameter p. Each window $\mathbf{X}$ is individually corrupted using random parameters $\sigma_n \sim \mathcal{U}(20, 50)\ \mu\text{V}$, $\eta \sim \mathcal{U}(0.5, 1)$, and a fixed p of 0.5.

3.2.3 Representation of spatial information in the DSF module

In this section, we briefly discuss different spatial representations of EEG and motivate our choice of the spatial covariance matrix for DSF.

Given some EEG signals $\mathbf{X} \in \mathbb{R}^{C \times T}$, where T is the number of time samples in $\mathbf{X}$, and which we assume to be zero-mean, an unbiased estimate of their covariance reads:

$$\boldsymbol{\Sigma}(\mathbf{X}) = \frac{\mathbf{X}\mathbf{X}^\top}{T} \in \mathbb{R}^{C \times C} \ . \tag{3.7}$$

The zero-mean assumption is justified after some high-pass filtering or simple baseline correction of the signals. To assess whether one channel is noisy or not, a human expert annotator will typically rely on the power of a signal and its similarity with the neighboring channels. This information is encoded in the covariance matrix.

Multiple well-established signal processing techniques rely on some estimate of $\boldsymbol{\Sigma}$. For instance, common spatial patterns (CSP) performs generalized eigenvalue decomposition of covariance matrices to identify optimal spatial filters for maximizing the difference between two classes (Koles et al., 1990). Riemannian geometry approaches to

[3] "Effective channel importance" measures how useful the data of a channel is. It is not to be confused with the theoretical importance of a channel, *i.e.,* the fact that in theory some channels (given good signal quality) might be more useful for some tasks than other channels.

EEG classification and regression instead leverage the geometry of the space of symmetric positive definite (SPD) matrices to develop geometry-aware metrics. They are used to average and compare covariance matrices, which has been shown to outperform other classical approaches (Congedo et al., 2017; Sabbagh et al., 2020). Artifact handling pipelines such as the Riemannian potato (Barachant et al., 2013a) and Artifact Subspace Reconstruction (Mullen et al., 2015) further rely on covariance matrices to identify bad epochs or attenuate noise.

The values in a covariance matrix often follow a heavy-tailed distribution. Therefore, knowing that neural networks are typically easier to train when the distribution of input values is fairly concentrated, it is helpful to standardize the covariance values before feeding them to the network. While scalar non-linear transformations (*e.g.,* logarithms) could help reduce the range of values and facilitate a neural network's task, the geometry of SPD matrices actually calls for metrics that respect the Riemannian structure of the SPD matrices' manifold (Lin, 2019). For instance, this means using the matrix logarithm instead of naively flattening the upper triangle and diagonal of the matrix (Sabbagh et al., 2020). For an SPD matrix $\mathbf{S}$, whose orthogonal eigendecomposition reads $\mathbf{S} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$, where $\mathbf{\Lambda} = \text{diag}(\lambda_1, ..., \lambda_n)$ contains its eigenvalues, the matrix logarithm $\log(\mathbf{S})$ is given by:

$$\log(\mathbf{S}) = \mathbf{U}\,\text{diag}(\log(\lambda_1), ..., \log(\lambda_n))\mathbf{U}^\top \quad . \tag{3.8}$$

The diagonal and upper-triangular part of $\log(\mathbf{S})$ can then be flattened into a vector with $C(C+1)/2$ values, which is then typically used with linear models, *e.g.,* support vector machines (SVMs) or logistic regression.

Other options to provide input values in a restricted range exist. For instance, one could simply use the element-wise logarithm of the diagonal of the covariance matrix, *i.e.,* the log-variance of the input signals. This is appropriate if pairwise inter-channel covariance information is deemed not critical down the line. Alternatively, Pearson's correlation matrix, which can be seen as the covariance matrix of the z-score normalized signals, could be used. It has the advantage that its values are already in a well-defined range (-1, 1), yet it is blind to channel variances. In our experiments, we focused on two spatial representations: the channel-wise variance obtained from the diagonal of $\mathbf{\Sigma}$, and the matrix logarithm of $\mathbf{\Sigma}$. Both helped improve robustness on the pathology detection and sleep staging tasks.

3.2.4 Computational considerations

We set the following hyperparameters when training deep neural networks: optimizer, learning rate schedule, batch size, regularization strength (number of training epochs, weight decay, dropout) and parameter initialization scheme. In all experiments, we used the AdamW optimizer (Loshchilov and Hutter, 2017) with $\beta_1 = 0.9$, $\beta_2 = 0.999$, a learning rate of 10^{-3} and cosine annealing. The parameters of all neural networks were randomly initialized using uniform He initialization (He et al., 2015). Dropout (Srivastava et al., 2014) was applied to $f_{\mathbf{\Theta}}$'s fully connected layer at a rate of 50% and weight decay was applied to the trainable parameters of all layers of both $f_{\mathbf{\Theta}}$ and $h_{\mathbf{\Theta}_{DSF}}$. Moreover, during training, the loss was weighted to optimize balanced accuracy. Weight decay and batch size hyperparameters were selected such that learning curves decreased steadily in the first 10 epochs of training on a subset of the training set of each dataset. Respectively for TUAB, PC18 and MSD (see Section 3.3.4) these were 256, 64 and 64 (batch size) and 0.01, 0.001 and 0.01 (weight decay).

Deep learning and baseline models were trained using a combination of the `braindecode` (Schirrmeister et al., 2017), `MNE-Python` (Gramfort et al., 2014), `PyTorch` (Paszke et al., 2019), `pyRiemann` (Barachant et al., 2013b), `mne-features` (Schiratti et al., 2018) and `scikit-learn` (Pedregosa et al., 2011) packages.[4] Finally, deep learning models were trained on 1 or 2 Nvidia Tesla V100 or P4 GPUs for anywhere from a few minutes to 7 hours, depending on the amount of data, early stopping and GPU configuration.

3.3 Experiments

3.3.1 Downstream tasks

We studied noise robustness through two common EEG classification downstream tasks: sleep staging and pathology detection (see Section 2.2.3 for a detailed description). First, sleep staging is a 5-class classification problem which consists of predicting which sleep stage an individual is in, in non-overlapping 30-s windows of overnight recordings. While a large number of machine learning approaches have been proposed to perform sleep staging (Motamedi-Fakhr et al., 2014; Chambon et al., 2018; Roy et al., 2019a; Phan et al., 2020), the handling of corrupted channels has not been addressed in a comprehensive manner yet, as channel corruption is less likely to occur in clinical and laboratory settings than in the real-world settings we consider here[5].

Second, the pathology detection task is a binary classification tasks aimed at detecting whether an individual's EEG is pathological or not (Smith, 2005; Micanovic and Pal, 2014). Such recordings are typically carried out in well-controlled settings (*e.g.,* in a hospital (Obeid and Picone, 2016)) where sources of noise can be monitored and mitigated in real-time by experts. To test pathology detection performance in the context of mobile EEG acquisition, we used a limited set of electrodes, in contrast to previous work (Lopez et al., 2015; Schirrmeister et al., 2017; Gemein et al., 2020).

3.3.2 Compared methods

We compared the performance of the proposed DSF and data augmentation method to other established approaches. In total, we contrasted combinations of three machine learning pipelines and three different noise-handling strategies.

We consider the following machine learning pipelines: 1) end-to-end deep learning (with and without the DSF module) from raw signals, 2) filter-bank covariance matrices with Riemannian tangent space projection and logistic regression (Barachant et al., 2013b; Congedo et al., 2017; Lotte et al., 2018; Sabbagh et al., 2020) (which we refer to as "Riemann"), and 3) handcrafted features and random forest (RF) (Gemein et al., 2020).

Deep learning pipelines used the same base CNN architectures as in our experiments on SSL (Section 2.2.4) for f_{Θ}. For pathology detection, we used the ShallowNet architecture from Schirrmeister et al. (2017) which parametrizes the FBCSP pipeline (Gemein et al., 2020). We used it without modifying the architecture from the original paper, yielding a total of 13,482 trainable parameters when $C = 6$. For sleep staging, we used a 3-layer CNN which takes 30-s windows as input (Chambon et al., 2018), with a total

[4] Code used for data analysis can be found at https://github.com/hubertjb/dynamic-spatial-filtering.

[5] A recent study reported training a neural network on artificially-corrupted sleep EEG data, with a goal similar to ours (Jónsson et al., 2020); however, this study only appears as a Supplement with little information on the methods and results.

of 18,457 trainable parameters when $C = 4$ and an input sampling frequency of 100 Hz. Finally, when evaluating DSF, we added modules m_{DSF} before the input layer of each neural network. The input dimensionality of m_{DSF} depended on the chosen spatial information extraction transform $\Phi(\mathbf{X})$: either C (log-variance) or $C(C+1)/2$ (vectorized covariance matrix). We fixed the hidden layer size of m_{DSF} to C^2 units, while the output layer size depended on the chosen C'. The DSF modules added between 420 and 2,864 trainable parameters to those of $f_{\boldsymbol{\Theta}}$ depending on the configuration.

The Riemann pipeline first applied a filter bank to the input EEG, yielding narrow-band signals in the 7 bands bounded by (0.1, 1.5, 4, 8, 15, 26, 35, 49) Hz. Next, covariance matrices were estimated per window and frequency band using the OAS algorithm (Chen et al., 2010). The covariance matrices were then projected into their Riemannian tangent space exploiting the Wasserstein distance to estimate the mean covariance used as the reference point (Sabbagh et al., 2019; Bhatia et al., 2018). The vectorized covariance matrices with dimensionality of $C(C+1)/2$ were finally z-score normalized using the mean and standard deviation of the training set, and fed to a linear logistic regression classifier.

The handcrafted features baseline, inspired by Gemein et al. (2020) and Engemann et al. (2018a), relied on 21 different feature types: mean, standard deviation, root mean square, kurtosis, skewness, quantiles (10, 25, 75 and 90th), peak-to-peak amplitude, frequency log-power bands between (0, 2, 4, 8, 13, 18, 24, 30, 49) Hz as well as all their possible ratios, spectral entropy, approximate entropy, SVD entropy, Hurst exponent, Hjorth complexity, Hjorth mobility, line length, wavelet coefficient energy, Higuchi fractal dimension, number of zero crossings, SVD Fisher information and phase locking value. This resulted in 63 univariate features per EEG channel, along with $\binom{C}{2}$ bivariate features, which were concatenated into a single vector of size $63 \times C + \binom{C}{2}$ (*e.g.,* 393 for $C = 6$). In the event of non-finite values in the feature representation of a window, we imputed missing values feature-wise using the mean of the feature computed over the training set. Finally, feature vectors were fed to a random forest model.

When applying traditional pipelines to pathology detection experiments, we aggregated the input representations recording-wise as each recording has a single label (*i.e.,* pathological or not). To do so, we used the geometric mean on covariance matrices and the median on handcrafted features. Deep learning models, on the other hand, were trained on non-aggregated windows, but their performance was evaluated recording-wise by averaging the predictions over windows within each recording. Hyperparameter selection for logistic regression and random forest models is described below (Section 3.3.3).

We combined the machine learning approaches described above with the following noise-handling strategies: 1) no denoising, *i.e.,* models are trained directly on the data without explicit or implicit denoising, 2) Autoreject (Jas et al., 2017), an automated correction pipeline, and 3) data augmentation, which randomly corrupts channels during training.

Autoreject is a denoising pipeline that explicitly handles noisy epochs and channels in a fully automated manner (Jas et al., 2017). First, using a cross-validation procedure, it finds optimal channel-wise peak-to-peak amplitude thresholds to be used to identify bad channels in each window separately. If more than κ channels are bad, the epoch is rejected. Otherwise, up to ρ bad channels are reconstructed using the good channels with spherical spline interpolation. In pathology detection experiments, we allowed Autoreject to reject bad epochs, as classification was performed recording-wise. For sleep staging experiments however, we did not reject epochs as one prediction per epoch

Table 3.2 – Selected hyperparameters for experiments on number of channels (Section 3.4.1).

Model	Hyperparameter	Number of channels 2	6	21
Random Forest (RF)	Number of trees	300	300	300
	Tree depth	17	21	19
	Criterion	entropy	Gini	entropy
	Features	all	all	all
Logistic regression (LR)	C	0.1	0.1	0.001

was needed, but still used Autoreject to automatically identify and interpolate bad channels. In both cases, we used default values for all parameters as provided in the Python implementation[6], except for the number of cross-validation folds, which we set to 5.

Finally, data augmentation consists of artificially corrupting channels during training to promote invariance to missing channels. When training neural networks, the data augmentation transform was applied on-the-fly to each batch. For feature-based methods, we instead precomputed augmented datasets by applying the augmentation multiple times to each window (10 for pathology detection, 5 for sleep staging), and then extracting features from the augmented windows.

3.3.3 Hyperparameter optimization of baseline models

A grid-search over hyperparameters of the random forest and logistic regression classifiers was performed with 3-fold cross-validation on combined training and validation sets. This search was performed for each reported experimental configuration: for each number of channels (for experiments in Section 3.4.1), each denoising strategy (no denoising, Autoreject and data augmentation) and each dataset (TUAB, PC18 and MSD, presented in the next section).

For all RF models, we used 300 trees. This turned out to be a good trade-off between model performance and computational costs. For each experiment, we selected by cross-validation the depth of the trees among {13,15,17,19,21,23,25}, the split criterion between Gini and entropy, and the fraction of selected features used in each tree among 'sqrt' (the square-root of the number of features is used) , 'log2' (the logarithm in base 2 of the number of features is used), and using all features. For logistic regression models, the regularization parameter C was chosen among $\{10^{-4}, 10^{-3}, \ldots, 10\}$. We expanded the search on MSD as performance did not peak in the ranges considered above by adding the following values to the search space: depth in {1,3,5,7,9,11} and C in $\{10^2, 10^3, 10^4, 10^5\}$.

The selected hyperparameter configurations are listed in Tables 3.2 and 3.3 for the experiments in Sections 3.4.1 and 3.4.2, respectively. Once the best hyperparameters for an experimental configuration were identified, the training and validation sets were combined into a single set on which the model with the best hyperparameters was finally trained.

[6] https://github.com/autoreject/autoreject

Table 3.3 – Selected hyperparameters for experiments on denoising strategies (Section 3.4.2).

			Denoising strategy		
Dataset	**Model**	**Hyperparameter**	**No denoising**	**Autoreject**	**Data augm.**
TUAB	RF	Number of trees	300	300	300
		Tree depth	21	13	17
		Criterion	Gini	entropy	entropy
		Features	all	all	all
	LR	C	0.1	0.1	0.01
PC18	RF	Number of trees	300	300	300
		Tree depth	15	15	17
		Criterion	entropy	Gini	entropy
		Features	sqrt	sqrt	sqrt
	LR	C	1	1	10
MSD	RF	Number of trees	300	300	300
		Tree depth	9	9	11
		Criterion	entropy	entropy	entropy
		Features	all	sqrt	sqrt
	LR	C	0.1	0.1	10^5

Table 3.4 – Summary of the datasets used in this study.

	TUAB	PC18 (train)	MSD
Recording settings	Hospital	Sleep clinic	At-home
# recordings	2,993	994	98
# unique subjects	2,329	994	67
Sampling frequency (Hz)	250, 256 or 512	200	256
# EEG channels	27 to 36	6	4
Reference	Common average	M1 or M2	Fpz
Labels	Normal, abnormal	W, N1, N2, N3, R	W, N1, N2, N3, R

3.3.4 Data

Approaches were compared on three datasets (Table 3.4): for pathology detection on the TUH Abnormal EEG dataset (TUAB) (Obeid and Picone, 2016) and for sleep staging on both the Physionet Challenge 2018 dataset (PC18) (Ghassemi et al., 2018; Goldberger et al., 2000) and the Muse Sleep Dataset (MSD), an internal dataset of mobile overnight EEG recordings. A description of the recordings and preprocessing methodology can be found in Section 2.2.7 for both TUAB and PC18.[7]

In addition to these two dataset, we also included MSD, a collection of real-world mobile EEG recordings in which channel corruption is likely to occur naturally. MSD contains overnight sleep recordings collected with the Muse S EEG headband from InteraXon Inc. (Toronto, Canada). This data was collected in accordance with the privacy policy (July

[7]The unique difference is that here, all six EEG channels of PC18 were considered in some of our experiments.

2020) users agree to when using the Muse headband[8] and which ensures their informed consent concerning the use of EEG data for scientific research purposes. The Muse S is a four-channel dry EEG device (TP9, Fp1, Fp2, TP10, referenced to Fpz), sampled at 256 Hz. The Muse headband has been previously used for event-related potentials research (Krigolson et al., 2017), brain performance assessment (Krigolson et al., 2021), research into brain development (Hashemi et al., 2016), sleep staging (Koushik et al., 2018), and stroke diagnosis (Wilkinson et al., 2020), among others. A total of 98 partial and complete overnight recordings (mean duration: 6.3 h) from 67 unique users were selected from InteraXon's anonymized database of Muse customers, and annotated by a trained scorer following the AASM manual. Despite the derivations being different from the common montage used in polysomnography, the typical microstructure necessary to identify sleep stages, *e.g.,* sleep spindles, k-complexes and slow waves, can be easily seen in all four channels. Therefore, sleep stage annotations were obtained from actual EEG activity rather than ocular or muscular artifacts. Mean age across all recordings is 37.9 years (min: 21, max: 74) and 45.9% of recordings are of female users.

Preprocessing of MSD data was the same as for PC18, with the following differences: (1) channels were downsampled to 128 Hz, (2) missing values (occurring when Bluetooth packets are lost) were replaced by linear interpolation using surrounding valid samples, (3) after filtering and downsampling, samples which overlapped with the original missing values were replaced by zeros, and (4) channels were zero-meaned window-wise. Moreover, since the input sampling rate was of 128 Hz instead of 100 Hz for MSD experiments, we adapted the temporal convolution and max pooling hyperparameters of our CNNs to cover approximately the same duration: filter size of 64 samples, padding size of 13 and max pooling size of 16 (vs. 50, 10 and 13, respectively), yielding a total of 21,369 parameters.

We split the available recordings from TUAB, PC18 and MSD into training, validation and testing, such that recordings used for testing were not used for training or validation. For TUAB, we used the provided evaluation set as the test set. The recordings in the development set were split 80-20% into a training and a validation set. Therefore, we used 2,171, 543 and 276 recordings in the training, validation and testing sets. For PC18, we used a 60-20-20% random split, meaning there were 595, 199 and 199 recordings in the training, validation and testing sets respectively. Finally, for MSD, we retained the 17 most corrupted recordings for the test set (see Section 3.3.5 below for a detailed description of this process) and randomly split the remaining 81 recordings into training and validation sets (65 and 16 recordings, respectively). This was done to emulate a situation where training data is mostly clean, and strong channel corruption occurs unexpectedly at test time. We performed hyperparameter selection on each of the three datasets using a cross-validation strategy on the combined training and validation sets.

We repeated training on different training-validation splits (two for PC18, three for TUAB and MSD). Neural networks and random forests were trained three times per split on TUAB and MSD (two times on PC18) with different parameter initializations. Training ran for at most 40 epochs or until the validation loss stopped decreasing for a period of a least 7 epochs on TUAB and PC18 (a maximum of 150 epochs with a patience of 30 for MSD, given the smaller size of the dataset).

[8] https://choosemuse.com/legal/privacy/

Finally, accuracy was used to evaluate model performance for pathology detection experiments, while balanced accuracy (bal), defined as the average per-class recall, was used for sleep staging due to important class imbalance (the N2 class is typically much more frequent than other classes).

3.3.5 Analysis of channel corruption in the Muse Sleep Dataset

In this section, we provide more detail about Muse Sleep Dataset (MSD) and its noise characteristics. The at-home overnight recordings of MSD were purposefully selected to evaluate sleep staging algorithms in challenging mobile EEG conditions and therefore include recordings with highly corrupted channels. Overall, noise is stronger and more prevalent in these recordings than in typical sleep datasets collected under controlled laboratory conditions (*e.g.,* PC18).

To characterize the prevalence of channel corruption in MSD recordings, we can inspect the variance and the slope of the power spectral density (PSD) of each EEG channel across 30-s windows. Variance is a good measure of signal quality, while the spectral slope is a global descriptor of the frequency content of a signal and allows distinguishing between channel corruption (which yields flatter spectra) and artifacts (often displaying strong low frequencies, *e.g.,* eye movements). Simple thresholds set empirically on these two markers allowed approximate detection of channel corruption events. Specifically, we flagged a channel in a window as "corrupted" if its $\log_{10}$-$\log_{10}$ spectral slope (Schiratti et al., 2018) between 0.1 and 30 Hz was above -0.5 (unitless) and its variance was above 1,000 μV^2. We then computed a recording-wise channel corruption metric by taking the percentage of bad windows for the most corrupted channel of each recording.

About two-thirds of the recordings had no channel corruption according to this metric, while the remaining had a value of up to 96.4% (Figure 3.2). In those recordings with channel corruption, half of the corruption events (defined as a continuous block of epochs flagged as corrupted) lasted for 1.5 minutes or less, suggesting a large portion of the corruption happened intermittently, *e.g.,* due to the temporary displacement of the electrodes relative to the head. Some corruption events however lasted much longer, for instance up to 88 minutes in one case. These longer corruption events are likely due to bad connection between the skin and the electrode or to problems with the instrumentation.

For our experiments on MSD, we therefore selected the 81 cleanest recordings (*i.e.,* with the lowest corruption fraction) for training and validation and kept the 17 noisiest recordings for testing. This procedure allowed testing whether a model trained on relatively clean data could perform well even when random channel corruption was introduced at inference time.

3.3.6 Evaluation under conditions of noise

The impact of noise on downstream performance and on the predicted DSF filters was evaluated in three steps. First, we artificially corrupted the input EEG windows of TUAB and PC18 by using a similar process to our data augmentation strategy (Equation 3.6). We used the same values for η, σ and p, but used a single mask $\boldsymbol{\nu}$ per recording, such that the set of corrupted channels remained the same across a recording. Before corrupting, we subsampled a few EEG channels to recreate the sparse montage settings of TUAB (Fp1, Fp2, T3, T4, Fz, Cz) and PC18 (F3-M2, F4-M1, O1-M2, O2-M1). We then analyzed downstream performance under varying noise level conditions.

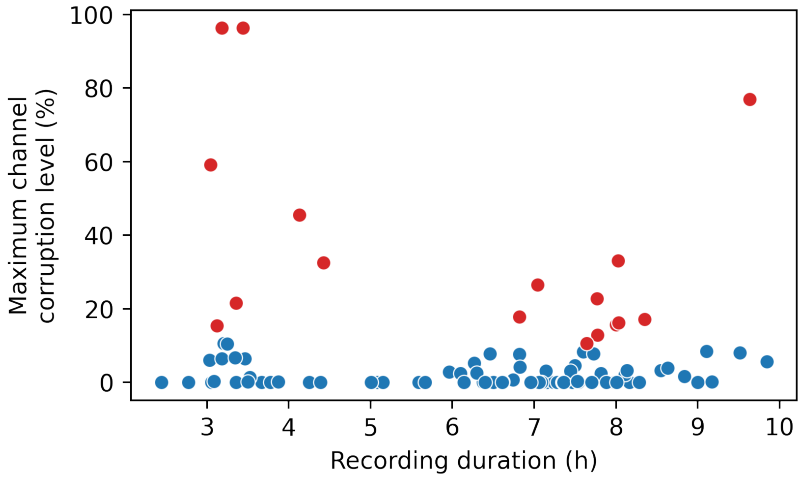


Figure 3.2 – Corruption percentage of the most corrupted channel of each of the 98 recordings of MSD. Each point represents a single recording. The 17 most corrupted recordings (red) were used as test set in our experiments of Section 3.4.2.

Second, we ran experiments on real corrupted data (MSD) by training our models on the cleanest recordings and evaluating their performance on the noisiest recordings. Finally, we analyzed the distribution of DSF filter weights predicted by a subset of the trained models.

3.4 Results

3.4.1 Performance of existing methods degrades under channel corruption

How do standard EEG classification methods fare against channel corruption? If channels have a high probability of being corrupted at test time, can noise be compensated for by adding more channels? To answer these questions, we measured the performance of three baseline approaches (Riemannian geometry, handcrafted features and a "vanilla" net, *i.e.,* ShallowNet without attention) trained on a pathology detection task on three different montages as channels were artificially corrupted. Results are presented in Figure 3.3.

All three baseline methods performed similarly and suffered considerable performance degradation as stronger noise was added (Figure 3.3A) and as more channels were corrupted (Figure 3.3B). First, under progressively noisier conditions, adding more channels did not generally improve performance. Strikingly, adding channels even hampered the ability of the models to handle noise. Indeed, the impact of noise was much less significant for 2-channel models than for 6- or 21-channel models. The vanilla net performed slightly better than the other methods in low noise conditions, however it was less robust to heavy noise when using 21 channels.

Second, when an increasing number of channels was corrupted (Figure 3.3B), using denser montages did improve performance, although by a much smaller factor than what might be expected. For instance, losing one or two channels with the 21-channel models only yielded a minor decrease in performance, while models trained on sparser montages

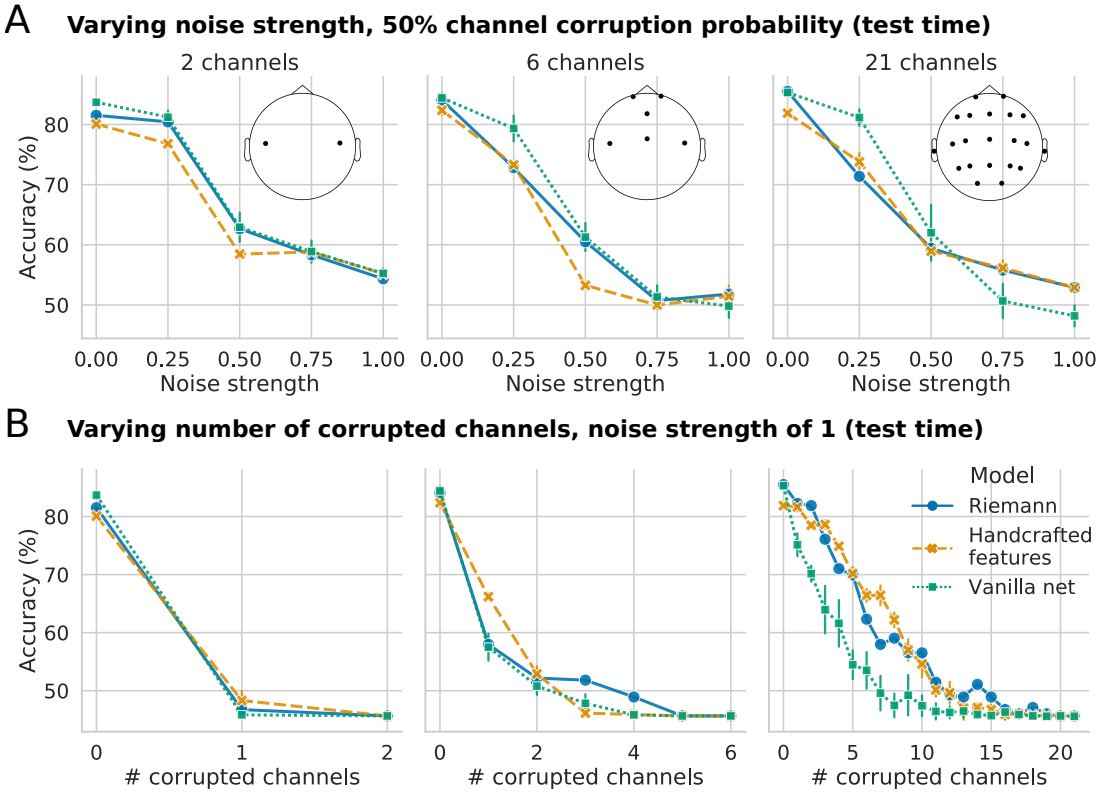


Figure 3.3 – Impact of channel corruption on pathology detection performance of standard models. We trained a filter-bank Riemannian geometry pipeline (blue), a random forest on handcrafted features (orange) and a standard ShallowNet architecture (green) on the TUAB dataset, given montages of 2 (T3, T4), 6 (Fp1, Fp2, T3, T4, Fz, Cz) or 21 (all available) channels. Performance was then evaluated on artificially corrupted test data under two scenarios: (A) the η noise strength parameter was varied given a constant channel corruption probability of 50%, and (B) the number of corrupted channels was varied given a constant noise strength of 1. Error bars show the standard deviation over 3 models for handcrafted features and 6 models for neural networks. While traditional feature-based models fared slightly better than a vanilla neural network in some cases (bottom right), adding noise predictably degraded the performance of all three models.

lost as much as 30% accuracy. However, even when as many as 15 channels were still available (*i.e.,* six corrupted channels), models trained on 21 channels performed worse than 2- or 6-channel models without any channel corruption, despite having access to much more spatial information on average. Interestingly, when models were trained on 21 channels, traditional feature-based methods were more robust to corruption than a vanilla net up to a certain point, however this did not hold for sparser montages.

These results suggest that standard approaches cannot handle significant channel corruption at a satisfactory level, even when denser montages are available. Therefore, better tools are necessary to train noise-robust models.

3.4.2 Attention and data augmentation mitigates performance loss under channel corruption

If including additional EEG channels does not by itself resolve performance degradation under channel corruption, what can be done to improve the robustness of standard EEG classification methods? We evaluated the performance of our models when combined with three denoising strategies (Section 3.3.2) for a fixed 6-channel montage[9]. Results on pathology detection (TUAB) are presented in Figure 3.4.

