Deep Learning for Electrophysiological Investigation and Estimation of Anesthetic-Induced Unconsciousness

Konstantinos Patlatzoglou

Supervisor: Dr. Srivas Chennu

Department of Computing

University of Kent

This dissertation is submitted for the degree of

Doctor of Philosophy

May 2022

I dedicate this thesis to the loving memory of my father…

Acknowledgements

I would like to acknowledge a number of people for their direct and indirect contribution to the creation of this thesis.

First and foremost, my supervisor Srivas Chennu, who gave me the opportunity to work on the field of computational neuroscience, and introduced me to the topics of consciousness and general anesthesia. I am grateful for his immense support over the academic and personal obstacles I encountered throughout my studies, his advice and guidance, and his openness to explore different research ideas. His contribution allowed me to expand my understanding on how to conduct research, and helped me grow into a better scientist. Similarly, I want to acknowledge Howard Bowman, for his valuable input, his insights on research and scientific writing, as well as for his guidance and encouragement throughout my PhD progression. My colleague Riku Ihalainen, for sharing our doctoral journey, and for all the wonderful discussions around consciousness and the brain. I am also grateful to Olivia Gosseries, Steven Laureys, and George Mashour, for their collaboration and sharing of clinical datasets, without which this work would not have been possible. To my friends and colleagues from the School of Computing, and to Sally Fincher and Rogerio de Lemos, who supported me and allowed me to study at the University of Kent. Finally, I want to thank Dr. Palaniappan Ramaswamy and Dr. Rosalyn Moran, for agreeing to examine this thesis.

I must also acknowledge several people outside my academic environment that greatly contributed to my personal life. My beloved Katerina, who shared with me this PhD journey, my highs and lows, the excitement and the concerns, over the years of my stay in UK. My friends Theo, Dimos, Nikos, Alex, Stefanos and Spyros for the great company and the wonderful discussions on science and philosophy. My past professors, Nikos Laskaris, Anastasios Tefas, Rafael Ramirez, and Paul Verschure, for inspiring me and helping me develop my interests and knowledge within neuroscience, cognitive science, and artificial intelligence. Last but not least, I want to thank my family for encouraging me to pursue this PhD, and for their support through all the years of my undergraduate and postgraduate studies.

Abstract

Neuroscience has made a number of advances in the search for the neural correlates of consciousness, but our understanding of the neurophysiological markers remains incomplete. In this work, we apply deep learning techniques to resting-state electroencephalographic (EEG) measures of healthy participants under general anesthesia, for the investigation and estimation of altered states of consciousness. Specifically, we focus on states characterized by different levels of unconsciousness and anesthetic depths, based on definitions and metrics from contemporary clinical practice.

Our experiments begin by exploring the ability of deep learning to extract relevant electrophysiological features, under a cross-subject decoding task. As there is no state-of-the-art model for EEG analysis, we compare two widely used deep learning architectures – convolutional neural networks (cNNs) and multilayer perceptrons (MLPs) – and show that cNNs perform effectively, using only one second of the raw EEG signals. Relying on cNNs, we derive a novel 3D architecture design and a standard preprocessing pipeline, which allows us to exploit the spatio-temporal structure of the EEG, as well as to integrate different acquisition systems and datasets under a common methodology. We then focus on the nature of different predictive tasks, by investigating classification and regression algorithms under a variety of clinical ground-truths, based on behavioral, pharmacological, and psychometrical evidence for consciousness. Our findings provide several insights regarding the interaction across the anesthetic states, the electrophysiological signatures, and the temporal dynamics of the models. We also reveal an optimal training strategy, based on which we can detect progressive changes in levels of unconsciousness, with higher granularity than current clinical methods. Finally, we test the generalizability of our deep learning-based EEG framework, across subjects, experimental designs, and anesthetic agents (propofol, ketamine and xenon). Our results highlight the capacity of our model to acquire appropriate, task-related, cross-study features, and the potential to discover common cross-drug features of unconsciousness.

This work has broader significance for discovering generalized electrophysiological markers that index states of consciousness, using a data-driven analysis approach. It also provides a basis for the development of automated, machine-learning driven, non-invasive EEG systems for real-time monitoring of the depth of anesthesia, which can advance patients' comfort and safety.

Contents

List of Abbreviations

AI – Artificial Intelligence

cNN – Convolutional Neural Networks

DL-EEG – Deep Learning-based Electroencephalography

DoA – Depth of Anesthesia

DoC – Disorders of Consciousness

EEG – Electroencephalography

GA – General Anesthesia

LOPOCV – Leave-One-Participant-Out Cross-Validation

PD – Pharmacodynamic

PK – Pharmacokinetic

TCI – Target-Controlled Infusion

Chapter 1 Introduction

1.1 Overview

In this chapter, we introduce our thesis with an overview of the main themes and fields of study that form the basis of our research, namely: the neurophysiological correlates of consciousness, human electroencephalography, deep learning, and general anesthesia. In the following section, we outline the structure of the thesis and summarize the content of each chapter.

1.1.1 The problem of Consciousness

Neuroscience has shown a number of advances in the search for the understanding of consciousness; a problem that started as one of the biggest mysteries in philosophy, but whose physical understanding relates directly to our everyday life, from our perception of reality, to the positive and negative experiences we have, to our own sense of self, and ultimately, to an essential part of what life and death is. Beyond our daily experiences, understanding consciousness can also have a tremendous impact in medicine, as altered states of consciousness and unconsciousness are voluntarily induced in patients for therapeutic purposes (e.g. under general anesthesia to perform a painful procedure, or as part of a psychiatric treatment), or have been involuntarily induced due to a medical condition, such as a disease or a brain injury (e.g. in patients with disorders of consciousness under a comatose, vegetative or locked-in state). For all these implications, whether philosophical and scientific, or social and ethical in nature, a neuroscientific study of consciousness has a great potential to improve the quality of life for many people within our society, and particularly for patients, through diagnosis, prognosis, and treatment.

1.1.2 Neurophysiological Signatures of Levels of Unconsciousness

During the past decades, and given the development of brain imaging technologies, a scientific endeavour for consciousness studies has emerged, which may allow us to better grasp and influence this reality. Due to the current lack of a scientific theory, many researchers have been driven to investigate the neural mechanisms underlying the various states of

consciousness, by combining behavioral, clinical, and brain imaging methodologies in humans and other animals. One of the main research questions, regards the identification of the full neural correlates of consciousness (full NCC) – defined as the neuronal substrates or mechanisms that correspond to our irrespective-of-content, general ability to experience (alternatively defined as the set of all content-specific NCC). However, while research has revealed a number of potentially associated structures and mechanisms responsible for the generation of our experiences, various proposed neurophysiological markers of consciousness have proved illusory (Koch *et al.* 2016).

In this work, we are focusing towards the engineering of a model for the investigation and estimation of states of consciousness, that reflect the electrophysiological signatures of the full NCC, and which allow the discovery of possible novel markers. Specifically, our exploration focuses on states characterized by different levels of unconsciousness, a concept aligned closely with the medical definition of consciousness (a continuum of states, from full alertness and comprehension, to drowsiness and disorientation, to delirium and loss of communication, and finally loss of movement and response to painful stimuli). Using state-based paradigms, and by incorporating minimal assumptions from clinical ground-truths, our research relies on a data-driven explorative study, rather than any prior hypotheses on the nature of consciousness and its possible mechanisms. The development of such a model has the potential to contribute not only to our theoretical understanding of the phenomenon, but also to the production of automated systems for clinical use (e.g. in cases such as monitoring the depth of anesthesia (DoA), or in diagnosis and prognosis of patients with disorders of consciousness (DoC)).

1.1.3 Electroencephalography under General Anesthesia

For such inquiry, experimental designs with electroencephalographic (EEG) measures of humans under various states and depths of anesthesia can provide us with a suitable set-up, and a reliable ground truth, given the desired association. General anesthesia (GA) provides a powerful paradigm to study altered states and levels of consciousness using a variety of anesthetics, which albeit having different molecular mechanisms, share many of their phenomenological effects (most evidently, the decrease in awareness and responsiveness of an individual, up to the eventual absence of any possible experience). While various anesthetic states and brain mechanisms are involved in different agents and doses (e.g. dream-like states of disconnected consciousness), the collection of data from multiple studies can help us resolve a number of open questions in GA research. Some of the major questions relate to our ability to sensitively distinguish the different anesthetic depths, the understanding of the dose-response relationship, the inter-individual variability of clinical outcomes, the transitional phases and asymmetries of anesthesia, as well as the discovery of any common cross-drug mechanisms of anesthetic-induced unconsciousness (all of which can help us better understand the full NCC) (Bonhomme *et al.* 2019).

When it comes to brain imaging, EEG provides a non-invasive way to record and analyze the electrophysiological signatures of the brain, of which we have acquired significant insights; from the underlying biophysics that generate the electrical fields, to the functional interpretation of the respective spatio-temporal dynamics (Buzsáki, Anastassiou and Koch 2012). It is also a valuable tool that can offer us a high temporal resolution of the brain's function, while being simple, convenient, and widely accessible in hospitals for clinical use (Chatelle *et al.* 2012). Nevertheless, there are several issues and constraints when trying to analyze EEG, due to the complex nature of the signals. Specifically, the contamination of the EEG with various sources of noise (biological or environmental), the significant intra-subject and inter-subject variability of the signals, and the lack of standardized processing methodologies, have all created problems of analysis and reproducibility (whilst human expertise and manual intervention are often required). Meanwhile, recent research has suggested that multivariate pattern analysis techniques can be very effective in detecting subtle changes within the otherwise rich electrophysiological signals, which allows for the decoding of more complex (and possibly hidden) brain states (Stokes, Wolff and Spaak 2015). Particularly during the past years, a need has emerged for the development of more sophisticated techniques of EEG analysis, beyond the traditional methodologies that have been used throughout research and clinical settings.

1.1.4 Deep Learning for EEG Decoding

Neuroscience and artificial intelligence (AI) have long benefited from strong mutual interactions. Acquiring knowledge from complex high-dimensional data, such as EEG, is a key challenge for biomedicine in general, with traditional approaches of hand-crafted feature engineering, alongside expert domain knowledge, reaching a limit (observed in contemporary systems' performance). Recent developments in the field of machine learning have shown its significant capabilities for EEG analysis, in all kinds of tasks and datasets (Roy *et al.* 2019; Pedregosa *et al.* 2012; Kammoun *et al.* 2022; Heilmeyer *et al.* 2019). Notably, during the past five years, deep learning, and particularly convolutional neural networks, have been successfully used for EEG decoding, by offering an end-to-end learning approach and by creating state-of-the-art models. These models are derived solely from the EEG data and improve their performance progressively with experience, as they discover generalized properties of the data. Trained under a specific task, artificial neural networks construct internal representations, which despite their predictive power, also open the possibility for a data-driven approach in acquiring (neuroscientific) knowledge.

Despite their strong advantages, there are several difficulties encountered in the creation, validation, and interpretability of deep learning models. These relate to their underlying mathematical properties and assumptions (through the architectural designs and imposed optimization problems), the limitations of collecting large, representative, and unbiased datasets, as well as our own cognition and disposition on how we understand and

assess complex phenomena. To work towards addressing these issues, different deep learning architectures, EEG representations, learning tasks, and anesthesia datasets are investigated through a series of research questions and experiments. Specifically, our investigation focuses on the appropriateness of the models to extract useful electrophysiological patterns, evaluated on their ability to impartially predict anesthetic-induced states and levels of unconsciousness.

1.1.5 Clinical Anesthesia and Contemporary Problems

From a clinical perspective, discovering robust neurophysiological markers that predict different states and levels of unconsciousness, can significantly improve problems found in contemporary anesthesia practice. To date, more than 200 million people around the globe have surgery every year, and mostly under GA (Weiser *et al.* 2008). An optimal anesthetic state for GA could be described as the minimum required medications that ensure unconsciousness, analgesia (calibrated as a function of noxious stimuli), amnesia, and most often immobility. Nevertheless, several problems arise from the significant inter-individual variability for the required anesthetic doses, due to a variety of pharmacological and biological factors (patient-specific characteristics) (Absalom *et al.* 2009).

Currently, levels of unconsciousness are assessed by various indirect physiological monitoring methods (e.g. ECG, blood pressure, respiratory rate, oxygen saturation, etc.), pharmacokinetic/pharmacodynamic (PK/PD) models, and the patient's ability to interact with the environment; none of which allow for an accurate estimation of the drug's effect in the brain. As a result, the administration of lower or higher doses creates several complications that emerge during or after the anesthetic procedure. Specifically, about 70% of patients are affected by under- and over-sedation in the ICUs (Kaplan and Bailey 2000). Over-sedation has been associated to hypo-perfusion of the heart and brain, prolonged ventilation/recovery, and rarely, even to induction of coma (Gunaydin and Babacan 1998). Under-sedation has been associated to pain, agitation, tachycardia, and arrhythmias. Most importantly, under-dosing can result in unintended intraoperative awareness, which occurs in up to 1% of all surgical patients (as indicated by implicit or explicit post-operative reports) (Avidan *et al.* 2011). In such cases, consciousness is preserved in a patient who is disconnected from the environment and cannot communicate (typically due to a neuromuscular blocking agent), yet still able to experience, causing confusion, potential pain, and likely post-traumatic stress disorder (PTSD) (Mashour and Avidan 2015). Finally, under- and over-sedation have both been linked to delirium and post-operative cognitive decline (which can last up to 6 months) (Eichhorn *et al.* 2019). For all these reasons, it is important to have a reliable measure of the depth of anesthesia (DoA), using a direct, non-invasive monitoring of the brain.

Despite the obvious fact that GA fundamentally modulates neuronal activity, brain monitoring is not routine practice in the operating room (3 - 4% usage in clinical settings (Pandit *et al.* 2014)), and is limited to proprietary systems that have produced mixed results (Avidan *et al.* 2011). Meanwhile, contemporary clinical scales are not able to reliably measure

states of consciousness (as they require behavior), or to continuously track the patient's status, which can be crucial for real-time management of patients. Ideally, by acquiring a personalized continuous brain-based estimate of the DoA, we can guide the anesthetic titration to a minimum-but-sufficient administration of the anesthetics, hence minimizing or even eliminating the above mentioned problems.

1.1.6 State-of-the-art and Learning-based EEG

The absence of a universally accepted methodology or device for tracking the brain under anesthesia is, in part of course, due to the lack of robust EEG markers. To date, there are several EEG-based devices that can be used to monitor the depth of anesthesia, using near real-time indices, such as the bispectral index (BIS) (Myles *et al.* 2004), Narcotrend (Kreuer *et al.* 2003), m-entropy (Bruhn *et al.* 2006), patient state index (PSI) (Drover *et al.* 2002), and WAVCNS (Zikov *et al.* 2006). Typically, for a given index it is assumed that a particular output value (or range of values) corresponds to a particular anesthetic state, independent of the administered agent. Nevertheless, there are several limitations with contemporary DoA monitors in their sensitivity, specificity, and value as diagnostic tools. For example, the BIS monitor – one of the commercial standards in DoA devices – has been shown to be unreliable under certain anesthetic agents (e.g. nitrous oxide, ketamine and dexmedetomidine (Barr *et al.* 1999)), anesthetic depths (e.g. burst suppression (Kearse *et al.* 1994)), and patient demographics (e.g. children (Khan *et al.* 2018)). It has also been shown to be significantly affected by signatures uncoupled from consciousness, such as the electromyographic (EMG) activity (Schuller *et al.* 2015). In general, there are known differences among these indices, and inconclusive studies about the improvement of patient outcomes under their use (Avidan *et al.* 2008; Chan *et al.* 2013; Wildes *et al.* 2019).

In spite of the limits of contemporary DoA systems, recent literature has highlighted the potential for improved brain monitoring, as we better understand the electrophysiological signatures of anesthetic states, which vary by agents, mechanisms of action and doses (Purdon *et al.* 2015; Brown, Pavone and Naranjo 2018). For example, studies using the isolated forearm technique allow us to differentiate signatures reflecting covert awareness from signatures of true unconsciousness (Linassi *et al.* 2018; Gaskell *et al.* 2017). Meanwhile, an increasing number of studies are introducing theoretically and empirically-based EEG metrics for discriminating states of consciousness, which have shown adequate generalization across a variety of anesthetic conditions (e.g. the perturbational complexity index (PCI) (Casali *et al.* 2013), the Lempel-Ziv complexity index (PLCZ) (Bai *et al.* 2015), entropy indices (Liang *et al.* 2015), slow-wave activity (SWA) (Ní Mhuircheartaigh *et al.* 2013), and others). However, many of the above metrics rely on several theoretical/mathematical assumptions, they have high computational requirements, and most often require data curation and manual intervention from experts, which makes them unsuitable for automated systems.

In this respect, deep learning has shown its strength and potential for improving healthcare, throughout a variety of diagnostic problems (and particularly within the area of medical imaging), as it offers high predictive accuracies, computational performance, and real-time automation, which is vital for clinical anesthesia. Of course, the performance of such models can be limited by the specific problem under analysis (e.g. the type of anesthetic agents used), the reliability of the ground truth (e.g. behavioral, clinical, or pharmacological evidence for levels of unconsciousness), and even our ability to interpret these systems for medical purposes (in terms of explainability, control, and potential improvement). This is particularly important, given the theoretical limitations in acquiring an electrophysiological understanding of the markers, under a learning-based approach. Therefore, it is important to consider all of these issues throughout our work, while focusing as much as possible on the aspects that our tools allow us to study.

1.2 Thesis Structure

In the previous section, we presented the conceptual framework of the fields and challenges related to our work, our general research goal, and our selected tools of investigation. Specifically, we discussed the scientific endeavour towards the identification of the neurophysiological signatures of consciousness, the selection of EEG under GA as a study paradigm, the rationale behind deep learning for EEG decoding, and finally, clinical anesthesia and contemporary problems related to DoA estimation.

In Chapter 2, we provide a short summary of the background literature regarding our main fields of study, namely: the problem of consciousness, electroencephalography, and deep learning. Each section introduces basic concepts that are used throughout our work, as well as supporting literature for our methodological underpinnings and prior assumptions (a reader familiar with any of the fields can skip the respective sections).