Without denoising, all methods showed a steep performance decrease as noise became stronger (Figure 3.4A) or more channels were corrupted (Figure 3.4B). Automated noise handling (second column) reduced differences between methods when noise strength was increased (Figure 3.4A), and helped marginally improve robustness when only one or two channels were corrupted (Figure 3.4B). However, it is only with data augmentation that clear performance improvements could be obtained, allowing all methods to perform considerably better in the noisiest settings (third column). Performance of traditional baselines was degraded however in low noise conditions. Neural networks, in contrast, saw their performance increase the most across noise strengths and numbers of corrupted channels. Whereas their performance decreased by at least 34.6% when going from no noise to strongest noise with the other strategies, training neural networks with data augmentation reduced performance loss to 5.3-10.5% on average. The DSF models improved performance further still over the vanilla ShallowNet by yielding an improvement of *e.g.,* 1.8-7.5% across noise strengths. Finally, adding the matrix logarithm and the soft-thresholding nonlinearity (DSFm-st, in magenta) yielded marginal improvements over DSFd. Under strong noise corruption ($\eta = 1$) our best performing model (DSFm-st + data augmentation) yielded an accuracy improvement of 29.4% over the vanilla net without denoising. Overall, this suggests that learning end-to-end to both predict and handle channel corruption at the same time is key to successfully improving robustness.

Next, we repeated this analysis on a sleep staging task using the PC18 dataset (Figure 3.5). As above, not using a denoising strategy led to a steep decrease in performance. Once more, Autoreject leveled out differences between the different methods and boosted performance under single-channel corruption, but otherwise did not improve or degrade performance as compared to training models without denoising. Data augmentation, in contrast, again helped improve the robustness of all methods. Interestingly, it benefited non-deep learning approaches more than in pathology detection, yielding

[9]This 6-channel montage (Fp1, Fp2, T3, T4, Fz, Cz) performed similarly to a 21-channel montage in no-corruption conditions (Figure 3.3) while being more representative of the sparse montages likely to be found in mobile EEG devices.

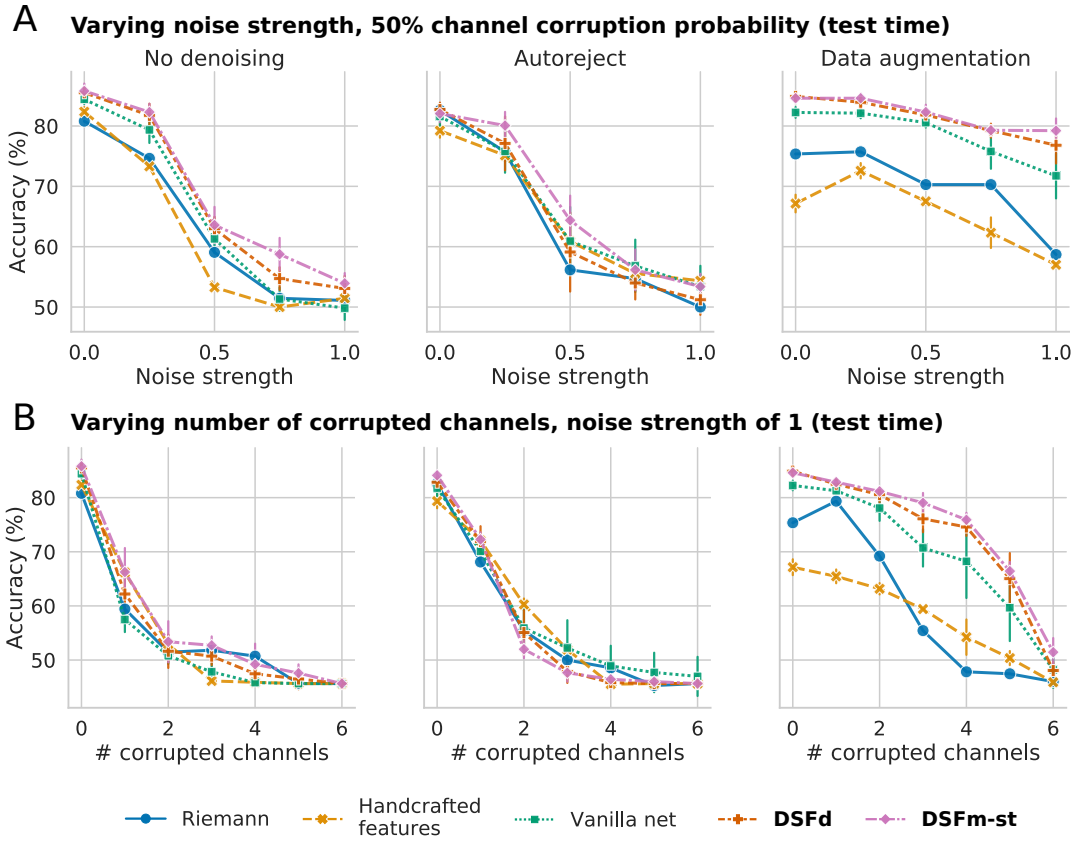


Figure 3.4 – Impact of channel corruption on pathology detection performance for models coupled with 1) no denoising strategy, 2) Autoreject and 3) data augmentation. We compared the per recording accuracy on the TUAB evaluation set (6-channel montage) as (A) the η noise strength parameter was varied given a constant channel corruption probability of 50%, and (B) the number of corrupted channels was varied given a constant noise strength of 1. Error bars show the standard deviation over 3 models for handcrafted features and 6 models for neural networks. Using an automated noise handling method (Autoreject; second column) provided some improvement in noise robustness over using no denoising strategy at all (first column). Data augmentation benefited all methods, but deep learning approaches and in particular DSF (third column, in red and magenta) yielded the best performance under channel corruption.

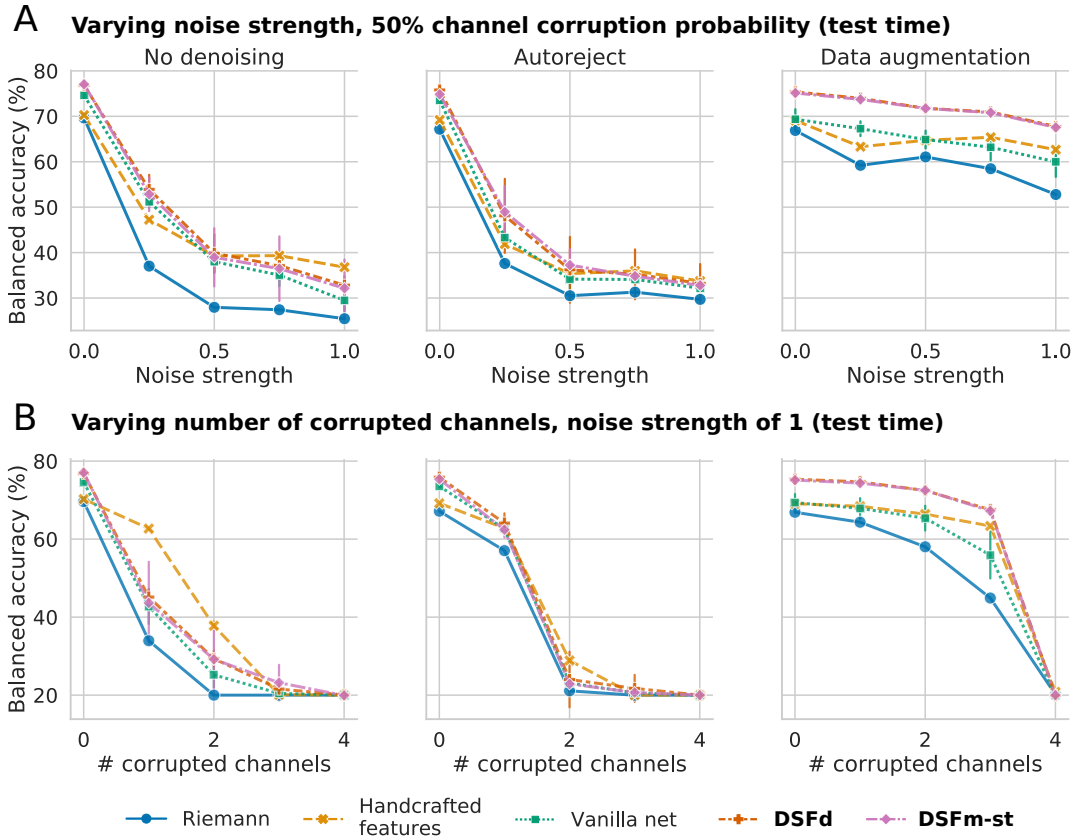


Figure 3.5 – Impact of channel corruption on sleep staging performance for models coupled with (1) no denoising strategy, (2) Autoreject and (3) data augmentation. We compared the test balanced accuracy on PC18 (4-channel montage) as (A) the η noise strength parameter was varied given a constant channel corruption probability of 50%, and (B) the number of corrupted channels was varied given a constant noise strength of 1. Error bars show the standard deviation over 3 models for handcrafted features and 4 models for neural networks. Similarly to Figure 3.4, automated noise handling provided a marginal improvement in noise robustness in some cases, data augmentation yielded a performance boost for all methods, while a combination of data augmentation and DSF (third column, red and magenta lines which overlap) led to the best performance under channel corruption.

for instance a similar performance for both handcrafted features and the vanilla StagerNet. DSF remained the most robust though with both DSFd and DSFm-st consistently outperforming all other methods. The performance of these two methods was highly similar, producing mostly overlapping lines (Figure 3.5).

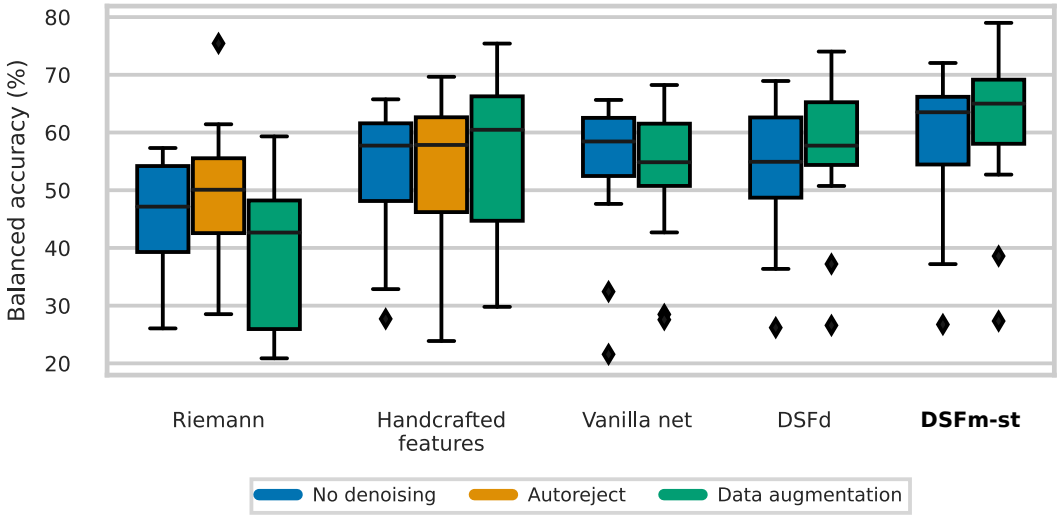


Figure 3.6 – Recording-wise sleep staging results on MSD. We show the distributions of balanced accuracy obtained by models with different random initializations (1, 3 and 9 initializations for Riemann, handcrafted features and deep learning models, respectively) on the test recordings of MSD. Noise handling with Autoreject had no clear impact on the performance of the handcrafted features, while data augmentation was detrimental to the Riemann model. The DSFm-st models reached the highest test performance when combined with data augmentation.

Finally, do these results hold under more intricate, naturally occurring corruption such as found in at-home settings? To verify this, we trained the same sleep staging models as above on the cleanest recordings of MSD (4-channel mobile EEG), and evaluated their performance on the 17 most corrupted recordings of the dataset. Results are presented in Figure 3.6. As above, the Riemann approach did not perform well, while the handcrafted features approach was more competitive with the vanilla StagerNet without denoising. However, contrary to the above experiments, noise handling alone did not improve the performance of our models. Data augmentation was even detrimental to the Riemann and vanilla net models on average. Combined with dynamic spatial filters (DSFd and DSFm-st) though, data augmentation helped improve performance over other methods. For instance, DSFm-st with data augmentation yielded a median balanced accuracy of 65.0%, as compared to 58.4% for a vanilla network without denoising. Performance improvements were as high as 14.2% when looking at individual sessions. Importantly, all recordings saw an increase in performance, showing the ability of our proposed approach to improve robustness in noisy settings.

Taken together, our experiments on simulated and natural channel corruption indicate that a strategy combining an attention mechanism and data augmentation yields higher robustness than traditional baselines and existing automated noise handling methods.

3.4.3 Attention weights are interpretable and correlate with signal quality

The DSF module was key to achieving high robustness to channel corruption on both pathology detection and sleep staging tasks. Can we explain the behavior of the module by inspecting its internal functioning? If so, in addition to improving robustness, DSF could also be used to monitor the effective importance of each incoming EEG channel, providing an interesting "free" insight into signal quality. To test this, we analyzed the effective channel importance ϕ_i of each EEG channel i to the spatial filters over the TUAB evaluation set. Results are shown in Figure 3.7.

Overall, the attention weights behaved as expected: the more usable (*i.e.,* noise-free) a channel was, the higher its effective channel importance ϕ_i was relative to those of other channels. For instance, without any additional corruption, the DSF module focused most of its attention on channels T3 and T4 (Figure 3.7A, first column), known to be highly relevant for pathology detection (Schirrmeister et al., 2017; Gemein et al., 2020). However, when channel T3 was replaced with white noise, the DSF module reduced its attention to T3 and instead further increased its attention on other channels (second column). Similarly, when both T3 and T4 were corrupted the module reduced its attention on both channels and leveraged the remaining channels instead, *i.e.,* mostly Fp1 and Fp2 (third column). Interestingly, this change is reflected by the topography of the predicted filters $\mathbf{W}_{\text{DSF}}$ (Figure 3.7B): for instance, some dipolar filters computing a difference between left and right hemispheres were dynamically adapted to rely on Fp1 or Fp2 instead of T3 or T4 (*e.g.,* filters 1, 3 and 5). Intuitively, the network has learned to ignore corrupted data and to focus its attention on the good EEG channels, and to do so in a way that preserves the meaning of each virtual channel.

To further verify the interpretability of DSF's attention weights on naturally-corrupted real-world EEG data, we visualized the normalized effective channel importance metric alongside a time-frequency representation of the raw EEG in Figure 3.8. As expected, the metric dropped to values close to zero when a channel suffered heavy corruption, *e.g.,* Fp1 throughout the recording (left column) and TP9 intermittently (right column). These results again illustrate the capacity of DSF to de-emphasize corrupted channels, but also highlight its capacity to dynamically adapt to changing noise characteristics.

3.4.4 Deconstructing the DSF module

What might explain the capacity of the DSF module to improve robustness to channel corruption and provide interpretable attention weights? By comparing DSF to simpler interpolation-based methods, DSF can be understood as a more complex version of a simple attention-based model that decides how much each input EEG channel should be replaced by its interpolated version (see detailed discussion in Section 3.5.6). With this connection in mind, we performed an ablation study to understand the importance of each additional mechanism leading to the formulation of the DSF module. Figure 3.9 shows the performance of the different attention module variations trained on the pathology detection task with data augmentation, under different noise strengths.

Naive interpolation of each channel based on the $C - 1$ others (orange) performed similarly to or worse than the vanilla ShallowNet model (blue) across noise strengths. Introducing a single attention weight (green) to control how much channels should be mixed with their interpolated version only improved performance for noise strengths above 0.5. Using one attention weight per channel (red) further improved performance,

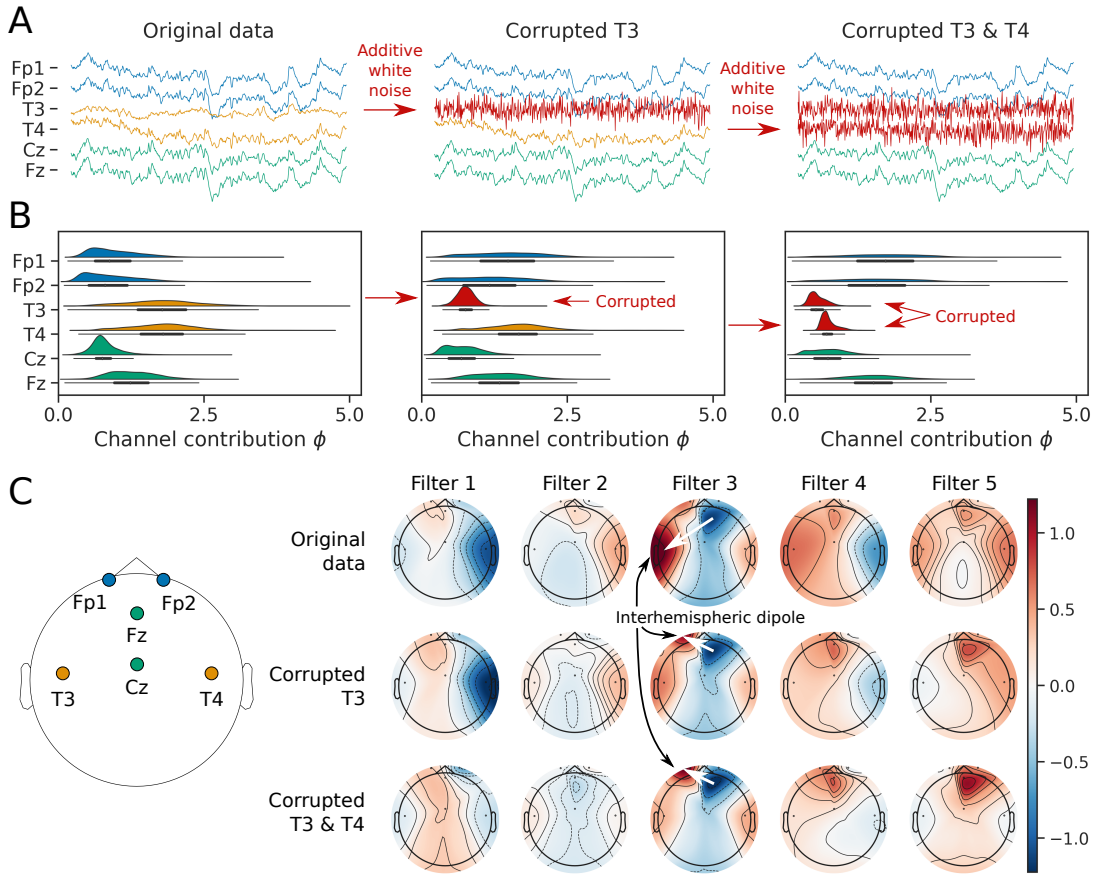


Figure 3.7 – Effective channel importance and spatial filters predicted by the DSF module trained on pathology detection. We compared three scenarios on the TUAB evaluation set: no added corruption, only T3 is corrupted and both T3 and T4 are corrupted. (A) The corruption process was carried out by replacing a channel with white noise ($\sigma \sim \mathcal{U}(20, 50)$ μV), as illustrated with a single 6-s example window (first row). (B) The distribution of effective channel importance values $\boldsymbol{\phi}$ is presented using density estimate and box plots. Corrupted channels are significantly down-weighted in the spatial filtering. (C) A subset of the spatial filters (median across all windows) are plotted as topomaps for the three scenarios. Corrupting T3 overall reduced the effective importance attributed to T3 and slightly boosted T4 values, while corrupting both T3 and T4 led to a reduction of $\boldsymbol{\phi}$ for both channels, but to an increase for the other channels. This change was also reflected in the overall topography: dipole-like patterns (indicated by white arrows) were dynamically modified to focus on clean channels (*e.g.*, Filter 3).

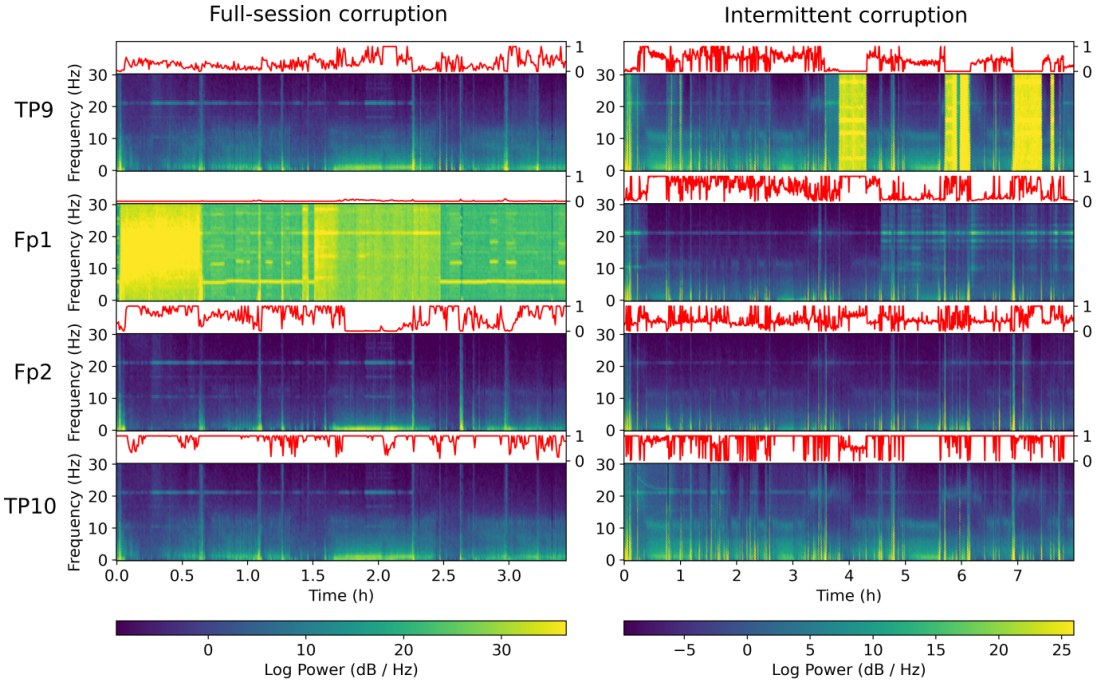


Figure 3.8 – Normalized effective channel importance $\hat{\phi}$ predicted by the DSF module on two MSD sessions with naturally-occurring channel corruption. Each column represents the log-spectrogram of the four EEG channels of one recording (Welch's periodogram on 30-s windows, using 2-s windows with 50% overlap). The red line above each spectrogram is the normalized effective channel importance $\hat{\phi}_i$ (see Eq. 3.5), between 0 and 1, computed using a DSFm-st model trained on MSD. When a channel is corrupted throughout the recording (left column, second row, as indicated by broad spectrum high power noise), DSF mostly "ignores" it by predicting small weights for that channel. This results in $\hat{\phi}_i$ values close to 0 for Fp1. When the corruption is intermittent (right column, first row), DSF dynamically adapts its spatial filters to only ignore important channels when they are corrupted. This is the case for channel TP9 around hours 4, 6, and 7, where $\hat{\phi}_i$ is again close to 0.

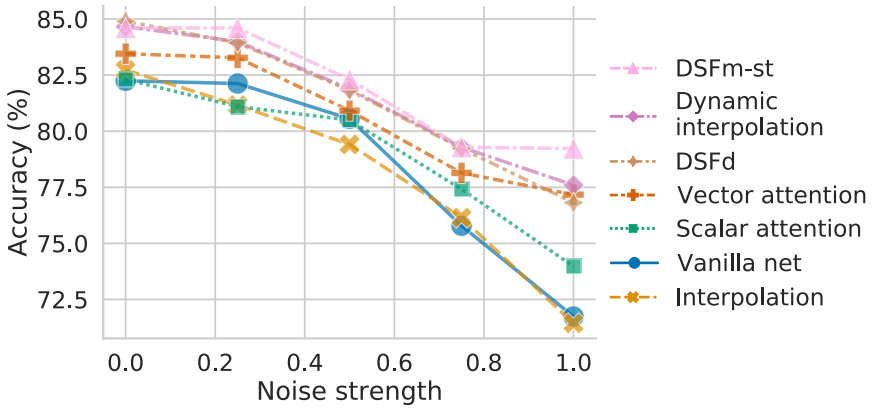


Figure 3.9 – Performance of different attention module architectures on the TUAB evaluation set under increasing channel corruption noise strength. Each line represents the average of 6 models (2 random initializations, 3 random splits). Models that dynamically generate spatial filters, such as DSF, outperform simpler architectures across noise levels.

this time across all noise strengths. The addition of dynamic interpolation (magenta), in which both the attention weights and an interpolation matrix are generated based on the input EEG window, yielded an additional substantial performance boost. Relaxing the constraints on the interpolation matrix and adding a bias vector to obtain DSFd (brown) led to very similar performance. Finally, the addition of the soft-thresholding non-linearity and the use of the matrix logarithm of the covariance matrix (DSFm-st, pink) further yielded performance improvements.

Together, these results show that combining channel-specific interpolation and dynamic prediction of interpolation matrices is necessary to outperform simpler attention module formulations. Performance can be further improved by providing the full covariance matrix as input to the attention module and encouraging the model to produce 0-weights with a nonlinearity.

3.5 Discussion

We introduced dynamic spatial filtering (DSF), a new method to handle channel corruption in EEG based on an attention mechanism architecture and a data augmentation transform. Plugged into a neural network whose input has a spatial dimension (*e.g.,* EEG channels), DSF predicts spatial filters that allow the model to dynamically focus on important channels and ignore corrupted ones. DSF shares links with interpolation-based methods traditionally used in EEG processing but in contrast does not require separate preprocessing steps that are often expensive with dense montages or poorly adapted to sparse ones. DSF outperformed feature-based approaches and automated denoising pipelines under simulated corruption on two large public datasets and in two different predictive tasks. Similar results were obtained on a smaller dataset of mobile sparse EEG with strong natural corruption, demonstrating the applicability of our approach to challenging at-home recording conditions. Finally, the inner functioning of DSF can easily be inspected using a simple measure of effective channel importance and topographical maps. Overall, DSF is computationally lightweight, easy to implement, and improves robustness to channel corruption in sparse EEG settings.

3.5.1 Handling EEG channel loss with existing denoising strategies

As opposed to the more general problem of "noise handling" (Table 3.1), we focused our experiments on the problem of channel corruption in sparse montages. In light of our results, we explain why existing strategies are not well suited for handling channel corruption, while DSF is.

Our first experiment (Section 3.4.1) demonstrated that adding more EEG channels does not necessarily make a classifier more robust to channel loss. In fact, we observed the opposite: a model trained on two channels can outperform 6- and 21-channel models under heavy channel corruption (Figure 3.3A). This can be explained by two phenomena. First, increasing the number of channels increases the input dimensionality of classifiers, making them more likely to overfit the training data. Tuning regularization hyperparameters can help with this, but does not solve the problem by itself. Second, in vanilla neural networks, the weights of the first spatial convolution layer, *i.e.,* the spatial filters applied to the input EEG, are fixed. If one of the spatial filter relies mostly on one specific (theoretically) important input channel, *e.g.,* T3, and this input channel is corrupted, all successive operations on the resulting virtual channel will carry noise as well. This highlights the importance of dynamic re-weighting: with DSF, we

can find alternative spatial filters when a theoretically important channel is corrupted, and even completely ignore a corrupted channel if it contains no useful information.

Since adding channels is not on its own a solution, can traditional EEG denoising techniques, *e.g.,* interpolation-based methods such as Autoreject (Jas et al., 2017), help handle the channel corruption problem? In our experiments, interpolation-based denoising did help but only marginally (middle column of Figure 3.4 and 3.5). The relative ineffectiveness of this approach can be explained by the very low number of available channels in our experiments (4 or 6) which likely harmed the quality of the interpolation. Our results therefore do not invalidate these kinds of methods (whose performance has been demonstrated multiple times on denser montages and in challenging noise conditions (Nolan et al., 2010; Bigdely-Shamlo et al., 2015; Jas et al., 2017)) but only expose their limitations when working with few channels. Still, there are other reasons why interpolation-based methods might not be optimal in settings like the ones studied in this chapter. For instance, completely replacing a noisy channel by its interpolated version means that any remaining usable information in this channel will be discarded and that any noise contained in the other (non-discarded) channels will end up in the interpolated channel.

Finally, an interesting case to consider is when tasks can be performed accurately with a single good channel, *e.g.,* sleep staging (Liang et al., 2012). In such a case, could a single-channel model perform as well as a multi-channel model, without the need to worry about the challenges discussed above? While this may be true if we have access to a reliably good channel, as soon as it is corrupted (*e.g.,* in real-world mobile EEG settings) it can no longer be used by the model. An ensemble of single-channel models might be an interesting solution; however this requires knowing both which channel to focus on and when, which is not trivial and requires additional logic and processing pipeline components. Moreover, to improve upon such a model by making use of spatial information (Chambon et al., 2018) the model should be trained on all possible combinations of good channels, which can quickly become prohibitive. DSF offers a compelling solution to the challenges encountered with single-channel models thanks to its end-to-end dynamic re-weighting capabilities.

3.5.2 Impact of the input spatial representation

The representation used by the DSF module constrains the types of patterns that can be leveraged to produce spatial filters. For instance, using the log-variance of each channel allows detecting large-amplitude corruption or artifacts, however this makes the DSF model blind to more subtle kinds of interactions between channels. These interactions can be very informative in certain cases, *e.g.,* when one channel is corrupted by a noise source which also affects other channels but to a lesser degree.