In Chapter 3, we begin by exploring whether deep learning is effective in extracting relevant electrophysiological features from resting-state EEG of healthy participants under GA. As there is no state-of-the-art model for EEG analysis, we compare two widely used deep learning architectures – convolutional neural networks (cNNs) and multilayer perceptrons (MLPs) – in their ability to discriminate anesthetic states, given a fully automated end-to-end learning approach, and under a cross-subject decoding task. We also investigate the effect of the models' input representation, by comparing the raw EEG time courses against a spectral representation, which is often used as an effective feature in many EEG decoding tasks. This work formed the basis of our publication in the 11th International Conference on Brain Informatics 2018 (Patlatzoglou *et al.* 2018).

In Chapter 4, we focus on the development of a cNN architecture and an EEG pre-processing pipeline, that will allow us to incorporate different EEG systems and datasets, under a common and consistent processing methodology (currently missing). For this purpose, we

explore several parameters of EEG representation, alongside a network design with the capacity to exploit the spatio-temporal structure of the EEG. Specifically, we perform a number of experiments and compare the performance of 2D and 3D cNN designs, with respect to some of the most influential EEG parameters, namely: the reference montage, the sample normalization method, the spatial and spectral resolution of the EEG, and the manipulation of EEG artifacts. We then derive a generic 3D cNN and a selected pre-processing pipeline, based on which we proceed with our research investigation.

In Chapter 5, we experiment with a number of optimization tasks undertaken by our cNN model, for the electrophysiological investigation and estimation of anesthetic-induced states of unconsciousness. Our aim here is to both understand the relationship between the EEG signatures and the various anesthetic states, but also to test the predictive power of classification and regression algorithms under particular learning tasks. To this end, we exploit the acquisition of datasets from experimental designs that control for several clinical variables, such as the anesthetic agent (propofol and ketamine) and the administration mode, which are used to target a particular behavioral, pharmacological, or psychometrical response. Based on this analysis, and in alignment with our research objective, we further derive an optimal training strategy that is shown to impartially track the depth of anesthesia. Part of this work formed the basis of our publication in the 42nd international conference of the IEEE Engineering in Medicine and Biology, 2020 (Patlatzoglou *et al.* 2020).

In Chapter 6, we test the generalizability and reproducibility of our findings within the framework of deep learning-based EEG, in order to assess the capacity of our model to impartially estimate levels of anesthetic-induced unconsciousness. Specifically, we evaluate the cross-study and cross-drug generalization performance of the model under two unseen experimental setups, that include a novel paradigm for measuring levels of unconsciousness, and two distinct anesthetic agents (ketamine and xenon). We also explore a cross-drug training strategy, with the potential to uncover common cross-drug features of unconsciousness, which have been hypothesized in GA research. Part of this work formed the basis of our publication in the 42nd international conference of the IEEE Engineering in Medicine and Biology, 2020 (Patlatzoglou *et al.* 2020).

Finally, in Chapter 7 we summarize the results and contributions of this thesis, and we assess our methodology – its strengths and its limitations – against current state-of-the-art models. Our methodological framework and findings are evaluated within the general literature of deep learning-based EEG decoding, but also with respect to other non-learning-based EEG methods for the analysis and estimation of levels of consciousness (or particularly the depth of anesthesia).

Chapter 2 Background Literature

2.1 The Problem of Consciousness

Throughout the human history, the origin of consciousness has been one of the three fundamental questions, along with the origin of life and the universe, which troubled philosophers and scientists. Within the last few centuries, the scientific revolutions of Newtonian mechanics and Darwinian evolution have led us to a profound understanding of how the universe and life emerged, for all practical and meaningful purposes. While these theories rely on a materialistic and mechanistic approach for explaining the world, born out of simple ideas and mathematical elegance, we are nowhere close to a theory of consciousness that fulfils such level of understanding yet; a theory that could explain the complexity of the phenomenon from non-complex principles. Nevertheless, during the past decades, there has been significant progress in forming and tackling the problem, with scientific models that can be tested experimentally, and empirical findings that guide our investigations. Besides the philosophical need for a formal explanation, our ability to assess consciousness and all its qualities has many implications for science and society, and particularly within the practice of medicine, where altered states of consciousness are part of diagnosis, prognosis, and treatment of patients.

2.1.1 Definition

The problem of consciousness starts with the definition itself. A definition based on the Webster's Third Dictionary states that "consciousness is the quality or state of being aware of an external object or something within oneself" (Merriam-Webster 2012). Although the term may have various meanings depending on the context, relevant definitions involve concepts such as awareness, perception, sensations, wakefulness, volition and thought; the ability to experience and feel, a sense of selfhood, and even the executive control of the mind. Despite the various definitions and debates around the topic (which remains controversial), most philosophers and scientists today agree that there is a shared intuition of what consciousness is (Honderich 2008). A simple but operational definition would be that consciousness is everything we experience, from images and sounds to emotions and thoughts. It is the most familiar and most mysterious aspect of our everyday life, as we lose it every night when we fall into dreamless sleep, yet returns when we wake up in the morning.

2.1.2 Scientific Study

For many decades the topic of consciousness was avoided by scientists, as the scientific method relies on objective measurements and experimentation, while most of consciousness definitions include subjective terms and private experiences. However, we have long known that mental processes – many of which characterize consciousness – causally relate to the physical processes of the brain. Due to the technological advancements of the past decades, and specifically the appearance of brain imaging technologies like EEG and fMRI, we have been able to systematically measure the function of the brain in individuals under various conditions of consciousness. To that end, an interdisciplinary study has grown within the cognitive sciences, with contributions from various fields, most notably psychology, neuroscience, and computer science.

The main objective of such study lies in the understanding of the connection between the behaviors (related to responsiveness, volition, and self-report) and perceptions (related to experiences and self-awareness) of individuals, with respect to their underlying biological mechanisms; namely, the connection between the behavioral/psychological and neural correlates of consciousness (Koch *et al.* 2016). An important clarification here is that the information present in consciousness cannot be ascribed directly as an emergent property (or epiphenomenon) of the neural states/activity itself, but rather to a specific mechanism or structure (notably, one of the earliest distinctions in psychology regard the limited, slow and serial conscious processes, against the fast and parallel unconscious ones). In other words, not all neuronal processes are part of our conscious experience (e.g. processes of the autonomic nervous system), while contemporary lesion studies have indicated particular brain regions that either have a direct causal relation to someone's ability to be conscious (Parvizi 2001), or have no effect at all (Lemon and Edgley 2010). Given these facts, the neural correlates of consciousness (NCC) can be defined as the minimum neuronal mechanisms jointly sufficient for a conscious percept (Koch 2004). This definition can be interpreted in two ways, depending on whether we focus on the substrates that support specific contents of experience (content-specific NCC), or the mechanisms that support an irrespective-of-content general ability to experience (full NCC). Typically, studies have been focusing on either of these two conceptual aspects, which are empirically separated into contents and levels of consciousness.

Parallel to these definitions, there are two types of experimental paradigms that have been used in research, in order to behaviorally assess contents and levels of consciousness. When it comes to identifying content-specific NCC, the majority of studies have relied on psychological investigations based upon verbal reports of experience, with no-report paradigms more recently developed, which focus on indirect physiological measures. Alternatively, the identification of the full NCC has been driven by paradigms employing medical definitions of consciousness, where clinicians and neurologists assess states and levels of consciousness by observing a patient's arousal, responsiveness, and volition. In both cases, there are various issues and phenomena of interest, including subliminal perception and

priming effects, blindsight, denial of impairment, as well as altered states of consciousness produced by sleep, meditation, trauma, illnesses, and drugs (altered states can refer to changes in thinking, sense of time, emotions, bodily image, or meaning, as in the case of dreaming). Overall, while we have several indications that these two aspects of consciousness can be partially dissociated from one another, it is important to consider their intertwined nature (i.e. contents of experience might change at different levels of consciousness, while a certain level of consciousness is required for any phenomenal content) (Mashour and Hudetz 2017).

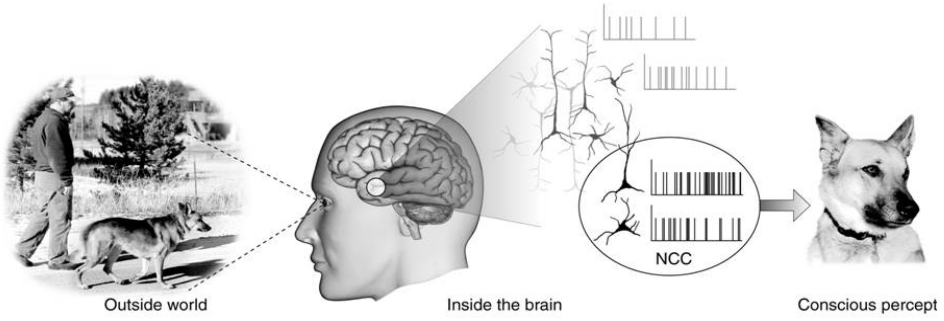

Fig. 2.1. Neural correlates of consciousness (Koch 2004). The conscious percept of an external stimulus is ascribed to a specific underlying neural mechanism or structure.

2.1.3 Behavioral Correlates of Consciousness

Due to the subjective nature of experience and the lack of a universally accepted definition, there are special difficulties in assessing the various states of consciousness. We normally infer that a person is conscious when they are awake and act purposefully, with contents of experience assessed and compared across individuals by the consistency of self-reports. For these reasons, there is a consensus among scientists regarding the assumptions made and the criterions used in experiments, with few approaches and methodologies chosen, depending on the suitability of the research interest and the phenomenon under study.

Explicit/Non-Explicit Report Paradigm. One of the most widely used methods to assess consciousness in humans is through verbal report. In this method, participants are asked to describe their experiences to a stimulus (such as visual or auditory), while the description is treated as an observation or measurement of consciousness (e.g. in a Necker cube, a participant may report alterations between the two 3D perceived configurations, despite the unchanged visual stimulus). Similar approaches have been used with button-press for simple 'yes' or 'no' questions, or other forced-choice procedures that evaluate the awareness of participants (responses may vary though, depending on the perceptual thresholds of each participant) (Kunimoto, Miller and Pashler 2001). For a more detailed way to characterize experience, perceptual awareness scales have been developed, as well as confidence ratings that may be used alongside (Sandberg *et al.* 2010). Response priming is a technique often used within this

paradigm, where the behavior or response of an individual is influenced under the presence of another (priming) stimulus (Schmidt and Vorberg 2006). All these methods provide a mean to access the content of experience, and although they are considered reliable indicators of consciousness, they raise a number of issues (Destrebecqz and Peigneux 2005). For example, measurements of consciousness could contain errors which are not detectable by the participant's behavior itself, unless third-person evidence have been acquired (such as physiological measurements) (Dennett 2003). Moreover, verbal reports are restricted to humans and exclude cases of people with language impairment, pre-linguistic children, or other animals that may be conscious, while language itself has been widely dissociated from consciousness and may mislead research investigation and results.

More recently, a no-report paradigm has been adopted by many researchers, as an alternative to the above mentioned approaches, which also tackles the issue of verbal validity. In this method, consciousness is assessed by trials of explicit report, along with trials of non-explicit report, where indirect physiological measures are used to infer the perception of an individual (Koch *et al.* 2016). For example, eye movements and pupil dilation have been shown to correlate highly with conscious reports in visual studies, and could be used as alternative measures to explicit behaviors (Frassle *et al.* 2014). Such approach allows for the differentiation of neural activity that is associated to events and processes related to the task given (or the report itself), and which may precede or follow conscious experience. As in most neuroscientific studies, the complete reduction or control of confounds is a major issue and concern for the experimental design, which can significantly influence the results, depending on its suitability for the phenomenon under study.

Medical Paradigm. Another way to assess consciousness, which better reflects our general and irrespective-of-content ability to experience, is based on the medical perspective of the concept. In clinical settings, the behavioral assessment focuses on a patient's arousal and responsiveness, where simple criteria are used to estimate the various states and levels of awareness (e.g. the ability of a patient to respond to a command, verbally or with a movement). Besides verbal report, a combination of physiological measures of arousal, brain activity and purposeful movement complement the clinical evaluation (Giacino and Smart 2007). Consciousness here can be thought as a continuum of states, from full alertness and comprehension, to drowsiness and disorientation, to delirium and loss of communication, and finally loss of movement and response to painful stimuli (Luby 1998). Each state is characterized by different levels of consciousness, measured by a standardized behavioral scale (most commonly the Glasgow Coma Scale) which typically consists of verbal, ocular, and motor response grading tests. Standardized scales have also been developed to assess levels of contents of experience, by assigning ratings to a patient's visual, auditory, verbal or motor function (Hoenig and Toakley 1959). Finally, other non-medical ways to assess levels of consciousness have been used in research settings, such as measuring the reaction times or the performance to a cognitive task.

As with self-report paradigms, the medical approach to consciousness also raises a number of issues, since arousal and purposeful movements cannot be considered as absolute indicators of consciousness. Extensive literature has shown that individuals may be wakeful and act purposefully, without reporting awareness of their actions (Schmidt and Vorberg 2006). Reports of experience can also be differentiated from actual behaviors in some cases, or even from patterns of brain activity (Haggard 2008). On the other hand, there are cases of patients with a complete absence of behavior (e.g. patients in vegetative states) which are misclassified as non-conscious, contrary to brain imaging evidence (Schnakers *et al.* 2009). For all these reasons we accept a strong dissociation, where consciousness does not require behavior, and vice versa. A variety of clinical states defined by the concepts of arousal/vigilance and awareness (levels of consciousness) are depicted in Fig. 2.2.

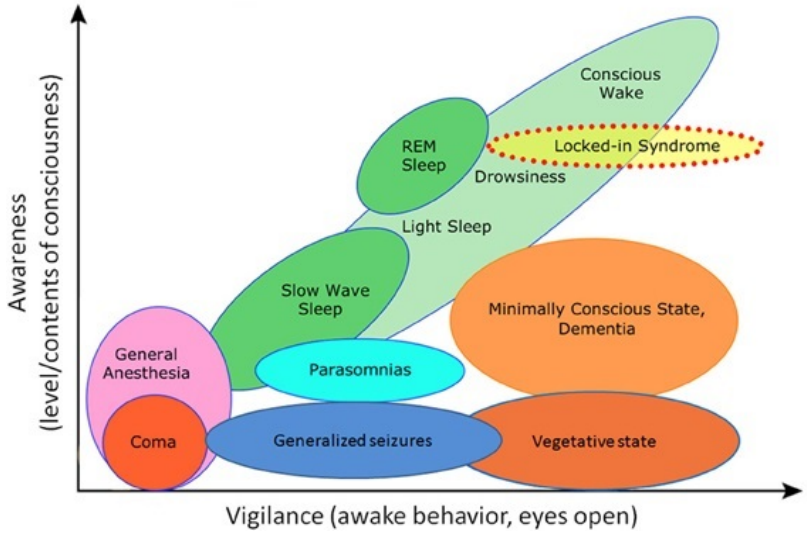

Fig. 2.2. Clinical states of consciousness, defined by the vigilance and awareness of an individual. The levels and contents of consciousness can be dissociated from the physiological arousal, which has been commonly used in medical assessment (Boly *et al.* 2013).

2.1.4 Neural Correlates of Consciousness

As already pointed out, science is primarily concerned with an explanation of what it means biologically (or physically, in general) for information to be present in consciousness. Neuroscientifically, this could either relate to a specific part of the brain, or a specific pattern of activity, where consciousness emerges from. The neural correlates of consciousness (NCC) are defined as the minimum neuronal mechanisms jointly sufficient for a conscious percept. To identify such mechanisms, different approaches and methodologies have been used in research (in parallel to the behavioral correlates of consciousness), depending on whether the investigation focuses on content-specific experiences, or the overall ability of an individual to be conscious. Here, we summarize our current understanding of the NCC, based on a number

of brain imaging studies (mostly EEG and fMRI) that have focused on a variety of phenomena and conditions.

Content-specific NCC. When it comes to the content-specific NCC, a minimum set of neurons (or neuronal mechanisms in general) is assumed to be responsible for the phenomenal distinction of a particular experience, such as the perception of a face. The activation or firing of such neurons is a necessary and unique condition for the experience of that percept, whether this activation is caused by an external stimulus, an internal process (e.g. a thought or a dream), or an artificial method (e.g. by electrical, magnetic or optogenetic stimulation) (Koch *et al.* 2016). This means that the perceptual experience can appear or disappear as a direct consequence of the activity or suppression of this mechanism, irrespective of the presence of a stimulus. To identify such neural sets, researchers use a variety of experimental designs that allow for the differentiation and comparison of brain activity, under conditions that only vary with respect to the conscious presence of a stimulus (the physical stimulus and the overall state of the participant are kept constant). For example, in report-based visual paradigms, there are several techniques to manipulate the perception of an image, such as binocular rivalry, interocular suppression, bi-stable perception and other masking techniques (similar psychological phenomena and techniques are used in audition as well). By contrasting the activity of perceived and non-perceived stimuli, a broad fronto-parietal network appears to be activated during visual-motor tasks (Koch *et al.* 2016).

Since most of these paradigms include events and processes related to the task, and which may precede or follow conscious experiences, neural activity which co-varies with the NCC is expected (Miller 2014; Jaan Aru *et al.* 2012; de Graaf, Hsieh and Sack 2012). Specifically, such activity may relate to the unconscious stimulus processing, or to a number of cognitive functions that appear during the task, such as selective attention, expectation, self-monitoring, task planning, and reporting. Although some of these functions correlate highly with perceptual experience (e.g. we often become conscious of what we attend to), consciousness does not require behavior, language, or long-term memory, and it has even been dissociated from processes such as attention (Jiang *et al.* 2006). A solution to this problem comes from the use of no-report paradigms (or other techniques such as matching performance and manipulating task relevance), which can reveal the NCC from other prerequisite brain activity (de Graaf, Hsieh and Sack 2012; J. Aru *et al.* 2012). These techniques have shifted our understanding of the content-specific NCC to a posterior cortical region, contrary to the prefrontal cortex, which is heavily involved in task monitoring and reporting (Koch *et al.* 2016).

Full NCC. The alternative interpretation of the NCC can be defined as the union of all the content-specific neural sets or mechanisms, and for all possible contents of experience; namely the full NCC. It is important to clarify here that we are only interested in neural mechanisms that contribute directly to contents of experience, rather than other biological factors that allow

these mechanisms to work properly. For example, background conditions such as the levels of oxygen and glucose in the blood, a neuromodulatory milieu and an adequate cortical excitability, are all necessary for being conscious (Koch *et al.* 2016). To identify the full NCC, state-based paradigms are used by researchers, taking advantage of the fact that altered states of consciousness (such as in sleep) change our overall ability to experience. As with the content-specific NCC, we want to differentiate and compare brain activity between states of consciousness (as seen in healthy wakeful individuals) and states of diminished consciousness, that appear in a variety of circumstances (e.g. sleep, general anesthesia, or patients with disorders of consciousness). Studies on the full NCC have revealed the role of the fronto-parietal network, in consistency to the previously mentioned content-specific approaches (Koch *et al.* 2016).

However, the problem of confounds remains when using between-states paradigms, since other cognitive functions also decline along with the levels of consciousness, when the physiology of the brain changes (e.g. levels of arousal-promoting neuromodulators will affect attention and vigilance). A solution to this problem comes from the use of within-state, no task paradigms, where spontaneous fluctuations of activity are expected to eliminate these confounds (e.g. same physiological states can be contrasted after a participant is asked whether he/she was dreaming or not, during a specific sleep phase). Notably, within-state sleep studies (in both NREM and REM states) have indicated that the full NCC appear to be localized within a posterior cortical region associated with perceptual experiences, while frontal areas activate in thought-like experiences. In addition, some contents of experience have also been associated to high-frequency activity from such posterior areas (Siclari *et al.* 2017).

Overall, both content-specific and full NCC studies provide evidence that converge to the same conclusion, namely a union of temporo-parieto-occipital cortical regions (the so called “hot zone”) responsible for consciousness.

2.1.5 Neurophysiological Markers of Consciousness

In this section, we present several candidate neurophysiological markers of consciousness that have been proposed in the literature over the past years. Specifically, we focus on markers that have been (or can be easily) detected by EEG, given its wide application as a functional neuroimaging method in research and clinical settings. As we further discuss in the next sections, the evaluation of EEG typically focuses on qualitative features of either event-related potentials (ERPs), or spontaneous recordings of activity (resting-state).

P3b. One of the earliest and well-studied electrophysiological signatures associated to consciousness, was the P3b ERP; a positive, fronto-parietal event-related potential that appears after the onset of a visual or auditory stimulus (~300 ms after the onset). This ERP component had been shown to correlate with the report of the stimulus detection, in several experimental

paradigms that incorporate masking, attentional blink and manipulations of stimulus strength (Sergent, Baillet and Dehaene 2005; Del Cul, Baillet and Dehaene 2007). However, subsequent studies have indicated that the P3b potential can be present without a conscious perception of the stimulus, whilst stimuli irrelevant to the task do not elicit a P3b response, even when participants are conscious of them (Silverstein *et al.* 2015). Similar findings have been found in vegetative and minimally conscious patients, with the presence of P3b being dissociated from their assessed state of consciousness (Sitt *et al.* 2014; Höller *et al.* 2011). Therefore, we cannot presume that P3b is a robust marker of consciousness (content-specific or general).

Gamma Synchrony. Another early discovered potential marker of consciousness was gamma synchrony; initially observed in animal studies as the synchronized neuronal discharges in the gamma range (30 – 70 Hz) located at the visual cortex, as a response to a number of visual stimuli (e.g. using luminance gratings or moving stimuli) (Gray *et al.* 1989). Long-distance gamma synchrony has also been correlated to visual consciousness in human EEG and MEG studies (Melloni *et al.* 2007), leading to the proposal that consciousness requires the synchronization of neural activity for the integration of visual features into a single experience. Nevertheless, not all perceived stimuli elicit such response, with several works showing gamma activity to be correlated to selective attention, even without the perception of the stimulus (attention can be dissociated from visual experience) (Wyart and Tallon-Baudry 2008). In addition, gamma synchrony can also persist during NREM sleep and anesthesia, where consciousness is absent (Pockett and Holmes 2009). Hence, we can infer that gamma synchrony is not a necessary condition for consciousness.

Activated EEG. One of the most useful electrophysiological markers of consciousness has been the low-amplitude/high frequency activity observed in EEG during wakefulness states, known as ‘desynchronized’ or ‘activated EEG’. This signature has been contrasted to the high-amplitude/low frequency activity observed mostly in the form of deep slow waves during physiological and pharmacological conditions of unconsciousness (i.e. sleep and GA) – also known as ‘EEG slowing’. The transition from activated EEG to EEG slowing towards the transition to unconsciousness has been further understood in animal studies, where an alteration of depolarization and hyperpolarization states of thalamic and cortical neurons, create a bursting firing mode in the delta and theta/spindle ranges, respectively (Steriade, Timofeev and Grenier 2001) (the detection of slow waves in particular has been used as one of the most effective ways to assess loss of consciousness (Ní Mhuircheartaigh *et al.* 2013)). Nonetheless, global patterns of activity are not always reliable for discriminating states of consciousness and unconsciousness. For example, slow waves have been observed in cortical regions of epileptic patients who are conscious, as well as during sleep stages that have been associated to reports of dreaming (Vuilleumier *et al.* 2000; Nobili *et al.* 2012). For these reasons, it might be important to focus on the role of localized activity changes, when assessing conditions of consciousness.

Neural Differentiation/Integration. A more recently proposed marker of consciousness is based on the notion of spatiotemporal differentiation and integration of cortical activity, as per the integrated information theory of consciousness (Oizumi, Albantakis and Tononi 2014). Similarly to 'activated EEG', this notion suggests that a large variety of activity patterns and connectivity configurations appear to be responsible for consciousness, as evident in many EEG and fMRI studies (this also explains the application of entropy and complexity measures for the assessment of the anesthetic depth, found in (Bai *et al.* 2015; Liang *et al.* 2015)). Moreover, it suggests that the synchronization of the brain as a single entity is an important requirement, given that several studies reveal a decrease in functional connectivity under states of unconsciousness, like sleep and anesthesia. On the other hand, indices of integration using EEG coherence and Granger causality can be increased in states of unconsciousness, while measures of differentiation have not always been useful at the individual level (such as the BIS index) (Koch *et al.* 2016). In general, despite some high-level evidence in favour of the theory, we still have mixed results on the ability of these measures to discriminate states of consciousness, on the basis of various prior hypothesis and post-hoc evaluations (Yaron *et al.* 2022).

Overall, our current understanding of the neurophysiological signatures of the full NCC remains incomplete, with previously proposed markers either proved illusory, or lacking the predictive power, specificity, and empirical basis for large-scale application.

2.2 Electroencephalography

The brain is by far the most remarkable organ within the human body, and the central part of our nervous system, which is responsible for a multitude of vital functions; from its primary purpose of movement and the regulation of bodily functions, to sensory information processing, to a number of complex cognitive functions, including our memories and emotions, our language and reasoning, and of course, consciousness itself. A product of millions of years of evolution, it is the most complex system that we know of, with a network of more than 86 billion neurons interconnected by trillions of synapses, alongside several biochemical mechanisms that allow its functionality. The modern scientific study of the brain has a history of about two centuries, as science and technology provided the means to investigate its structure and function. Extensive literature on neuroanatomy and neuroscience has revealed a hierarchy of structures and their associated functions, with higher – more complex – functioning happening to outer – more recently developed – layers of the brain, where consciousness also seems to reside (Koch *et al.* 2016). Although most neuroscience and cognitive science research is focusing on a specific level of analysis, a proper understanding of any complex phenomenon may require an understanding of the mechanisms involved over multiple levels (from single neurons to large neural networks, and eventually, to behavior). Of course, the brain/mind is far

from fully understood yet, with ongoing research on a variety of methodologies and spatial/temporal scales. One of such levels of analysis regards the electrophysiological signatures of the brain, which were known to exist already from the 19th century.

Electroencephalography (EEG) is one of the most used techniques in research and medicine, which provide access to the brain's function and its underlying information processing. It is a non-invasive neurophysiological monitoring method that allows the macroscopic recording of the electrical activity of the brain, as it appears externally on the scalp of an individual (electrocorticography-EcoG and intracranial-EEG, are the alternative invasive methods). The measurements reflect voltage fluctuations of ionic currents induced by neuronal activity, which are captured by multiple electrodes placed along the scalp, resulting in a spatio-temporal image of the overall brain activity. In clinical contexts, recordings of spontaneous activity take place usually within several minutes (20-30 minutes, plus preparation time), with diagnosis traditionally focusing on event-related potentials (ERPs) and the spectral content of the signals (Drazkowski 2011). ERPs refer to time-locked averaged fluctuations induced by an external event/stimulus (e.g. visual or auditory) or an internal process. The spectral content of the EEG refers to specific frequency bands of neural oscillations (brain waves), classified into five main rhythms, namely: delta, theta, alpha, beta, and gamma waves.

EEG has been used in the clinic to diagnose conditions such as epilepsy, sleep disorders, encephalopathies, and the depth of anesthesia (DoA). Although more recent imaging techniques, such as the magnetic resonance imaging (MRI) and computed tomography (CT), have replaced it in several domains due to their high spatial resolution (e.g. in diagnosis of tumors or stroke), EEG has been proven very valuable to this day, given its high temporal resolution (in the orders of milliseconds), and the ease of access to the required hardware equipment. Nevertheless, a number of issues and constraints appear when trying to analyze EEG signals (due to the complex nature of the signals and the noise contamination from various sources), which in turn have given rise to the appearance of more and more sophisticated techniques, including electrophysiological modelling, connectivity/causality analysis, Bayesian inference and machine/deep learning techniques.

2.2.1 Mechanism

The electrical charge of the brain is maintained by billions of neurons, as the whole network exchanges ions through the neurons' membranes with the extracellular milieu, during resting and action potentials. When ions are pushed through a large number of neighboring neurons, a wave of ions propagates and reaches the electrodes on the scalp (volume conduction), since the brain's tissue, meninges, skull and skin, act as conductors. The pushing and pulling of ions on the metal part of the EEG electrodes, creates voltage differences between two electrodes that can be measured over time, resulting in the EEG signal (Tatum 2014). These voltage changes are mostly sensitive to activity that reflects the summation of synchronous activation of thousands or even millions of neurons, as the activity of a single neuron would be

far too weak to be captured. Moreover, they are primarily sensitive to post-synaptic potentials rather than action potentials, and to neurons with similar spatial orientation (such as the pyramidal neurons), which tend to produce stronger waves and thus, stronger EEG signals. A large number of aligned cells can be thought as a dipole that generates an electric field, and whose location and angle determines the projected spatial distribution over the scalp surface. As the voltage field gradients fall off with the square of the distance, activity in regions located deeper within the brain are harder to capture, in contrast to cortical areas.

EEG signals show oscillations at various frequencies with characteristic ranges and spatial distributions, some of which have been associated to different states of brain functioning (e.g. wakefulness vs sleep). Although some of these oscillations are understood in terms of their underlying network function (e.g. the sleep spindles due to thalamocortical resonance (Piantoni, Halgren and Cash 2016)), the relationship between the two remains generally unknown, given the complexity of the spiking networks and the nature of the EEG measurements. From a structural perspective, a large variety of functions have been associated with one of the four main lobes, which conventionally divide each cerebral hemisphere; namely the frontal, parietal, temporal, and occipital lobes (Fig. 2.3). Of course, functional overlap between these regions is expected in many cases, as complex cognition emerges from the integration of many simpler, but distributed processing mechanisms.

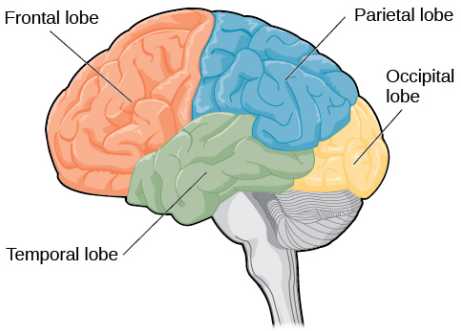

Fig. 2.3. The four cerebral lobes of the brain (image source: (*Lobes of the Brain | Introduction to Psychology* n.d.)).

2.2.2 Methodological Advantages and Limitations

A substantial proportion of studies in neuroscience, cognitive science, and psychology research have used EEG for the investigation of all kinds of simple and complex phenomena within the brain. Given the variety of the available methods for functional neuroimaging – such as magnetoencephalography (MEG), functional magnetic resonance imaging (fMRI), positron emission tomography (PET) and others – a comparison to EEG is natural in terms of the advantages and disadvantages that it offers over the alternatives. Some of the most notable ones are mentioned below.

Advantages:

- Non-invasive, simple technique
- Compact, silent, and easy to use
- Inexpensive hardware
- EEG systems are portable can be used in many places, without special requirements (e.g. it can be used in the presence of metallic implants, or with patients incapable of motor response)
- Widely available for immediate care in hospitals (Schultz 2012), which can be deployed at the bedside
- Better suited for repetitive assessment in patients with fluctuating vigilance (Chatelle *et al.* 2012)
- Very high temporal resolution, in the order of milliseconds (sampling rates between 250 and 2000 Hz are common)
- More resilient to noise artifacts generated by frequent, uncontrollable physical movements (which can moreover minimized or eliminated)
- EEG signals can be physically interpreted with respect to the brain's neuronal activity (in contrast to indirect measures of blood flow and metabolic activity, found in other techniques)
- There is no exposure to intense magnetic fields or radioligands

Disadvantages:

- Preparation time is higher than other methods (in case of a large number of electrode placement, use of gels, saline solutions, etc.)
- Low spatial resolution constrained by the scalp (further localization of current sources requires source reconstruction techniques, which are based on hypothesized estimates (Nunez 1988))
- Brain activity from deep structures below the cortex may be poorly measured
- Cannot identify specific neurotransmitter-activated regions
- Poor Signal-to-noise (SNR) ratio, which in turn requires sophisticated and often manual analysis, in order to extract useful information

EEG has also been combined with other neuroimaging techniques (e.g. MEG, fNIRS, fMRI or PET) and brain stimulation techniques (e.g. TMS or tES), taking advantage of the various information that each method provides about the brain, albeit there can be technical difficulties (e.g. the presence MRI pulse artifacts). For example, simultaneous recordings of EEG and MEG allow for the correction of aspects required for EEG analysis (e.g. information about skull radius and conductivities), and the improvement of the quality of the signals, due

to differentiation of errors exhibited by each method (deep signals from EEG can also be better isolated) (Huizenga *et al.* 2001). Furthermore, EEG has been used alongside fMRI and PET, in order to acquire high-spatial resolution data, or to trace a drug's action in a specific region of the brain, respectively (Laufs *et al.* 2003; Schreckenberger *et al.* 2006).

While high-temporal and high-spatial resolutions are usually mutually exclusive for most non-invasive neuroimaging techniques, recent research on EEG and MEG suggests that there is adequately rich spatiotemporal information that can be used to discriminate a number of complex brain states. Using theoretical modelling and empirical measures, (Stokes, Wolff and Spaak 2015) showed that EEG contains information that can differentiate spatially overlapping states, as even small differences in the angle of neighbouring dipoles/sources (a fair assumption, given the irregular surface of the cortex), can produce statistically separable field patterns. In general, the localization of the source of activity given the fields measured at the scalp, is a hard and ambiguous inverse problem, without a unique solution (although, several techniques have been developed that probabilistically constrain the solutions). Nevertheless, if the exact localization of the sources is not a primary purpose of the analysis, multivariate pattern analysis (MVPA) has been shown to be very effective in decoding subtle changes in neural states, within or across subjects (in parallel to MVPA in fMRI, where subtle differences in the distribution of neurons accumulate over a number of voxels, which can be used to decode complex activity patterns).

Overall, EEG alone is a tool with several strong points for functional neuroimaging. Its simplicity, portability, low cost and wide availability, makes it particularly effective for clinical research and applications, in comparison to all other methods. Although the methodology used in research can vary significantly across studies, and have not been standardised sufficiently for clinical use, there are several steps conventionally followed when acquiring and analysing EEG signals. In the following sections, we discuss some of these steps in detail.

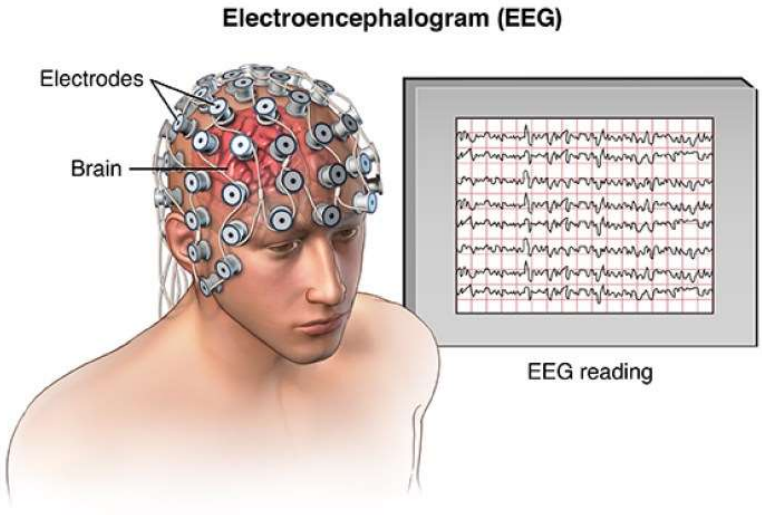

Fig. 2.4. EEG measurements of brain activity using a net of electrodes (Shen 2020)

2.2.3 Data Acquisition

Electrodes and Amplifier. The first step in conventional EEG recording regards the placement of the sensors on the scalp (Fig. 2.4), as individual electrodes or embedded within a net/cap (in case of high-density array systems), usually with the help of a conductive gel or paste (sometimes light abrasion to reduce impedance on the scalp area may precede). Over the past years, a variety of developments in EEG sensing have tried to deal with several limitations, such as setup time, maintenance, poor electrode contact, and its use in challenging environments, as required in specialized applications (e.g. flexibly dry electrodes, semi-dry, hydrogels, conductive EEG sponges, and other technologies). In general, the scalp is covered uniformly by a number of electrodes, from a few dozen to a few hundreds, depending on the application's demand in spatial resolution (19 electrodes, plus ground and system reference are often used in the clinic (Acharya *et al.* 2016), with high-density arrays containing up to 256 mostly found in research settings).