Our experiments suggested that models based on log-variance (DSFd) or vectorized covariance matrices (DSFm-st) were roughly equivalent in simulated noise conditions (Figure 3.4-3.5). This is likely because the additive white noise we used was not spatially correlated and therefore no spatial interactions could be leveraged by the DSF modules to identify noise. On naturally corrupted data however, using the full spatial information along with soft-thresholding was critical to outperforming other methods (Figure 3.6). This is likely because the noise in at-home recordings was often correlated spatially and because corrupted channels, often containing mostly noise (Section 3.3.5), could be completely ignored by DSF.

Related attention block architectures have used average-pooling (Hu et al., 2018) or a combination of average- and max-pooling (Woo et al., 2018) to summarize channels. Intuitively, average pooling should not yield a useful representation of the input, as EEG channels are often assumed to have zero-mean, or are explicitly highpass filtered to remove their DC offset. Max-pooling, on the other hand, does capture amplitude information that overlaps with second-order statistics, however it does not allow differentiating between large transient artifacts and more temporally consistent corruption. Experiments on TUAB (not shown) confirmed this: a combination of min- and max-pooling was less robust to noise than covariance-based models. From this perspective, vectorized covariance matrices or similar representations (Section 3.2.3) are an ideal choice of spatial representation. Ultimately, DSF could be fed with any learned representations with a spatial dimension, *e.g.,* filter-bank representations.

3.5.3 Impact of the data augmentation transform

Data augmentation was critical to developing invariance to corruption (Section 3.4.2). For instance, under simulated corruption, a vanilla neural network trained with our data augmentation transform gained considerable robustness, even without an attention mechanism. Does this mean that data augmentation is the key ingredient to DSF? In fact, our results on naturally corrupted data (Figure 3.6) showed that data augmentation without attention negatively impacted performance and that adding an attention mechanism was necessary to improve performance. Moreover, traditional pipelines generally did not benefit from data augmentation as much as neural networks did, and even saw their performance degrade considerably in certain cases, *e.g.,* in low noise conditions in pathology detection experiments and on the real-world data for the Riemann models.

Nonetheless, these results highlight the role of data augmentation transforms in developing robust representations of EEG. Recently, work in self-supervised learning for EEG (Banville et al., 2021a; Cheng et al., 2020; Mohsenvand et al., 2020) has further suggested the importance of well-characterized data augmentation transforms for representation learning. Importantly though, the motivation behind the use of data augmentation in our experiments was not primarily to reduce overfitting due to limited sample sizes like commonly done in deep learning, but rather to evaluate methods under controlled corruption of experimental data. Ultimately, our additive white noise transform could be combined with channel masking and shuffling (Saeed et al., 2020) and other potential corruption processes such as those described in (Cheng et al., 2020; Mohsenvand et al., 2020).

3.5.4 Interpreting dynamic spatial filters to measure effective channel importance

The results in Figure 3.7 demonstrated that visualizing the spatial filters produced by the DSF module can reveal the spatial patterns a model has learned to focus on (Section 3.4.3). As observed in our experiments, a higher $\boldsymbol{\phi}$ indicates higher effective importance of a channel for the downstream task. For instance, temporal channels were given a higher importance in the pathology detection task, as suggested by previous work (Schirrmeister et al., 2017; Gemein et al., 2020). Similarly, in real-world data, low $\boldsymbol{\phi}$ values were given to a channel whenever it was corrupted (Figure 3.8).

However, $\boldsymbol{\phi}$ is not a strict measure of signal quality but more of channel usefulness: there could be different reasons behind the boosting or attenuation of a channel by the DSF module. Naturally, if a channel is particularly noisy, its contribution might be brought down to zero to avoid contaminating virtual channels with noise. Conversely though, if the noise source behind a corrupted channel is also found (but to a lesser degree) in other channels, the corrupted channel could also be used to regress out noise and recover clean signals (Haufe et al., 2014). In other words, $\boldsymbol{\phi}$ reflects the importance of a channel conditionally to others.

Finally, using DSF to obtain a measure of channel usefulness actually opens the door to DSF being used in non-machine learning settings. For instance, once a neural network is trained with DSF, its effective channel importance values can be reused as an indicator of signal quality on similar data (*e.g.,* , data collected with the same or similar hardware). Such a signal quality metric can be helpful during data collection, or to know which parts of the recording should be kept for analysis.

3.5.5 Practical considerations

When faced with channel corruption in a predictive task, which modelling and denoising strategies should be preferred? This choice should depend on the number of available channels, as well as on assumptions about the stationarity of the noise. When using sparse montages, as in this chapter, different solutions can lead to good results. For instance, handcrafted features with random forests can perform well when spatial information is not critical (*e.g.,* sleep staging, Section 3.4.2) or noise is stationary (Engemann et al., 2018a), although they require a non-trivial feature engineering step. However, when less can be assumed about the predictive task, *e.g.,* corruption might be non-stationary or spatial information is likely important, DSF with data augmentation is an effective way to make a neural network noise-robust. Although we did not test denoising approaches on dense montages, we can expect different methods to work well in these settings. For instance, under stationary noise, Riemmanian geometry-based approaches were shown to be robust to the lack of preprocessing in MEG data (Sabbagh et al., 2020). If, on the other hand, noise is not stationary and the computational resources allow it, interpolation-based methods might be used to impute missing channels before applying a predictive model (*e.g.,* Jas et al. (2017)). In cases where introducing a separate preprocessing step is not desirable, DSF with data augmentation might again be a promising end-to-end solution.[10]

3.5.6 From simple interpolation to Dynamic Spatial Filtering

In this section, we establish an interesting conceptual link between DSF and noise handling pipelines such as Autoreject (Section 3.2.1) which rely on an interpolation step to reconstruct channels that have been identified as bad. Specifically, these pipelines use head geometry-informed interpolation methods (based on the 3D coordinates of EEG electrodes and spline interpolation) to compute the weights necessary to interpolate each channel using a linear combination of the $C-1$ other channels (Perrin et al., 1989). From this perspective, a naive method of handling corrupted channels might be to always replace each input EEG channel by its interpolated version based on the other $C-1$ channels. An "interpolation-only" module m_{interp} could be written as:

[10] In this case, the number of parameters of the module can be controlled by *e.g.,* selecting log-variance as the input representation or reducing dimensionality by using fewer spatial filters than there are input channels.

$$m_{\text{interp}}(\mathbf{X}) = \mathbf{W}_{\text{interp}}\mathbf{X} \ , \tag{3.9}$$

where $\mathbf{W}_{\text{interp}}$ is a $C \times C$ real-valued matrix with a 0-diagonal[11]. The limitation of this approach is that given at least one corrupted channel in the input $\mathbf{X}$, the interpolated version of all non-corrupted channels will be reconstructed in part from corrupted channels. This means noise will still be present, however given enough clean channels, its impact might be mitigated.

Improving upon the naive interpolation-only approach, we might add the ability for the model to decide whether (and to what extent) channels should be replaced by their interpolated version. For instance, if the channels in a given window are mostly clean, it might be desirable to keep the initial channels; however, if the window is overall corrupted, it might instead be better to replace channels with their interpolated version. This leads to a "scalar-attention" module m_{scalar}:

$$m_{\text{scalar}}(\mathbf{X}) = \alpha_{\mathbf{X}}\mathbf{X} + (1 - \alpha_{\mathbf{X}})\mathbf{W}\mathbf{X} \ , \tag{3.10}$$

where $\alpha_{\mathbf{X}} \in [0, 1]$ is the attention weight predicted by an MLP conditioned on $\mathbf{X}$ (*e.g.,* on its covariance matrix) and $\mathbf{W}$ is the same as for the interpolation-only module. While this approach is more flexible, it still suffers from the same limitation as before: there is a chance interpolated channels will be reconstructed from noisy channels. Moreover, the fact that the attention weight is applied globally, *i.e.,* a single weight applies to all C channels, limits the ability of the module to focus on reconstructing corrupted channels only.

Instead, the "vector attention" module m_{vector} introduces channel-wise attention weights, so that the interpolation can be independently controlled for each channel:

$$m_{\text{vector}}(\mathbf{X}) = \text{diag}(\boldsymbol{\alpha}_{\mathbf{X}})\mathbf{X} + (\mathbf{I} - \text{diag}(\boldsymbol{\alpha}_{\mathbf{X}}))\mathbf{W}\mathbf{X} \ , \tag{3.11}$$

where $\boldsymbol{\alpha}_{\mathbf{X}} \in [0, 1]^C$ is again obtained with an MLP and $\mathbf{W}$ is as above. Although more flexible, this version of the attention module still faces the same problem caused by static interpolation weights.

To solve this issue, we build on the previous approach by both predicting an attention vector $\boldsymbol{\alpha}_{\mathbf{X}}$ as before and dynamically interpolating with a matrix $\mathbf{W}_{\mathbf{X}} \in \mathbb{R}^{C \times C}$ (with a 0-diagonal) predicted by another MLP:

$$m_{\text{dynamic}}(\mathbf{X}) = \text{diag}(\boldsymbol{\alpha}_{\mathbf{X}})\mathbf{X} + (\mathbf{I} - \text{diag}(\boldsymbol{\alpha}_{\mathbf{X}}))\mathbf{W}_{\mathbf{X}}\mathbf{X} \ . \tag{3.12}$$

In practice, a single MLP can output $C \times C$ real values, which are then reorganized into a 0-diagonal interpolation matrix $\mathbf{W}$ and a C-length vector whose values are passed through a sigmoid nonlinearity to obtain the attention weights $\boldsymbol{\alpha}_X$. An interesting property of this formulation which holds for m_{vector} too is that $\boldsymbol{\alpha}_{\mathbf{X}}$ can be directly interpreted as the level to which each channel is replaced by its interpolated version. However, in contrast to m_{vector} the interpolation filters can dynamically adapt to focus on the most informative channels.

Finally, we observe that Eq. (3.12) can be rewritten as a single matrix product:

$$m_{\text{dynamic}}(\mathbf{X}) = \Big(\text{diag}(\boldsymbol{\alpha}_{\mathbf{X}}) + (\mathbf{I} - \text{diag}(\boldsymbol{\alpha}_{\mathbf{X}}))\mathbf{W}_{\mathbf{X}}\Big)\,\mathbf{X} = \Omega_{\mathbf{X}}\mathbf{X} \ , \tag{3.13}$$

[11] $\mathbf{W}_{\text{interp}}$ can be set or initialized using head geometry information (Perrin et al., 1989) or can be learned from the data end-to-end.

where, denoting the element i, j of matrix $\mathbf{W}_\mathbf{X}$ as W_{ij},

$$\mathbf{\Omega}_\mathbf{X} = \begin{bmatrix} \alpha_1 & (1-\alpha_1)W_{12} & \dots & (1-\alpha_1)W_{1C} \\ (1-\alpha_2)W_{21} & \alpha_2 & \dots & (1-\alpha_2)W_{2C} \\ \vdots & \vdots & \ddots & \vdots \\ (1-\alpha_C)W_{C1} & (1-\alpha_C)W_{C2} & \dots & \alpha_C \end{bmatrix} . \tag{3.14}$$

The matrix $\mathbf{\Omega}_\mathbf{X}$ contains C^2 free variables, that are all conditioned on $\mathbf{X}$ through an MLP. We can then relax the constraints on $\mathbf{\Omega}_\mathbf{X}$ to obtain a simple matrix $\mathbf{W}_{\text{DSF}}$ where there are no dependencies between the parameters of a row and the diagonal elements are allowed to be real-valued. This new unconstrained formulation can be interpreted as a set of spatial filters that perform linear combinations of the input EEG channels. We can further introduce an additional bias term to recover the DSF formulation introduced in Section 3.2.2:

$$m_{\text{DSF}}(\mathbf{X}) = \mathbf{W}_{\text{DSF}}(\mathbf{X})\mathbf{X} + \mathbf{b}_{\text{DSF}}(\mathbf{X}) \ . \tag{3.15}$$

This bias term can be interpreted as a dynamic re-referencing of the virtual channels. In contrast to the interpolation-based formulations, DSF allows controlling the number of "virtual channels" C' to be used in the downstream neural network in a straightforward manner (*e.g.,* enabling the use of montage-specific DSF heads that could all be plugged into the same $f_\mathbf{\Theta}$ with fixed input shape). As shown in Section 3.4.4, DSF also outperformed interpolation-based formulations in our experiments.

3.5.7 Related work

Deep learning and noise robustness for audio data Noise robustness is of particular interest to the speech recognition community. For example, "noise-aware training" was proposed to train deep neural networks on noisy one-channel speech signals by providing an estimate of the noise level as input to the network (Seltzer et al., 2013). Noise-invariant representations of speech signals were also developed by training a classifier to perform well on the speech recognition task but badly on signal quality classification (Serdyuk et al., 2016) or by penalizing the distance between the internal representations of clean and noisy signals (Liang et al., 2018; Salazar et al., 2018). Methods have also been designed to leverage the spatial information of multiple audio channels similarly to our proposed DSF approach. Deep beamforming networks were used to dynamically re-weight different audio channels to improve robustness to noise, for instance with filter prediction subnetworks (Li et al., 2016; Xiao et al., 2016b,a). In a fashion similar to ours, recent work also used spatial attention to re-weight beamformed input speech signals to decide which filters to focus on (He et al., 2020).

Attention mechanisms for EEG processing Recent efforts in the deep learning and EEG community have led to various applications of attention mechanisms to end-to-end EEG processing. First, some studies used attention to improve performance on a specific task by focusing on different dimensions of an EEG representation. For instance, NLP-inspired attention modules were used in sleep staging architectures to improve processing of temporal dependencies (Phan et al., 2019; Yuan et al., 2019; Guillot et al., 2020; Phan et al., 2020; Guillot and Thorey, 2021). Attention was also applied in the spatial dimension to dynamically combine information from different EEG channels (Yuan et al., 2018; Yuan and Jia, 2019) or even from heterogeneous channel types (Yuan et al., 2019). In one case, spatial and temporal attention were used simultaneously in

a BCI classification task (Huang et al., 2019). Second, attention mechanisms have been used to enable transfer learning between different datasets with possibly different montages. In Nasiri and Clifford (2020), two parallel attention mechanisms allowed a neural network to focus on the channels and windows that were the most transferable between two datasets. Combined with an adversarial loss, this approach improved domain adaptation performance on a cross-dataset sleep staging task. Similarly to DSF, a spatial attention block was used in Guillot and Thorey (2021) to recombine input channels into a fixed number of virtual channels and allow models to be transferred to different montages. A Transformer-like spatial attention module was also proposed to dynamically re-order input channels (Saeed et al., 2020). In contrast to DSF, though, these approaches used attention weights in the $[0, 1]$ range, breaking the conceptual connection between channel recombination and spatial filtering.

3.5.8 Limitations

Our experiments on sleep data focused on window-wise decoding, *i.e.,* we did not aggregate larger temporal context but directly mapped each window to a prediction. However, modeling these longer-scale temporal dependencies was recently shown to help sleep staging performance significantly (Supratak et al., 2017; Chambon et al., 2018; Phan et al., 2019; Yuan et al., 2019; Guillot et al., 2020; Phan et al., 2020; Guillot and Thorey, 2021). Despite a slight performance decrease, window-wise decoding offered a simple but realistic setting to test robustness to channel corruption, while limiting the number of hyperparameters and the computational cost of the experiments. In practice, the effect of data corruption by far exceeded the drop in performance caused by using slightly simpler architectures.

The data augmentation and the noise corruption strategies exploited in this work employ additive Gaussian white noise. While this approach helped develop noise robust models, spatially non-correlated additive white noise represents an "adversarial scenario". Indeed, under strong white noise, the information in higher frequencies is more likely to be lost than with *e.g.,* pink or brown noise. Additionally, the absence of spatial noise correlation means that spatial filtering can less easily leverage multi-channel signals to regress out noise (Section 3.5.4). Exploring more varied and realistic types of channel corruption could further help clarify the ability of DSF to work under different conditions. Despite this, our experiments on naturally corrupted sleep data showed that additive white noise as a data augmentation does help improve noise robustness.

Finally, we focused our empirical study of channel corruption on two clinical problems that are prime contenders for mobile EEG applications: pathology screening and sleep monitoring. Interestingly, these two tasks have been shown to work well even with limited spatial information (*i.e.,* single-channel sleep staging (Liang et al., 2012)) or to be highly correlated with simpler spectral power representations (Schirrmeister et al., 2017). Therefore, future work will be required to validate the use of DSF on tasks where fine-grained spatial patterns might be critical to successful prediction, *e.g.,* brain age estimation (Engemann et al., 2020) as presented in the next chapter. Other common EEG-based prediction tasks such as seizure detection might benefit from DSF and will require further validation.

3.6 Conclusion

We presented dynamic spatial filtering (DSF), an attention mechanism architecture that improves robustness to channel corruption in EEG prediction tasks. Combined with a data augmentation transform, DSF outperformed other noise handling procedures under simulated and real channel corruption on three datasets. Altogether, DSF enables efficient end-to-end handling of channel corruption, works with few channels, is interpretable and does not require expensive preprocessing. We hope that our method can be a useful tool to improve the reliability of EEG processing in challenging non-traditional settings such as user-administered, at-home recordings.

4

Brain age as a proxy measure of neurophysiological health using low-cost mobile EEG

Contents

The novel methodology presented in Chapters 2 and 3 facilitated the use of real-world EEG data. Here, in the final chapter of this thesis, we take a step back and consider a promising application of large-scale cross-sectional datasets such as those created by real-world applications of EEG.

As we showed in Chapter 2, unlabelled EEG data can be harnessed for common downstream tasks like sleep staging and pathology detection thanks to new approaches like self-supervision. However, in many cases, we do actually have access to *some* labels which, though they might not be as informative as carefully collected labels, contain sufficient information to be used for a related learning task. For instance, as compared to the labels a trained neurologist can provide (*e.g.,* the onset and duration of patho-

logical EEG events or an overall diagnosis), *weak labels* like age and gender are widely accessible since they can be obtained without clinical expertise or careful methodological planning. Age, specifically, is a valuable piece of information to have access to, as aging and health are so intricately linked.

One compelling use of simple age labels is the *brain age* framework proposed by Franke et al. (2010) for building biomarkers of neurophysiological health. By measuring how someone's chronological age deviates from the age predicted based on their brain anatomy or physiology (*i.e.,* the brain age Δ), it is possible to identify individuals whose neurological characteristics look different from those of healthy people in the same age range. For example, a large brain age Δ could reflect accelerated aging or neurological pathologies. Although this framework has been studied for more than a decade, most research on brain age still relies on expensive neuroimaging modalities, such as structural magnetic resonance imaging (MRI), in order to derive predictions. Low-cost mobile EEG, because of its ease of use and affordability, is an obvious contender for translating brain age monitoring to real-world settings and making it a more practical framework.

In this chapter, we validate the idea of an EEG-based brain age metric through experiments on hundreds of real-world meditation and sleep EEG recordings. Our validation analysis relies on well-known biomarkers of sleep and aging to which we compare the predicted brain age of our models. Deep learning models, including the DSF attention module described in Chapter 3, are shown to outperform other approaches to age regression. We also investigate the variability of such an EEG-based brain age metric by analyzing longitudinal data in subjects with up to 220 recordings spread out over more than a year. Overall, our findings open the door to widespread, regular neurophysiological health monitoring in at-home settings.

This chapter is based on a manuscript in preparation for submission to the *Journal of Sleep Research*:

- **Hubert Banville**, Sean UN Wood, Maurice Abou Jaoude, Chris Aimone, Alexandre Gramfort, and Denis-Alexander Engemann. Brain age as a proxy measure of neurophysiological health using low-cost mobile EEG. Manuscript in preparation, 2021b

4.1 Introduction

The accessibility and affordability of low-cost mobile EEG devices, as well as their ease of use, open the door to the possibility of tracking neurophysiological health on a day-to-day basis. For instance, the regular screening of brain health might support early identification and prevention or treatment of various neurological pathologies. However, a major challenge to the realization of this objective is the need for automated EEG analysis. Relying on trained experts to inspect recordings visually is not scalable, given the amount of data that is already being produced. Unfortunately, building machine learning models to detect or predict pathologies from EEG data often requires large amounts of labelled data, which again requires expert inspection. Moreover, this labelling process needs to be repeated for each pathology of interest. Instead, a recently proposed alternative to address this problem is to use proxy measures of health (Dadi et al., 2021). For instance, *brain age* prediction is a versatile approach to performing brain health monitoring without having access to expert labels (Cole et al., 2019).

While chronological age, *i.e.,* the number of years elapsed since birth, is a useful metric to monitor the actual biological aging process, it does not provide a complete picture of this complex phenomenon. For instance, age does not affect everyone uniformly: some individuals tend to lose cognitive abilities earlier than others (LaPlume et al., 2021). As the brain changes through development and adulthood, the characteristics of EEG signals evolve as well (Feinberg et al., 1967; Mourtazaev et al., 1995; Landolt and Borbély, 2001; Chiang et al., 2011; Voytek et al., 2015; Hashemi et al., 2016). Rather than chronological age then, a data-driven measure might reflect the process of aging more faithfully. This is the objective behind the brain age framework proposed by Franke et al. (2010, 2012). Specifically, the *brain age* Δ is defined as the difference between an age estimate (*i.e.,* the age predicted by a model trained on a healthy population) and chronological age.[1] Positive values indicate someone's brain looks "older" than it normally does for healthy individuals in the same age group, suggesting this individual's brain is aging prematurely or that it presents a pathology that makes it look older[2] (Cole and Franke, 2017; Cole et al., 2018).

Most research on the topic of brain age has focused on structural MRI, as this technique allows the quantitative analysis of anatomical changes such as reduced grey matter volume and increased cerebrospinal fluid volume (Franke et al., 2010; Cole et al., 2019). For instance, a positive MRI-based brain age Δ has been shown to be correlated with Alzheimer's disease (Franke and Gaser, 2012), sleep pathologies such as sleep apnea (Weihs et al., 2021), and even to predict mortality (Cole et al., 2018). Considering the compelling results obtained with anatomical brain information only, the question arises as to whether functional information, *i.e.,* obtained through functional neuroimaging modalities, could provide similar, or even complementary, information on brain age. This hypothesis was tested in Engemann et al. (2020), where it was shown that adding fMRI and MEG modalities significantly improve brain age prediction performance.

In light of these findings, and as EEG is portable and significantly more affordable than MRI, functional information available through EEG emerges as a prime contender for estimating brain age more practically. This has already been the topic of a few studies. For instance, in Al Zoubi et al. (2018), a classifier ensemble learned to predict age from 31-channel eyes-open resting state EEG with a MAE of 6.87 years. On a more restricted population of children and adolescents, a similar handcrafted feature-based approach yielded low MAEs in the 1.5 to 2.1 year range using 14-channel eyes-closed resting state EEG recordings (Vandenbosch et al., 2019). An approach relying on filter-bank covariance matrices and Riemannian geometry was also shown to yield good brain age prediction performance on the TUH Abnormal Dataset (Sabbagh et al., 2020). A series of papers further showed how a similar brain age measure could be derived from sleep EEG data. For instance, an MAE of 7.6 years was obtained by training a linear regression (with a softplus link function) on handcrafted features capturing sleep micro- and macrostructure obtained with six EEG electrodes (Sun et al., 2019). Building on this approach, it was further shown that this sleep EEG-based brain age Δ was associated with life expectancy (Paixao et al., 2020) and dementia (Ye et al., 2020).

[1] Here, we use the term "brain age" to refer to the age predicted by the normative model, and "brain age Δ" to refer to the difference between chronological age and brain age. The brain age Δ is also sometimes called BrainAGE (Franke et al., 2010) or "brain age index" (Sun et al., 2019).

[2] Generally speaking, a deviation, be it positive or negative, from the chronological age, might indicate abnormal aging of the brain.

More recently, feature-based approaches to brain age prediction have been improved upon using deep learning. In Cole et al. (2017), CNNs were used to predict brain age from structural MRI, improving upon the results obtained with a Gaussian process regression model. Again on structural MRI, Peng et al. (2021) have shown state-of-the-art results by combining a CNN and various performance improving techniques (including data augmentation and pre-training). Similar approaches have been proposed for EEG data, for example in Brink-Kjaer et al. (2020), where a MobileNetv2 architecture was trained on multimodal sleep recordings (including 2-channel EEG, but also EMGs and ECGs, among others). In Nygate et al. (2021), a large dataset of 134,000 sleep recordings was used to train and evaluate a deep learning-based brain age prediction model, whose predictions were correlated to various clinical conditions such as epilepsy and seizure disorders, stroke and an elevated apnea-hypopnea index. Finally, deep learning architectures have also been used, although on a more limited dataset of young children, to predict brain age from auditory event-related potentials (ERPs) captured with 30-channel EEG (Bruns, 2021).

These previous studies have laid the groundwork for neuroimaging-based, and specifically, EEG-based, brain age prediction to provide a valuable proxy measure of brain health and aging. However, to maximize the impact brain age could have as a neurophysiological health screening tool, it should be made available to as many people as possible as part of a regular screening procedure. To this end, could low-cost mobile EEG be used to reliably predict brain age from real-world, out-of-the-lab recordings?

In this chapter, we aim to answer this question by presenting experiments on at-home EEG recordings from more than 1,000 subjects which address the following points: 1) how can machine learning be used to build brain age prediction models from low-cost mobile EEG?, 2) do the resulting brain age predictions contain health-related information not available in the chronological age? and 3) how variable are the brain age predictions over long periods of time?

The rest of the chapter is organized as follows. In Section 4.2, we introduce the machine learning problem under study, along with the datasets, machine learning pipelines and training procedure. The extraction of biomarkers from sleep recordings, as well as their analysis is also described. Next, Section 4.3 presents the results of three experiments designed to explore the feasibility and meaningfulness of brain age prediction from low-cost mobile EEG. Lastly, in Section 4.4, we discuss the results and propose directions for future research.

4.2 Methods

In this section, we describe how we train machine learning models to predict age from low-cost mobile EEG recordings and how the resulting brain age metric is evaluated (Figure 4.1). We first define the machine learning problem under study. We then present the different EEG datasets used in our experiments. The extraction of sleep biomarkers, central to the validation of the brain age metric, is then motivated. Next, we describe the different machine learning models considered, as well as their training and performance evaluation procedures, and describe how their brain age predictions will be validated against sleep biomarkers. Finally, we list some computational considerations for our experiments.

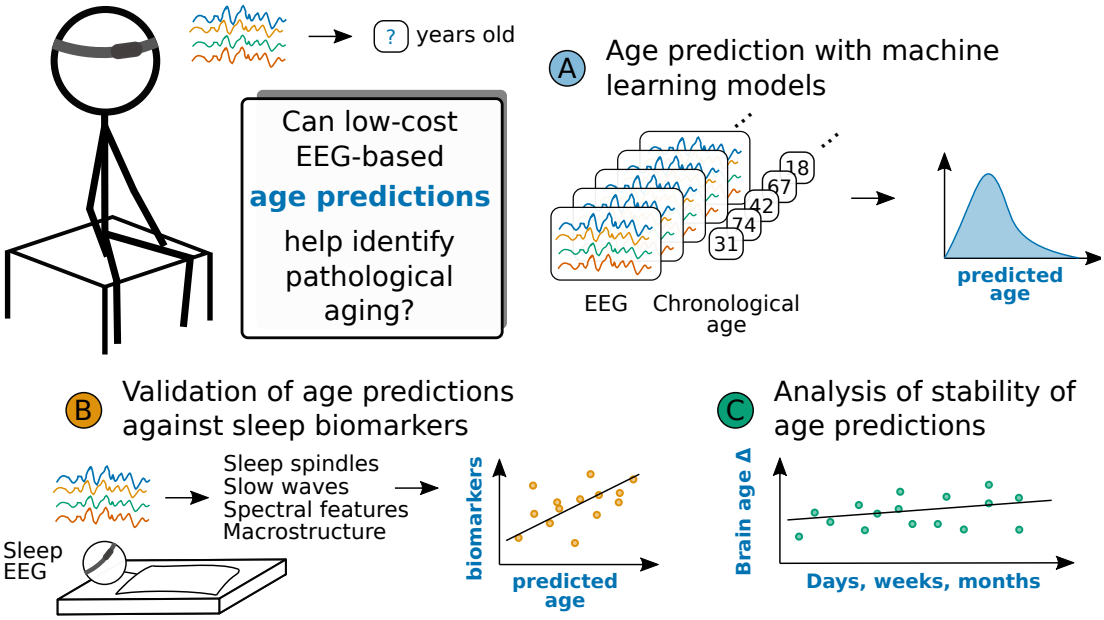


Figure 4.1 – Overview of our approach and experiments. We seek to predict age from low-cost, mobile EEG to provide health-related information to complement chronological age. (A) Machine learning models were trained to predict age from 5-8 minutes of eyes-closed meditation data from 971 unique subjects. (B) The predictions of the best machine learning models were compared to known biomarkers of sleep in order to assess whether brain age contains information that is not contained in chronological age. (C) The stability of the brain age predictions was assessed longitudinally using data from a few subjects with recordings collected over the span of more than one year.

4.2.1 Problem definition

The brain age prediction problem is a supervised learning regression problem of the form:

$$\hat{f}_{\boldsymbol{\Theta}} = \arg\min_{\boldsymbol{\Theta}} \mathbb{E}_{\mathbf{X},y\in\mathcal{X}\times\mathcal{Y}}[\mathcal{L}(f_{\boldsymbol{\Theta}}(\mathbf{X}), y)] \ , \tag{4.1}$$

where a model $f_{\boldsymbol{\Theta}} : \mathcal{X} \to \mathcal{Y}$ with parameters $\boldsymbol{\Theta}$ is trained to predict the age y of a subject whose EEG has been recorded. Here, $\mathcal{Y}$ is a positive integer (representing the age in years). It ranged from 18 and 81 in the datasets we used (Section 4.2.2). $f_{\boldsymbol{\Theta}}$ can be implemented, for instance, as a convolutional neural network. The model might receive a single EEG window $\mathbf{X}^{(i)}$, a sequence of consecutive windows, or even an entire recording $\mathbf{S}^{(i)}$ as input.[3] Here, our deep learning models used single EEG windows as input, while baseline "shallow" models used aggregated features over entire recordings (Section 4.2.5). We train $f_{\boldsymbol{\Theta}}$ to minimize the loss $\mathcal{L}$, *e.g.,* the mean absolute error (MAE) (*i.e.,* the L1 loss) between the true target age $y^{(i)}$ and the predicted age $\hat{y}^{(i)} = f_{\boldsymbol{\Theta}}(X^{(i)})$ over a training set of N examples:

$$\mathcal{L}_{MAE}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{1}{N}\sum_{i=1}^{N}\left|y^{(i)} - \hat{y}^{(i)}\right| \ , \tag{4.2}$$

[3] Age is considered to be fixed over the course of a single recording. In other words, all samples from a recording share the same target.

where $|\cdot|$ is the element-wise absolute value. We directly optimize the L1 loss as it is the most common metric used to evaluate performance in brain age studies (Franke et al., 2012; Cole and Franke, 2017; Sun et al., 2019; Sabbagh et al., 2020).

4.2.2 Data

Our experiments make use of real-world mobile EEG datasets collected from users of the Muse S headband (InteraXon Inc., Toronto, Canada). This data was collected in accordance with the privacy policy (July 2020) users agree to when using the Muse headband[4] and which ensures their informed consent concerning the use of EEG data for scientific research purposes. As previously described in Section 3.3.4, the Muse headband has been already used in many different areas of research, such as research into brain development (Hashemi et al., 2016), sleep staging (Koushik et al., 2018), and stroke diagnosis (Wilkinson et al., 2020), among others.[5]

Data collection The Muse S is a wireless, dry EEG device with four channels (TP9, Fp1, Fp2, TP10, referenced to Fpz). EEG is sampled at 256 Hz and sent over Bluetooth to a nearby mobile device (*e.g.,* phone or tablet) for real-time analysis and saving to file. Two types of recordings were collected to build the datasets used in this study: 1) meditation recordings with neurofeedback and 2) overnight sleep recordings (Table 4.1). In both cases, recordings were done by the users through the *Muse* application on iOS or Android mobile operating systems.[6] The application first guided the users through the setup of their EEG device, then through a signal quality check and finally through the recording, as described next.

Meditation recordings Awake continuous EEG was collected during meditation recordings in which users performed a focused attention task. First, an eyes-closed resting state period of 30-s to more than a minute was performed for the purpose of calibrating the neurofeedback algorithm and allowing signal quality to improve. Users then moved on to a neurofeedback exercise, in which auditory feedback was provided to them in real-time. The length of the neurofeedback exercise (between 1 minute and 3 hours) was selected beforehand by the users. During the exercise, users were instructed to stay still, keep their eyes closed and focus their attention on their breathing. As the exercise proceeded, users were then provided with auditory feedback on their mental state derived from their own real-time EEG, through a proprietary algorithm using machine learning and spectral analysis to map EEG into sound.

Overnight sleep recordings In overnight sleep recordings, users had the option to listen to a choice of audio content and/or auditory feedback while falling asleep. On-device sleep staging was then performed to provide the users with a hypnogram-based analysis of their night when they woke up the next morning.

Next, we describe the three datasets collected by selecting recordings from InteraXon Inc.'s anonymized database of Muse users.

[4] https://choosemuse.com/legal/privacy/
[5] https://choosemuse.com/muse-research/
[6] https://choosemuse.com/muse-app/

Muse Meditation Dataset (MMD)

Meditation recordings of 5 minutes and more were selected from Muse S meditation recordings collected between October 2019 and October 2021. Recordings were then filtered to only keep users whose age was between 18 and 81 years at the time of recording. A single recording was sampled per user, such that the age distribution across all sampled recordings was approximately uniform and the dataset was balanced for male and female users. This yielded a total of 3,667 candidate recordings. From this set, only recordings with excellent signal quality, based on basic signal statistics defined as follows, were retained. First, recordings for which more than 5% of samples were missing for any of the EEG channels (caused by Bluetooth packet loss during transmission from the headband to the mobile device) were rejected. The variance of the signals bandpass-filtered between 2 and 26 Hz was also computed for non-overlapping 1-s windows. Recordings for which, for any of the four channels, 25% or more of the windows had a variance above a threshold of 100 μV^2 were rejected. This finally yielded a subset of 971 recordings (mean duration: 15.1 ± 6.9 minutes) with excellent signal quality from 971 unique individuals. Mean age across recordings was 47.7 ± 16.4 years old (min: 18, max: 81) and 35.9% of recordings were of female users.

Overlapping Meditation-Sleep Dataset (OMSD)

Separately, we sought to collect a dataset of recordings where users performed both a meditation recording and a sleep recording in close temporal proximity. Being able to analyze both types of recordings alongside one another is key to assessing how the predicted brain age obtained with models trained on meditation recordings relates to known sleep EEG biomarkers. To identify candidate recordings, we first sampled sleep recordings, following a similar procedure to the one described above for MMD. Recordings that lasted between 5 and 11 hours, from users between 18 and 81 years of age at the time of recording, and which were started between 5 PM and 5 AM, local time, were selected. A maximum of two recordings per user were selected, yielding a total of 1046 candidate recordings, that were then screened for signal quality. Signal quality screening was performed on non-overlapping 30-s windows (*i.e.,* the standard window length in sleep recording analysis). Recordings for which more than 25% of the windows had a variance above 1,000 μV^2 were rejected. Finally, the recording with the highest number of good windows was kept for each user. This yielded a total of 312 overnight sleep recordings from 312 unique users.

Next, we searched for meditation recordings performed by the subjects selected above. Meditation recordings that were recorded up to 24h before or after the beginning of a sleep recording were selected. This yielded a subset of 98 meditation recordings (mean duration: 14.9 ± 9.1 minutes) with good signal quality from 98 unique individuals, collected between October 2020 and October 2021. Of these, nine recordings were from subjects who also had different recordings in MMD, and one recording was already contained in MMD (and was therefore removed). These meditation recordings occurred, on average, 5.4 hours before (63.3% of recordings) or 13.9 hours after (36.7%) the closest sleep recording. The matching sleep recordings had a mean duration of 488.8 ± 58.6 minutes. Mean age across recordings was 45.1 ± 13.6 years (min: 20, max: 74) and 23.5% of recordings were of female users.

Table 4.1 – Description of the datasets used in this study.

	MMD	**OMSD**		**MMD-long**
Recording type	Meditation	Meditation	Sleep	Meditation
# recordings	971	98	98	497
# subjects	971	98	98	4
Average duration (min)	15.1 ± 6.9	14.9 ± 9.1	488.8 ± 58.6	23.1 ± 8.7
Age range	18-81	20-74	20-74	-

Longitudinal Muse Meditation Dataset (MMD-long)

Finally, we explored the long-term dynamics of the predicted brain age measure using another subset of Muse meditation recordings, this time focusing on users with multiple consecutive recordings. The search criteria used to sample MMD recordings were reused to identify users with at least 100 meditation recordings. After filtering recordings for excellent signal quality following the same procedure as was used for MMD, we selected four subjects (three male, one female) of different age groups.[7] In total, this yielded 497 recordings (mean duration: 23.1 ± 8.7 minutes).

4.2.3 Extraction of sleep EEG biomarkers

To assess whether a low-cost mobile EEG-based brain age metric could be used as a proxy measure of health, we need to demonstrate that the brain age predictions contain information about aging that is not already found in the chronological age. To do so, we extracted a variety of biomarkers from the overnight sleep recordings of OMSD, to which the predicted age on corresponding meditation recordings could be compared. These biomarkers, which we describe in detail in this Section (see Table 4.2), are widely accepted indicators of health and aging in the sleep research literature (Purcell et al., 2017; Muehlroth and Werkle-Bergner, 2020) and as such provide us with a data-driven view on the neurophysiological changes that occur with aging.

We grouped the extracted sleep biomarkers into four categories: sleep spindle characteristics, slow wave characteristics, spectral features and macrostructure-related features. The different features are described in Table 4.2. In order to extract these features, we first obtained sleep stage predictions using a sleep staging classifier trained on labelled Muse S data with classes W (wake), N1, N2, N3 (non-REM sleep) and R (rapid eye movement sleep). This also required careful handling of bad windows to avoid introducing noise in our measures. We next provide a detailed description of these sleep staging and signal quality estimation procedures, followed by a description of the different biomarker categories we included in our analysis.

Automatic sleep staging

A convolutional and recurrent neural network architecture inspired by Abou Jaoude et al. (2020) was trained on private Muse S sleep recordings annotated according to the AASM guidelines (Berry et al., 2012) by a sleep technician (see Section 3.3.4 for a description of the dataset). In contrast to the original architecture, the model 1) used unidirectional, rather than bidirectional, LSTM layers, to allow its use in a real-time processing context and 2) had different input layer and maxpooling kernel sizes,

[7] We do not disclose these subjects' ages in order to minimize identifiable information.

in order to accommodate signals sampled at 256 Hz (vs. 200 Hz in Abou Jaoude et al. (2020)). The model was trained to predict which sleep stage (W, N1, N2, N3 or R) a 30-s EEG window corresponded to. Moreover, the data was preprocessed in a similar manner to Section 4.2.4: 1) linear interpolation of missing values, 2) downsampling to 128 Hz, 3) bandpass filtering between 1 and 40 Hz, and 4) channel-wise zero-meaning of each window. The trained model was deployed as-is on the sleep recordings of OMSD to produce hypnograms, *i.e.,* sequences of overnight sleep stage predictions, with 30-s resolution.

Signal quality estimation

In order to extract sleep spindle, slow wave and spectral feature biomarkers, bad windows must be rejected, as noise can heavily impact parameter estimation for these biomarkers. To do so, we used an approach heavily inspired by Muehlroth and Werkle-Bergner (2020), where we divided each recording into non-overlapping 1-second windows and computed the peak-to-peak amplitude of each channel in each window. Channels inside a window were marked as bad if their peak-to-peak amplitude was below 3 μV or above 250 μV. Additionally, entire 1-s windows were marked as bad if, for at least one of their channels, the absolute z-scored peak-to-peak amplitude was higher than 2.5. Finally, the longer time windows used to extract the other biomarkers were dropped if they overlapped with too many 1-s bad windows (the exact number varies between the biomarkers and is described below).

Sleep spindle characteristics

Sleep spindles are phasic waves between 11 and 16 Hz with a characteristic increasing-then-decreasing morphology that typically last between 0.5 and 2 seconds (De Gennaro and Ferrara, 2003; Purcell et al., 2017; Muehlroth and Werkle-Bergner, 2020). Being a highly recognizable sleep microstructure event, sleep spindles are the main feature used to score N2 sleep according to the AASM guidelines (Berry et al., 2012). Sleep spindles are known to originate from thalamocortical sources, although their function remains unclear (De Gennaro and Ferrara, 2003; Purcell et al., 2017). Two distinct types of sleep spindles are now commonly recognized: slow spindles (between 9 and 12.5 Hz, prevalent in frontal regions) and fast sleep spindles (between 12.5 and 16 Hz, prevalent in parieto-central regions) (Muehlroth and Werkle-Bergner, 2020). While they share a common generating mechanism, these two spindle types have distinct generators in the brain and have been shown to be genetically uncorrelated (De Gennaro and Ferrara, 2003; Purcell et al., 2017; Muehlroth and Werkle-Bergner, 2020).

The characteristics of sleep spindles are known to vary with age in multiple ways which are thought to reflect the maturation of thalamocortical mechanisms (Purcell et al., 2017). For instance, their amplitude decreases with age (De Gennaro and Ferrara, 2003; Purcell et al., 2017; Muehlroth and Werkle-Bergner, 2020). Their frequency has been shown to vary differently with age for slow and fast spindles, with fast spindles seeing an increase in frequency (Muehlroth and Werkle-Bergner, 2020) and a plateau around adulthood (Purcell et al., 2017), and slow spindles seeing a decrease in frequency (Muehlroth and Werkle-Bergner, 2020). The spatial distribution of sleep spindles also varies in a frequency-dependent manner: while slow spindles remain frontal throughout aging, fast spindles move posterially as people age (Muehlroth and Werkle-Bergner, 2020). Finally, spindle density, *i.e.,* the number of spindles that occur per unit time, increases with age until adolescence and then decreases (Purcell et al., 2017). Interest-

Table 4.2 – Description of the extracted sleep biomarkers.

Category	Name	Description	References
Spindles	Slow spindle frequency	Average slow spindle frequency (Hz) during N2 sleep	(De Gennaro and Ferrara, 2003; Purcell et al., 2017; Muehlroth and Werkle-Bergner, 2020)
	Fast spindle frequency	Average fast spindle frequency (Hz) during N2 sleep	(De Gennaro and Ferrara, 2003; Purcell et al., 2017; Muehlroth and Werkle-Bergner, 2020)
Slow waves	Slow wave frequency	Average slow wave frequency (Hz) during N3 sleep	(Carrier et al., 2011; Schwarz et al., 2017; Ujma et al., 2019; Timofeev et al., 2020; Muehlroth and Werkle-Bergner, 2020)
Spectral	N3 δ log-power	Log-power of the EEG in the δ band (0.5-4.5 Hz) during N3 sleep	(Mourtazaev et al., 1995; Muehlroth and Werkle-Bergner, 2020)
Macro-structure	Wake after sleep onset (WASO)	Total awake time between the first and last sleep windows, in minutes	(Sun et al., 2019; Muehlroth and Werkle-Bergner, 2020)
	% N1	Percentage of total sleep time spent in N1	(Schwarz et al., 2017; Sun et al., 2019; Muehlroth and Werkle-Bergner, 2020)
	% N2	Percentage of total sleep time spent in N2	(Schwarz et al., 2017; Sun et al., 2019; Muehlroth and Werkle-Bergner, 2020)
	% N3	Percentage of total sleep time spent in N3	(Mourtazaev et al., 1995; Schwarz et al., 2017; Sun et al., 2019; Muehlroth and Werkle-Bergner, 2020)

ingly, multiple other variables appear to influence sleep spindle characteristics, *e.g.,* the circadian rhythm (De Gennaro and Ferrara, 2003), intelligence and cognitive abilities (Ujma et al., 2014; Hoedlmoser et al., 2014) and dementia (D'Atri et al., 2021). We focused on slow and fast spindle frequencies for the analyses in Section 4.3.2, as this measure is more robust to higher proportions of bad windows (*e.g.,* as compared to spindle density measures).

Sleep spindles were extracted using the open-source `YASA` toolbox (Vallat and Walker, 2021). First, the `spindles_detect()` function was used on the raw data, separately focusing on the slow spindles (9-12.5 Hz) or fast spindles (12.5-16 Hz) frequency ranges (Muehlroth and Werkle-Bergner, 2020). Spindles lasting between 0.5 and 2.5 seconds were detected in the N2 windows of each recording using a combination of thresholds on 1) the relative σ power (11-16 Hz), 2) Pearson's correlation between the σ-filtered signal and the raw signal, and 3) the root mean square (RMS) of the σ-filtered signal. This yielded a list of potential channel-specific sleep spindle events, with multiple parameters for each detected event, including median instantaneous frequency, peak-to-peak amp-

litude, and duration. As explained above, we only considered the frequency parameter in our experiments. A spindle event was rejected if it overlapped with a 1-s window that was previously marked as bad. Finally, all values that passed signal quality evaluation were aggregated channel-wise using a trimmed mean with a cut proportion of 25%, yielding one frequency value per recording per channel for both slow and fast spindles.

Slow wave characteristics

Slow waves are low frequency biphasic waves consisting of a positive and a negative deflection (Timofeev et al., 2020). Following to the AASM guidelines (Berry et al., 2012), slow waves are detected using an amplitude threshold of 75 μV on oscillations in the 0.5-2 Hz range. As the name suggests, slow waves are the defining characteristic of slow wave sleep: a 30-s window will be labelled as "N3" if it is made out of more than 20% slow waves (Malhotra and Avidan, 2013). Slow waves are known to play a critical role in memory consolidation (Bellesi et al., 2014) although the mechanism by which they are involved is still debated (Léger et al., 2018).

Multiple characteristics of slow waves are known to evolve with age (Carrier et al., 2011; Ujma et al., 2019). For instance, a decrease in slow wave amplitude is a consistent finding across sleep and aging studies (Schwarz et al., 2017; Muehlroth and Werkle-Bergner, 2020). Similarly, the density of slow waves, *i.e.,* the number of occurrences per unit of time, is known to decrease with age (Carrier et al., 2011; Muehlroth and Werkle-Bergner, 2020). An overall slowing of low frequency components with age has also been reported (Carrier et al., 2011; Muehlroth and Werkle-Bergner, 2020). Finally, the spatial distribution of slow wave activity changes with age as well (Timofeev et al., 2020; Muehlroth and Werkle-Bergner, 2020). We focused our analysis on the slow wave frequency, as this value is less sensitive to the confounding effect of fixed amplitude thresholds that might bias slow wave detection in different age groups (Muehlroth and Werkle-Bergner, 2020). Amplitude information was instead captured using a spectral power analysis, as described in the next section.

Slow waves were also extracted using the `YASA` toolbox (Vallat and Walker, 2021). The default frequency range of the `sw_detect()` function was kept (0.3-1.5 Hz). Internally, the raw data was first bandpass filtered in the 0.3-2 Hz range. Peak detection was then applied to find negative peaks in a given range. A peak-to-peak amplitude threshold, as well as a duration criteria were finally applied to identify potential slow wave events in the N3 windows of each recording. Again, multiple parameters per event were computed, however we only considered frequency in our analyses. A slow wave event was rejected if it overlapped with more than one 1-s window that had previously been marked as bad. Finally, all values that passed signal quality evaluation were aggregated channel-wise using a trimmed mean with a cut proportion of 25%, yielding one frequency value per recording per channel.

Spectral features

To mitigate the confounding effect of fixed amplitude thresholding on slow wave detection, it has been proposed that one should instead rely on slow-wave activity (SWA), *i.e.,* the spectral power in the δ band (0.5-4.5 Hz), to study slow-wave-related changes in aging (Muehlroth and Werkle-Bergner, 2020). With this in mind, and given its simplicity, and the wide use of spectral band powers as features in EEG-based machine

learning pipelines, we included the log-power in the δ band during N3 sleep in our analyses.

Frequency band power was computed window-wise using the `mne-features` package (Schiratti et al., 2018). Welch's method was applied to 30-s windows, using 1-s non-overlapping Hamming windows. Channel-wise power in the δ frequency bands (0.5-4.5 Hz) was then averaged and log-transformed for each window labelled as N3, yielding one value per window and channel. A window-channel combination was rejected if it overlapped with more than nine 1-s sub-windows marked as bad by the signal quality estimation procedure described above. Finally, all values that passed signal quality evaluation were aggregated channel-wise using a trimmed mean with a cut proportion of 25%, yielding one value per recording per channel.

Macrostructure-derived features

Sleep macrostructure, *i.e.,* the overall structure of the different sleep stages across the night, is also known to vary with age, though not as much as the microstructure components described above (Schwarz et al., 2017; Sun et al., 2019; Muehlroth and Werkle-Bergner, 2020). Indeed, the capacity of the brain to generate sleep, as well as the prevalence of each sleep stage change throughout the lifetime. For instance, the percentages of N1 and N2 stages across an overnight recording tend to increase with age, with the percentage of N3 tending to decrease with age (Schwarz et al., 2017; Muehlroth and Werkle-Bergner, 2020). Similarly, the time spent awake between the first and last sleep event of a night, called "wake after sleep onset" (WASO), is also known to increase with age (Muehlroth and Werkle-Bergner, 2020). We included these four measures (%N1, %N2, %N3 and WASO) as biomarkers for the analysis of the predicted brain age.

In order to compute these features, we directly processed the hypnograms obtained by the automatic sleep stager. The ratio of each non-REM sleep stage was computed by dividing the number of windows of a given sleep stage by the total number of sleep windows (*i.e.,* all labels except W). The "wake after sleep onset" (WASO) metric was computed by evaluating the total time (in minutes) spent awake between the first and last sleep events of a night. Of note, the extraction of the other sleep biomarkers relied on the predicted hypnogram as well, so as to enable focusing on specific sleep stages, *i.e.,* N2 for sleep spindles, and N3 for spectral features and slow waves.

4.2.4 Preprocessing

EEG data was minimally preprocessed before being passed to brain age prediction models. In the case of meditation recordings, the meditation exercise part of the recording was cropped to remove the first minute of the recording (in which signal quality might still be settling), and retain as many as eight of the following minutes of data. Next, missing values (which can occur if the wireless connection is weak and Bluetooth packets are lost) were replaced through linear interpolation using surrounding valid samples.

The remaining steps differed for filterbank models, where the preprocessing was inspired by Engemann et al. (2021), and for deep learning models, where the preprocessing was instead similar to Section 2.2.7. For filterbank models, a zero-phase FIR band-pass filter between 0.1 and 49 Hz was applied. Non-overlapping 10-s windows were then extracted. Windows for which peak-to-peak amplitude exceeded a value of 250 μV were rejected. For deep learning models, a zero-phase FIR low-pass filter with a cutoff

frequency of 40 Hz was instead applied to the data, followed by resampling to 128 Hz. Non-overlapping 15-s windows were extracted, and no rejection criterion was applied. Finally, channels were zero-meaned window-wise before being fed to the neural networks.

Sleep recordings were preprocessed within the sleep biomarker extraction pipelines which implemented specific signal quality and filtering steps for each biomarker category (see Section 4.2.3).

4.2.5 Machine learning pipelines

To predict age from mobile EEG, we compared two common methods that have been shown to perform well on different EEG-based learning tasks: using filter-bank covariance matrices with log-diagonal vectorization or with Riemannian tangent space projection (Barachant et al., 2013b; Congedo et al., 2017; Lotte et al., 2018; Sabbagh et al., 2020) and using end-to-end deep learning on raw signals (Roy et al., 2019a).

The filter-bank covariance pipeline first applied a filter bank to the input EEG, yielding narrow-band signals in the nine following bands: low frequencies (0.1-1 Hz), δ (1-4 Hz), θ (4-8 Hz), α_{low} (8-10 Hz), α_{mid} (10-12 Hz), α_{high} (12-15 Hz), β_{low} (15-26 Hz), β_{mid} (26-35 Hz) and β_{high} (35-49 Hz). Next, covariance matrices were estimated in each frequency band from non-overlapping 10-s window using the OAS algorithm (Chen et al., 2010), yielding a set of nine covariance matrices per recording. For the "diag" variation, the log-diagonal of each covariance matrix was extracted and concatenated into a single feature vector of dimension $C \times 9$, z-score normalized using the mean and standard deviation of the training set, then fed to a linear regression model with an L2 penalty ("Ridge regression"). For the "Riemann" variation, covariance matrices were instead projected into their Riemannian tangent space, exploiting the Wasserstein distance to estimate the mean covariance used as the reference point (Sabbagh et al., 2019; Bhatia et al., 2018). The vectorized covariance matrices with dimensionality $C(C+1)/2$ were finally z-score normalized, concatenated, and fed to a linear regression model with an L2 penalty, just as for the "diag" model. Of note, the Ridge regression models were trained by optimizing the mean squared error instead of the mean absolute error, following previous work (Sun et al., 2019; Engemann et al., 2020; Sabbagh et al., 2020)

The deep learning pipelines used the ShallowNet base architecture (Schirrmeister et al., 2017) which parametrizes the FBCSP pipeline (Gemein et al., 2020). We first used it without modifying the architecture, yielding a total of 11,641 trainable parameters with $C = 4$. We then tested a variation of ShallowNet, prepended with a DSF attention module, as presented in Chapter 3. We set the number of virtual channels to the number of input channels, *i.e.*, $C' = C$, fixed the hidden layer size to $C^2 = 16$ and used the matrix logarithm of the covariance matrix as the input spatial representation. We used a soft-thresholding non-linearity on the output spatial filtering matrix, since it was shown in Chapter 3 to provide an additional performance boost on real-world Muse S data. The DSF module added 516 trainable parameters to those of ShallowNet. Deep learning models were trained on individual 15-s windows, but their performance was evaluated recording-wise by averaging the predictions over windows within each recording. Finally, an element-wise logistic non-linearity was added to the output of the network to facilitate the prediction of values that fell within a plausible range of age:

$$\sigma(x) = a_{low} + \frac{a_{high} - a_{low}}{1 + e^{-rx}} + bx \ , \tag{4.3}$$

where $a_{high} = 0$ and $a_{low} = 98$ are upper and lower asymptotes, $r = 0.1$ is the growth rate, and $b = 0.1$ is the slope of an additional linear trend added to avoid getting stuck in regions with zero-gradient during training.

4.2.6 Training and performance evaluation

Models were first trained using 10-fold cross-validation. The training folds were further split to perform a hyperparameter search (leave-one-out cross validation for selecting the L2 regularization strength[8] in the filterbank models) or early stopping (80-20% random split for deep learning models). The examples from each subject were always restricted to only one of the training, validation or testing sets.

Neural networks were trained three times per split with different random parameter initializations. Training ran for at most 70 epochs or until the validation loss stopped decreasing for a period of a least 20 epochs. In all experiments, we used the AdamW optimizer (Loshchilov and Hutter, 2017) with $\beta_1 = 0.9$, $\beta_2 = 0.999$, a learning rate of 5×10^{-4} and cosine annealing. The parameters of all neural networks were randomly initialized using uniform He initialization (He et al., 2015). Dropout was applied to $f_{\boldsymbol{\Theta}}$'s fully connected layers at a rate of 50%. Training hyperparameters for deep learning models were selected such that learning curves decreased steadily in the first 10 epochs of training on a subset of a training fold of MMD.

Predictions on the test folds were then used to evaluate the performance of the different models, as measured with MAE (see Equation 4.2) and the coefficient of determination R^2, *i.e.,* the percentage of chronological age variance explained by the predicted brain age:

$$R^2(\boldsymbol{y}, \hat{\boldsymbol{y}}) = 1 - \frac{\sum_{i=1}^{N} \left(y^{(i)} - \hat{y}^{(i)}\right)^2}{\sum_{i=1}^{N} \left(y^{(i)} - \overline{y}\right)^2} \ , \tag{4.4}$$

where $\overline{y} = \frac{1}{N} \sum_{i=1}^{N} y^{(i)}$.

Finally, models were retrained using the entire MMD (with the same validation set splitting strategies) and evaluated on the meditation recordings of OMSD and MMD-long.

4.2.7 Analysis of predicted brain age

We studied the added value of predicted brain age for modelling health and aging through the analysis of linear regression coefficients (Dadi et al., 2021). First, we modelled each of the 20 (z-score normalized) extracted sleep biomarkers $y_j \in \mathbb{R}$, where $j \in [\![20]\!]$, using independent univariate linear regression models of the form

$$y_j \approx \beta_{j,k}^{\text{univar}} \times x_k \ , \tag{4.5}$$

where $k \in \{\text{age}, \text{predicted age}\}$ such that $x_k \in \mathbb{R}$ is either the chronological age (x_{age}) or predicted age ($x_{\text{predicted age}}$) normalized to have zero-mean and unit-standard deviation. Once the models were fit, the sign and magnitude of the estimated coefficients $\hat{\beta}_{j,k}^{\text{univar}}$ indicated the direction and strength of the relationship between each of the two age measures and the sleep biomarkers.

[8] We searched through 100 values spaced log-uniformly between 10^{-5} and 10^{10}.

Next, we modelled each sleep biomarker using both age measures at once using bivariate linear regression models:

$$y_j \approx \beta^{\text{bivar}}_{j,\text{age}} \times x_{\text{age}} + \beta^{\text{bivar}}_{j,\text{predicted age}} \times x_{\text{predicted age}} \ . \tag{4.6}$$

Again, the estimated coefficients $\hat{\beta}^{\text{bivar}}_{j,k}$ reflected the explanatory power of each predictor, however this time taking both predictors into account simultaneously. By comparing the goodness of fit (*i.e.,* the likelihood) of the univariate and bivariate models of the same sleep biomarker, we could then estimate how much additional explanatory power is obtained by adding brain age to chronological age. This can be quantified using a likelihood ratio test where the test statistic is computed as:

$$\lambda^{\text{LR}}_j = -2\left[\ell(\beta^{\text{univar}}_{j,\text{age}}, 0) - \ell(\beta^{\text{bivar}}_{j,\text{age}}, \beta^{\text{bivar}}_{j,\text{predicted age}})\right] \ , \tag{4.7}$$

where $\ell(\beta_{\text{age}}, \beta_{\text{predicted age}})$ is the log-likelihood of the model. Under the null hypothesis $H_0 : \beta^{\text{bivar}}_{j,\text{predicted age}} = 0$, λ^{LR}_j then follows a χ^2 distribution with $K = 1$ degree of freedom. As a result, p-values can be obtained for each biomarker j. After applying Bonferroni corrections to control the family-wise error rate (FWER), biomarkers with a significant p-value (*e.g.,* with $\alpha = 0.01$) were finally identified. Biomarkers for which the null hypothesis could be rejected are interpreted as being better explained when brain age is included in the model.

4.2.8 Computational considerations

A combination of the `MNE-Python` (Gramfort et al., 2014), `PyTorch` (Paszke et al., 2019), `braindecode` (Schirrmeister et al., 2017), `pyRiemann` (Barachant et al., 2013b), `coffeine` (Sabbagh et al., 2020), `mne-features` (Schiratti et al., 2018), `scikit-learn` (Pedregosa et al., 2011), `statsmodels` (Seabold and Perktold, 2010), `YASA` (Vallat and Walker, 2021), `mne-bids` (Appelhoff et al., 2019) and `mne-bids-pipeline` (Jas et al., 2018) packages were used to carry out our experiments. Deep learning models were trained on a machine with a single Nvidia P4 GPU.