The exact location and name of each electrode is specified by a standard placement method, such as the International 10-20 system (Fig. 2.5), which ensures that the recorded activity can be consistently compared across individuals and studies (Towle *et al.* 1993). The name 10-20 derives from the choice of 10% or 20% of the front-to-back or right-to-left distance of the skull (e.g. inion – nasion), as the distance between adjacent electrodes. A letter and a number are assigned to each electrode, which indicate their position relative to the lobes and hemispheres of the brain. Specifically, 'F', 'T', 'C', 'P' and 'O' refer to frontal, temporal, central, parietal and occipital areas (or lobes, with the exception of central), while 'z' refers to the midline that divides the two hemispheres. Even numbers refer to electrodes of the right hemisphere, while odd numbers refer to the left hemisphere. 'A', 'Pg' and 'Fp' refer to the ear area, the nasopharynx and the prefrontal areas, respectively. When high-density EEG systems are used, extra electrodes are placed in intermediate sites halfway between the existing locations of the 10-20 system.

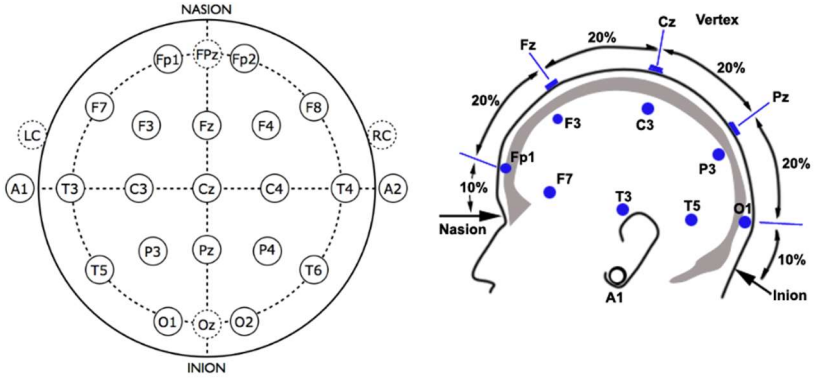

Fig. 2.5. The International 10-20 System (image sources: (*Montages* n.d.; *EEG* n.d.)).

The next step regards the amplification of the signals, since EEG measurements are weak as received from the scalp electrodes (about 10 to 100 μV in amplitude) (Aurlien *et al.* 2004). A pair of electrodes is connected to a differential amplifier that amplifies the voltage by 1,000 to 100,000 times (typically 60-100 dB of voltage gain). Any interference that is present on the pair is rejected (e.g. ground noise), while a combination of amplifiers is used to accurately measure the signals of interest. Moreover, an anti-aliasing filter is used before the analog-to-digital (AD) conversion, as most contemporary systems digitize and store the signals in an electronic format, for further visualization and analysis. Typical AD sampling rates range from 256 to 512 Hz in clinical settings, and up to 20 kHz within research applications.

Montage. As already mentioned, EEG measures the relative voltage differences between two electrodes, the selection of which determines the representation of the EEG signals/channels. This selection is known as the 'montage', and is important to understand in order to have a deeper insight on its effects on the signals, as well as on the interpretation of the EEG. Electrodes placed over areas that are expected to show brain activity correspond to online positions, while inactive areas correspond to offline positions. Very often, an online position is selected for a reference channel, which defines the zero-level voltage and determines the amplitude and polarity of all other channels (channel – reference). Depending on the spatial distance of an electrode from the reference, the resulting signal could show low amplitudes for neighboring electrodes, or high amplitudes for distant electrodes, as neighboring locations would normally share more common sources (brain or noise sources) and thus, signals. While the relative distribution of the EEG topography is preserved, the temporal structure of the signals can vary, depending on the nature of the reference signal. Finally, the selection of the montage can be determined by the presence of local noise (which can have global effects), or the location of the activity of interest (which can be highlighted or distorted as a function of distance), resulting in a resolution trade-off and hence, to no optimal approach. When choosing a montage, there are few available approaches:

- ***Sequential/Bipolar montage***: Each channel represents the difference between two adjacent online electrodes (e.g. Fp1-F3). This can be achieved either as anterior-posterior, or transverse chains of electrode pairs. It is a versatile montage, but not the best at detecting either focal or diffuse abnormalities.
- ***Referential montage***: Each channel represents the difference between an online electrode and a common reference electrode. In general, a symmetric position of the online/offline reference with the respect to the brain areas is preferred, to avoid representation bias of voltage distribution. This montage is useful for broadly distributed abnormalities, but not for focal discharges. Although there is no standard position for the common reference electrode, there are several choices found in clinical and research settings:

  - Midline position: an online electrode placed over the midline, which ensures that neither hemisphere is explicitly amplified. Cz is a common reference choice.
  - REST reference: offline computational reference at infinity where the potential is assumed zero
  - Linked ears: a physical or mathematical average of offline electrodes attached to both earlobes or mastoids

- ***Average reference montage***: the average signal of all electrodes is used as a common reference channel. Frontal and occipital electrodes may be excluded in summation, due to eye and head movement artifacts, respectively. It is a versatile montage, with no theoretical bias towards any particular location of the head, but susceptible to reference contamination (e.g. in case where one or more electrodes have high amplitude noise). A large number of evenly-spaced electrodes is preferred for average reference, for better unbiased representation.
- ***Laplacian montage***: a weighted average of the surrounding electrodes is selected as the reference channel. This montage is good for focal discharges, but not for broadly distributed abnormalities. Given the assumption that nearby electrode locations share activity from common sources, it could improve the spatial SNR of the EEG (Sanei and Chambers 2007). Nevertheless, it is the least often used, since it is the hardest to physically conceptualize.

While different montages highlight different features, EEG is usually stored and used with a referential montage (most systems have a pre-defined reference channel), or an average montage (often found in research). However, any particular representation can be mathematically constructed subsequently from any other (offline re-referencing). This can be easily seen from the subtraction of the new reference, which contains the original reference representation, resulting in (channel – online reference) – (new reference – online reference) = (channel – new reference). This is important in terms of the comparability and consistency across results and studies that use different EEG systems and montages. As there is no optimal referencing method, and given that different approaches have been shown to profoundly affect further analysis, (e.g. with power spectra, connectivity measures or machine learning analysis (Yao *et al.* 2005; Sanqing Hu *et al.* 2010; Bastos and Schoffelen 2016; Lopez *et al.* 2017)), methodological experimentation and interpretation of the results can be essential.

2.2.4 EEG Signals and Pre-processing

EEG has several advantages compared to other functional neuroimaging methods, but there are specific limitations and challenges when it comes to its analysis. As we mentioned, one of the main disadvantages regards its low signal-to-noise ratio (SNR), which often creates

the need for data curation and manipulation. During the recording of EEG, signals typically exhibit a variety of fluctuations induced by normal and abnormal activity, which is related to neuronal or other physiological and non-physiological processes. Activity that is not directly induced by neuronal processes contributes to noise (artifacts) that often needs to be handled, prior to the analysis and interpretation of EEG (typically by removing or cleaning the contaminated signals). EEG is also a non-stationary signal, as its statistical properties vary with time. Last but not least, EEG can show significant variability across acquisition systems, recording sessions and most importantly, individuals (Melnik *et al.* 2017). Inter-subject variability results from the physiological differences of individuals, which is most evidently observed in the distinct electrophysiological signatures of different age groups.

Of course, many of these challenges can be tackled by appropriate experimental setups and methodological approaches. In general, a number of pre-processing steps are applied to the data, in order to improve their quality (SNR), before any further processing analysis.

Wave Patterns. EEG signals are generally described in terms of transient features (e.g. vertex or spindle waves, seen during sleep) and rhythmic activity. Rhythmic activity is typically classified into five frequency bands of neural oscillations, known as delta, theta, alpha, beta and gamma waves. Although there is no standard definition for the exact frequency ranges (which can vary across the literature), activity within these bands has been roughly associated to certain functions and spatial distribution over the scalp.

- **Delta Waves (< 4 Hz):** delta waves have been associated to conditions such as deep sleep and general anesthesia, most prominently exhibited in frontal brain regions
- **Theta Waves (4 – 8 Hz):** theta waves have been associated to drowsiness, and inhibition of elicited responses
- **Alpha Waves (8 – 12 Hz):** alpha waves have been associated to relaxed resting-states, most prominently exhibited in the posterior brain regions (increased with eyes closed)
- **Beta Waves (12 – 30 Hz):** beta waves have been associated to high alertness, stress and active focus
- **Gamma Waves (> 30 Hz):** gamma waves have been associated to perceptual sensory processing, most prominently exhibited in the parietal lobe

EEG Artifacts. Signals that are produced by sources other than the brain and contribute to noise contamination, are referred as EEG artifacts. Artifacts are normally present in all recordings, with amplitudes that are significantly larger relative to brain signals (thus, contributing to the low SNR). Therefore, artifact handling – either in the form or rejection or reconstruction of the signals – can be an important step for EEG analysis and interpretation. The morphology, duration and frequency of each artifact can vary depending on its source and the recording environment (e.g. the acquisition system). For this reason, their detection has traditionally required visual inspection and manual annotation by experts trained in the field.

In general, artifacts can be categorized into biological (physiological) and environmental (non-physiological). Some of the most common biological artifacts are induced by bodily movements, ocular (e.g. eye movements and blinks), cardiac (ECG), muscle (EMG), electrodermal and glossokinetic activity. Common environmental artifacts are produced by changes in electrode impedance (e.g. due to bad electrode contact or movement) and grounding noise from the power line.

EEG Pre-processing. As we discussed, a number of pre-processing steps can be applied for the curation of the data. Although EEG pre-processing pipelines are not as standardized as ones found in other neuroimaging techniques, particular processes are typically found in the majority of research and clinical setups. We briefly describe some of these processing steps below:

- **Channel selection:** a number of EEG channels can be selected for further analysis, depending on the task or application demands
- **High-pass filtering:** high-pass filtering is used to remove slow artifacts, such as electrodermal (due to sweating) and movement activity (typical settings: 0.1 – 1 Hz)
- **Low-pass filtering:** low-pass filtering is used to remove high-frequency artifacts, such as muscle activity (EMG). Cerebral signals are mostly observed within the 1-30 Hz range (typical settings: 30 – 70 Hz)
- **Notch filtering:** notch filtering is used to remove electrical (power line) noise (typical settings: 50/60 Hz, depending on the mains electricity)
- **Epoching:** EEG data can be segmented into temporal windows (epochs), allowing us to individually process different events and brain states/response
- **Resampling:** signals can be down-sampled without information loss (as per the Nyquist theorem), usually for storage and computational purposes
- **Re-referencing:** re-referencing is used to obtain alternative representations of the signals (discussed in the previous section)
- **Artifact Handling:** EEG artifacts can be rejected by removing specific segments of the signals (bad epochs), or even specific contaminated channels, often by applying a peak-to-peak amplitude threshold (activity higher than the threshold is considered contaminated). Alternatively, artifacts can be removed by spatial or temporal filtering, or more sophisticated techniques like independent component analysis (ICA), which attempt to unmix the underlying signal components. During the past years, different approaches have been developed for an automated handling of artifacts.
- **Frequency band extraction:** a Fourier-type method is implemented to obtain the spectral representation of the signals, from which the frequency bands can be extracted (e.g. Welch method)

2.2.5 Quantitative Analysis

During the past decades, quantitative analysis techniques – such as signal processing and pattern recognition algorithms – have transformed our ability to decode and interpret EEG signals. These techniques have allowed researchers to better understand the underlying processes and mechanisms of the brain, whilst associating them to particular mental processes. Besides research, quantitative analysis has also assisted physicians in clinical assessment (diagnosis), as prior evaluation involved the subjective interpretation of visually inspected signals. Over the years, a large variety of linear and non-linear methods have been developed, which can be largely categorized by their domain of analysis; most notably, time domain methods (such as autoregressive models), frequency domain methods (such as Fourier transforms), and time-frequency domain methods (such as wavelet transforms). In addition, a number of nonlinear methods have been explored, given the nature of EEG and mental processes, such as entropy and complexity measures.

More recently, machine learning techniques have shown their increasing strength in decoding neural activity, with a variety of algorithms developed for different kinds of tasks, such as end-to-end processing, feature extraction, and predictive modelling. These algorithms rely on a data-driven investigation of the data, which allows for the identification of novel patterns, and the creation of automated systems for clinical use. As we discussed in section 2.2.2, multivariate pattern analysis techniques have shown their ability to decode complex brain states, which would otherwise be hard to assess. Their capacity to find non-linear structure in the data, contrary to the constrains and assumptions of previously employed methods, is often attributed behind their success (King *et al.* 2017). More specifically though, artificial neural networks are a class of machine learning algorithms that have been largely successful in EEG analysis. In the next section, we further dive into the topics of artificial neural networks and the advancements of deep learning.

2.3 Deep Learning

Despite its short history, computer science has shown an ever-increasing significance and improvement of its usability and effectiveness, in every aspect of human activity and for a vast variety of problems, over the past decades. In this revolutionary age of digital information, its contribution to science and society is fundamental, as the computational sciences and the availability of large amounts of data allow for the practical evaluation of processes and phenomena of great complexity. By constructing mathematical models, computers have enabled us to analyze and solve scientific problems, many of which have advanced our current understanding of the human brain.

Notably, the applied field of artificial intelligence (AI) has become one of the most intriguing and important areas, as we require complex but efficient methods to synthesize goal-oriented processes such as learning, adaptation, knowledge representation and decision-making. Although there is no established unified theory or paradigm in AI research, a variety of approaches have been explored, from the traditional symbolic AI to the more recent and sophisticated advances in statistical learning methods (i.e. machine learning), which do not require a semantic understanding/representation of the data under analysis. Within machine learning, knowledge and tools from a variety of other fields, such as mathematics (e.g. information theory, optimization theory, statistics and probability theory), philosophy, electrical engineering (e.g. signal processing techniques), psychology, neurophysiology, and others, have been used to tackle the nature of such processes. As a general purpose technology, it has been used in a variety of domains, mostly in the form of an embedded component within hardware and software (e.g. in medicine, it has been used in cases such as medical imaging diagnosis, or for optimization of drugs and dosages). Overall, the automation of evaluative and predictive tasks through computational algorithms has been increasingly successful, as a substitute for human monitoring and intervention. This is particularly the case for specialized domains that involve complex real-world data, an area that machine learning and deep learning specifically, have shown prominence (Pennachin and Goertzel 2007).

2.3.1 Machine Learning

Machine learning, a subfield of AI and concept following the approach of statistical learning principles, regards a family of algorithms used to solve a task by progressively improving the performance of a model, as it acquires more experience. This mathematical model is derived from a sample of data, with an aim to analyze and discover unknown properties of the data (data mining), or most often, to make predictions about new unseen data, without being explicitly programmed for the task (we can consider this paradigm as an alternative way from the traditional imperative programming, where we manually create an algorithm for a given problem). The model can be usually thought as a function approximator of an unknown process (usually a non-trivial task), which optimally tries to determine an output for each input (Bishop 2006). These input and output data samples (instances) are represented by vectors/arrays of values (features), the totality of which determines the distribution of a dataset. Some of the most prominent families of machine learning models are regression analysis, Bayesian networks, decision trees, support vector machines, genetic algorithms and artificial neural networks, the latter of which form the basis of deep learning.

2.3.2 Supervised and Unsupervised Learning

Most algorithms can be divided into two main types, depending on the nature of the task and the data provided. Supervised learning refers to models where we provide both input

data and the corresponding desired outputs (targets), for the purpose of classification or regression tasks, depending on whether the output is restricted to a discrete number of values (classes), or being a continuous variable, respectively. Unsupervised learning refers to models constructed solely on input data, as a way to find patterns and structure within the data, for the purpose of knowledge extraction, data compression/denoising, grouping or clustering tasks. In general, these models can be predictive, if they make predictions for new data, descriptive, if they try to gain knowledge and represent the data, or both. The evaluation of performance for each approach is ultimately determined by whether the main goal is to reproduce existing knowledge, or to acquire new knowledge.

Within supervised learning, data can be divided into three main categories, depending on their role in creating and evaluating the model, namely: training dataset, validation dataset and test dataset (or holdout set). The training dataset is used to optimally fit (train) the parameters of the model, given the predictions from the inputs and the desired (labeled) outputs. The validation dataset is used to evaluate the performance of the model in an unbiased way, in cases where possible hyper-parameters of the model need to be searched and adjusted (tuning). Finally, the test dataset is used to provide a final evaluation of the model, with data samples that have never been used during training or tuning (generalization test). The division into training/validation or training/testing datasets, is often done with a cross-validation paradigm, where the data repeatedly split into training/validation or training/testing partitions, and the overall performance calculated as the average performance of all validation/test sets, for all possible partition configurations. A depiction of the whole supervised learning paradigm can be seen in Fig. 2.6.

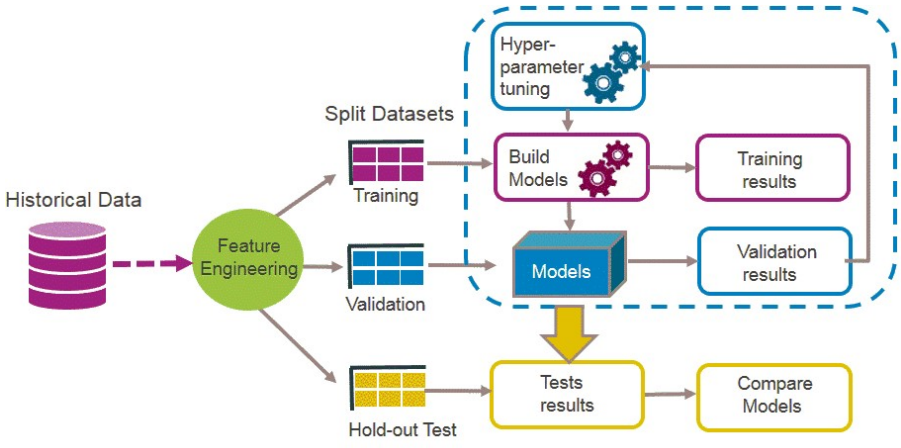

Fig. 2.6. Supervised Learning Paradigm. The data, with or without feature engineering, are split into training, validation, and test sets. Training data are used for model creation, while validation data measure its performance, in order to tune its hyper-parameters. Test set is finally used to objectively evaluate the performance of the model (image source: (*Machine Learning Has Transformed Many Aspects of Our Everyday Life, Can It Do the Same for Public Services? – Capgemini Worldwide* n.d.)).

Some of the most widely used supervised models in cognitive neuroscience and EEG analysis for medical diagnosis, are linear discriminant analysis (LDA), logistic regression (LR), common spatial filters (CSP), support vector machines (SVM), and artificial neural networks (ANN). Unsupervised models are also used in EEG data for compression, denoising and representation learning, such as principal component analysis (PCA), independent component analysis (ICA), convolutional sparse coding (SCS), and spatio-spectral decomposition (SSD) (King *et al.* 2017).

2.3.3 Model Training, Assessment and Limitations

As we mentioned, machine learning models are derived from a sample of data by trying to optimally associate specific inputs to outputs. This optimization can be expressed with respect to the parameters of the model and an objective function (or cost function, where the cost can be thought as the average distance between a predicted and a desired output value), while many learning problems can be formulated as the minimization of such function, through iteration (model training). Although for some models it is possible to theoretically approximate any input-to-output function accurately, time constraints, the nature and the amount of data, do not usually allow it in practice.

Since training data are finite and there is uncertainty about possible future data, a core objective of these algorithms is to discover a model that can generalize as much as possible (that is to say, to minimize a quantifiable generalization error, or the cost function on unseen test samples (Bishop 2006)). Creating a model that is less or more complex than the function underlying the data can lead to underfitting or overfitting, which compromise the errors in the training and test datasets, respectively (Fig. 2.7). In general, simple models are preferred to complex ones, as they are easier to interpret and understand (similarly from a scientific point of view, simple theories are preferred to complex ones by Occam's razor).

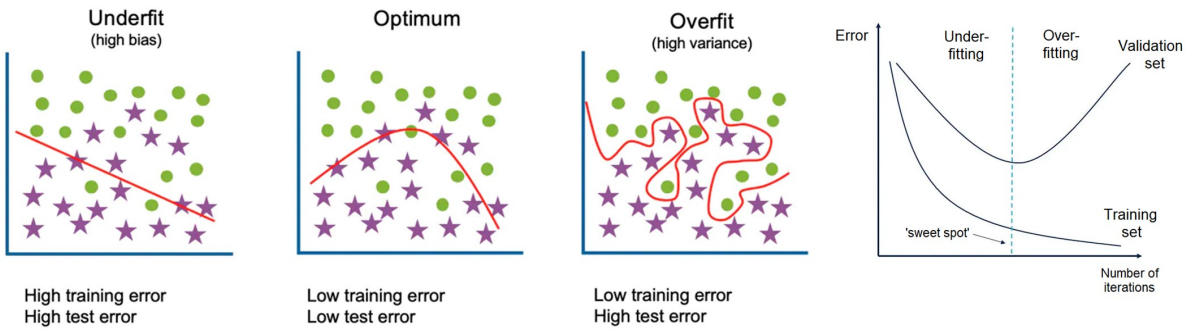

Fig. 2.7. Visualizations of model underfitting and overfitting, resulting in different training and test errors. Ideally, we want to restrain the complexity of the model to a point where the validation or test error is minimized ('sweet spot'). One way this is achieved is by pausing the training process after a number of iterations (image source: (*What Is Overfitting?* | *IBM* n.d.)).

In this view, optimization and generalization are not always aligned, as optimization adjusts the model to perform well on the training data, while generalization is measured by its performance on unseen test data. During training, optimization and generalization are initially correlated, with the respective errors decreasing up to the point of overfitting, where training and validation/test errors begin to diverge (Fig. 2.7.). Notably, overfitting is more likely to occur with noisy, mislabelled, and ambiguous data, as well as by the inclusion of rare features and spurious correlations to the task under analysis.

Given all the above, the performance of a model depends highly on the nature of the task and the data itself (whether data are representative enough for the task under analysis), the size of the dataset, as well as characteristics such as the input representation (features), dimensionality, statistical distribution and noise. For these reasons, various methods can be applied to the data, prior to these computational models and as a way to enhance their performance, such as feature learning, feature selection and engineering, outlier detection, and other preprocessing steps for data curation (Bengio, Courville and Vincent 2013). Beyond the task and data, model performance can also depend on the nature of the learning algorithm, as different algorithms and learning techniques have different limitations. Other issues related to model assessment and their limitations regard the computational complexity of the model (e.g. number of trainable parameters) – which reflect memory, time and data requirements for training – as well as its relation to the suitability of the data, possible biases, data access/resources, etc. (Jordan and Mitchell 2015). Finally, predictive accuracy, speed, scalability and interpretability are all issues of interest, which can vary across models.

2.3.4 Artificial Neural Networks

Artificial neural networks are computational models that were inspired by the architecture of biological neural networks in the brain. A single neuron can be thought as a basic computational unit, which – after a number of mathematical approximations – acts as a simple function. Similarly to biological neurons, an artificial neuron gets inputs from other neurons and produces an output (activation), when the weighted sum of its inputs exceeds a certain threshold (bias). These weights and biases determine the sensitivity of a neuron to particular patterns of activity (in analogy to biological synapses), which characterize its properties. By applying a nonlinear function to the sum of its inputs, the output of a neuron can have a wide spectrum of activation values (in analogy to neuronal firing rates).

Mathematically, this process can be expressed as $\hat{y} = g(w^T x + b)$, where x is the input data (features), w is the weights vector, b is a bias, g is an activation function, and $\hat{y}$ is the output of the neuron (Fig. 2.8). Depending on the task and the underlying data distribution, different activation functions might behave more or less appropriately. Some examples of activation functions used in neural networks are the sigmoid, tanh, ReLU, ELU, softmax, and others.

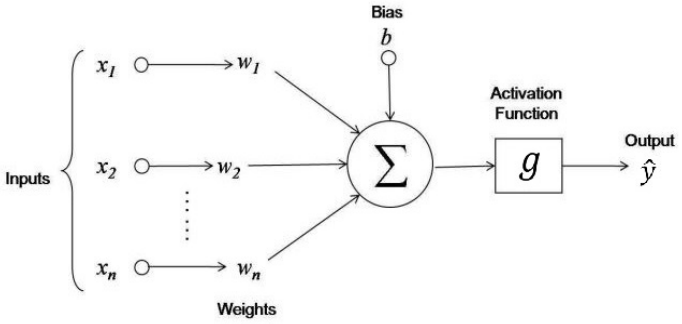

Fig. 2.8. A mathematical depiction of an artificial neuron. The weighted sum of the input features (*x* vector), calculated by multiplying with the weights (*w* vector) and adding the bias (*b*), is passed through a nonlinear activation function (*g*), which produces the output ($\hat{y}$) (image source: (*Artificial Intelligence - Accelerate the Power with Neural Networks | Blog* n.d.)).

An assembly of many interconnected neurons creates a neural network (NN) that can model complex relationships between the inputs and outputs (Fig. 2.9). Each neuron in the network codes for a particular feature in the data, as it connects with other neurons through the weights that increase or decrease their connection strengths. The units are organized in layers, with neurons in a given layer typically receiving inputs only from neurons of the previous layer. Such architecture is characterized as 'feedforward', given that the activation signals pass in one direction, from the initial input layer to the final output layer. The nonlinearities introduced by the activation functions of the units is what allows NNs to approximate a variety of complex functions. In fact, given a large enough model, NNs can act as a universal function approximator (proven by the universal approximation theorem (Csanád Csáji 2001)).

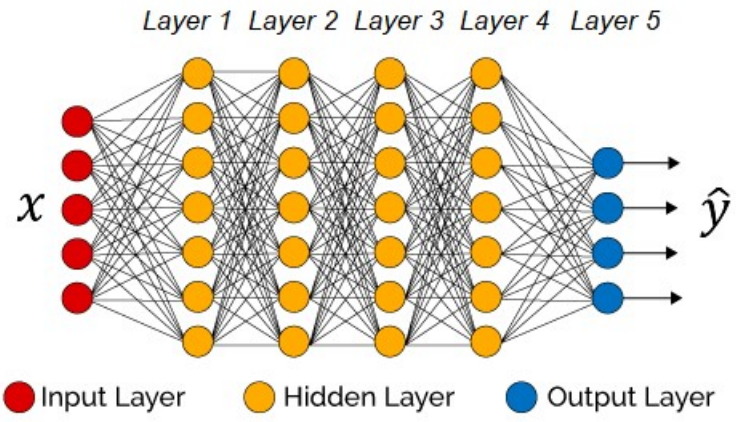

Fig. 2.9. A fully-connected neural network architecture with 5 layers (input is excluded). Red nodes represent the input features, while yellow and blue nodes represent hidden and output neurons. The input vector (input layer) is progressively processed through the subsequent layers of neurons (hidden layer), producing the output vector (output layer). Intermediate neurons in the hidden layers of the network code for features that are discovered during training (image source: (*Deep Learning Made Easy with Deep Cognition | by Favio Vázquez | Becoming Human: Artificial Intelligence Magazine* n.d.)).

During the past years, deep learning architectures – which use several layers of neurons between the input and output layers – have been used to accomplish sophisticated models of data and predictive tasks across all kinds of application domains (scientific or otherwise). The main idea behind deep learning is that the networks extract a hierarchical representation of the data, by composing low-level to higher-level features, as layers progress (this idea was inspired by architectures of sensory information in the brain, as found in the visual system). This process can be alternatively thought as a series of non-linear transformations applied to the original data space, towards a more task-relevant feature coding. Empirically, research has shown that there are particular types of functions that can be learned from deeper architectures, that shallower models could not easily approximate (or would require an exponentially larger number of neurons).

Model Training and Evaluation. Training the network requires the adjustment of its parameters, so that the performance of the model improves over time – i.e. producing outputs increasingly closer to the provided targets (model fitting). This is typically performed by gradient descent over the multi-dimensional space provided by the cost function and the trainable parameters of the model (weights and biases). More formally, given a training set of m samples (instances), with input size n_x (no. of features) and output size n_y, where:

- $x^{(i)}$ is the input vector (features) of the i^{th} sample
- $y^{(i)}$ is the target output vector of the i^{th} sample
- $\hat{y}^{(i)}$ is the predicted output vector of the i^{th} sample
- X is the input matrix of all training samples ($X \in R^{n_x \times m}$)
- Y is the label matrix of all training samples ($Y \in R^{n_y \times m}$) (ground-truth)

and a neural network with L number of layers with $n_h^{[l]}$ number of hidden units in the l^{th} layer, where:

- $W^{[l]}$ is the weight matrix in the l^{th} layer ($W \in R^{n_h^{[l]} \times n_h^{[l-1]}}$)
- $b^{[l]}$ is the bias vector in the l^{th} layer ($b \in R^{n_h^{[l]}}$)

the optimization of the model can be expressed as the minimization of a cost function $J(X, Y; W, b)$ with respect to W and b. The cost function reflects the average loss across all training samples, as $J(W, b) = \frac{1}{m}\sum_1^m L(\hat{y}^{(i)}, y^{(i)})$, where $L(\hat{y}, y)$ measures the distance between the predicted ($\hat{y} = P(y|x)$) and the desired output vectors (loss). A variety of loss functions can be used for classification and regression tasks, such as the categorical cross entropy (CCE) or the mean-squared error (MSE). While the cost function is not generally convex and does not have a unique optimal solution, model convergence can be affected by the network architecture and the selected optimizer.

The minimization of the cost function is performed iteratively using the gradient descent algorithm, which comprises of the following steps:

1. The weights W and biases b of the model are initialized to random non-zero values, often by applying a predefined probability distribution (e.g. normal, uniform, Xavier, etc.). For learning to be achieved, it is important for values to be close to, but not zero.

2. For a number of iterations (training epochs):
   a) The activation matrices A for all layers and m training instances are calculated as: $Z^{[l]} = W^{[l]} A^{[l-1]} + b^{[l]}$, $A^{[l]} = g^{[l]}(Z^{[l]})$, where $A^{[l]}$ is the activation matrix of the l^{th} layer for all training samples ($A^{[0]} = X$, $A^{[L]} = \hat{Y}$) (Forward propagation)

   b) The partial derivatives of the cost function with respect to the W, b parameters are calculated (gradient), given the loss from the predicted and targeted output values (Backpropagation). The following formulas can be derived using the chain rule:

$$\frac{\partial J}{\partial W^{[l]}} = \frac{1}{m}\frac{\partial J}{\partial Z^{[l]}} A^{[l]T} \ , \ \frac{\partial J}{\partial b^{[l]}} = \frac{1}{m}\sum \frac{\partial J}{\partial Z^{[l]}} \ , \ \frac{\partial J}{\partial Z^{[l-1]}} = \frac{\partial J}{\partial W^{[l]T}}\frac{\partial J}{\partial Z^{[l]}} g^{[l-1]'}(Z^{[l-1]})$$

   c) The weights W and biases b of the network are adjusted using the update rule: $W^{[l]} \leftarrow W^{[l]} - \alpha \frac{\partial J}{\partial W^{[l]}}$, $b^{[l]} \leftarrow b^{[l]} - \alpha \frac{\partial J}{\partial b^{[l]}}$

   where α is a learning rate (defining the size of the gradient descent steps)

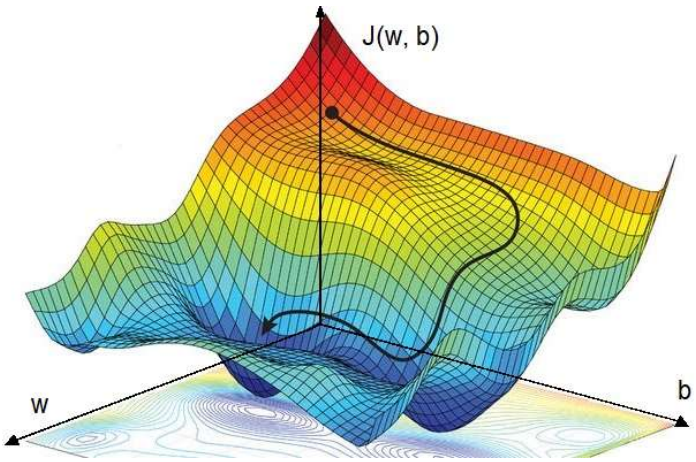

Fig. 2.10. Gradient descent visualization (Amini *et al.* 2018). At each training step (epoch), the gradient of the cost function J is calculated with respect to the W and b parameters. The parameters are adjusted by following the direction of the gradient's steepness, towards a local minimum.

The training data are often split into several batches for calculating the average gradient across *m* samples, the size of which can affect the computational speed and nature of convergence (typically a batch size of 32 or 64 is used). The training process terminates after a number of iterations of processing the whole training dataset, where typically the network can no longer learn from the data (the model converges to a loss value – see Fig. 2.7). Several variations of the gradient descent algorithm have been devised over the years, in order to improve the solutions and the convergence speed, such as the stochastic gradient descent (SGD), RMSprop, Adam, Adadelta, and others (Choi *et al.* 2019). Besides loss, the performance of the model can be measured by a selected metric, such as the accuracy in classification tasks, or the MSE and MAE (mean-absolute-error) for regression tasks.

Beyond the trainable parameters of the model, other hyperparameters – such as the number of hidden layers, the number of units in each layer, and others – are configured before training and remain constant (many of these hyperparameters determine the architecture of the network and ultimately, the training process). Of course, there is a large number of settings that can be tested and tuned, with respect to model performance. Nevertheless, this requires an iterative and empirical process, given that there is no theoretical basis for the design of a network, with different architectures and configurations working better for different problems and tasks (good hyperparameter configurations do not often transfer to other tasks, even within the same domain of applications).

Importantly, despite the strengths of neural networks, there are special difficulties in deciphering how these models process information, which still remain a significant constrain of the algorithms (interpretability problem). The ideas behind neural networks emerged from the connectionism theory in cognitive science – which is based on the concept of distributed processing of information through communication nodes – as an attempt to explain the processes and mental phenomena in the brain. Yet, whether complex biological systems (such as the brain) can be understood in terms of simple descriptive models, or whether the inherit complexity of a biological and artificial neural network is necessary for the expression and generation of robust solutions, is open to debate.

2.3.5 Deep Learning Architectures

Within the past decade, deep learning has revolutionized the field of machine learning and AI, with many models that have evolved into a broad family of neural network architectures. Such models have advanced the state-of-the-art across a large variety of problems and applications – including biomedical engineering (e.g. in problems such as drug design and medical image analysis). The success behind deep learning can be attributed to two main reasons. The first one, concerns the increasing availability of vast amounts of data, alongside the computational speed from hardware (e.g. GPUs/TPUs) and algorithmic advances, which allow us to train large NN models. The second reason, regards the development of more sophisticated processing architectures that resemble and exploit the structure of the data,

allowing the models to extract useful, task-relevant information. Contrary to traditional machine learning models, which relied more heavily on domain knowledge and manual feature engineering, deep learning provided the capacity for an end-to-end learning approach (tackling both representation learning and task decoding problems, concurrently).

As we already mentioned, the core objective of such models is to maximize their generalization power to novel, unseen data. The ability of deep learning to successfully generalize in various tasks has been based on the manifold hypothesis, which posits that natural data lie on a low-dimensional manifold (latent space), which is encoded within a high-dimensional original data space. This implies that deep learning fits relatively simple, structure subspaces, upon which it interpolates (or generalizes) its input samples. Deep learning architectures are suitable to learn such manifolds as they implement smooth, continuous functions, whilst constrained in a way that mirrors the information within the data (via architecture priors). Given this context, a successful use of the models require the understanding of the data under analysis, which can guide model design and the selection of an appropriate learning algorithm.

Processing Architectures. A large variety of processing architectures and layers have been developed over the past years, with various properties and connectivities across neurons. One of the main breakthroughs, regards the convolutional neural networks (cNNs), which although were initially designed for tasks within computer vision, have been recently proven to be highly suitable in many areas, including EEG analysis (Roy *et al.* 2019). Some of the most prominent layer types are described below.