4.3 Results

4.3.1 Predicting age from low-cost mobile EEG

Can low-cost mobile EEG be used to predict someone's age, despite sparse spatial information and recordings made in uncontrolled environments? If so, what kind of machine learning approach is the most accurate? To answer these questions, we trained four brain age prediction pipelines and compared their performance on a dataset of 971 recordings of unique users. Results are shown in Figure 4.2.

All four models outperformed a dummy regressor that predicted the median age of the training set (MAE=14.02, R^2=-0.04). Predictably, the model based on simple log-powers ("diag", in yellow) performed the worst out of the considered models, likely due to its limited ability to leverage spatial information. The Riemann model (green), which had access to full covariance information in each frequency band, performed better, with a mean MAE across folds of 11.33 years, and a mean R^2 of 0.23. Deep learning models (red and magenta) yielded better mean MAE and R^2 than the Riemann model, with a combination of ShallowNet and DSF yielding the best performance (MAE=10.99, R^2=0.31). Interestingly, although MAE values were similar for Riemann and deep

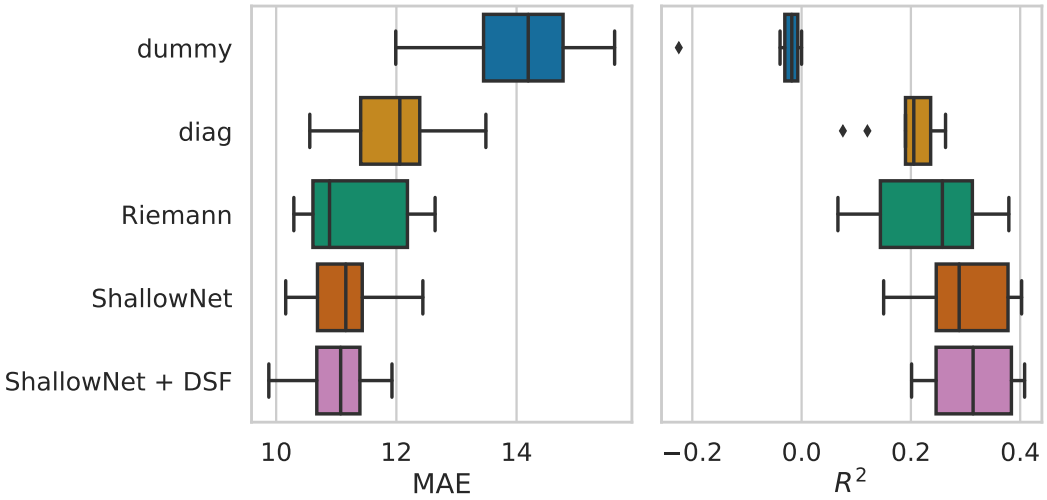


Figure 4.2 – Brain age prediction performance on MMD. The mean absolute error (MAE) and ratio of explained variance (R^2) obtained across 10 cross-validation folds are shown for different models: simply returning the median of the training set ("dummy"), Ridge regression on log-power in different frequency bands ("diag"), Riemannian geometry-based pipeline on filter-bank covariance matrices ("Riemann"), and deep learning models ("ShallowNet" and "ShallowNet-DSF" with dynamic spatial filtering). While all models performed better than a dummy regressor, the combination of ShallowNet with DSF yielded the lowest mean MAE and highest mean R^2.

learning models, R^2 values were a lot more spread out for the Riemann model across cross-validation folds, likely due in part to the use of the sigmoidal non-linearity in deep learning models which restricts their output range.

Overall, these results suggest that it is indeed possible to predict age from low-cost mobile EEG, with performance significantly above that of a dummy regressor.

4.3.2 Brain age as a complementary source of information to chronological age

To be useful as a health biomarker, a brain age measure needs to encode information that is complementary to the chronological age itself, *i.e.,* which is not available when only the chronological age is known. If this is the case, brain age could be a valuable tool, in combination with chronological age, for quickly screening individuals for pathological aging based on only a few minutes of their EEG. To test this, we used linear regression to model known sleep biomarkers (see Section 4.2.3 for a detailed description of the analysis) based on chronological age, predicted brain age, or a combination of both. Results of this analysis are presented in Figure 4.3.

First, as shown through univariate analysis (first column), most models yielded non-zero coefficients for both chronological and brain age variables. Exceptions include *e.g.,* slow wave frequency and percentage of N2 sleep, for which the two variables were not useful predictors. Interestingly, chronological age was generally more correlated with the biomarkers than predicted age was. An exception to this however is N3 δ log-power, for which chronological age was a bad predictor (only one non-zero coefficient),

as opposed to predicted age (three non-zero coefficients). Overall, as expected (see Section 4.2.3), most sleep biomarkers were significantly correlated with both measures independently.

Next, we combined both chronological and brain age within bivariate models to explore their associations with sleep biomarkers (Figure 4.3, second column). Again, each biomarker category (N3 δ log-power, fast spindle frequency, slow spindle frequency, slow wave frequency and hypnogram-derived markers) had biomarkers for which both chronological and brain age had non-zero coefficients. Moreover, as shown with the p-value of likelihood ratio tests (third column), most of these biomarkers were accordingly easier to model once predicted age was included in the linear regression. Notably, models of channel-wise N3 δ log-power had larger coefficients for both variables (as compared to their univariate versions), and all had significantly better fits once predicted age was included in the model. This demonstrates that the two age measures contain complementary information that helps model a known biomarker of sleep and aging, *i.e.,* δ power or slow-wave activity. Similarly, the fast spindle frequency was significantly easier to model (for all channels except TP10) when brain age was made available to the linear regression. Interestingly, for other biomarkers, *e.g.,* slow spindle frequency and percentage of N3 sleep, the sign of the predicted age coefficient flipped when going from a univariate to a bivariate model. This suggests that the residual information contained in the brain age is actually anticorrelated to the information contained in the chronological age, or in other words that the correlation between the two is negative conditional on age but positive otherwise.

These results suggest that the proposed brain age metric - which is predicted from short five to eight-minute segments of non-sleep data - is meaningfully correlated to known biomarkers of sleep and aging and contains additional information not contained in chronological age, supporting the potential use of brain age derived from mobile EEG as a proxy measure of health.

4.3.3 Variability of brain age over multiple days

To confirm the usability of our mobile EEG-based brain age metric, we must consider its variability over medium- to long-term periods. Indeed, stable values over weeks and months would indicate actual trait-like information on the subject is being captured. On the other hand, significant variability at smaller scales would suggest the metric is also influenced by momentary factors and therefore captures "brain states" as well. To answer these questions, we computed the recording-by-recording brain age Δ of four subjects with a large number of recordings and looked at the characteristics of their predicted values. The results of this analysis are presented in Figure 4.4.

First, while there is substantial variability across recordings from a same subject, the predicted values remain fairly stable across longer-term periods. Indeed, despite across-recording standard deviations of 2.7 to 4.6 years, the average predictions over longer periods (*e.g.,* months) do not vary substantially. This is seen in both a subject whose chronological and brain age match closely (sub-002) and in one for whom there is a large difference between the two measures (sub-004). Interestingly, sub-001 saw a short-lived decrease in their brain age Δ about 100 days after their first recording, but otherwise had stable predictions on average.

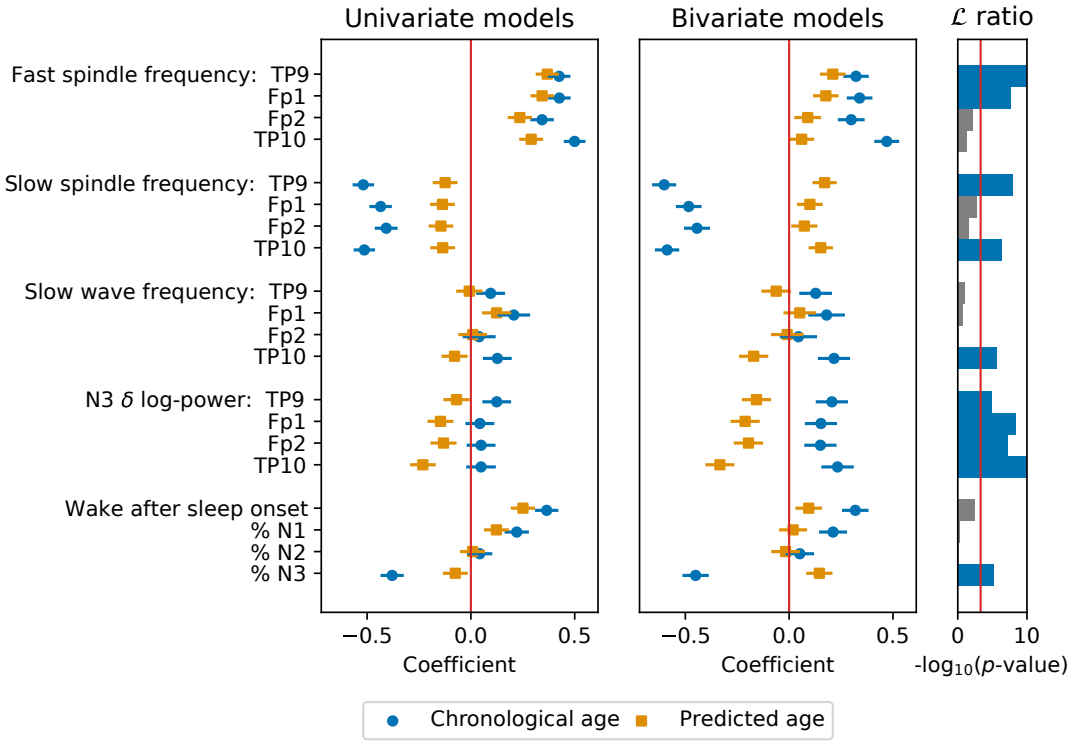


Figure 4.3 – Analysis of the relationship between sleep biomarkers and age measures. Linear regression was used to model each of 20 sleep biomarkers using age, predicted brain age, or a combination of both. In the first column, the coefficients of univariate models that were trained either on chronological age (blue) or on predicted age (orange) are shown for each biomarker. In the second column, the coefficients of a single bivariate model trained on both chronological and predicted age are shown. In both cases, 95% confidence intervals are depicted by horizontal bars around markers. Finally, the third column shows the negative log p-value of the likelihood ratio test comparing a univariate model with chronological age only, and a bivariate model with both chronological and predicted age. The vertical red line indicates a threshold of $p = 0.01$ with Bonferroni correction, and grey bars indicate biomarkers for which the null hypothesis cannot be rejected. While the coefficients of chronological age are generally larger than those of predicted age, multiple sleep biomarkers are significantly easier to model once predicted age is made available to the linear regression, suggesting brain age contains information that is not found in the chronological age alone.

Next, to understand what might explain this variability, we looked at different factors that are likely to have an influence on our EEG-based predicted age. Specifically, we analyzed the relationship between predicted age and time of day, on the larger population of the MMD dataset (Figure 4.5). A multiple linear regression model with continuous factor "time of day" and categorical factor "gender", along with a "time of day × gender" interaction term was fitted to the cross-validated predictions obtained on MMD. The coefficients of the model were not significantly different from zero, suggesting neither variable is driving the variability seen in our results. Therefore, while brain age Δ is on average slightly higher earlier in the day, the relationship with time of day is not significant. Likewise, our analysis shows that gender does not explain the observed variability either. Therefore, other factors not readily available in our datasets are likely the cause of the variability seen in the brain age Δ over time.

Moreover, thanks to the availability of longitudinal recordings, "brain aging" (*i.e.,* the rate of change of the brain age Δ) can be estimated by fitting a linear model to the brain age Δ values (red lines in Figure 4.4). Here, a positive slope indicates a person's brain age is increasing faster than their chronological age, while a negative slope indicates that they are "aging" more slowly than what is expected biologically. While subjects 1 and 3 had positive slopes, sub-002's slope was close to zero, and sub-003's slope was in fact negative. More recordings over longer time periods are needed to enable further investigation into the meaning of these slopes for brain health.

Overall, these last results suggest that our mobile EEG-based brain age metric captures both "trait"-like information (*e.g.,* related to aging and, potentially, pathological aging), and shorter-term "state"-like information, which may reflect a subject's state at the time of the recording, and be related to factors like *e.g.,* cognitive fatigue, food and drink intake, medication, etc.

4.4 Discussion

In this chapter, we showed that low-cost mobile EEG-based brain age predictions can be used as a proxy measure of aging. We presented results on over 1,500 at-home EEG recordings from 1,073 unique individuals, combining real-world data collected during meditation recordings and overnight sleep. While methods based on filter-bank covariance matrices already performed better than chance, we found deep learning methods performed the best on the brain age prediction task. Importantly, our experiments showed that the brain age metric contains information that is not found in the chronological age alone, as evidenced by correlating chronological age and brain age with various commonly studied sleep biomarkers. Finally, we looked at the stability of brain age over more than a year for four individuals and, despite short-term recording-to-recording variability, we showed that brain age predictions were stable across longer-term periods. These results provide a strong foundation for the development of low-cost mobile EEG-based brain age measurement.

4.4.1 What does EEG-based brain age capture?

A majority of studies focused on predicting age from EEG did so using sleep EEG (Sun et al., 2019; Paixao et al., 2020; Ye et al., 2020; Brink-Kjaer et al., 2020; Nygate et al., 2021). Given the large literature on sleep EEG and age, this choice makes a lot of sense: many biomarkers, both on the microstructure (*i.e.,* related to the patterns in the EEG time series) and macrostructure (*i.e.,* related to the structure of the different

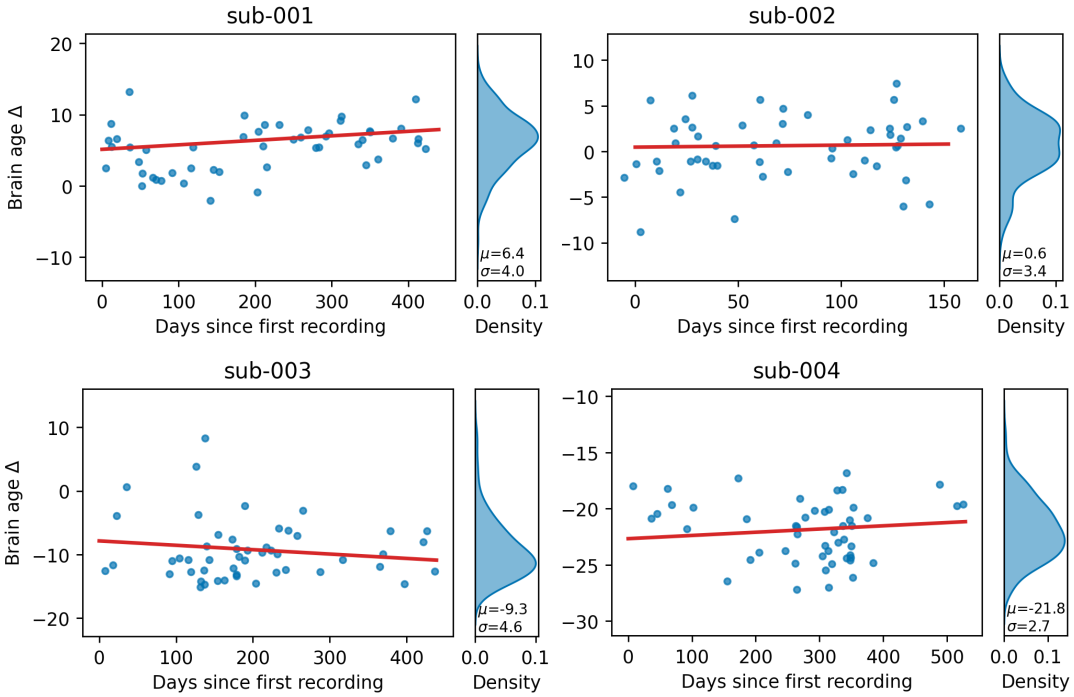


Figure 4.4 – Longitudinal brain age Δ predictions for four subjects with multiple consecutive recordings. Four subjects with more than 50 recordings with good signal quality were selected, and their brain age predicted using a DSFm-st model trained on MMD. Each blue point of the scatterplots represents the brain age Δ predicted for a single recording. In order to ensure the anonymity of the subjects, we only show 50 randomly sampled recordings for each subject and add random jitter to the recording dates. However, we use all available sessions to fit a linear model (red line) which shows the trend for each subject, and to summarize the distribution of predicted age using density plots (in blue, next to each scatterplot). Despite significant variability visible in all four subjects, the average measure remains stable across longer periods of time, suggesting the proposed brain age metric captures both "trait"- and "state"-like information.

sleep stages) timescales are known to vary with age (see Section 4.2.3). We also sought to leverage this well-studied connection in our analysis, but in order to validate, rather than build, our brain age prediction models. This was done to avoid the circularity of building a brain age model and validating the age-related information content of its predictions on the same kind of data, with the same underlying mechanisms. Instead, we used EEG collected during meditation recordings to predict age. Crucially, since our models worked on short (*e.g.,* fewer than 10 minutes) awake recordings rather than hours-long sleep recordings, our approach is a lot more time- and data-efficient than other work (Sun et al., 2019; Nygate et al., 2021).

This elicits the following question: what information is there in awake continuous EEG that enables the prediction of someone's age? Previous work on very similar cross-sectional data but from a descriptive, rather than predictive, point-of-view has shown that the characteristics of the α band vary significantly throughout adulthood (Hashemi et al., 2016). For instance, the α peak frequency decreases with age (a widely reported finding, see *e.g.,* Klimesch (1999); Chiang et al. (2011); Scally et al. (2018); Knyazeva et al. (2018); Tröndle et al. (2021)). Additionally, a small but significant increase in α

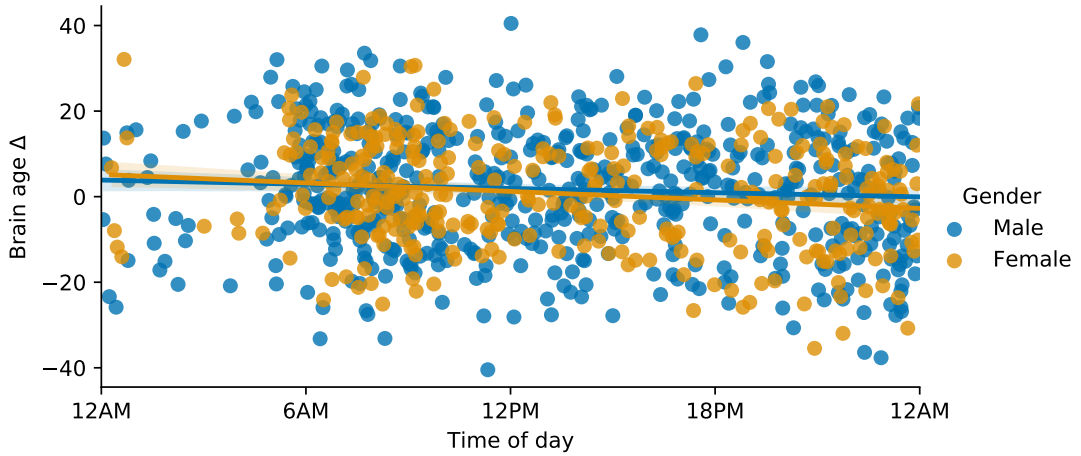


Figure 4.5 – Effects of time of day and gender on brain age Δ. The cross-validated age predictions obtained on MMD are shown as a function of the time at which recordings were started. Subject gender is color-coded (blue for male, yellow for female). Linear fits display the gender-specific impact of time of day on the brain age metric. While a small negative trend is visible for both groups, it is not significant, suggesting these two variables do not have an impact on brain age Δ in this dataset.

and β power with age has also been reported in conjunction with a decrease in δ and θ power (Hashemi et al., 2016). As expected, and as shown by a permutation importance analysis of the filter-bank Riemann model (Figure 4.6), our model did in fact rely heavily on the information contained in the mid- and high-α bands, *i.e.,* between 10-12 and 12-15 Hz. Since ShallowNet is essentially a trained, parametrized version of a filterbank spatial patterns pipeline, it is likely that the same information was important for the deep learning models under study. Visualization of the spectrum around the α band in MMD similarly confirmed that the peak α frequency decreased across age groups (Figure 4.7). In fact, both the α peak frequency, *i.e.,* the frequency of maximum power in the α band, as well as the α power, varied across age groups. Indeed, similarly to Hashemi et al. (2016), power also appeared to increase at this peak frequency in our sample. Therefore, these features were likely leveraged by the model to distinguish age groups in our meditation datasets.

Why, specifically, is the α band useful in predicting age? Age-related decline in α power has previously been linked to increased excitability of thalamo-cortical and cortico-cortical pathways, due to a gradual loss of cholinergic function in the basal forebrain (Tröndle et al., 2021). The more recent "neural noise" hypothesis posits instead that the decrease in α power is caused by the flattening of the 1/f spectral profile of EEG, itself caused by reduced coupling between brain regions (Voytek and Knight, 2015; Tröndle et al., 2021). The α peak frequency is also known to decrease in some pathological conditions, *e.g.,* schizophrenia, Alzheimer's diseases and major depression (Christie et al., 2017). In either case, combined with the results of our own experiments, this supports the idea that by focusing on α band information, our brain age models may indeed be accessing information about the aging of the human brain. Of note, other bands, *e.g.,* β_{low} (15-26 Hz), α_{low} (8-10 Hz) and low frequencies (0.1-1 Hz) also contributed to prediction performance (Figure 4.6), suggesting that more than just mid/high α information was used by the model.

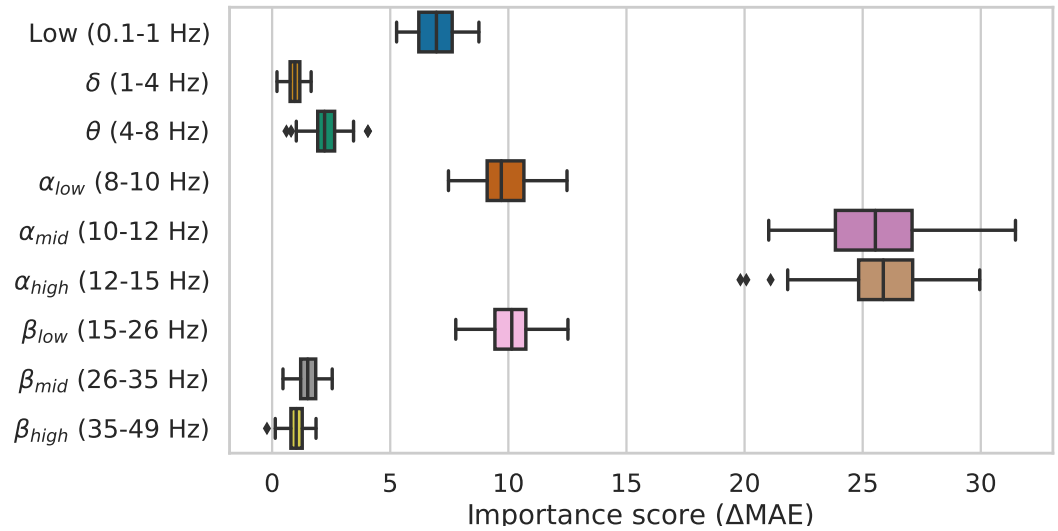


Figure 4.6 – Permutation analysis with the filter-bank Riemann model on OMSD meditation recordings. We focus on the filter-bank Riemann model, since it yielded high performance in our experiments (Section 4.3.1) and its input is already divided into different spectral bands. We evaluated permutation importance (Breiman, 2001) on the meditation recordings of OMSD to identify the features whose random shuffling causes the largest drop in performance. The x-axis indicates by how much the MAE increases when the values in a specific frequency band are randomly permuted (100 repetitions). The features in the α band, and more precisely the α_{mid} (10-12 Hz) and α_{high} (12-15 Hz) bands, were the most useful features for the Riemann model. Other bands also carried some predictive power, although less so: α_{low} (8-10 Hz), β_{low} (15-26 Hz) and low frequencies (0.1-1 Hz).

4.4.2 What causes brain age variability?

Longitudinal experiments on four subjects with more than 100 recordings each showed that while the average brain age Δ predictions were mostly stable across longer periods of time (*e.g.,* months), there was significant variability from one day to the next. What does this variability reflect? The analysis presented in Figure 4.5 showed that time of day, at least cross-sectionally, did not have a significant effect on the predicted age for male or female individuals in our sample.

Following the discussion above, it is clear that the proposed brain age metric is largely influenced by the information contained in the α band. Interestingly, core characteristics of the α band, *i.e.,* its peak frequency, are also known to comprise both a "trait" and a "state" component (Christie et al., 2017). For instance, it has been suggested that a higher peak α frequency is associated with a higher level of "cognitive preparedness", *i.e.,* reflecting how ready an individual is to perform a task (Angelakis et al., 2004; Christie et al., 2017). Research on "mental fatigue" has also revealed that the spectral properties of EEG change as fatigue increases, with, for instance, power in θ, α and β bands increasing (Arnau et al., 2017). Overall, this suggests that cognitive factors might play a role in the observed variability.

An important factor to consider as well is the nature of the exercise undertaken during the recordings used for training the brain age models. In our dataset, subjects were expected to perform a focused attention meditation exercise. Interestingly, multiple

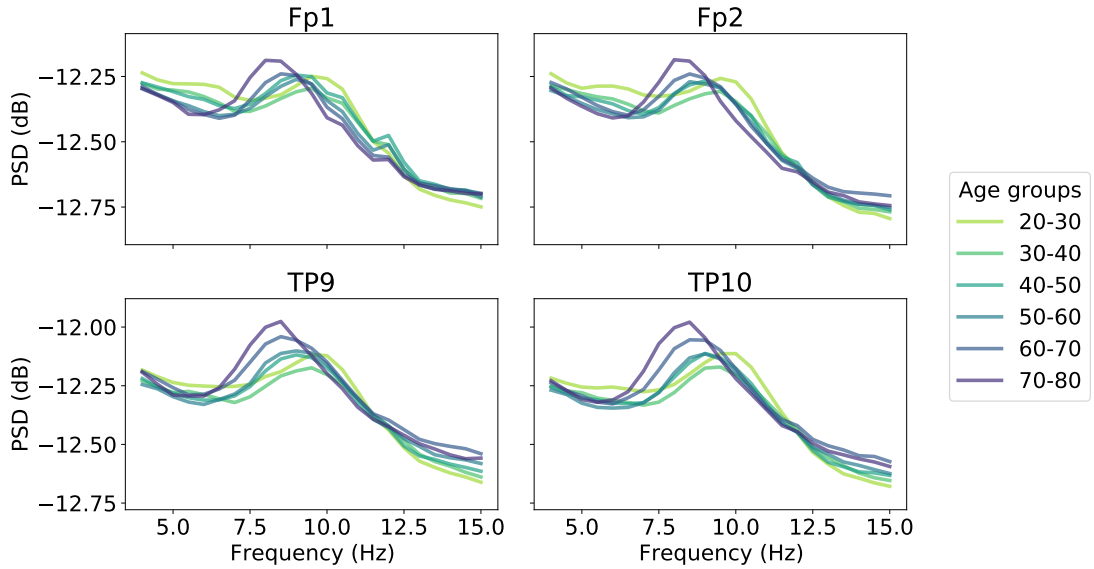


Figure 4.7 – Visualization of the spectrum around the α band for MMD recordings. Welch's periodogram with non-overlapping 4-s segments and median aggregation was computing from the 10-s windows preprocessed for use with filter-bank models. The PSDs were then $\log_{10}$ −transformed and averaged within each 10-year slice of the subjects. Older individuals tended to have a lower α peak frequency, as well as higher α peak power, in all four electrodes.

studies on meditation and EEG reported decreased α peak frequency and increased alpha power, as well as theta bursts, during meditation exercises (West, 1980; Cahn and Polich, 2006; Lomas et al., 2015). Given the relationship between both α peak frequency and α power, and aging (see previous section), this would suggest that meditating may effectively make one's EEG look older than it would at rest. In this scenario, the observed variability might arise from the quality of the meditation or the specific type of meditation exercise carried out during the recording. However, according to previous reports on brain age in meditators, we could actually expect brain age Δ to decrease with long-term meditation practice (Luders et al., 2016). Therefore, the impact of the meditation exercise on the subject's state, as well as of extended meditation practice on their EEG traits, is an important factor to take into account and should be investigated in greater detail in the future.

Of note, one other study looked at EEG-based brain age and its variability across different days (Hogan et al., 2021). On a dataset of 86 patients sleeping overnight in an epilepsy monitoring unit, this study found a night-to-night standard deviation in brain age Δ of 7.5 years. This standard deviation could be further decreased to 3 years when averaging four nights. In comparison, our experiments of Section 4.3.3 showed a standard deviation of 2.7 to 4.6 years when considering four subjects with above 50 recordings each. Additional experiments on more subjects are required to validate these results, however this suggests that similar variability is observed even with longer sleep EEG recordings.

4.4.3 Interpretation of coefficients in the sleep biomarkers analysis

Many biomarkers were significantly easier to model once the brain age predictions were made available to the linear regression in addition to the chronological age (Section 4.3.2). For almost all of these significant biomarkers, the sign of the brain age coefficient $\beta_{\text{predicted age}}$ stayed the same when using univariate or bivariate models (Figure 4.3). This suggests that the correlation between these biomarkers and brain age remains the same, regardless of whether chronological age is considered, as would be expected, intuitively. This was not the cause though for slow spindle frequency and percentage of N3 sleep, whose sign went from negative to positive. What could explain such an effect? A straightforward interpretation is that given a fixed age group, brain age actually varies in the opposite direction to the direction it would normally vary in, were all age groups to be considered. In other words, the residual information contained in the brain age (once the chronological age has been regressed out) is negatively correlated to these biomarkers. For instance, while the slow spindle frequency at temporal locations (TP9 and TP10) decreased when brain age increased, if chronological age was also known, then slow spindle frequencies actually increased with brain age. While the mechanism by which this phenomenon occurs will require more investigation, this further demonstrates that our brain age measure contains information that is not present when only the chronological age is available.

4.4.4 Limitations

We identify three principal limitations to this work. First, the datasets that we used in our experiments were collected in uncontrolled, at-home conditions, *i.e.,* no expert or technician could monitor the subjects to ensure good signal quality and compliance with the recording instructions. This contrasts with existing work on brain age prediction, where datasets are usually collected in clinical or research settings. Similarly, only minimal metadata was available in the datasets. As a result, the datasets did not contain the necessary information to separate healthy individuals (used for training) from pathological individuals (used for testing), as is commonly done in brain age studies. Instead, we leveraged the information contained in sleep data, where well-known biomarkers of health and aging can be extracted, to relate our predictions to aging. Moreover, since the datasets were not collected in the controlled conditions typical of laboratory- or clinical-based studies, there is a possibility that some individuals did not enter their actual age when creating their profile. However, the type of sample size enabled by mobile EEG devices, much larger than traditional studies, means our models will be less sensitive to the occasional incorrect label. For instance, a study using similar datasets produced results that agreed with earlier lab-based studies (Hashemi et al., 2016), supporting the validity of our dataset for brain age analysis.

Second, as discussed above, the interaction between meditation and α activity means it is possible that during meditation, a person's EEG appears to be from an older individual. However, since every subject contained in our dataset was instructed to perform a meditation exercise during the recording, this phenomenon should impact all subjects similarly. Nonetheless, further work will be necessary to assess whether the quality of the meditation, or the meditation experience of a subject, could explain some of the variability in brain age Δ across subjects and across the recordings of individual subjects. Nevertheless, our validation using sleep EEG biomarkers indicates that this effect is small enough that our brain age metric could still provide complementary information towards the prediction of known biomarkers of aging.

Finally, the relationship between chronological age and predicted age was evaluated on recordings that were as much as 24 hours away from each other. This may have drowned out interesting information about shorter-time variations. In a future iteration, this could be avoided by evaluating brain age on the wake portion of sleep recordings instead.

Overall, despite these limitations, our experiments demonstrated that brain age can be predicted from challenging real-world EEG data, providing first baseline results to which future research can be compared.

4.4.5 Future directions

Multiple research directions would be interesting to explore in order to improve upon the results presented in this chapter. First, a much larger sample size could be considered. In this work, we limited the sample size to focus on recordings with high signal quality only, as a first step toward harnessing low-cost mobile EEG for measuring brain age. With tools like DSF (Chapter 3) which were shown to improve robustness to corrupted channels, future attempts should also include any recording with viable EEG. Using larger datasets will likely improve the quality of the models and the stability of predictions over time. Similarly, one could look into averaging multiple recordings to provide a more robust brain age prediction (Hogan et al., 2021). Next, as sleep EEG data is also becoming increasingly available thanks to consumer-focused technology, our results could be extended to model brain age on sleep rather than awake data, such as in Sun et al. (2019); Brink-Kjaer et al. (2020); Nygate et al. (2021). In this case however another approach would be required to validate the information content of the age predictions, *e.g.,* via access to clinical labels (Sun et al., 2019; Nygate et al., 2021). Generally speaking, validation with clinical labels such as the results of cognitive tests will be necessary to advance our understanding of proxy measures of health (Dadi et al., 2021). Finally, deeper neural network architectures with temporal aggregation mechanisms, such as the one proposed in Brink-Kjaer et al. (2020), could be used to improve the modelling, and consequently, the brain age predictions.

4.5 Conclusion

In this chapter, we presented results on brain age prediction with low-cost mobile EEG and showed that this metric provides information that is complementary to chronological age. This is additional evidence that unlabelled or weakly labelled EEG is in fact highly valuable. In all likelihood, expert labels will remain critical to understanding the physiological processes which influence EEG signals, and to building high-performance predictive models. However, our results demonstrate that raw EEG data, for instance as obtained in large-scale real-world applications of EEG, has the potential to play an increasingly important role in supporting and enabling the creation of new EEG-based technologies. As predictive models become increasingly resilient to the difficult signal quality conditions encountered in the real world conditions that such applications are subject to - for instance, through the use of attention modules like DSF - we can begin to envision a new era where the century-old neuroimaging modality that is EEG becomes a widely-adopted personal brain health monitoring tool.

Conclusion

When work on this thesis started, the traditional machine learning paradigm based on feature engineering and shallow modelling overwhelmingly dominated applications of machine learning to EEG time series. As EEG datasets - and neuroimaging datasets in general - are typically characterized by small sample sizes and high input dimensionality, feature-based approaches were an obvious choice and predictably led to high performances on many tasks. However, as demonstrated by our comprehensive literature review on deep learning and EEG (Roy et al., 2019a), the number of studies that had begun to leverage deep learning to improve the performance and flexibility of their models was growing exponentially. In fact, many of these studies found that notoriously data-hungry deep learning models could work as well as, if not better than, traditional machine learning approaches on typical small-scale EEG datasets. Now that much larger datasets are being collected and shared openly by various research consortia (Shafto et al., 2014; Obeid and Picone, 2016; Zhang et al., 2018; Bycroft et al., 2018), we can only expect deep learning to become an even more productive method for modelling EEG data.

At the same time, the availability of new EEG hardware that can be used out-of-the-lab has been growing steadily. Low-cost mobile hardware, such as the Muse headband (InteraXon Inc., Toronto, Canada), already make it possible to monitor EEG on a regular basis and remotely. Because they allow EEG to be monitored virtually anywhere and anytime, and enable a plethora of real world applications, these mobile devices promise to generate an unprecedented volume of data. This data, however, tends to be unlabelled, noisier, and sparser than traditional EEG data.

In this thesis, we sought to demonstrate how novel deep learning methodology can be used to facilitate and enable new large-scale and real-world applications of EEG. Notably, we provided innovative solutions to the problem of extracting information from unlabelled or weakly labelled data. We also showed how neural network modules can be designed to adapt to the particularly challenging noise characteristics of real-world EEG. Finally, we extended the brain age framework, originally limited to research and clinical settings, to real-world contexts.

Previously, most applications of SSL were limited to the fields of computer vision and natural language processing. We adapted and extended this paradigm to learn representations from multivariate EEG time series, leveraging the temporal dependencies in the signal. Our work, which has already inspired studies from other groups, has laid the foundation for translating SSL to biosignal data. Since then, significant advances have been made on both the theoretical (Hyvärinen et al., 2019; Roeder et al., 2021) and applied (Chen et al., 2020a; Brown et al., 2020; Goyal et al., 2021) fronts, providing further evidence that SSL is a powerful framework for learning generalizable representations in an unsupervised manner. Considering these recent advances, we can expect to see the development of improved self-supervised techniques, that learn even more useful representations, *e.g.,* through the data augmentation-invariance framework (Chen et al., 2020a; Rommel et al., 2021). Ultimately, substantial benefits could be derived from learning a common representation space that captures the many intricacies of EEG

signals, analogous to the embeddings widely used in language tasks, and which would enable the transfer of electrophysiological representations across individuals, montages, hardware, and recording paradigms. Additionally, in the future, representation learning techniques such as SSL will likely heavily influence the field of neuroimaging, such that larger, unlabelled datasets, are increasingly preferred over smaller, painstakingly labelled datasets, the latter of which may not contain enough samples to build robust and generalizable models (Jas, 2018).

The last few years have also seen the development of many new deep learning architectures. Notably, the flexibility and representational capacity of neural networks has improved greatly, thanks to the development of attention mechanisms (Bahdanau et al., 2014; Luong et al., 2015; Vaswani et al., 2017; Hu et al., 2018; Woo et al., 2018). In this thesis, we drew on this framework to provide neural networks with robustness to noise, which is a key requirement for practical real-world EEG applications. This approach also opened an interesting window onto the functioning of our deep learning models, allowing us to visualize how useful different EEG channels were to the predictions, on a window-by-window basis. As hardware and sensor technology continues to improve, noise and artifacts may become less of a challenge. Even in this scenario, however, attention modules such as DSF will help improve the generalizability and transferability of EEG representations, for instance by facilitating the sharing of a common core neural network between different montages or hardware devices.

As neuroimaging technology has matured and larger datasets have become more commonplace, new paradigms have also been developed to make use of this data. The brain age framework, for instance, is a compelling approach to understanding and monitoring brain health, however it has almost exclusively been studied using high-end expensive modalities, such as structural MRI. The results presented in this thesis are the first to show the applicability of the brain age framework in real-world settings. Interestingly, this framework can be extended to encompass various other proxy measures of health (Dadi et al., 2021) and, as such, is a promising approach to extracting health-related insights from large datasets of EEG (or of any neuroimaging modality) data, alongside SSL, for instance. Ultimately, larger datasets, reflecting increasingly diverse demographics, will continue to be collected, making it possible to train progressively more accurate and generalizable models.

Numerous other challenges will have to be tackled to fully unleash the potential of neuroimaging outside of the lab. Notably, the ability to perform on-device federated learning, *i.e.,* to train models directly on personal computing devices, will be key to enabling model finetuning and ensuring high standards of data privacy. Fusing EEG with complementary data streams, such as other biosignals (photoplethysmography (PPG), ECG, etc.), neuroimaging modalities (fNIRS), or even behavioral measures (*e.g.,* user interface interactions) will also improve the accuracy and flexibility of EEG-based applications. Looking forward, concurrent developments in real-world neuroimaging technology and artificial intelligence will set the stage for exciting new applications of neurotechnology and lead us to a deeper understanding of the brain.

Bibliography

Pierre Ablin, Jean-François Cardoso, and Alexandre Gramfort. Faster independent component analysis by preconditioning with Hessian approximations. *IEEE Transactions on Signal Processing*, 66(15):4040–4049, 2018. page 28

Khald Aboalayon, Miad Faezipour, Wafaa Almuhammadi, and Saeid Moslehpour. Sleep stage classification using EEG signal analysis: a comprehensive survey and new investigation. *Entropy*, 18(9):272, 2016. pages 5, 38

Maurice Abou Jaoude, Haoqi Sun, Kyle R Pellerin, Milena Pavlova, Rani A Sarkis, Sydney S Cash, M Brandon Westover, and Alice D Lam. Expert-level automated sleep staging of long-term scalp electroencephalography recordings using deep learning. *Sleep*, 43(11):zsaa112, 2020. pages 92, 93

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. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2013.02.014. URL https://www.sciencedirect.com/science/article/pii/S0950705113000798. pages 5, 24

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. ISSN 2076-3417. doi: 10.3390/app7121239. URL http://www.mdpi.com/2076-3417/7/12/1239. page 6

Obada Al Zoubi, Chung Ki Wong, Rayus T. Kuplicki, Hung-wen Yeh, Ahmad Mayeli, Hazem Refai, Martin Paulus, and Jerzy Bodurka. Predicting age from brain EEG signals—a machine learning approach. *Frontiers in Aging Neuroscience*, 10:184, 2018. ISSN 1663-4365. doi: 10.3389/fnagi.2018.00184. URL https://www.frontiersin.org/article/10.3389/fnagi.2018.00184. pages 5, 87, 140

Bruce M Altevogt and Harvey R Colten. *Sleep disorders and sleep deprivation: an unmet public health problem*. National Academies Press, 2006. page 28

Kai Keng Ang, Zheng Yang Chin, Haihong Zhang, and Cuntai Guan. Filter bank common spatial pattern (FBCSP) in brain-computer interface. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 2390–2397, 2008. doi: 10.1109/IJCNN.2008.4634130. page 10

Efthymios Angelakis, Joel F Lubar, Stamatina Stathopoulou, and John Kounios. Peak alpha frequency: an electroencephalographic measure of cognitive preparedness. *Clinical Neurophysiology*, 115(4):887–897, 2004. ISSN 1388-2457. doi: https://doi.org/10.1016/j.clinph.2003.11.034. URL https://www.sciencedirect.com/science/article/pii/S1388245703004632. page 106

Stefan Appelhoff, Matthew Sanderson, Teon L Brooks, Marijn van Vliet, Romain Quentin, Chris Holdgraf, Maximilien Chaumon, Ezequiel Mikulan, Kambiz Tavabi, Richard Höchenberger, et al. MNE-BIDS: Organizing electrophysiological data into

the BIDS format and facilitating their analysis. *The Journal of Open Source Software*, 4(44), 2019. page 99

Stefan Arnau, Tina Möckel, Gerhard Rinkenauer, and Edmund Wascher. The interconnection of mental fatigue and aging: An EEG study. *International Journal of Psychophysiology*, 117:17–25, 2017. ISSN 0167-8760. doi: https://doi.org/10.1016/j.ijpsycho.2017.04.003. URL https://www.sciencedirect.com/science/article/pii/S0167876016308790. page 106

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. ISSN 10870547. doi: 10.1177/1087054712460087. page 5

Frederico A.C. Azevedo, Ludmila R.B. Carvalho, Lea T. Grinberg, José Marcelo Farfel, Renata E.L. Ferretti, Renata E.P. Leite, Wilson Jacob Filho, Roberto Lent, and Suzana Herculano-Houzel. Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. *Journal of Comparative Neurology*, 513(5):532–541, 2009. doi: https://doi.org/10.1002/cne.21974. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cne.21974. page 3

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. page 33

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014. pages 15, 112

Sylvain Baillet. Magnetoencephalography for brain electrophysiology and imaging. *Nature neuroscience*, 20(3):327–339, 2017. page 2

Sylvain Baillet, John C Mosher, and Richard M Leahy. Electromagnetic brain mapping. *IEEE Signal Processing Magazine*, 18(6):14–30, 2001. page 4

Hubert Banville, Graeme Moffat, Isabela Albuquerque, Denis-Alexander Engemann, Aapo Hyvärinen, and Alexandre Gramfort. Self-supervised representation learning from electroencephalography signals. In *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2019. pages 20, 24, 26, 48

Hubert Banville, Omar Chehab, Aapo Hyvärinen, Denis-Alexander Engemann, and Alexandre Gramfort. Uncovering the structure of clinical EEG signals with self-supervised learning. *Journal of Neural Engineering*, 18(4):046020, 2021a. pages 21, 24, 26, 79

Hubert Banville, Sean UN Wood, Chris Aimone, Denis-Alexander Engemann, and Alexandre Gramfort. Robust learning from corrupted EEG with dynamic spatial filtering. *arXiv preprint arXiv:2105.12916*, 2021b. page 21

Alexandre Barachant, Anton Andreev, and Marco Congedo. The Riemannian Potato: an automatic and adaptive artifact detection method for online experiments using Riemannian geometry. In *TOBI Workshop lV*, pages 19–20, 2013a. page 61

Alexandre Barachant, Stéphane Bonnet, Marco Congedo, and Christian Jutten. Classification of covariance matrices using a Riemannian-based kernel for BCI applications. *Neurocomputing*, 112:172–178, 2013b. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2012.12.039. URL http://www.sciencedirect.com/science/article/pii/S0925231213001574. Advances in artificial neural networks, machine learning, and computational intelligence. pages 36, 62, 97, 99

Pouya Bashivan, Irina Rish, Mohammed Yeasin, and Noel Codella. Learning representations from EEG with deep recurrent-convolutional neural networks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL http://arxiv.org/abs/1511.06448. pages 55, 58

Christina Jayne Bathgate and Jack D Edinger. Diagnostic criteria and assessment of sleep disorders. In *Handbook of Sleep Disorders in Medical Conditions*, pages 3–25. Elsevier, 2019. pages 5, 32

Suzanna Becker. Learning to categorize objects using temporal coherence. In *Advances in Neural Information Processing Systems*, pages 361–368, 1993. pages 27, 28, 140

Suzanna Becker and Geoffrey E Hinton. Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 355(6356):161–163, 1992. page 27

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. URL http://arxiv.org/abs/1711.06068. page 16

Michele Bellesi, Brady A. Riedner, Gary N. Garcia-Molina, Chiara Cirelli, and Giulio Tononi. Enhancement of sleep slow waves: underlying mechanisms and practical consequences. *Frontiers in Systems Neuroscience*, 8:208, 2014. ISSN 1662-5137. doi: 10.3389/fnsys.2014.00208. URL https://www.frontiersin.org/article/10.3389/fnsys.2014.00208. page 95

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. page 19

Hans Berger. Über das Elektroenkephalogramm des Menschen. *Archiv für psychiatrie und nervenkrankheiten*, 87(1):527–570, 1929. pages 2, 3

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. page 6

Richard B Berry, Rita Brooks, Charlene E Gamaldo, Susan M Harding, Carole L Marcus, Bradley V Vaughn, et al. The AASM manual for the scoring of sleep and associated events. *Rules, Terminology and Technical Specifications, American Academy of Sleep Medicine*, 176, 2012. pages 5, 37, 49, 56, 92, 93, 95

Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. On the Bures–Wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, 2018. pages 63, 97

Andrea Biasiucci, Benedetta Franceschiello, and Micah M Murray. Electroencephalography. *Current Biology*, 29(3):R80–R85, 2019. page 5

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. pages 56, 58, 78

Benjamin Blankertz, Ryota Tomioka, Steven Lemm, Motoaki Kawanabe, and Klaus-Robert Muller. Optimizing spatial filters for robust EEG single-trial analysis. *IEEE Signal Processing Magazine*, 25(1):41–56, 2007. pages 13, 57

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001. page 106

Andreas Brink-Kjaer, Emmanuel Mignot, Helge BD Sorensen, and Poul Jennum. Predicting age with deep neural networks from polysomnograms. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 146–149. IEEE, 2020. pages 88, 103, 109

Michael M. Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: Going beyond Euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017. doi: 10.1109/MSP.2017.2693418. page 56

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. page 111

BMA Bruns. Predicting developmental age in young children by applying deep learning approaches to EEG data. Master's thesis, Utrecht University, 2021. page 88

Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562 (7726):203–209, 2018. pages 26, 111

B Rael Cahn and John Polich. Meditation states and traits: EEG, ERP, and neuroimaging studies. *Psychological bulletin*, 132(2):180, 2006. page 107

Julie Carrier, Isabelle Viens, Gaétan Poirier, Rébecca Robillard, Marjolaine Lafortune, Gilles Vandewalle, Nicolas Martin, Marc Barakat, Jean Paquet, and Daniel Filipini. Sleep slow wave changes during the middle years of life. *European Journal of Neuroscience*, 33(4):758–766, 2011. doi: https://doi.org/10.1111/j.1460-9568.2010.07543.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1460-9568.2010.07543.x. pages 94, 95

Alexander J Casson. Wearable EEG and beyond. *Biomedical Engineering Letters*, 9(1): 53–71, 2019. pages 7, 24

Richard Caton. Electrical currents of the brain. *The Journal of Nervous and Mental Disease*, 2(4):610, 1875. pages 1, 3

Augustin Cauchy et al. Méthode générale pour la résolution des systemes d'équations simultanées. *Comp. Rend. Sci. Paris*, 25(1847):536–538, 1847. page 16

Stanislas Chambon, Mathieu N Galtier, Pierrick J Arnal, Gilles Wainrib, and Alexandre Gramfort. A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(4):758–769, 2018. pages 16, 32, 33, 35, 38, 51, 55, 62, 78, 83

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020a. pages 27, 111

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020b. page 27

Yilun Chen, Ami Wiesel, Yonina C Eldar, and Alfred O Hero. Shrinkage algorithms for MMSE covariance estimation. *IEEE Transactions on Signal Processing*, 58(10): 5016–5029, 2010. pages 63, 97

Joseph Y Cheng, Hanlin Goh, Kaan Dogrusoz, Oncel Tuzel, and Erdrin Azemi. Subject-aware contrastive learning for biosignals. *arXiv preprint arXiv:2007.04871*, 2020. pages 24, 79

AKI Chiang, CJ Rennie, PA Robinson, SJ Van Albada, and CC Kerr. Age trends and sex differences of alpha rhythms including split alpha peaks. *Clinical Neurophysiology*, 122(8):1505–1517, 2011. pages 87, 104

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014. page 15

Sommer Christie, Selenia di Fronso, Maurizio Bertollo, and Penny Werthner. Individual alpha peak frequency in ice hockey shooting performance. *Frontiers in Psychology*, 8:762, 2017. ISSN 1664-1078. doi: 10.3389/fpsyg.2017.00762. URL https://www.frontiersin.org/article/10.3389/fpsyg.2017.00762. pages 105, 106

Yaqi Chu, Xingang Zhao, Yijun Zou, Weiliang Xu, Jianda Han, and Yiwen Zhao. A decoding scheme for incomplete motor imagery EEG with deep belief network. *Frontiers in Neuroscience*, 12:680, 2018. pages 56, 58

Maureen Clerc, Laurent Bougrain, and Fabien Lotte. *Brain-Computer Interfaces 1: Foundations and Methods*. Wiley, 2016. ISBN 9781119144991. URL https://books.google.de/books?id=STgZDQAAQBAJ. page 6

David Cohen. Magnetoencephalography: evidence of magnetic fields produced by alpha-rhythm currents. *Science*, 161(3843):784–786, 1968. page 2

James H. Cole and Katja Franke. Predicting age using neuroimaging: Innovative brain ageing biomarkers. *Trends in Neurosciences*, 40(12):681–690, 2017. ISSN 0166-2236. doi: https://doi.org/10.1016/j.tins.2017.10.001. URL https://www.sciencedirect.com/science/article/pii/S016622361730187X. pages 5, 87, 90, 140

James H. Cole, Rudra P.K. Poudel, Dimosthenis Tsagkrasoulis, Matthan W.A. Caan, Claire Steves, Tim D. Spector, and Giovanni Montana. Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. *NeuroImage*, 163:115–124, 2017. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2017.07.059. URL https://www.sciencedirect.com/science/article/pii/S1053811917306407. page 88

James H Cole, Stuart J Ritchie, Mark E Bastin, MC Valdés Hernández, S Muñoz Maniega, Natalie Royle, Janie Corley, Alison Pattie, Sarah E Harris, Qian Zhang, et al. Brain age predicts mortality. *Molecular psychiatry*, 23(5):1385–1392, 2018. pages 5, 87, 141

James H. Cole, Katja Franke, and Nicolas Cherbuin. *Quantification of the Biological Age of the Brain Using Neuroimaging*, pages 293–328. Springer International Publishing, Cham, 2019. ISBN 978-3-030-24970-0. doi: 10.1007/978-3-030-24970-0_19. URL https://doi.org/10.1007/978-3-030-24970-0_19. pages 86, 87

Scott R Cole and Bradley Voytek. Cycle-by-cycle analysis of neural oscillations. *bioRxiv*, 2018. doi: 10.1101/302000. URL https://www.biorxiv.org/content/early/2018/04/16/302000. page 6

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. pages 36, 61, 62, 97

Isaac A Corley and Yufei Huang. Deep EEG super-resolution: Upsampling EEG spatial resolution with generative adversarial networks. In *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pages 100–103. IEEE, 2018. pages 56, 58

Kamalaker Dadi, Gaël Varoquaux, Josselin Houenou, Danilo Bzdok, Bertrand Thirion, and Denis Engemann. Population modeling with machine learning can enhance measures of mental health. *GigaScience*, 10(10), 10 2021. ISSN 2047-217X. doi: 10.1093/gigascience/giab071. URL https://doi.org/10.1093/gigascience/giab071. giab071. pages 86, 98, 109, 112

Alain de Cheveigné and Jonathan Z Simon. Denoising based on spatial filtering. *Journal of Neuroscience Methods*, 171(2):331–339, 2008. pages 13, 57

Luigi De Gennaro and Michele Ferrara. Sleep spindles: an overview. *Sleep Medicine Reviews*, 7(5):423–440, 2003. ISSN 1087-0792. doi: https://doi.org/10.1053/smrv.2002.0252. URL https://www.sciencedirect.com/science/article/pii/S1087079202902522. pages 93, 94

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. page 8

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. pages 15, 25, 27

Kiret Dhindsa. Filter-bank artifact rejection: High performance real-time single-channel artifact detection for EEG. *Biomedical Signal Processing and Control*, 38:224–235, 2017. pages 7, 55, 58

Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, pages 1422–1430, 2015. pages 27, 30, 50

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. page 16

Aurora D'Atri, Maurizio Gorgoni, Serena Scarpelli, Susanna Cordone, Valentina Alfonsi, Camillo Marra, Michele Ferrara, Paolo Maria Rossini, and Luigi De Gennaro. Relationship between cortical thickness and EEG alterations during sleep in the Alzheimer's disease. *Brain Sciences*, 11(9):1174, 2021. page 94

Heba El-Fiqi, Kathryn Kasmarik, Anastasios Bezerianos, Kay Chen Tan, and Hussein A Abbass. Gate-layer autoencoders with application to incomplete EEG signal recovery. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019. pages 57, 58

Denis A Engemann, Federico Raimondo, Jean-Rémi 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, 2018a. pages 5, 56, 58, 63, 80

Denis A Engemann, Federico Raimondo, Jean-Rémi King, Benjamin Rohaut, Gilles Louppe, Frédéric Faugeras, Jitka Annen, Helena Cassol, Olivia Gosseries, Diego Fernandez-Slezak, Steven Laureys, Lionel Naccache, Stanislas Dehaene, and Jacobo D Sitt. Robust EEG-based cross-site and cross-protocol classification of states of consciousness. *Brain*, 141(11):awy251, 2018b. doi: 10.1093/brain/awy251. URL http://dx.doi.org/10.1093/brain/awy251. pages 25, 140

Denis A Engemann, Oleh Kozynets, David Sabbagh, Guillaume Lemaître, Gael Varoquaux, Franziskus Liem, and Alexandre Gramfort. Combining magnetoencephalography with magnetic resonance imaging enhances learning of surrogate-biomarkers. *eLife*, 9:e54055, 2020. pages 26, 83, 87, 97

Denis A. Engemann, Apolline Mellot, Richard Höchenberger, Hubert Banville, David Sabbagh, Lukas Gemein, Tonio Ball, and Alexandre Gramfort. A reusable benchmark of brain-age prediction from M/EEG resting-state signals. *bioRxiv*, 2021. doi: 10.1101/2021.12.14.472691. URL https://www.biorxiv.org/content/early/2021/12/16/2021.12.14.472691. page 96

Mary Jane England, Catharyn T Liverman, Andrea M Schultz, and Larisa M Strawbridge. Epilepsy across the spectrum: Promoting health and understanding.: A summary of the institute of medicine report. *Epilepsy & Behavior*, 25(2):266–276, 2012. page 50

Irwin Feinberg, Richard L Koresko, and Naomi Heller. EEG sleep patterns as a function of normal and pathological aging in man. *Journal of psychiatric research*, 5(2):107–144, 1967. page 87

Marco Ferrari and Valentina Quaresima. A brief review on the history of human functional near-infrared spectroscopy (fNIRS) development and fields of application. *NeuroImage*, 63(2):921–935, 2012. ISSN 1053-8119. doi: https://doi.org/10.