- **Fully-connected Layers.** In a fully-connected layer, neurons receive inputs from the activations of every neuron in the previous layer. In this setup, layers do not consider the structure of the input data, making the network more prone to overfitting. Fully-connected layers can be impractical for inputs with large dimensionality, as a large number of weights is harder to regularize (multilayer perceptrons – MLPs comprise of fully-connected layers).

- **Convolutional layers.** Convolutional layers create a specialized type of network that imposes a sparse connectivity, by exploiting the operation of convolution. In this setup, neurons receive inputs only from a local subset of neurons in the previous layer – typically arranged within a particular structure that mirrors the data (e.g. in image processing, the input data can be a 2D array). By using convolutional windows (kernels), this layer forces the model to learn invariant representations of the data, as kernel weights are shared for all neurons in the layer. This can be interpreted as a sliding filter that detects a particular type of activation pattern, convolving across the preceding data, and providing translation-equivariant responses (features) (Fig. 2.11). The output values of a given kernel/filter create a feature map, whilst the network can implement

many feature maps in parallel, the outputs of which can be aggregated into higher layers of the model. A general approach to convolutional layers is to explore abstract patterns of increasing complexity, by employing layer-wise increasing kernel windows. Although, this organization was inspired by the architecture of the visual cortex and the concept of the receptive field, it can be useful for any type of data with a grid-like topology (e.g. data that have spatially or temporally local relations). Some of the main hyperparameters of convolutional layers are:

- **Kernel size**: the dimensions of the convolution window applied to the preceding input layer
- **Stride**: the dimensions that control the sliding steps of the convolution window (kernels can typically overlap across features)
- **Number of filters**: a pre-defined number of filters can be selected for each convolutional layer, producing a stack of multiple feature/activation maps

The size of the output map in a given dimension can be determined by the input (W), kernel (K) and stride (S) sizes, as $\frac{W-K}{S} + 1$.

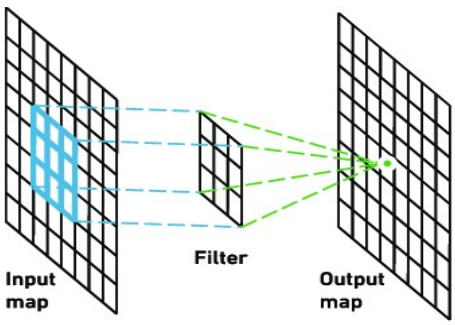

Fig. 2.11. The convolutional layer (output map) performs a dot product (weighted sum) between the convolution kernel (filter) and the preceding input layer (input map), before applying an activation function. The kernel convolves across the input layer, producing different outputs in the activation map (image source: (Yakura *et al.* 2018)).

- **Pooling layers.** Pooling layers are used to reduce the dimensionality of the preceding inputs by aggregating the activations into a single value for the next layer. This type of layer can be used for feature down-sampling, and to a degree, for imposing representation invariance to slight translations of the input. Pooling can be performed globally, or locally over small clusters, whilst often follows convolutional layers. There are two common types of pooling implementation:

  - **Max pooling**: uses the maximum value of the inputs as the output value
  - **Average pooling**: uses the average activation across the cluster.

- **Recurrent layers.** In recurrent layers, connections between neurons form a directed or undirected graph, typically along a temporal sequence. Most often, neurons are connected with others within the same or previous layers, combining both current and preceding activations. Recurrent layers are designed to exploit the temporal structure of the data, with units that can process sequences of inputs with variable lengths. Two main architectures that have been particularly successful over the past years – and often used alongside convolutional neural networks – are long short-term memory layers (LSTM) and gated recurrent units (GRUs).

Model Regularization. Contemporary deep learning architectures have a significant amount of trainable parameters – from thousands to millions – which essentially allow them to learn any kind of input-to-output mapping. For this reason, deep learning models can be susceptible to overfitting, given the many layers of abstraction which allow them to detect rare dependencies in the data (high-dimensional coincidences such as noise or confounds). To tackle these challenges, several approaches have been recognized and proposed for the improvement of model generalization.

Some of the most impactful strategies for dealing with overfitting are related to the acquisition of large, representative training datasets and the application of robust cross-validation approaches. The choice of training/validation/test sets can have a significant effect on the assessment of the model's performance, and the respective meaning we can acquire from it. Over the past years, and as we better understand the limitations of neural networks, an increasing trend for using mismatched training and validation/test set distributions is commonly found. By separating the training set distribution from the data found in validation/test sets, we can assess whether the network has generalized across particular desired features of the data. For example, in the case of clinical datasets, these distributions may vary across measurement configurations or participants. Nonetheless, it is the validation/test set distribution that eventually determines the learning process, and the overall performance of the model.

Another way to deal with model overfitting is by the employment of regularization techniques, which tend to make the model's function simpler, and more generic. Some of the most common regularization techniques are reducing the network size (usually by reducing the number of neurons in the hidden layer), early stopping (during training), adding weight regularization (e.g. L1/L2 regularization), and applying Dropout layers. Dropout in particular has been empirically proven as one of the most effective techniques available. The core idea here is that by introducing noise to the network during training (by randomly dropping neurons/features with a given probability), we can break up spurious patterns that might potentially memorized by the model. Of course, a combination of all of the above techniques can be used in parallel, in order to ensure robust training and evaluation.

Overall, beyond the success of deep learning, criticism remains with respect to the problem of network interpretability – as models are perceived as black boxes which can be empirically, rather than theoretically, evaluated. Over the past few years, researchers have put significant effort into developing algorithms and models towards explainable AI, following the principles of transparency, interpretability and explainability (Goebel *et al.* 2018). Of course, understanding the basis behind knowledge representation and the decision making of neural networks in terms of human appreciation, may not be always possible for many scientific problems. Nevertheless, a variety of algorithms have been developed to either provide inherently interpretable models, or to allow us to apply post-hoc explanations on existing trained models.

Chapter 3 Deep Learning for EEG Decoding of Anesthetic-Induced Unconsciousness

3.1 Introduction

3.1.1 Overview

In this chapter, we show – as a proof of principle – the ability of deep learning models to discriminate multiple levels of unconsciousness, given a fully automated end-to-end learning approach, and within a cross-subject decoding task. Deep learning has shown promising results in all biomedical domains, and particularly with convolutional neural networks, which use less parameters by computing convolutions over small regions of data (Miotto *et al.* 2018). Despite providing the opportunity for real-time applications, deep learning can also extract novel patterns and hierarchical representations, opening a data-driven approach for acquiring neuroscientific knowledge. As there is no state-of-the-art model or reference dataset for EEG classification, we compare the performance of two widely used architectures – multilayer perceptrons (MLP) and convolutional neural networks (cNN) – in their ability to discriminate three anesthetic states defined by clinical assessments of unconsciousness. Moreover, we investigate and compare the effect of the input representation, by using either the raw EEG time courses, or a spectral decomposition of the given time window, which has been shown to be an effective feature in many EEG decoding tasks (Liu *et al.* 2019). Using a leave-one-participant-out-cross-validation paradigm, we show that cNNs achieve 86% accuracy and significantly outperform MLPs, using only 1 second segments of the raw signals and with minimal preprocessing. We also show that both models perform equally well using the spectral feature extraction, which nevertheless is not needed to be relied on or used, given the ability of the networks to learn optimal parameters from the original and free-from-constraints data space. Finally, we discuss the implications of a cross-subject decoding analysis for model robustness, the representation efficiency of the two architectures, and the appropriateness of the chosen EEG window of analysis.

3.1.2 Background

Our aim here is to investigate the ability of deep learning to perform a classification task, in which resting-state EEG measures of healthy participants under anesthesia are used, in

order to differentiate and predict the various brain states characterized by decreasing levels of consciousness.

Literature review on EEG under general anesthesia has revealed a variety of reproducible findings (e.g. changes in delta power and coherence (Purdon *et al.* 2013; Mhuircheartaigh *et al.* 2013a), or changes in functional/effective connectivity and complexity (Hudetz 2012; Alkire 2008; Casali *et al.* 2013)), some of which appear to be shared among various anesthetics, yet missing to provide a unitary and reliable marker for levels of consciousness. Based on these findings, and depending on the nature of investigation, several techniques and metrics have been developed, which can be used to either measure states and levels of unconsciousness, or particularly the depth of anesthesia (e.g. the PCI index (Casali *et al.* 2013), signal complexity measures (Schartner *et al.* 2015; Hudetz *et al.* 2016), SWAS (Warnaby *et al.* 2017), connectivity measures (Juel *et al.* 2018), and others (Choi, Noh and Shin 2020; Colombo *et al.* 2019)).

Nevertheless, the majority of these techniques rely on several mathematical and theoretical assumptions about the nature of EEG and/or the nature of consciousness. For example, the perturbational complexity index (PCI) – one of the most successful indices of the past years – is based on approximating a measure of differentiation and integration of cortical activity, which is assumed to be required for consciousness. As these assumptions are open, with stronger or weaker scientific basis, and given that these techniques can significantly limit a data-driven investigation of the otherwise rich neurophysiological signals, machine learning offers a potential approach for end-to-end feature learning. Moreover, machine learning allows automation for clinical applications (and has a crucial role in many BCI systems), where current methodological practices in EEG analysis are constrained from long repeated measurements, they require extensive data processing tools, and often intervention from human expertise, which is unsuitable for real-time monitoring of patients.

Deep learning techniques in particular have already revolutionized many domains with their performance, including biomedical imaging and engineering (Miotto *et al.* 2018). Over the past years, there is an increased interest to use deep learning, and especially convolutional neural networks, as a way to decode the brain for EEG research and applications (Roy *et al.* 2019). Of course, as a new field of study, questions related to deep learning architecture, design and other methodological concerns are still open for investigation.

3.1.3 Related Work

Over the past years, an increasing number of studies appear to use deep learning models for EEG decoding, with all kinds of data and tasks, reporting a significant success (Schirrmeister *et al.* 2017; Heilmeyer *et al.* 2019; Lawhern *et al.* 2018; Roy *et al.* 2019). When it comes to clinical applications, there are already several efforts towards improving predictive models for diagnosis and prognosis, in cases such as epileptic seizures (Korshunova *et al.* 2017; Shamim Hossain *et al.* 2019), sleep stage classification (Chambon *et al.* 2018), and many

others. Specifically for general anesthesia, a variety of studies have tried to analyze EEG using deep learning techniques, for the purpose of discriminating states of unconsciousness.

Traditionally, such studies have focused on either spontaneous recordings of EEG (resting-state EEG) or event-related potentials (ERPs), for monitoring the depth of anesthesia and helping clinicians to guide the anesthetic procedure (Robert *et al.* 2002). For studies focusing on ERPs, the mid-latency auditory evoked potentials (MLAEP) have been used alongside neural networks, where MLAEP peak latencies have been correlated to the hypnotic component of anesthesia (Zhang *et al.* 2001). In more recent years, researchers have focused on different features of spontaneous EEG (e.g. autoregressive parameters, wavelet analysis, bispectral analysis, and others), as input to neural networks, while in some cases other physiological measures have also been used as additional input features (e.g. drug concentrations, electromyography (EMG), electrocardiogram (ECG), electrodermal activity (EDA), blood pressure (BP), respiratory rate (RR), hemodynamic parameters, or other signals which are traditionally and qualitatively assessed in clinical practice). However, here we focus on a number of studies that aim to analyze resting state EEG (rather than ERPs or other physiological measures), given that 1) the main action site of general anesthetics and the neural correlates of consciousness, reside in the brain and 2) we aim to investigate novel features from the rich spatio-temporal signals of the EEG, as traditional monitoring methods have been shown to be unreliable in measuring the depth of anesthesia (e.g. due to significant variability across drugs, patients, etc. (Lan 2012)).

Under this given scope, there is a limited number of works that have tackled this problem, with a variety of methodologies in anesthesia and EEG processing. In one of the earliest studies we are aware, (Roy and Sharma 1994) used autoregressive modelling and neural network analysis of EEG in dogs under halothane, where they predicted 3 depths of anesthesia with an accuracy of 85%. (Krkic 1996) reported high performance in humans (15 patients) under propofol and desflurane anesthesia using the spectral entropy of EEG as input to a neural network, under a 2-state classification problem, with up to 98% accuracy. (Lalitha and Eswaran 2007) showed almost perfect performance in 5 propofol patients, using non-linear chaotic features and multilayer perceptrons in a 3-state classification problem. More recently, (Liu *et al.* 2015) and (Jiang *et al.* 2015) compared the BIS index (a commercial standard in US healthcare for monitoring DoA) with their approach using multivariate empirical mode decomposition (MEMD) and sample entropy, alongside neural networks, reporting higher than BIS precision in a regression task against expert assessment of conscious levels (EACL). In most of these studies, results were encouraging under an intra-subject decoding task, with high performances obtained using a small number of electrodes (e.g. in BIS-like systems, 1-3 frontal electrodes are often found) and multilayer perceptrons (usually 1-3 layers, with deeper architectures being more successful). Table 3.1 references some of the works within this framework, albeit not exhaustively.

Table 3.1. Studies assessing anesthetic-induced unconsciousness using EEG and Neural Networks

Study	Subjects	Anesthetic	EEG Features	Architecture	Task	Validation	Performance
Roy et al, 1994	Dogs (10)	Halothane	AR Parameters	MLP	3-state classification	Intra-subject	85% accuracy
Krkic et al, 1996	Humans (15)	Propofol, Desflurane	NSV, Spectral Entropy	RBF	2-state classification	Intra-subject	98% accuracy
Kangas et al, 2000	Humans (7)	Isoflurane	EEG spectra	MLP	2-state regression	N/M	N/M
Lalitha et al, 2007	Humans (5)	Propofol	Non-linear chaotic features	MLP	3-state classification	Intra-subject	99% accuracy
Liu et al, 2015	Humans (26)	Propofol, Sevoflurane, Desflurane	MEMD, Sample Entropy	MLP	regression to EACL	Intra-subject	0.83 correlation coef.
Our study	Humans (9)	Propofol	Raw EEG/EEG spectra	MLP/CNN	3-state classification	Inter-subject	86% accuracy

In spite of some promising results, the comparison and evaluation of the findings across these studies is limited, due to the large variation in methodology. Regarding general anesthesia, the subjects under analysis (human/non-human), the type of anesthetic agents and procedures (e.g. the co-administration of other substances which might influence brain activity in patients or healthy subjects), the definition of anesthetic states and the clinical ground truth (e.g. states characterized based on loss of consciousness, recall, or reflex to noxious stimuli) are some important factors to consider, as there is no specific clinical/experimental setup and standardized way to assess levels of consciousness. Most importantly, the majority of the studies have used data from patients in real clinical scenarios, which introduce significant limitations for EEG recordings and significant variability due to patient-individualized comorbidities and treatments (e.g. a particular selection of medications and doses).

From an engineering point of view, the EEG preprocessing (e.g. artifact cleaning), the deep learning architecture, the optimization task (classification vs regression), the amount of data and the validation approach can also be important factors of variation in these findings. In general, most of the studies have used EEG with a number of preprocessing steps (e.g. MEMD, independent component analysis – ICA, or other cleaning techniques), in conjunction with a feature extraction technique as input to the deep learning model, and under a classification task. The feature extraction technique is often based on spectral features of the EEG, which have been heavily used throughout the literature. The variation of deep learning architecture is also important to consider, given the lack of agreement even in basic design principles, which heavily affects performance, such as network connectivity and organization, hidden layer depth, etc.

One of the major problems shared among these studies though, is reflected in the inconsistency of model evaluation, and most importantly in the lack of predictive generalization to unseen subjects (accuracy >90% has been achieved with cross-subject validation only for 2-state classification). This is particularly important given the inter-subject variability of EEG signals and the vulnerabilities of neural networks, which will be further discussed in the next sections.

3.2 Methods

3.2.1 Dataset Collection

The data used in this work were acquired from a propofol anesthesia study (Murphy *et al.* 2011), in which the experimental design is described in detail. Briefly, the study was approved by the Ethics Committee of the Faculty of Medicine of the University of Liege, with participants giving written informed consent. Moreover, physical examination and medical history were obtained, in order to assure of no potential issues during anesthesia (e.g. pregnancy, head trauma, surgery, mental illness, drug addiction, asthma, motion sickness).

Fifteen-minute spontaneous high-density electroencephalography (hd-EEG, 257 channel EGI Hydrocel GSN) was recorded from 9 healthy participants (mean age = 22± 2 y; 4 male, 5 female) during propofol anesthesia, at three different levels of consciousness: from fully awake (*Wakefulness*), to moderate sedation (*Sedation*), and finally loss of consciousness (*LOC*), as depicted in Fig. 3.1. The levels of consciousness were assessed using a behavioral scale (Ramsay score), after two consecutive verbal commands to squeeze the hand of the investigator (clear response to command in *Wakefulness* – Ramsay 2, slow response in *Sedation* – Ramsay 3, and no response in *LOC* – Ramsay 5-6). Sedation procedure was continuously monitored (electrocardiogram, blood pressure, SpO_2, etc.) and additional oxygen (5 liters/min) was provided to the participants. Computer-controlled intravenous infusion was used to estimate effect-site concentrations of propofol (3-compartment pharmacokinetic Marsh model), while a 5-minute equilibration period was allowed after reaching the desired Ramsay state, to ensure steady-state recordings. Average levels of propofol were 1.91 ± 0.52 for *Sedation* and 3.87 ± 1.39 for *LOC*, as measured by arterial blood samples before and after each anesthetic state (Murphy *et al.* 2011).