1016/j.neuroimage.2012.03.049. URL https://www.sciencedirect.com/science/article/pii/S1053811912003308. page 3

Sean Ferrell, Vineetha Mathew, Matthew Refford, Vincent Tchiong, Tameem Ahsan, Iyad Obeid, and Jospeh Picone. The Temple University Hospital EEG Corpus: Electrode location and channel labels. Technical report, The Neural Engineering Data Consortium, Temple University, Philadelphia, Pennsylvania, 7 2019. page 45

Peter Földiák. Learning invariance from transformation sequences. *Neural Computation*, 3(2):194–200, 1991. pages 28, 140

Jean-Yves Franceschi, Aymeric Dieuleveut, and Martin Jaggi. Unsupervised scalable representation learning for multivariate time series. In *Advances in Neural Information Processing Systems*, pages 4650–4661, 2019. page 31

Katja Franke and Christian Gaser. Longitudinal changes in individual BrainAGE in healthy aging, mild cognitive impairment, and Alzheimer's disease. *GeroPsych*, 2012. page 87

Katja Franke, Gabriel Ziegler, Stefan Klöppel, and Christian Gaser. Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: Exploring the influence of various parameters. *NeuroImage*, 50(3): 883–892, 2010. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2010.01.005. URL https://www.sciencedirect.com/science/article/pii/S1053811910000108. pages 5, 86, 87, 140

Katja Franke, Eileen Luders, Arne May, Marko Wilke, and Christian Gaser. Brain maturation: Predicting individual BrainAGE in children and adolescents using structural mri. *NeuroImage*, 63(3):1305–1312, 2012. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2012.08.001. URL https://www.sciencedirect.com/science/article/pii/S105381191200794X. pages 5, 87, 90, 140

Lukas AW Gemein, Robin T Schirrmeister, Patryk Chrabąszcz, Daniel Wilson, Joschka Boedecker, Andreas Schulze-Bonhage, Frank Hutter, and Tonio Ball. Machine-learning-based diagnostics of EEG pathology. *NeuroImage*, page 117021, 2020. pages 20, 33, 35, 38, 50, 51, 58, 62, 63, 74, 79, 97, 140

Mohammad M Ghassemi, Benjamin E Moody, Li-Wei H 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. In *2018 Computing in Cardiology Conference (CinC)*, volume 45, pages 1–4. IEEE, 2018. pages 5, 24, 32, 37, 65, 140

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 arXiv:1805.11704*, 2018. page 20

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):99–114, 2014. page 5

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet: components of a new

research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000. pages 23, 32, 37, 65, 140

Ian Goodfellow. NIPS 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016. page 19

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org. pages 9, 10, 11, 14, 15, 17, 139

Priya Goyal, Mathilde Caron, Benjamin Lefaudeux, Min Xu, Pengchao Wang, Vivek Pai, Mannat Singh, Vitaliy Liptchinsky, Ishan Misra, Armand Joulin, et al. Self-supervised pretraining of visual features in the wild. *arXiv preprint arXiv:2103.01988*, 2021. page 111

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. page 6

Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A Engemann, Daniel Strohmeier, Christian Brodbeck, Lauri Parkkonen, and Matti S Hämäläinen. MNE software for processing MEG and EEG data. *NeuroImage*, 86:446–460, 2014. pages 36, 62, 99

Antoine Guillot and Valentin Thorey. RobustSleepNet: Transfer learning for automated sleep staging at scale. *arXiv preprint arXiv:2101.02452*, 2021. pages 82, 83

Antoine Guillot, Fabien Sauvet, Emmanuel H During, and Valentin Thorey. Dreem open datasets: Multi-scored sleep datasets to compare human and automated sleep staging. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28 (9):1955–1965, 2020. pages 58, 82, 83

Juan Lorenzo Hagad, Kenichi Fukui, and Masayuki Numao. Deep visual models for EEG of mindfulness meditation in a workplace setting. In *International Workshop on Health Intelligence*, pages 129–137. Springer, 2019. pages 56, 58

S. Hagihira. Changes in the electroencephalogram during anaesthesia and their physiological basis. *British Journal of Anaesthesia*, 115(suppl_1):i27–i31, 2015. ISSN 14716771. doi: 10.1093/bja/aev212. page 5

Jinpei Han, Xiao Gu, and Benny Lo. Semi-supervised contrastive learning for generalizable motor imagery EEG classification. In *2021 IEEE 17th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, pages 1–4. IEEE, 2021. page 24

Sangjun Han, Moonyoung Kwon, Sunghan Lee, and Sung Chan Jun. Feasibility study of EEG super-resolution using deep convolutional networks. In *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1033–1038. IEEE, 2018. page 58

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. page 19

Riitta Hari and Aina Puce. *MEG-EEG Primer*. Oxford University Press, 2017. pages 2, 3, 5, 6

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)*, pages 1–6. IEEE, 2018. pages 20, 49

Ali Hashemi, Lou J Pino, Graeme Moffat, Karen J Mathewson, Chris Aimone, Patrick J Bennett, Louis A Schmidt, and Allison B Sekuler. Characterizing population EEG dynamics throughout adulthood. *ENeuro*, 3(6), 2016. pages 66, 87, 90, 104, 105, 108

Stefan Haufe, Frank Meinecke, Kai Görgen, Sven Dähne, John-Dylan Haynes, Benjamin Blankertz, and Felix Bießmann. On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, 87:96–110, 2014. page 80

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. doi: 10.1146/annurev-bioeng-062117-120853. URL https://doi.org/10.1146/annurev-bioeng-062117-120853. PMID: 29494213. page 4

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, pages 1026–1034, 2015. pages 34, 61, 98

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019a. pages 27, 33

Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking ImageNet pre-training. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4918–4927, 2019b. page 25

Weipeng He, Lu Lu, Biqiao Zhang, Jay Mahadeokar, Kaustubh Kalgaonkar, and Christian Fuegen. Spatial attention for far-field speech recognition with deep beamforming neural networks. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7499–7503. IEEE, 2020. page 82

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, 2018. pages 55, 58

Olivier J Hénaff, Ali Razavi, Carl Doersch, SM Eslami, and Aaron van den Oord. Data-efficient image recognition with contrastive predictive coding. *arXiv preprint arXiv:1905.09272*, 2019. pages 27, 48

Sepp Hochreiter and Jürgen Schmidhuber. LSTM can solve hard long time lag problems. *Advances in Neural Information Processing Systems*, pages 473–479, 1997. page 15

Kerstin Hoedlmoser, Dominik PJ Heib, Judith Roell, Philippe Peigneux, Avi Sadeh, Georg Gruber, and Manuel Schabus. Slow sleep spindle activity, declarative memory, and general cognitive abilities in children. *Sleep*, 37(9):1501–1512, 2014. page 94

Jacob Hogan, Haoqi Sun, Luis Paixao, Mike Westmeijer, Pooja Sikka, Jing Jin, Ryan Tesh, Madalena Cardoso, Sydney S. Cash, Oluwaseun Akeju, Robert Thomas, and M. Brandon Westover. Night-to-night variability of sleep electroencephalography-based brain age measurements. *Clinical Neurophysiology*, 132(1):1–12, 2021. ISSN 1388-2457. doi: https://doi.org/10.1016/j.clinph.2020.09.029. URL https://www.sciencedirect.com/science/article/pii/S1388245720305204. pages 107, 109

Matthias R Hohmann, Lisa Konieczny, Michelle Hackl, Brian Wirth, Talha Zaman, Raffi Enficiaud, Moritz Grosse-Wentrup, and Bernhard Schölkopf. MYND: Unsupervised evaluation of novel BCI control strategies on consumer hardware. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, pages 1071–1084, 2020. pages 7, 139

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989. ISSN 00485756. doi: 10.1016/0893-6080(89)90020-8. URL https://pdfs.semanticscholar.org/f22f/6972e66bdd2e769fa64b0df0a13063c0c101.pdf. page 9

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018. pages 16, 55, 57, 79, 112, 140

Yen-Cheng Huang, Jia-Ren Chang, Li-Fen Chen, and Yong-Sheng Chen. Deep neural network with attention mechanism for classification of motor imagery EEG. In *2019 9th International IEEE/EMBS Conference on Neural Engineering (NER)*, pages 1130–1133. IEEE, 2019. page 83

Aapo Hyvärinen and Hiroshi Morioka. Nonlinear ICA of temporally dependent stationary sources. In *AISTATS*, 2017. pages 27, 50

Aapo Hyvärinen, Hiroaki Sasaki, and Richard E Turner. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *AISTATS*, 2019. pages 27, 28, 51, 111

Costantino Iadecola. The neurovascular unit coming of age: a journey through neurovascular coupling in health and disease. *Neuron*, 96(1):17–42, 2017. page 3

Mainak Jas. *Contributions pour l'analyse automatique de signaux neuronaux*. PhD thesis, Paris, ENST, 2018. page 112

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. pages 56, 58, 63, 78, 80, 140

Mainak Jas, Eric Larson, Denis A Engemann, Jaakko Leppäkangas, Samu Taulu, Matti Hämäläinen, and Alexandre Gramfort. A reproducible MEG/EEG group study with the MNE software: recommendations, quality assessments, and good practices. *Frontiers in neuroscience*, 12:530, 2018. pages 7, 99

Herbert Jasper and Wilder Penfield. Electrocorticograms in man: effect of voluntary movement upon the electrical activity of the precentral gyrus. *Archiv für Psychiatrie und Nervenkrankheiten*, 183(1):163–174, 1949. page 2

Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):4037–4058, 2021. doi: 10.1109/TPAMI.2020.2992393. pages 18, 139

Kristina T Johnson and Rosalind W Picard. Advancing neuroscience through wearable devices. *Neuron*, 108(1):8–12, 2020. pages 7, 139

S Æ Jónsson, E Gunnlaugsson, E Finssonn, DL Loftsdóttir, GH Ólafsdóttir, H Helgadóttir, and JS Ágústsson. 0447 ResTNet: A robust end-to-end deep learning approach to sleep staging of self applied somnography studies. *Sleep*, 43(Supplement_1):A171–A171, 2020. page 62

Tzyy-Ping Jung, Colin Humphries, Te-Won Lee, Scott Makeig, Martin J McKeown, Vicente Iragui, and Terrence J Sejnowski. Extended ICA removes artifacts from electroencephalographic recordings. In *Advances in Neural Information Processing Systems*, pages 894–900, 1997. pages 28, 56, 58

Frans F. Jöbsis. Noninvasive, infrared monitoring of cerebral and myocardial oxygen sufficiency and circulatory parameters. *Science*, 198(4323):1264–1267, 1977. doi: 10.1126/science.929199. URL https://www.science.org/doi/abs/10.1126/science.929199. page 3

Anthony Kales, Allan Rechtschaffen, Los Angeles. Brain Information Service University of California, and NINDB Neurological Information Network (U.S.). *A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects: Allan Rechtschaffen and Anthony Kales, Editors*. NIH publication. U. S. National Institute of Neurological Diseases and Blindness, Neurological Information Network, 1968. page 49

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. pages 18, 33

Wolfgang Klimesch. EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain Research Reviews*, 29(2):169–195, 1999. ISSN 0165-0173. doi: https://doi.org/10.1016/S0165-0173(98)00056-3. URL https://www.sciencedirect.com/science/article/pii/S0165017398000563. page 104

Maria G. Knyazeva, Elham Barzegaran, Vladimir Y. Vildavski, and Jean-François Demonet. Aging of human alpha rhythm. *Neurobiology of Aging*, 69:261–273, 2018. ISSN 0197-4580. doi: https://doi.org/10.1016/j.neurobiolaging.2018.05.018. URL https://www.sciencedirect.com/science/article/pii/S0197458018301799. page 104

Zoltan J Koles, Michael S Lazar, and Steven Z Zhou. Spatial patterns underlying population differences in the background EEG. *Brain Topography*, 2(4):275–284, 1990. page 60

Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1920–1929, 2019. page 51

Demetres Kostas, Stéphane Aroca-Ouellette, and Frank Rudzicz. BENDR: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data. *Frontiers in Human Neuroscience*, 15:253, 2021. ISSN 1662-5161.

doi: 10.3389/fnhum.2021.653659. URL https://www.frontiersin.org/article/10.3389/fnhum.2021.653659. page 24

Christian Andreas Edgar Kothe and Tzyy-Ping Jung. Artifact removal techniques with signal reconstruction, April 28 2016. US Patent App. 14/895,440. page 56

Abhay Koushik, Judith Amores, and Pattie Maes. Real-time sleep staging using deep learning on a smartphone for a wearable EEG. *arXiv preprint arXiv:1811.10111*, 2018. pages 66, 90

Mark A Kramer. Nonlinear principal component analysis using autoassociative neural networks. *AIChE journal*, 37(2):233–243, 1991. page 35

Matthias Kreuzer. EEG based monitoring of general anesthesia: taking the next steps. *Frontiers in Computational Neuroscience*, 11:56, 2017. pages 7, 139

Olave E Krigolson, Chad C Williams, Angela Norton, Cameron D Hassall, and Francisco L Colino. Choosing MUSE: Validation of a low-cost, portable EEG system for ERP research. *Frontiers in Neuroscience*, 11:109, 2017. pages 7, 66, 139

Olave E Krigolson, Mathew R Hammerstrom, Wande Abimbola, Robert Trska, Bruce W Wright, Kent G Hecker, and Gordon Binsted. Using Muse: Rapid mobile assessment of brain performance. *Frontiers in Neuroscience*, 15, 2021. pages 7, 66, 139

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. ISSN 19326203. doi: 10.1371/journal.pone.0172578. page 16

Moonyoung Kwon, Sangjun Han, Kiwoong Kim, and Sung Chan Jun. Super-resolution for improving EEG spatial resolution using deep convolutional neural network—feasibility study. *Sensors*, 19(23):5317, 2019. page 58

Hans-Peter Landolt and Alexander A Borbély. Age-dependent changes in sleep EEG topography. *Clinical Neurophysiology*, 112(2):369–377, 2001. page 87

Annalise A LaPlume, Nicole D Anderson, Larissa McKetton, Brian Levine, and Angela K Troyer. When I'm 64: Age-related variability in over 40,000 online cognitive test takers. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 2021. page 87

Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, 15(5):056013, 2018. pages 55, 58

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. pages 11, 13

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436, 2015. pages 2, 8, 9, 11, 139

Alexander LeNail. NN-SVG: Publication-ready neural network architecture schematics. *Journal of Open Source Software*, 4(33):747, 2019. page 12

Javier León, Juan José Escobar, Andrés Ortiz, Julio Ortega, Jesús González, Pedro Martín-Smith, John Q Gan, and Miguel Damas. Deep learning for EEG-based motor imagery classification: Accuracy-cost trade-off. *Plos one*, 15(6):e0234178, 2020. page 48

Bo Li, Tara N Sainath, Ron J Weiss, Kevin W Wilson, and Michiel Bacchiani. Neural network adaptive beamforming for robust multichannel speech recognition. *Interspeech 2016*, pages 1976–1980, 2016. page 82

Junhua Li, Zbigniew Struzik, Liqing Zhang, and Andrzej Cichocki. Feature learning from incomplete EEG with denoising autoencoder. *Neurocomputing*, 165:23–31, 2015. pages 56, 58

Yang Li, Xian-Rui Zhang, Bin Zhang, Meng-Ying Lei, Wei-Gang Cui, and Yu-Zhu Guo. A channel-projection mixed-scale convolutional neural network for motor imagery EEG decoding. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(6):1170–1180, 2019. page 58

Davis Liang, Zhiheng Huang, and Zachary C Lipton. Learning noise-invariant representations for robust speech recognition. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 56–63. IEEE, 2018. page 82

Sheng-Fu Liang, Chin-En Kuo, Yu-Han Hu, Yu-Hsiang Pan, and Yung-Hung Wang. Automatic stage scoring of single-channel sleep EEG by using multiscale entropy and autoregressive models. *IEEE Transactions on Instrumentation and Measurement*, 61 (6):1649–1657, 2012. doi: 10.1109/TIM.2012.2187242. pages 78, 83

Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013. page 14

Zhenhua Lin. Riemannian geometry of symmetric positive definite matrices via Cholesky decomposition. *SIAM Journal on Matrix Analysis and Applications*, 40(4):1353–1370, 2019. page 61

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. page 15

Nikos K Logothetis, Jon Pauls, Mark Augath, Torsten Trinath, and Axel Oeltermann. Neurophysiological investigation of the basis of the fMRI signal. *Nature*, 412(6843): 150–157, 2001. page 3

Tim Lomas, Itai Ivtzan, and Cynthia H.Y. Fu. A systematic review of the neurophysiology of mindfulness on EEG oscillations. *Neuroscience & Biobehavioral Reviews*, 57:401–410, 2015. ISSN 0149-7634. doi: https://doi.org/10.1016/j.neubiorev.2015.09.018. URL https://www.sciencedirect.com/science/article/pii/S0149763415002511. page 107

Alfred L Loomis, E Newton Harvey, and Garret A Hobart. Cerebral states during sleep, as studied by human brain potentials. *Journal of experimental psychology*, 21(2):127, 1937. page 49

S Lopez, G Suarez, D Jungreis, I Obeid, and J Picone. Automated identification of abnormal adult EEGs. In *2015 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pages 1–5. IEEE, 2015. pages 33, 62

Silvia López, I Obeid, and J Picone. Automated interpretation of abnormal adult electroencephalograms. *MS Thesis, Temple University*, 2017. pages 32, 37, 140

M. A. Lopez-Gordo, D. Sanchez-Morillo, and F. Pelayo Valle. Dry EEG electrodes. *Sensors*, 14(7):12847–12870, 2014. ISSN 1424-8220. doi: 10.3390/s140712847. URL https://www.mdpi.com/1424-8220/14/7/12847. page 5

Beth A Lopour, Savas Tasoglu, Heidi E Kirsch, James W Sleigh, and Andrew J Szeri. A continuous mapping of sleep states through association of EEG with a mesoscale cortical model. *Journal of computational neuroscience*, 30(2):471–487, 2011. pages 42, 49

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. pages 61, 98

Fabien Lotte and Cuntai Guan. Regularizing common spatial patterns to improve BCI designs: unified theory and new algorithms. *IEEE Transactions on Biomedical Engineering*, 58(2):355–362, 2010. pages 13, 57

Fabien Lotte, Marco Congedo, Anatole Lécuyer, Fabrice Lamarche, and Bruno Arnaldi. A review of classification algorithms for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, 4(2):R1, 2007. page 54

Fabien Lotte, Laurent Bougrain, and Maureen Clerc. *Electroencephalography (EEG)-Based Brain-Computer Interfaces*, pages 1–20. American Cancer Society, 2015. ISBN 047134608X. doi: 10.1002/047134608X.W8278. URL http://doi.wiley.com/10.1002/047134608X.W8278. page 6

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. pages 7, 24, 36, 54, 62, 97

Eileen Luders, Nicolas Cherbuin, and Christian Gaser. Estimating brain age using high-resolution pattern recognition: Younger brains in long-term meditation practitioners. *NeuroImage*, 134:508–513, 2016. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2016.04.007. URL https://www.sciencedirect.com/science/article/pii/S1053811916300404. page 107

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015. pages 15, 112

Damien Léger, Eden Debellemaniere, Arnaud Rabat, Virginie Bayon, Karim Benchenane, and Mounir Chennaoui. Slow-wave sleep: From the cell to the clinic. *Sleep Medicine Reviews*, 41:113–132, 2018. ISSN 1087-0792. doi: https://doi.org/10.1016/j.smrv.2018.01.008. URL https://www.sciencedirect.com/science/article/pii/S1087079217300059. page 95

Emanuele Maiorana. Deep learning for EEG-based biometric recognition. *Neurocomputing*, 410:374–386, 2020. page 49

Scott Makeig, Anthony J Bell, Tzyy-Ping Jung, Terrence J Sejnowski, et al. Independent component analysis of electroencephalographic data. *Advances in Neural Information Processing Systems*, pages 145–151, 1996. pages 13, 57

Scott Makeig, Tzyy-Ping Jung, Anthony J Bell, Dara Ghahremani, and Terrence J Sejnowski. Blind separation of auditory event-related brain responses into independent components. *Proceedings of the National Academy of Sciences*, 94(20):10979–10984, 1997. page 28

Raman K Malhotra and Alon Y Avidan. Sleep stages and scoring technique. *Atlas of Sleep Medicine*, pages 77–99, 2013. pages 25, 32, 95

Jaakko Malmivuo, Robert Plonsey, et al. *Bioelectromagnetism: principles and applications of bioelectric and biomagnetic fields*. Oxford University Press, USA, 1995. page 5

Nadia Mammone, Fabio La Foresta, and Francesco Carlo Morabito. Automatic artifact rejection from multichannel scalp EEG by wavelet ICA. *IEEE Sensors Journal*, 12 (3):533–542, 2011. pages 56, 58

Bryce A Mander, Joseph R Winer, and Matthew P Walker. Sleep and human aging. *Neuron*, 94(1):19–36, 2017. page 43

Ran Manor and Amir B Geva. Convolutional neural network for multi-category rapid serial visual presentation BCI. *Frontiers in Computational Neuroscience*, 9:146, 2015. pages 16, 55, 58

Dennis J. McFarland and Jonathan R. Wolpaw. EEG-based brain–computer interfaces. *Current Opinion in Biomedical Engineering*, 4:194–200, 2017. ISSN 2468-4511. doi: https://doi.org/10.1016/j.cobme.2017.11.004. URL https://www.sciencedirect.com/science/article/pii/S246845111730082X. Synthetic Biology and Biomedical Engineering / Neural Engineering. page 6

Dennis J McFarland, Lynn M McCane, Stephen V David, and Jonathan R Wolpaw. Spatial filter selection for EEG-based communication. *Electroencephalography and Clinical Neurophysiology*, 103(3):386–394, 1997. pages 13, 57

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. page 41

Christina Micanovic and Suvankar Pal. The diagnostic utility of EEG in early-onset dementia: a systematic review of the literature with narrative analysis. *Journal of Neural Transmission*, 121(1):59–69, 2014. pages 5, 32, 62

F.M. Miezin, L. Maccotta, J.M. Ollinger, S.E. Petersen, and R.L. Buckner. Characterizing the hemodynamic response: Effects of presentation rate, sampling procedure, and the possibility of ordering brain activity based on relative timing. *NeuroImage*, 11(6):735–759, 2000. ISSN 1053-8119. doi: https://doi.org/10.1006/nimg.2000.0568. URL https://www.sciencedirect.com/science/article/pii/S1053811900905688. page 3

Vojkan Mihajlović, Bernard Grundlehner, Ruud Vullers, and Julien Penders. Wearable, wireless EEG solutions in daily life applications: what are we missing? *IEEE Journal of Biomedical and Health Informatics*, 19(1):6–21, 2014. pages 7, 24, 139

Kaare B Mikkelsen, Huy Phan, Mike L Rank, Martin C Hemmsen, Maarten de Vos, and Preben Kidmose. Light-weight sleep monitoring: electrode distance matters more than placement for automatic scoring. *arXiv preprint arXiv:2104.04567*, 2021. pages 7, 139

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. pages 25, 27

Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *ECCV*, pages 527–544. Springer, 2016. pages 27, 30, 50

Mostafa Neo Mohsenvand, Mohammad Rasool Izadi, and Pattie Maes. Contrastive representation learning for electroencephalogram classification. In *Machine Learning for Health*, pages 238–253. PMLR, 2020. pages 24, 79

Doris Moser, Peter Anderer, Georg Gruber, Silvia Parapatics, Erna Loretz, Marion Boeck, Gerhard Kloesch, Esther Heller, Andrea Schmidt, Heidi Danker-Hopfe, et al. Sleep classification according to AASM and Rechtschaffen & Kales: Effects on sleep scoring parameters. *Sleep*, 32(2):139–149, 2009. page 49

Shayan Motamedi-Fakhr, Mohamed Moshrefi-Torbati, Martyn Hill, Catherine M. Hill, and Paul R. White. Signal processing techniques applied to human sleep EEG signals—a review. *Biomedical Signal Processing and Control*, 10:21–33, 2014. ISSN 1746-8094. doi: https://doi.org/10.1016/j.bspc.2013.12.003. URL http://www.sciencedirect.com/science/article/pii/S174680941300178X. pages 32, 62

MS Mourtazaev, B Kemp, AH Zwinderman, and HAC Kamphuisen. Age and gender affect different characteristics of slow waves in the sleep EEG. *Sleep*, 18(7):557–564, 1995. pages 87, 94

Beate E Muehlroth and Markus Werkle-Bergner. Understanding the interplay of sleep and aging: Methodological challenges. *Psychophysiology*, 57(3):e13523, 2020. pages 92, 93, 94, 95, 96

Tim R Mullen, Christian AE Kothe, Yu Mike Chi, Alejandro Ojeda, Trevor Kerth, Scott Makeig, Tzyy-Ping Jung, and Gert Cauwenberghs. Real-time neuroimaging and cognitive monitoring using wearable dry EEG. *IEEE Transactions on Biomedical Engineering*, 62(11):2553–2567, 2015. page 61

Samaneh Nasiri and Gari D Clifford. Attentive adversarial network for large-scale sleep staging. In *Machine Learning for Healthcare Conference*, pages 457–478. PMLR, 2020. page 83

Neha Nayak, Gabor Angeli, and Christopher D Manning. Evaluating word embeddings using a representative suite of practical tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 19–23, 2016. page 27

Vadim V Nikulin, Guido Nolte, and Gabriel Curio. A novel method for reliable and fast extraction of neuronal EEG/MEG oscillations on the basis of spatio-spectral decomposition. *NeuroImage*, 55(4):1528–1535, 2011. pages 13, 57

Zhaoyang Niu, Guoqiang Zhong, and Hui Yu. A review on the attention mechanism of deep learning. *Neurocomputing*, 452:48–62, 2021. pages 11, 15, 20

Hugh Nolan, Robert Whelan, and Richard B Reilly. FASTER: fully automated statistical thresholding for EEG artifact rejection. *Journal of Neuroscience Methods*, 192 (1):152–162, 2010. pages 56, 58, 78

Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, pages 69–84. Springer, 2016. pages 25, 27

Yoav Nygate, Sam Rusk, Chris Fernandez, Nick Glattard, Jessica Arguelles, Jiaxiao Shi, Dennis Hwang, and Nathaniel Watson. EEG-based deep neural network model for brain age prediction and its association with patient health conditions. *Sleep*, 44 (Supplement 2):A214–A214, 05 2021. ISSN 0161-8105. doi: 10.1093/sleep/zsab072.541. URL https://doi.org/10.1093/sleep/zsab072.541. pages 88, 103, 104, 109

Iyad Obeid and Joseph Picone. The temple university hospital EEG data corpus. *Frontiers in Neuroscience*, 10:196, 2016. pages 23, 26, 37, 62, 65, 111

Seiji Ogawa, Tso-Ming Lee, Alan R Kay, and David W Tank. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *proceedings of the National Academy of Sciences*, 87(24):9868–9872, 1990. page 3

Alison M Pack. Epilepsy overview and revised classification of seizures and epilepsies. *CONTINUUM: Lifelong Learning in Neurology*, 25(2):306–321, 2019. page 50

Luis Paixao, Pooja Sikka, Haoqi Sun, Aayushee Jain, Jacob Hogan, Robert Thomas, and M Brandon Westover. Excess brain age in the sleep electroencephalogram predicts reduced life expectancy. *Neurobiology of aging*, 88:150–155, 2020. pages 87, 103

James Pardey, Stephen Roberts, Lionel Tarassenko, and John Stradling. A new approach to the analysis of the human sleep/wakefulness continuum. *Journal of sleep research*, 5(4):201–210, 1996. pages 42, 49

Lucas C Parra, Clay D Spence, Adam D Gerson, and Paul Sajda. Recipes for the linear analysis of EEG. *NeuroImage*, 28(2):326–341, 2005. pages 13, 28, 57

Josef Parvizi and Sabine Kastner. Human intracranial EEG: promises and limitations. *Nature neuroscience*, 21(4):474, 2018. page 2

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf. pages 36, 62, 99

Avijit Paul. Prediction of missing EEG channel waveform using LSTM. In *2020 4th International Conference on Computational Intelligence and Networks (CINE)*, pages 1–6. IEEE, 2020. page 57

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(Oct):2825–2830, 2011. pages 36, 62, 99

Wilder Penfield and Theodore Rasmussen. The Cerebral Cortex of Man: A Clinical Study of Localization of Function. *Journal of the American Medical Association*, 144 (16):1412–1412, 12 1950. ISSN 0002-9955. doi: 10.1001/jama.1950.02920160086033. URL https://doi.org/10.1001/jama.1950.02920160086033. page 1

Han Peng, Weikang Gong, Christian F. Beckmann, Andrea Vedaldi, and Stephen M. Smith. Accurate brain age prediction with lightweight deep neural networks. *Medical Image Analysis*, 68:101871, 2021. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2020.101871. URL https://www.sciencedirect.com/science/article/pii/S1361841520302358. page 88

François Perrin, J Pernier, O Bertrand, and JF Echallier. Spherical splines for scalp potential and current density mapping. *Electroencephalography and Clinical Neurophysiology*, 72(2):184–187, 1989. pages 59, 80, 81, 140

Gert Pfurtscheller and Christa Neuper. Motor imagery activates primary sensorimotor area in humans. *Neuroscience letters*, 239(2-3):65–68, 1997. page 8

Huy Phan, Fernando Andreotti, Navin Cooray, Oliver Y Chén, and Maarten De Vos. SeqSleepNet: end-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(3):400–410, 2019. pages 58, 82, 83

Huy Phan, Oliver Y Chén, Philipp Koch, Alfred Mertins, and Maarten De Vos. XSleepNet: Multi-view sequential model for automatic sleep staging. *arXiv preprint arXiv:2007.05492*, 2020. pages 58, 62, 82, 83

SM Purcell, DS Manoach, C Demanuele, BE Cade, S Mariani, R Cox, G Panagiotaropoulou, R Saxena, JQ Pan, JW Smoller, et al. Characterizing sleep spindles in 11,630 individuals from the national sleep research resource. *Nature communications*, 8(1): 1–16, 2017. pages 43, 92, 93, 94

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. page 15

AG Ramakrishnan and JV Satyanarayana. Reconstruction of EEG from limited channel acquisition using estimated signal correlation. *Biomedical Signal Processing and Control*, 27:164–173, 2016. pages 57, 58

Santiago Ramón y Cajal. The Croonian lecture — la fine structure des centres nerveux. *Proceedings of the Royal Society of London*, 55(331-335):444–468, 1894. page 1

Geoffrey Roeder, Luke Metz, and Durk Kingma. On linear identifiability of learned representations. In *International Conference on Machine Learning*, pages 9030–9039. PMLR, 2021. page 111

Cédric Rommel, Thomas Moreau, Joseph Paillard, and Alexandre Gramfort. CADDA: Class-wise automatic differentiable data augmentation for EEG signals. *arXiv preprint arXiv:2106.13695*, 2021. page 111

F Rosenblatt. The perceptron : a probabilistic model for information storage and organization. *Psychological Review*, 65(6):386–408, 1958. page 8

Ivana Rosenzweig, András Fogarasi, Birger Johnsen, Jørgen Alving, Martin Ejler Fabricius, Michael Scherg, Miri Y Neufeld, Ronit Pressler, Troels W Kjaer, Walter van Emde Boas, et al. Beyond the double banana: improved recognition of temporal lobe seizures in long-term EEG. *Journal of Clinical Neurophysiology*, 31(1):1–9, 2014. page 56

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, 2019a. pages vii, 9, 16, 17, 19, 20, 26, 32, 48, 49, 54, 55, 62, 97, 111

Yannick Roy, Hubert Banville, Isabela Albuquerque, Alexandre Gramfort, Tiago H Falk, and Jocelyn Faubert. We need to go DEEPER: An online portal to improve reproducibility and accelerate research in deep learning for EEG analysis. Workshop on Multimodal Brain/Body-Machine Interfaces for "In-the-Wild" Experiments at IEEE SMC 2019, 2019b. page 20

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986. pages 11, 17

David Sabbagh, Pierre Ablin, Gaël Varoquaux, Alexandre Gramfort, and Denis A Engemann. Manifold-regression to predict from MEG/EEG brain signals without source modeling. In *Advances in Neural Information Processing Systems*, pages 7323–7334, 2019. pages 63, 97

David Sabbagh, Pierre Ablin, Gaël Varoquaux, Alexandre Gramfort, and Denis A Engemann. Predictive regression modeling with MEG/EEG: from source power to signals and cognitive states. *NeuroImage*, page 116893, 2020. pages 5, 48, 56, 58, 61, 62, 80, 87, 90, 97, 99, 140

Aaqib Saeed, David Grangier, Olivier Pietquin, and Neil Zeghidour. Learning from heterogeneous EEG signals with differentiable channel reordering. *arXiv preprint arXiv:2010.13694*, 2020. pages 79, 83

Simanto Saha and Mathias Baumert. Intra-and inter-subject variability in EEG-based sensorimotor brain computer interface: a review. *Frontiers in computational neuroscience*, 13:87, 2020. page 7

Julian Salazar, Davis Liang, Zhiheng Huang, and Zachary C Lipton. Invariant representation learning for robust deep networks. In *Workshop on Integration of Deep Learning Theories, NeurIPS*, 2018. page 82

Pritam Sarkar and Ali Etemad. Self-supervised ECG representation learning for emotion recognition. *arXiv preprint arXiv:2002.03898*, 2020. pages 26, 28, 47, 51

Brian Scally, Melanie Rose Burke, David Bunce, and Jean-Francois Delvenne. Resting-state EEG power and connectivity are associated with alpha peak frequency slowing in healthy aging. *Neurobiology of Aging*, 71:149–155, 2018. ISSN 0197-4580. doi: https://doi.org/10.1016/j.neurobiolaging.2018.07.004. URL https://www.sciencedirect.com/science/article/pii/S0197458018302550. page 104

J-B Schiratti, Jean-Eudes Le Douget, Michel Le van Quyen, Slim Essid, and Alexandre Gramfort. An ensemble learning approach to detect epileptic seizures from long intracranial EEG recordings. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 856–860. IEEE, 2018. pages 62, 67, 96, 99

R. Schirrmeister, L. Gemein, K. Eggensperger, F. Hutter, and T. Ball. Deep learning with convolutional neural networks for decoding and visualization of EEG pathology. In *2017 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pages 1–7, 2017. doi: 10.1109/SPMB.2017.8257015. pages 58, 62, 74, 79, 83

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*, 8 2017. ISSN 1097-0193. doi: 10.1002/hbm.23730. URL http://dx.doi.org/10.1002/hbm.23730. pages 16, 20, 24, 28, 33, 36, 51, 55, 58, 62, 97, 99

Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015. ISSN 18792782. doi: 10.1016/j.neunet.2014.09.003. pages 11, 139

Donald L. Schomer and Fernando Lopes Da Silva. *Niedermeyer's electroencephalography: basic principles, clinical applications, and related fields*. Lippincott Williams & Wilkins, 2012. page 5

Hartmut Schulz. Rethinking sleep analysis comment on the AASM manual for the scoring of sleep and associated events. *Journal of Clinical Sleep Medicine*, 4(02): 99–103, 2008. page 49

Johanna F. A. Schwarz, Torbjörn Åkerstedt, Eva Lindberg, Georg Gruber, Håkan Fischer, and Jenny Theorell-Haglöw. Age affects sleep microstructure more than sleep macrostructure. *Journal of Sleep Research*, 26(3):277–287, 2017. doi: https://doi.org/10.1111/jsr.12478. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/jsr.12478. pages 94, 95, 96

Skipper Seabold and Josef Perktold. statsmodels: Econometric and statistical modeling with Python. In *9th Python in Science Conference*, 2010. page 99

Michael L Seltzer, Dong Yu, and Yongqiang Wang. An investigation of deep neural networks for noise robust speech recognition. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7398–7402. IEEE, 2013. page 82

Dmitriy Serdyuk, Kartik Audhkhasi, Philémon Brakel, Bhuvana Ramabhadran, Samuel Thomas, and Yoshua Bengio. Invariant representations for noisy speech recognition. *arXiv:1612.01928*, 2016. page 82

Meredith A Shafto, Lorraine K Tyler, Marie Dixon, Jason R Taylor, James B Rowe, Rhodri Cusack, Andrew J Calder, William D Marslen-Wilson, John Duncan, Tim Dalgleish, et al. The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) study protocol: a cross-sectional, lifespan, multidisciplinary examination of healthy cognitive ageing. *BMC neurology*, 14(1):204, 2014. pages 26, 111

SJM Smith. EEG in the diagnosis, classification, and management of patients with epilepsy. *Journal of Neurology, Neurosurgery & Psychiatry*, 76(suppl 2):ii2–ii7, 2005. pages 5, 32, 62

Jordi Sole-Casals, Cesar Federico Caiafa, Qibin Zhao, and Adrzej Cichocki. Brain-computer interface with corrupted EEG data: a tensor completion approach. *Cognitive Computation*, 10(6):1062–1074, 2018. pages 57, 58

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014. page 61

Haoqi Sun, Luis Paixao, Jefferson T. Oliva, Balaji Goparaju, Diego Z. Carvalho, Kicky G. van Leeuwen, Oluwaseun Akeju, Robert J. Thomas, Sydney S. Cash, Matt T. Bianchi, and M. Brandon Westover. Brain age from the electroencephalogram of sleep. *Neurobiology of Aging*, 74:112–120, 2019. ISSN 0197-4580. doi: https://doi.org/10.1016/j.neurobiolaging.2018.10.016. URL https://www.sciencedirect.com/science/article/pii/S0197458018303804. pages 87, 90, 94, 96, 97, 103, 104, 109

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. pages 51, 58, 83

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112, 2014. page 15

Mats Svantesson, Hakan Olausson, Anders Eklund, and Magnus Thordstein. Virtual EEG-electrodes: Convolutional neural networks as a method for upsampling or restoring channels. *bioRxiv*, 2020. pages 56, 58

Jafar Tanha, Maarten van Someren, and Hamideh Afsarmanesh. Semi-supervised self-training for decision tree classifiers. *International Journal of Machine Learning and Cybernetics*, 8(1):355–370, 2017. page 41

Samu Taulu, Matti Kajola, and Juha Simola. Suppression of interference and artifacts by the signal space separation method. *Brain Topography*, 16(4):269–275, 2004. page 58

Pierre Thodoroff, Joelle Pineau, and Andrew Lim. Learning robust features using deep learning for automatic seizure detection. In *Machine learning for Healthcare Conference*, pages 178–190, 2016. pages 55, 58

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. URL http://stacks.iop.org/1741-2552/8/i=2/a=025005. page 6

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning, 2020. page 48

Tim M. Tierney, Niall Holmes, Stephanie Mellor, José David López, Gillian Roberts, Ryan M. Hill, Elena Boto, James Leggett, Vishal Shah, Matthew J. Brookes, Richard Bowtell, and Gareth R. Barnes. Optically pumped magnetometers: From quantum origins to multi-channel magnetoencephalography. *NeuroImage*, 199:598–608, 2019. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2019.05.063. URL https://www.sciencedirect.com/science/article/pii/S1053811919304550. page 3

Igor Timofeev, Sarah F Schoch, Monique K LeBourgeois, Reto Huber, Brady A Riedner, and Salome Kurth. Spatio-temporal properties of sleep slow waves and implications for development. *Current Opinion in Physiology*, 15:172–182, 2020. ISSN 2468-8673. doi: https://doi.org/10.1016/j.cophys.2020.01.007. URL https://www.sciencedirect.com/science/article/pii/S2468867320300134. Physiology of sleep. pages 94, 95

Marius Tröndle, Tzvetan Popov, Andreas Pedroni, Christian Pfeiffer, Zofia Barańczuk-Turska, and Nicolas Langer. Decomposing age effects in EEG alpha power. *bioRxiv*, 2021. pages 104, 105

Joseph Turian, Lev Ratinov, and Yoshua Bengio. Word representations: a simple and general method for semi-supervised learning. In *Association for Computational Linguistics Annual Meeting*, pages 384–394. Association for Computational Linguistics, 2010. page 27

Péter P Ujma, Boris Nikolai Konrad, Lisa Genzel, Annabell Bleifuss, Péter Simor, Adrián Pótári, János Körmendi, Ferenc Gombos, Axel Steiger, Róbert Bódizs, et al. Sleep spindles and intelligence: evidence for a sexual dimorphism. *Journal of Neuroscience*, 34(49):16358–16368, 2014. page 94

Péter P. Ujma, Péter Simor, Axel Steiger, Martin Dresler, and Róbert Bódizs. Individual slow-wave morphology is a marker of aging. *Neurobiology of Aging*, 80:71–82, 2019. ISSN 0197-4580. doi: https://doi.org/10.1016/j.neurobiolaging.2019.04.002. URL https://www.sciencedirect.com/science/article/pii/S0197458019301083. pages 94, 95

Mikko A. Uusitalo and Risto J. Ilmoniemi. Signal-space projection method for separating MEG or EEG into components. *Medical & Biological Engineering & Computing*, 35(2):135–140, 1997. doi: 10.1007/BF02534144. page 56

Raphael Vallat and Matthew P Walker. An open-source, high-performance tool for automated sleep staging. *eLife*, 10:e70092, 10 2021. ISSN 2050-084X. doi: 10.7554/eLife.70092. pages 94, 95, 99

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. pages 25, 27, 31, 32, 33, 34, 47, 48, 50, 51, 140

Marjolein MLJZ Vandenbosch, Dennis van't Ent, Dorret I Boomsma, Andrey P Anokhin, and Dirk JA Smit. EEG-based age-prediction models as stable and heritable indicators of brain maturational level in children and adolescents. *Human brain mapping*, 40(6):1919–1926, 2019. page 87

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017. pages 15, 59, 112

Bradley Voytek and Robert T Knight. Dynamic network communication as a unifying neural basis for cognition, development, aging, and disease. *Biological psychiatry*, 77 (12):1089–1097, 2015. page 105

Bradley Voytek, Mark A Kramer, John Case, Kyle Q Lepage, Zechari R Tempesta, Robert T Knight, and Adam Gazzaley. Age-related changes in 1/f neural electrophysiological noise. *Journal of Neuroscience*, 35(38):13257–13265, 2015. page 87

Glenn D Walters. Dementia: Continuum or distinct entity? *Psychology and aging*, 25 (3):534, 2010. page 50

Zitong Wan, Rui Yang, Mengjie Huang, Nianyin Zeng, and Xiaohui Liu. A review on transfer learning in EEG signal analysis. *Neurocomputing*, 421:1–14, 2021. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2020.09.017. URL https://www.sciencedirect.com/science/article/pii/S0925231220314223. page 7

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 *International Conference on Multimedia Modeling*, pages 82–93. Springer, 2018. pages 55, 58

Antoine Weihs, Stefan Frenzel, Katharina Wittfeld, Anne Obst, Beate Stubbe, Mohamad Habes, András Szentkirályi, Klaus Berger, Ingo Fietze, Thomas Penzel, et al. Associations between sleep apnea and advanced brain aging in a large-scale population study. *Sleep*, 44(3):zsaa204, 2021. page 87

Michael A West. Meditation and the EEG. *Psychological medicine*, 10(2):369–375, 1980. page 107

Cassandra M Wilkinson, Jennifer I Burrell, Jonathan WP Kuziek, Sibi Thirunavukkarasu, Brian H Buck, and Kyle E Mathewson. Predicting stroke severity with a 3-min recording from the muse portable EEG system for rapid diagnosis of stroke. *Scientific Reports*, 10(1):1–11, 2020. pages 66, 90

Irene Winkler, Stefan Haufe, and Michael Tangermann. Automatic classification of artifactual ICA-components for artifact removal in EEG signals. *Behavioral and Brain Functions*, 7(1):30, 2011. pages 56, 58

Laurenz Wiskott and Terrence J Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, 14(4):715–770, 2002. pages 28, 140

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. pages 16, 55, 57, 79, 112, 140

Xiong Xiao, Shinji Watanabe, Eng Siong Chng, and Haizhou Li. Beamforming networks using spatial covariance features for far-field speech recognition. In *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pages 1–6. IEEE, 2016a. page 82

Xiong Xiao, Shinji Watanabe, Hakan Erdogan, Liang Lu, John Hershey, Michael L Seltzer, Guoguo Chen, Yu Zhang, Michael Mandel, and Dong Yu. Deep beamforming networks for multi-channel speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5745–5749. IEEE, 2016b. page 82

Junjie Xu, Yaojia Zheng, Yifan Mao, Ruixuan Wang, and Wei-Shi Zheng. Anomaly detection on electroencephalography with self-supervised learning. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 363–368. IEEE, 2020a. page 24

Lichao Xu, Minpeng Xu, Yufeng Ke, Xingwei An, Shuang Liu, and Dong Ming. Cross-dataset variability problem in EEG decoding with deep learning. *Frontiers in Human Neuroscience*, 14, 2020b. page 49

David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd annual meeting of the association for computational linguistics*, pages 189–196, 1995. pages 18, 36

Elissa Ye, Haoqi Sun, Michael J Leone, Luis Paixao, Robert J Thomas, Alice D Lam, and M Brandon Westover. Association of sleep electroencephalography-based brain age index with dementia. *JAMA network open*, 3(9):e2017357–e2017357, 2020. pages 87, 103

Jianan Ye, Qinfeng Xiao, Jing Wang, Hongjun Zhang, and Jiaoxue Deng. CoSleep: A multi-view representation learning framework for self-supervised learning of sleep stage classification. *IEEE Signal Processing Letters*, 2021. page 24

Dani Yogatama, Chris Dyer, Wang Ling, and Phil Blunsom. Generative and discriminative text classification with recurrent neural networks. *arXiv preprint arXiv:1703.01898*, 2017. page 14

Magdy Younes, Jill Raneri, and Patrick Hanly. Staging sleep in polysomnograms: Analysis of inter-scorer variability. *Journal of Clinical Sleep Medicine*, 12(06):885–894, 2016. doi: 10.5664/jcsm.5894. URL https://jcsm.aasm.org/doi/abs/10.5664/jcsm.5894. pages 25, 32, 49

Ye Yuan and Kebin Jia. FusionAtt: Deep fusional attention networks for multi-channel biomedical signals. *Sensors*, 19(11):2429, 2019. page 82

Ye Yuan, Guangxu Xun, Qiuling Suo, Kebin Jia, and Aidong Zhang. Wave2Vec: Learning deep representations for biosignals. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 1159–1164, 11 2017. doi: 10.1109/ICDM.2017.155. pages 26, 28, 47

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. In *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pages 206–209. IEEE, 2018. page 82

Ye Yuan, Kebin Jia, Fenglong Ma, Guangxu Xun, Yaqing Wang, Lu Su, and Aidong Zhang. A hybrid self-attention deep learning framework for multivariate sleep stage classification. *BMC Bioinformatics*, 20(16):586, 2019. pages 82, 83

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. doi: 10.6084/m9.figshare.4751320. page 16

Dalin Zhang, Kaixuan Chen, Debao Jian, and Lina Yao. Motor imagery classification via temporal attention cues of graph embedded eeg signals. *IEEE Journal of Biomedical and Health Informatics*, 2020. page 48

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, 05 2018. ISSN 1527-974X. doi: 10.1093/jamia/ocy064. URL https://doi.org/10.1093/jamia/ocy064. pages 19, 23, 26, 111

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. page 14

Xiaojin Zhu and Zoubin Ghahramani. Learning from labeled and unlabeled data with label propagation. 2002. page 18

A

Sommaire récapitulatif en français

Au cours des dernières décennies, les avancées révolutionnaires en neuroimagerie ont permis de considérablement améliorer notre compréhension du cerveau. Aujourd'hui, avec la disponibilité croissante des dispositifs personnels de neuroimagerie portables, tels que l'EEG mobile « à bas prix », une nouvelle ère s'annonce où cette technologie n'est plus limitée aux laboratoires de recherche ou aux contextes cliniques. Dans le cadre de cette thèse, nous appelons ce nouveau paradigme « EEG en conditions réelles », afin de souligner le fait que la technologie n'est plus confinée à ces contextes traditionnels et peut désormais être utilisée pour diverses applications dans la vie courante.

Ces appareils EEG grand public ne disposent généralement que de quelques canaux, sont sans fil, utilisent des électrodes sèches et sont disponibles à un prix largement inférieur à celui d'appareils EEG traditionnels. Grâce à ce prix abordable et à leur facilité d'utilisation, ils permettent d'enregistrer l'activité cérébrale à domicile ou là où l'infrastructure médicale ou de recherche est absente, en vue d'applications telles que le monitorage du sommeil, le dépistage de pathologies, le neurofeedback et l'interfaçage cerveau-ordinateur (Mihajlović et al., 2014; Kreuzer, 2017; Krigolson et al., 2017; Johnson and Picard, 2020; Hohmann et al., 2020; Krigolson et al., 2021; Mikkelsen et al., 2021). Cette technologie permet ainsi de collecter des quantités sans précédent de données d'EEG auprès de populations variées à travers le monde, ouvrant la voie à de nouvelles manières de mesurer et de faire le suivi de la santé neurophysiologique de ses utilisateurs.

Les applications de l'EEG en conditions réelles présentent cependant leur lot de défis, de la rareté des données étiquetées à la qualité imprévisible des signaux et leur résolution spatiale limitée. Dans cette thèse, nous nous appuyons sur l'apprentissage profond (LeCun et al., 2015; Schmidhuber, 2015; Goodfellow et al., 2016), un sous-domaine de l'apprentissage automatique qui étudie l'apprentissage de représentations hiérarchiques avec des réseaux de neurones artificiels, afin de transformer cette modalité d'imagerie cérébrale centenaire, axée sur la recherche et les applications en clinique, en une technologie pratique qui peut bénéficier à l'individu au quotidien. Trois contributions principales sont présentées, portant sur 1) l'apprentissage auto-supervisé pour l'entraînement de modèles prédictifs avec des données d'EEG non étiquetées, 2) la robustesse au bruit et à la corruption de canaux d'EEG dans des modèles prédictifs avec des mécanismes attentionnels et 3) la mesure de la santé neurophysiologique avec l'EEG en conditions réelles grâce au concept d'âge cérébral.

Tout d'abord, nous étudions comment les données d'EEG non étiquetées peuvent être mises à profit via l'apprentissage auto-supervisé (Jing and Tian, 2021) pour améliorer la performance d'algorithmes d'apprentissage entraînés sur des tâches cliniques courantes. Nous présentons trois approches auto-supervisées qui s'appuient sur la structure temporelle des données elles-mêmes, plutôt que sur des étiquettes souvent difficiles à obtenir, pour apprendre des représentations pertinentes à des tâches de classification des stades du sommeil et la détection de pathologies : le positionnement relatif ("re-

lative positioning"), le brassage temporel ("temporal shuffling"), et le codage prédictif contrastif ("contrastive predictive coding") (van den Oord et al., 2018). Ces trois approches reposent intuitivement sur le fait qu'une bonne représentation de l'EEG devrait changer *lentement* (Földiák, 1991; Becker, 1993; Wiskott and Sejnowski, 2002), puisque des fenêtres d'EEG proches dans le temps partagent normalement la même étiquette. Nous étudions les propriétés de ces méthodes auto-supervisées par le biais d'expériences sur deux grands ensembles de données publiques, contenant des milliers d'enregistrements (López et al., 2017; Ghassemi et al., 2018; Goldberger et al., 2000) nous permettant d'effectuer des comparaisons avec des approches purement supervisées ou basées sur l'extraction manuelle de traits caractéristiques (Gemein et al., 2020). Premièrement, nous démontrons l'utilité de l'apprentissage auto-supervisé lorsque peu d'étiquettes sont disponibles : en effet, la performance de modèles de classification linéaires entraînés sur les représentations apprises par auto-supervision est supérieure à celle de réseaux de neurones profonds purement supervisés. Deuxièmement, nous explorons les représentations apprises avec les méthodes auto-supervisées, qui démontrent des structures latentes liées à des phénomènes physiologiques et cliniques, tels que les stades du sommeil et le vieillissement. Nos résultats suggèrent que l'auto-supervision peut ouvrir la voie à une utilisation plus répandue des modèles d'apprentissage profond sur les données d'EEG dans un scénario semi-supervisé.

Ensuite, nous explorons des techniques pouvant améliorer la robustesse des réseaux de neurones aux fortes sources de bruit souvent présentes dans l'EEG enregistré en conditions réelles. Nous présentons le Filtrage Spatial Dynamique ("Dynamic Spatial Filtering"), un mécanisme attentionnel (Hu et al., 2018; Woo et al., 2018) qui permet à un réseau de neurones de dynamiquement concentrer son traitement sur les canaux EEG les plus instructifs tout en minimisant l'apport des canaux corrompus. Cette approche peut être intuitivement interprétée comme une manière de faciliter la détection et l'interpolation de canaux corrompus par un réseau de neurones, à l'image des techniques traditionnelles de traitement de bruit en EEG (Perrin et al., 1989; Jas et al., 2017), mais de manière entièrement guidée par les données. Des expériences sur des ensembles de données d'EEG à peu de canaux comprenant des milliers d'enregistrements, ainsi que sur des données du monde réel, démontrent que notre module attentionnel, en combinaison avec la corruption de canaux aléatoires pendant l'entraînement (une forme d'augmentation de données), gère mieux la corruption que des approches de classification robustes au bruit (Engemann et al., 2018b; Sabbagh et al., 2020) avec ou sans traitement automatisé du bruit (Jas et al., 2017). De plus, notre module attentionnel permet d'inspecter le fonctionnement de réseaux de neurones en évaluant l'importance de chacun des canaux d'EEG aux prédictions du modèle à partir des cartes d'attention prédites. Globalement, ces résultats attestent de l'utilité du Filtrage Spatial Dynamique pour l'apprentissage sur des données d'EEG en conditions réelles.

Enfin, nous explorons l'utilisation d'étiquettes faibles afin de développer un biomarqueur de la santé neurophysiologique à partir d'EEG collecté dans le monde réel. Pour ce faire, nous transposons à ces données d'EEG le principe d'*âge cérébral*, originellement développé avec l'imagerie par résonance magnétique en laboratoire et en clinique (Franke et al., 2010, 2012; Al Zoubi et al., 2018; Sabbagh et al., 2020). En comparant l'âge chronologique d'un individu à l'âge prédit par un modèle recevant uniquement des données de neuro-imagerie et entraîné sur une population saine, il est possible d'identifier les individus dont le cerveau semble « plus vieux » que celui d'autres personnes du même groupe d'âge, ce qui suggère un vieillissement prématuré ou pathologique (Cole

and Franke, 2017; Cole et al., 2018). Nous validons ce concept par le biais d'expériences sur l'EEG de plus d'un millier d'individus enregistré pendant un exercice d'attention focalisée ou le sommeil nocturne. Nos résultats démontrent non seulement que l'âge peut être prédit à partir de l'EEG portable collecté en conditions réelles, mais aussi que ces prédictions encodent des informations contenues dans des biomarqueurs de santé neurophysiologique extraits à partir de l'EEG du sommeil, mais absentes dans l'âge chronologique. Ceci suggère que l'âge cérébral peut être utilisé pour évaluer et faire le suivi de la santé neurophysiologique d'un individu. Sur un sous-ensemble d'enregistrements longitudinaux, nous explorons de plus la stabilité de l'âge cérébral sur une échelle de plusieurs mois, et identifions un niveau de variabilité qui pourrait ouvrir la porte à une analyse plus précise de la santé cérébrale.

Dans l'ensemble, cette thèse franchit un pas de plus vers l'utilisation de l'EEG pour le suivi neurophysiologique en dehors des contextes de recherche et cliniques traditionnels. Tout particulièrement, les résultats présentés dans cette thèse démontrent que l'apprentissage profond est une approche prometteuse pour faciliter le développement de nouvelles applications de l'EEG en condtions réelles, permettant à la fois de mettre à profit de grandes quantités de données non étiquetées, de favoriser la robustesse au bruit, et d'améliorer la performance sur diverses tâches prédictives.

université PARIS-SACLAY

ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre : Apprentissage profond pour la mise en application de l'EEG en conditions réelles

Mots clés : Apprentissage profond, apprentissage de représentations, apprentissage auto-supervisé, électroencéphalographie, neuroimagerie, neurotechnologie portable

Résumé : Au cours des dernières décennies, les avancées révolutionnaires en neuroimagerie ont permis de considérablement améliorer notre compréhension du cerveau. Aujourd'hui, avec la disponibilité croissante des dispositifs personnels de neuroimagerie portables, tels que l'EEG mobile « à bas prix », une nouvelle ère s'annonce où cette technologie n'est plus limitée aux laboratoires de recherche ou aux contextes cliniques. Les applications de l'EEG dans le « monde réel » présentent cependant leur lot de défis, de la rareté des données étiquetées à la qualité imprévisible des signaux et leur résolution spatiale limitée. Dans cette thèse, nous nous appuyons sur le domaine de l'apprentissage profond afin de transformer cette modalité d'imagerie cérébrale centenaire, purement clinique et axée sur la recherche, en une technologie pratique qui peut bénéficier à l'individu au quotidien.

Tout d'abord, nous étudions comment les données d'EEG non étiquetées peuvent être mises à profit via l'apprentissage auto-supervisé pour améliorer la performance d'algorithmes d'apprentissage entraînés sur des tâches cliniques courantes. Nous présentons trois approches auto-supervisées qui s'appuient sur la structure temporelle des données elles-mêmes, plutôt que sur des étiquettes souvent difficiles à obtenir, pour apprendre des représentations pertinentes aux tâches cliniques étudiées. Par le biais d'expériences sur des ensembles de données à grande échelle d'enregistrements de sommeil et d'examens neurologiques, nous démontrons l'importance des représentations apprises, et révélons comment les données non étiquetées peuvent améliorer la performance d'algorithmes dans un scénario semi-supervisé.

Ensuite, nous explorons des techniques pouvant assurer la robustesse des réseaux de neurones aux fortes sources de bruit souvent présentes dans l'EEG hors laboratoire. Nous présentons le Filtrage Spatial Dynamique, un mécanisme attentionnel qui permet à un réseau de dynamiquement concentrer son traitement sur les canaux EEG les plus instructifs tout en minimisant l'apport des canaux corrompus. Des expériences sur des ensembles de données à grande échelle, ainsi que des données du monde réel démontrent qu'avec l'EEG à peu de canaux, notre module attentionnel gère mieux la corruption qu'une approche automatisée de traitement du bruit, et que les cartes d'attention prédites reflètent le fonctionnement du réseau de neurones.

Enfin, nous explorons l'utilisation d'étiquettes faibles afin de développer un biomarqueur de la santé neurophysiologique à partir d'EEG collecté dans le monde réel. Pour ce faire, nous transposons à ces données d'EEG le principe d'âge cérébral, originellement développé avec l'imagerie par résonance magnétique en laboratoire et en clinique. À travers l'EEG de plus d'un millier d'individus enregistré pendant un exercice d'attention focalisée ou le sommeil nocturne, nous démontrons non seulement que l'âge peut être prédit à partir de l'EEG portable, mais aussi que ces prédictions encodent des informations contenues dans des biomarqueurs de santé cérébrale, mais absentes dans l'âge chronologique.

Dans l'ensemble, cette thèse franchit un pas de plus vers l'utilisation de l'EEG pour le suivi neurophysiologique en dehors des contextes de recherche et cliniques traditionnels, et ouvre la porte à de nouvelles applications plus flexibles de cette technologie.

Title : Enabling real-world EEG applications with deep learning

Keywords : Deep learning, representation learning, self-supervised learning, electroencephalography, neuroimaging, wearable neurotechnology

Abstract : Our understanding of the brain has improved considerably in the last decades, thanks to groundbreaking advances in the field of neuroimaging. Now, with the invention and wider availability of personal wearable neuroimaging devices, such as low-cost mobile EEG, we have entered an era in which neuroimaging is no longer constrained to traditional research labs or clinics. "Real-world" EEG comes with its own set of challenges, though, ranging from a scarcity of labelled data to unpredictable signal quality and limited spatial resolution. In this thesis, we draw on the field of deep learning to help transform this century-old brain imaging modality from a purely clinical- and research-focused tool, to a practical technology that can benefit individuals in their day-to-day life.

First, we study how unlabelled EEG data can be utilized to gain insights and improve performance on common clinical learning tasks using self-supervised learning. We present three such self-supervised approaches that rely on the temporal structure of the data itself, rather than onerously collected labels, to learn clinically-relevant representations. Through experiments on large-scale datasets of sleep and neurological screening recordings, we demonstrate the significance of the learned representations, and show how unlabelled data can help boost performance in a semi-supervised scenario.

Next, we explore ways to ensure neural networks are robust to the strong sources of noise often found in out-of-the-lab EEG recordings. Specifically, we present Dynamic Spatial Filtering, an attention mechanism module that allows a network to dynamically focus its processing on the most informative EEG channels while de-emphasizing any corrupted ones. Experiments on large-scale datasets and real-world data demonstrate that, on sparse EEG, the proposed attention block handles strong corruption better than an automated noise handling approach, and that the predicted attention maps can be interpreted to inspect the functioning of the neural network.

Finally, we investigate how weak labels can be used to develop a biomarker of neurophysiological health from real-world EEG. We translate the brain age framework, originally developed using lab and clinic-based magnetic resonance imaging, to real-world EEG data. Using recordings from more than a thousand individuals performing a focused attention exercise or sleeping overnight, we show not only that age can be predicted from wearable EEG, but also that age predictions encode information contained in well-known brain health biomarkers, but not in chronological age.

Overall, this thesis brings us a step closer to harnessing EEG for neurophysiological monitoring outside of traditional research and clinical contexts, and opens the door to new and more flexible applications of this technology.