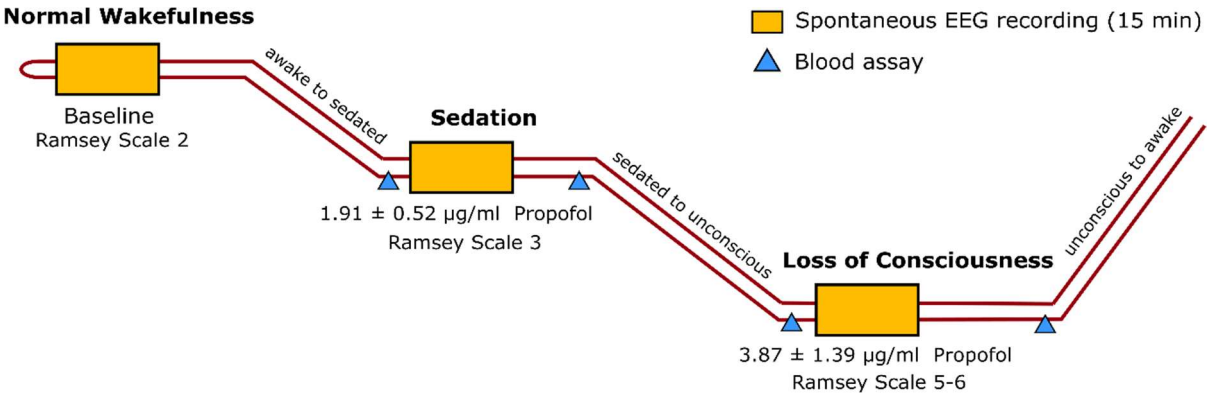

Fig. 3.1. Experimental design of the propofol anesthesia study. Nine participants underwent anesthetic induction into progressively deeper states of unconsciousness, measured by a clinical scale (Ramsay score).

An illustration of the raw EEG data can be seen in Fig. 3.2. Steady-state recordings were acquired after the Ramsay assessment and the drug's equilibration period, which ensured the exclusion of confounding noise and task-related EEG activity from the behavioral scoring (auditory/motor activity).

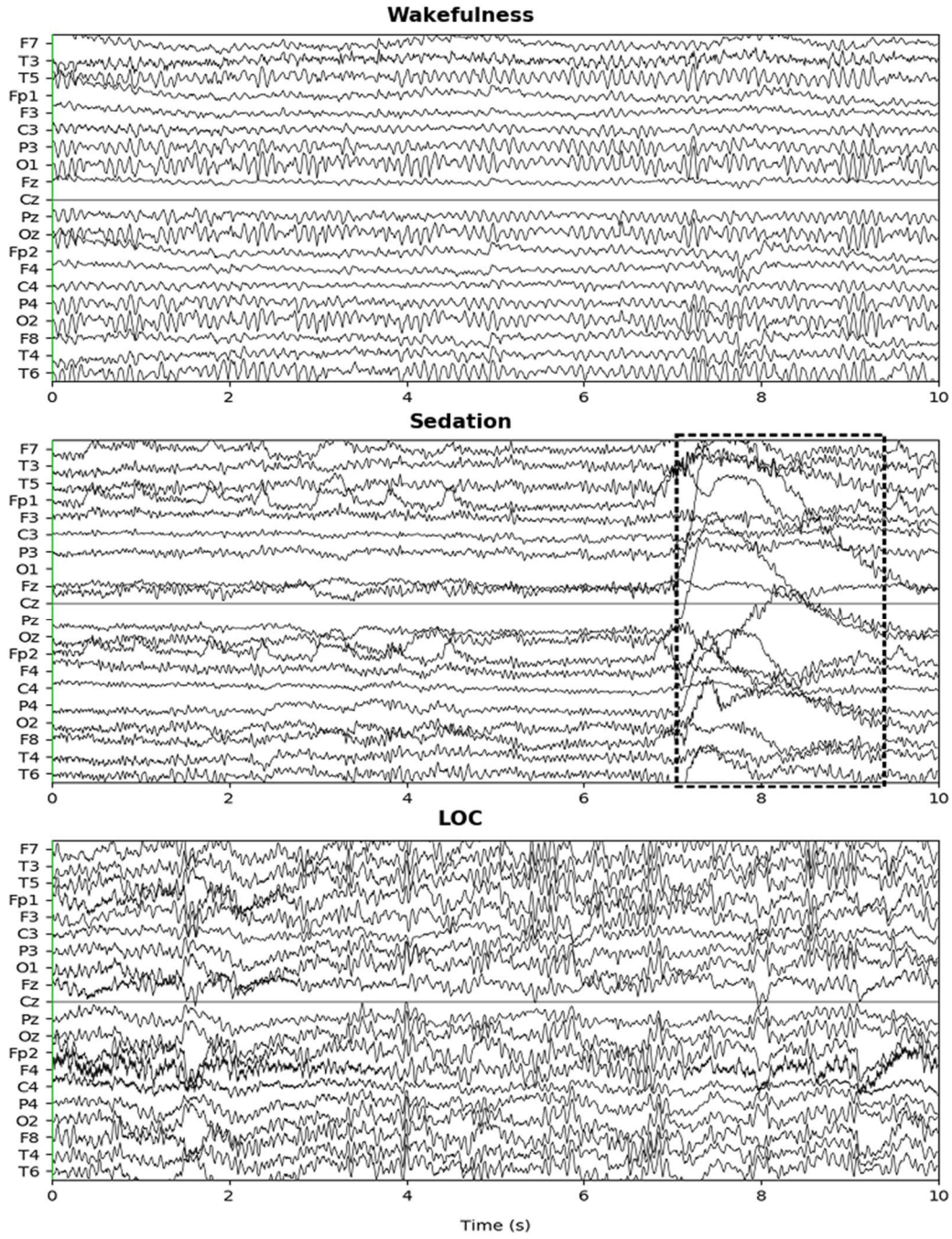

Fig. 3.2. Resting-state EEG recordings of one participant, during the states of *Wakefulness*, *Sedation* and *LOC*. Twenty channels are depicted based on the 10-20 system. An example of a prominent movement artifact is highlighted in the dashed box (Duration=10 sec, Scale=50 μV).

3.2.2 EEG Pre-processing

Minimal pre-processing steps were applied to the original data, in order to simulate a real-world scenario where deep learning could be applied to EEG data in real-time. Although raw EEG recordings tend to be noisy, the selection of the workflow was based on the notion of an automated feature extraction done by deep learning. Such implementation has a potential practical value within a clinical context, as manual intervention and *a priori* knowledge of the signals would be infeasible (albeit often used in preprocessing methods for artifact rejection).

Two different representations were extracted from the data, to compare the effect of using the raw time series versus a spectral representation. The latter has often been used in similar studies as a useful feature for EEG classification (Schirrmeister *et al.* 2017; Stober, Cameron and Grahn 2014; Howbert *et al.* 2014; Park *et al.* 2011).

Raw Data Representation. To reduce the dimensionality and computational complexity of the deep learning pipeline, 20 electrodes were selected from the EEG, located as per the standard international 10-20 system, namely: Fp1, Fp2, F7, F3, Fz, F4, F8, T3, C3, Cz, C4, T4, T5, P3, Pz, P4, T6, O1, Oz, and O2. Data were segmented into 1 second non-overlapping epochs (windows) and band-pass filtered between 0.5-40 Hz (as per the original study) using a window FIR design (*firwin*, *scipy*). The vertex (Cz) electrode was the online reference, with its activity replaced by the average activity across all 19 channels. Finally, the time series were down-sampled to 100 Hz, resulting in 100 samples per epoch. No artifact or bad channel rejection was performed, other than the removal of the first 10 seconds of recording which contained large unstable drifts. All pre-processing steps were implemented using the *MNE-python* library with default settings (*ver. 0.15*), unless specified otherwise.

Power Spectral Density Representation. To generate the spectral representation of the EEG, the raw data processed as above were submitted to the periodogram function (*scipy*), in order to obtain the power spectral density (PSD) of each channel and epoch. 201 points were used to compute the PSD, which resulted in 100 frequency bins (one-sided spectrum, dc coefficient removed). Importantly, this ensured that the dimensionality of the data was identical in both raw and PSD representations.

In both representations, the resulting dimension of each epoch instance was a 20 x 100 2D-array (channels x time samples/frequency bins). These data were normalized epoch-wise using the *StandardScaler* (*Scikit-learn*), before feeding them into the deep learning networks. This can be thought as standardizing (z-score) the whole spatio-temporal activity for each epoch and participant independently. Although there are many ways to normalize the data (e.g. by time sample or by channel), this way can be considered more appropriate in terms of its physical interpretation, but also from a practical perspective, as only data from the current epoch are required for applying the normalization.

3.2.3 Deep Learning Architectures

Two deep learning architectures were compared, in order to investigate the suitability of the algorithms in extracting relevant electrophysiological features. Convolutional neural networks (cNN) are a class of feed-forward networks that have shown immense success, initially within computer vision problems, but more recently in other fields as well (e.g. natural language processing, healthcare, and others). Within EEG research, cNNs are becoming particularly attractive during the last few years, with an increasing number of studies that employ them in all kinds of tasks (Schirrmeister *et al.* 2017). Interestingly, there are two basic ideas behind their architectural design. The first one relates to the extraction of local patterns that are invariant across the data, by using convolutional kernels with shared parameters. As with many natural signals, the assumption here is that spatially or temporally nearby features are more likely to share mutual information, thus making the sparse connectivity of the networks more efficient (this assumption appears to be, for the most part, true for EEG). The second idea relates to the composition of local low-level patterns, into more global higher-level patterns, through layers of abstraction which create a hierarchical representation of features (Krizhevsky, Sutskever and Hinton 2017). Contrary to this architecture, the Multilayer perceptron (MLP) network is a naïve implementation of a fully-connected neural network, which has been previously used in similar studies and can serve as a baseline for comparison (a cNN can be thought as an MLP with a specialized structure).

Our aim here was not to optimize each network for the given task, but rather to compare them fairly, in revealing the computational advantages of each design. Hence the two models were compared with respect to their architectural sizes, which can be thought as the number of neurons or trainable parameters, within each functional layer.

Convolutional Neural Network. The architecture of the cNN is a sequential model based on a simple convolutional design that has been used in many computer vision tasks (e.g. mnist classification). The input samples to the network are the EEG epochs structured as a 2D matrix ($X_i \in R^{c\ x\ l}$), where rows correspond to channels (c) and columns correspond to either time samples (l), or the spectral coefficients (l), depending on the selected representation. The first functional layer (feature extraction/data compression) is a sequence of two convolutional layers, followed by a max-pooling and a dropout layer. The second functional layer (classification), consists of a fully connected layer, followed by a dropout layer and three softmax units (one for each conscious state). As a reference size, the original number of feature maps and hidden neurons were used, namely 32 for the 1st convolutional layer, 64 for the 2nd convolutional layer and 128 neurons for the 3rd dense layer. The patch window for max pooling was 2x2. Dropout rates were 0.25 and 0.5, respectively. Convolution windows were chosen with kernels 1x5 and 5x10 (1x1 strides), with the first layer only extracting temporal information (no padding used). Finally, all activation functions were ReLU units (except output layer). The cNN architecture is summarized in Fig. 3.3.

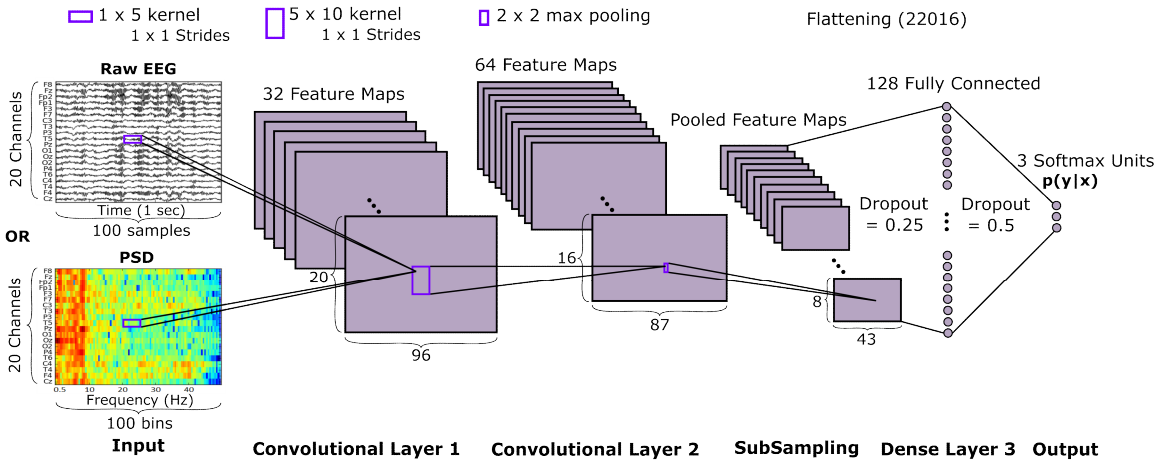

Fig. 3.3. Convolutional neural network architecture (reference size) for the 3-state classification task (*Wakefulness*, *Sedation*, *Loss of Consciousness*). The raw EEG or the PSD epochs are used as input tensors to the network.

Multilayer Perceptron. The architecture of the MLP is also a sequential model, with the input represented as a vector (2D matrix flattening). We employed a model designed to match the number of output neurons in each functional layer of the cNN (rather than equalising network layers). This ensured that the computational cost of each design was comparable in terms of training time. Both layers of the MLP consist of fully connected layers, followed by a dropout layer, with the 2nd layer including the three softmax output units. The number of hidden units for the 1st layer was based on the number of neurons after the flattening of the 1st functional layer of the cNN architecture (22016 for the reference size), while for the 2nd layer was kept the same. Activation functions, dropout rates and other model parameters during training were also kept similar to the cNN design. The MLP architecture is summarized in Fig. 3.4.

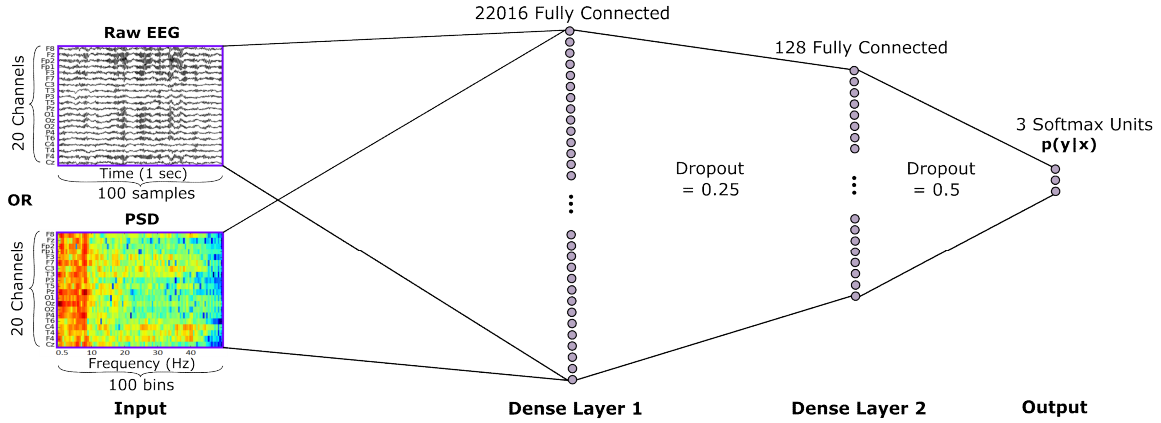

Fig. 3.4. Multilayer perceptron network architecture (reference size) for the 3-state classification task (*Wakefulness*, *Sedation*, *Loss of Consciousness*). The raw EEG or the PSD epochs are used as input tensors, after flattening the 2D-array into a 2000-dimensional vector.

3.2.4 Experiments

Twelve experiments were performed in total for the 2 x 2 x 3 combinations of input data representations (Raw/PSD), deep learning architectures (MLP/cNN) and three different network sizes – small, reference and large – in order to compare the performance of the models under all possible configurations (network sizes were included to ensure that the observed effects were not size-depended). The number of feature maps and the number of neurons for each architecture and network size can be seen in detail in Table 3.2.

Table 3.2. CNN and MLP Network Sizes – Small, Reference, Large

Network Size	cNN	MLP
Small	(**16**, **32**, 64)	(11008, 64)
Reference	(**32**, **64**, 128)	(22016, 128)
Large	(**64**, **128**, 256)	(44032, 256)

Number of feature maps are denoted in bold

To evaluate model performance, the EEG data need to be divided into training and test sets. Previous studies have used a training/test set split which incorporates EEG data proportionally across participants (Juel *et al.* 2018; Korshunova *et al.* 2017; Stober, Cameron and Grahn 2014). However, a harder but ideal goal would be to generalize the prediction of consciousness states in unseen participants. With this goal in mind, a leave-one-participant-out cross validation (LOPOCV) paradigm was used for training and testing the models, with each participant contributing on average ~2700 instances (9 participants x 3 states x 15x60 1-sec epochs ≈ 24300 total instances). Each instance was labeled with one-hot encoding as the target vector, indicating one of the three anesthetic states. The model was trained using the categorical cross-entropy loss function and the Adadelta optimizer. Initialization of network weights was done with the Xavier uniform initializer. A batch size of 100 was used, and for 10 runs of the data (training epochs). Models were evaluated by their accuracy, computed as the percentage of correctly predicted epochs in the left-out participant. All experiments were implemented in Python 3 using the *Keras*/*Tensorflow* library and a CUDA NVIDIA GPU (Tesla P100).

3.3 Results

3.3.1 Architecture Comparison

The results of our 2 x 2 experimental design (Raw/PSD *X* cNN/MLP) were consistent for all three network sizes and are summarized below. Reported figures and accuracies are for the reference size networks, depicted in Figs. 3.2 and 3.3.

Raw Data. Using the raw EEG input, the MLP achieved an average accuracy of 75.45% across participants, with the cNN reaching an average of 86.05% (Fig. 3.5). These accuracies are significantly higher than the chance level accuracy of 33.33%, expected for the 3-state problem. Cross-entropy loss on the test sets did not significantly decrease after the first few epochs for both implementations, while it showed an increase after the 5th epoch for the cNN. Despite this, the cNN was able to achieve better accuracies for each state of consciousness, as observed from the confusion matrices (Fig. 3.5).

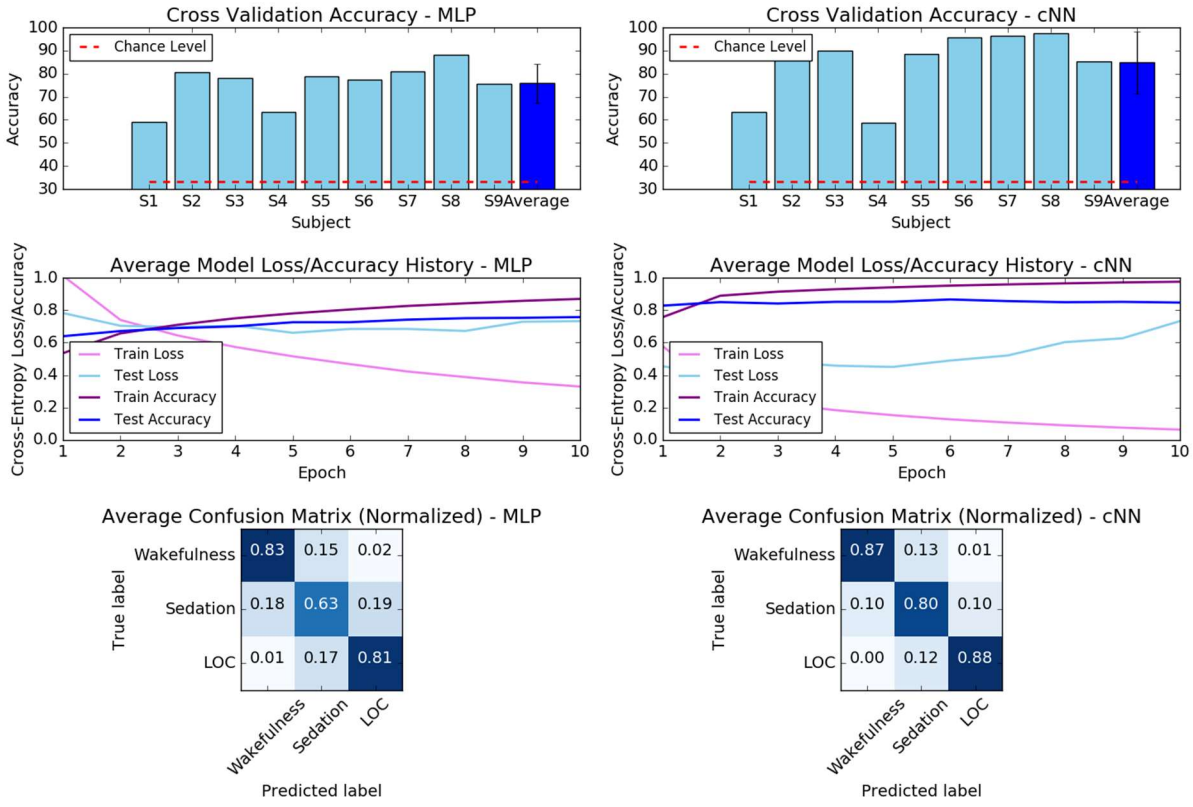

Fig. 3.5. MLP vs cNN (reference size) comparison of the raw EEG classification for the 3 anesthetic states. Cross-validation accuracies, average model loss, and confusion matrices are shown for each architecture.

Power Spectral Density. Using the PSD input, the two architectures were more similarly capable in classifying the 3 anesthetic states, with reported accuracies of 83.4% for MLP and 87.35% for cNN (Fig. 3.6). Notably, we observe a significant accuracy increase for the MLP compared to the corresponding raw input configuration (~12% increase), but not for the cNN. Moreover, cross-entropy loss curves revealed that the models converged faster using the PSD representation. This can be understood under the expectation of the networks to require more training iterations (epochs) before an adequate feature learning state is obtained, compared to an already compressed representation (such as the PSD).

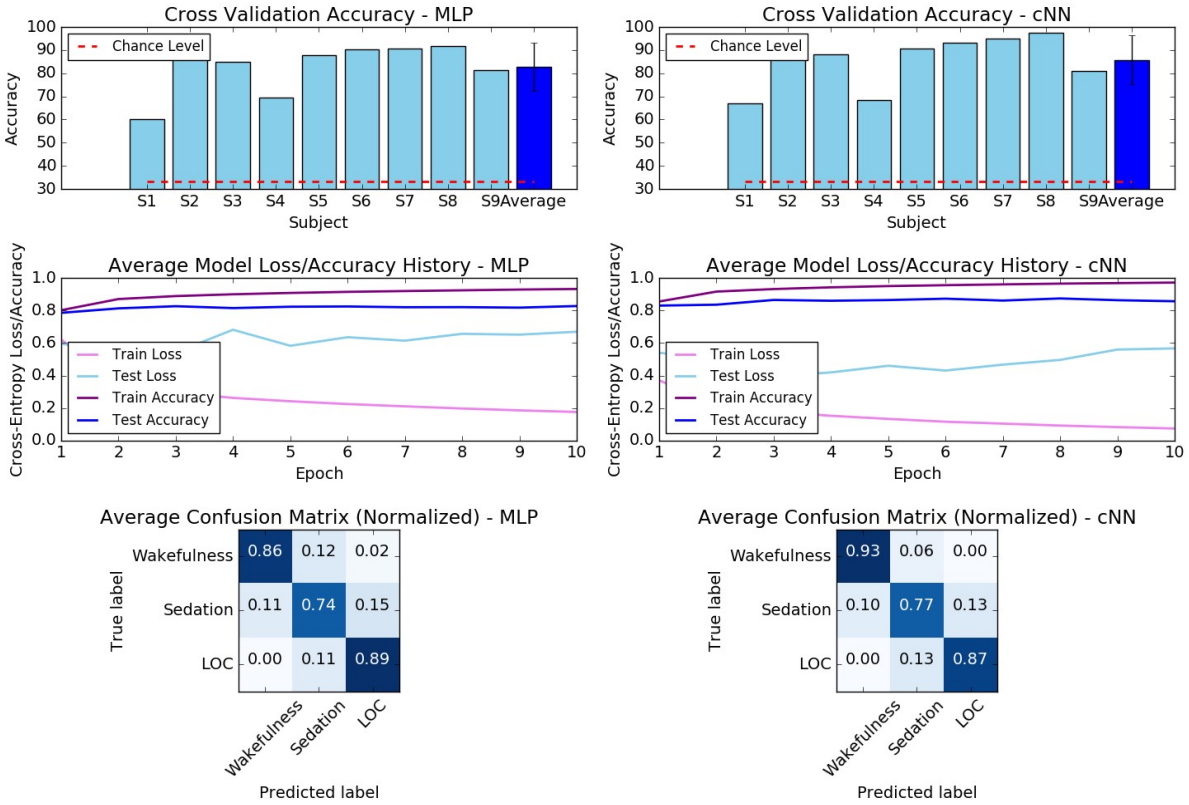

Fig. 3.6. MLP vs cNN (reference size) comparison of EEG classification for the 3 anesthetic states, using the PSD representation. Cross-validation accuracies, average model loss, and confusion matrices are shown for each architecture.

Overall, average model loss shows that the 10 training epochs were adequate for a stable convergence of the models, in all configurations. Figures 3.4 and 3.5 reveal that the models obtained consistently high performance for six testing participants (S2, S3, S5, S6, S7, S8), with the remaining three (S1, S4, S9) resulting in lower performances (65-85% accuracy, for cNN). Moreover, confusion matrices show that *Wakefulness* and *LOC* were not misclassified to one another. On the other hand, the intermediate state of *Sedation* was the hardest to predict, possibly due to a transitional nature of the EEG signature and the imposed classification task. Another possibility for such effect could relate to the inter-individual variability in response to propofol, which has been documented in (Chennu *et al.* 2016), and which could manifest in the electrophysiological signatures of *Sedation*. In this regard, the cNN using the raw input data had the most balanced performance, with *Sedation* class reaching 80% accuracy.

To better understand the changes of the underlying EEG signals that drive these accuracies, we visualized the PSDs in each state of consciousness (Fig. 3.7). As previously reported, we observe a decrease in alpha oscillations for *Sedation*, followed by the emergence of high-alpha and delta oscillations during *LOC* (Patrick L. Purdon *et al.* 2015).

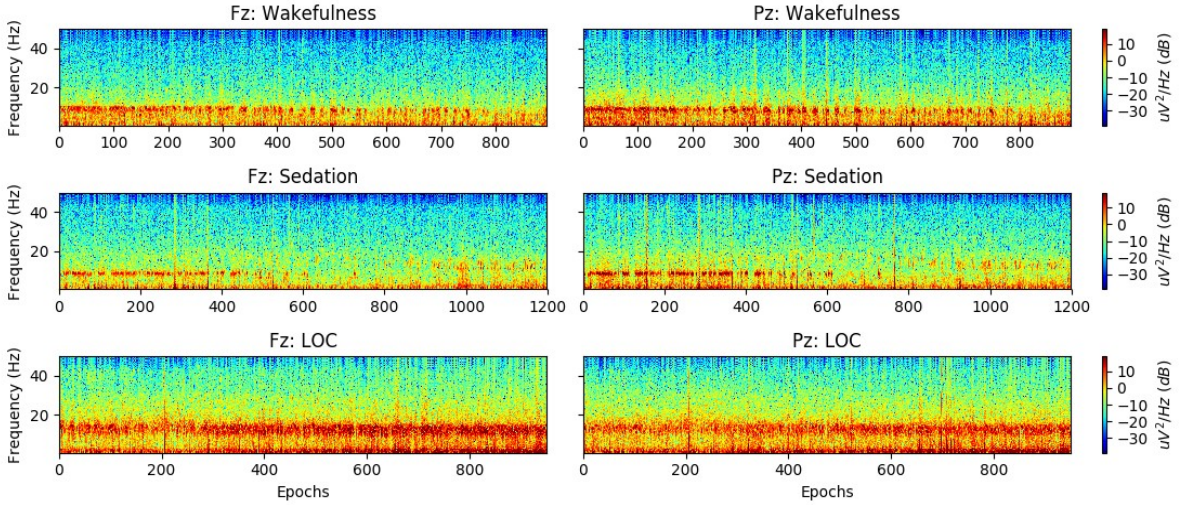

Fig. 3.7. Power spectral density representation (uV^2 /Hz, dB) of the EEG epochs, for each of the 3 anesthetic states (*Wakefulness*, *Sedation*, *LOC*). A representative frontal (Fz) and parietal (Pz) electrode is shown for one subject.

3.3.2 Statistical Analysis – ANOVA Model

As a final step to this analysis, a three-way ANOVA (type 2) was performed on the accuracies obtained by our 2 x 2 x 3 experimental design (architecture x input representation x model size), which is depicted and summarized in Fig. 3.8 and Table 3.3.

The results of the ANOVA indicate that the network architecture (cNN/MLP) was the strongest contributor to model performance (F = 10.6, p=0.0015), with the input representation (Raw/PSD) also having a significant but weaker effect (F = 5.34, p=0.0229), driven by the improved accuracy of MLPs using the PSD data (p=0.08 interaction effect). As we already mentioned, model size had no contribution to overall performance in any configuration.

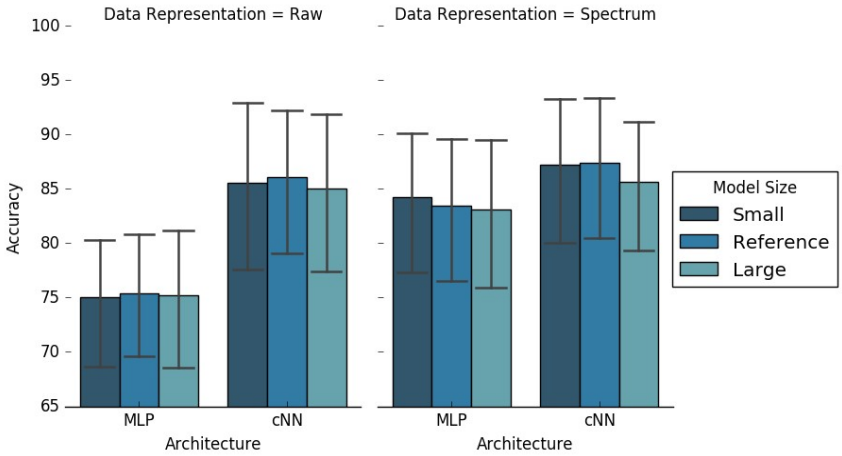

Fig. 3.8. Model accuracies for each configuration of our 2 x 2 x 3 experimental design (architecture x input representation x network size). Error bars indicate the 95% confidence interval.

Table 3.3. ANOVA Table of Model Comparison

	Sum_sq	df	F	Pr(>F)
Architecture	1230.94	1	10.601	**0.0015**
Data Representation	620.06	1	5.340	**0.0229**
Model Size	14.34	2	0.061	0.9401
Architecture-Data Representation	351.13	1	3.024	0.0852
Architecture-Model Size	6.21	2	0.026	0.9735
Data Representation-Model Size	6.18	2	0.026	0.9737
Architecture-Data Representation-Model Size	0.85	2	0.003	0.9963
Residual	11146.71	96		

3.4 Discussion

3.4.1 Cross-subject Generalization

Our findings have highlighted the capabilities and potential of deep learning to discover and utilize generalizable features from human EEG, given the task of automatic identification of multiple anesthetic states characterized by decreasing levels of consciousness. As generalization performance is often missed in EEG studies, this work provides a solid foundation for further investigation. Training and testing on different participants (with leave-p-subjects-out CV) can have a profound effect when creating and validating the models, as EEG signals can show significant variability across subjects (Melnik *et al.* 2017), and deep learning is particularly vulnerable in identifying false patterns (e.g. within-subject confounds or noise). While many previous studies have used an intra-subject decoding methodology, which can be useful for certain tasks (although it can require training time for real-time BCI applications), there is a general paradigm shift over the past years, and particularly within machine learning, for more appropriate and systematic evaluation towards model robustness. Besides its relevance for finding universal patterns of neuroscientific value, clinical models are most often used and evaluated in new patients, while in particular for anesthetic procedures, subject-specific training would be infeasible in clinical settings, given the importance of time management and expenses in hospitals (i.e. patient time under GA for surgery).

3.4.2 Architecture Comparison and Representation Efficiency

Apart from cross-subject generalization, we have also shown that modern cNN architectures significantly outperform fully connected MLPs (in agreement with the current literature (Hernandez 2017)), potentially due to their ability to extract more effective spatio-temporal features from the raw signals. This notion is supported by the fact that MLPs

performed as well as cNNs when given the PSD data as an input. Nevertheless, the cNN with raw input representation configuration was optimal, as it obtained the most balanced within-class error, which indicates better generalization. Similar findings have been observed in other studies as well, and in particular for cNN analysis of EEG under anesthesia, using raw or spectral/band power features, with the raw signal having the best results (Sun *et al.* 2019a; Truong *et al.* n.d.).

Using the raw input data is also important for the identification of novel task-relevant biomarkers, as feature-engineering based on expert knowledge and prior assumptions on the signals, can neglect information of interest. Specifically, the raw input representation is preferable to PSD in the sense that the model's parameter optimization happens within the original data space, which is not constrained within a Fourier-type spectral decomposition. Although EEG signals show a variety of frequencies (which have already been studied extensively and have revealed a variety of signatures for different anesthetics (Patrick L. Purdon *et al.* 2015)), other recurring and shift-invariant waveform patterns might be better suited for a cognitive and clinical understanding of the underlying neurophysiology (Jas, La Tour, *et al.* 2017). The distortion of the signals due to the often-used Fourier/Morlet methods, has been referred to as a 'Fourier fallacy', given the ad hoc assumption of a sinusoidal nature of the neuronal activity in the brain (Jasper 1948).

In terms of resource utilization, the cNN was also better than the MLP, as the latter had a significantly larger number of parameters to learn (i.e. 46,872,579 in MLP vs 2,921,219 in cNN, for the reference network size), by a factor of 16. CNN was also faster to train by ~35%. Furthermore, a repetition of the above experiments with an alternative comparison using the same number of trainable parameters (rather than the same number of neurons) in each architecture, gave a much more prominent difference in the results, with the MLP performing much worse and having higher computational demands. The number of model parameters is an important factor for machine learning algorithms, as a large number of trainable parameters can easily lead to overfitting, if the model is not provided with enough samples (although there is no mathematical relationship between the two, we know empirically that given enough parameters a model can learn individualized samples). Nonetheless, several deep learning techniques have been developed over the past years to effectively avoid overfitting (e.g. L1/L2 regularization (Neyshabur, Tomioka and Srebro 2015), dropout (Srivastava *et al.* 2014), etc.), one of which is the dropout layers used in our architectures.

3.4.3 Window of Analysis

Further experimentation with the EEG epoch size (2, 5 and 10 sec lengths, with or without 50% window overlap) did not show any significant changes in model performance. This could be partly justified given that the sizes of the cNN kernels remain fixed, which results in specific constrains for feature representation through most of the layers in the network. When using a 10 second window, performance showed a minor increase, which is nevertheless

expected as a consequence of the decrease in the total number of training instances (error averaging).

Previous studies on deep learning-based assessment of the depth of anesthesia have used a variety of input windows, ranging from 1 to 30 sec. From a theoretical standpoint, conscious experiences can be thought as either a continuous stream of percepts, or, as most experimental evidence support, percepts that are discrete in nature. Even in this latter case, experiences range from few milliseconds to few hundreds milliseconds at maximum, depending on the modality of the experience (Herzog, Kammer and Scharnowski 2016). Given this fact, a window of 1 second can be considered long enough to capture the changes of the underlying electrophysiology of interest. This idea is also supported in anesthesia studies focusing on EEG microstates, where the observed microstate duration ranges from 20 – 120 ms (Comsa, Bekinschtein and Chennu 2019; Shi *et al.* 2020), making our temporal cNN kernel sizes, sensible. From a clinical application perspective, a small window size is desirable considering that the temporal resolution for a DoA monitor is critical (typically, anesthesiologists have to respond rapidly to a change during surgery).

Beyond the input window of analysis, changes in brain activity that take place over several seconds or minutes, can also contribute significantly to the performance of our model. As our current analysis is based on a steady-state classification within an experimental anesthetic setup, and clinical anesthetic induction/emergence is gradual and rapid (Juel *et al.* 2018), the difficulty of classifying intermediate sub-anesthetic states may not present as a problem here. This can be further studied and determined, depending on whether the problem is posed as a classification problem (assuming a discrete transition to unconsciousness), or a regression problem, with a gradual change in levels of consciousness. In any case, our selected architecture presumes static vector-based inputs, which do not consider the large-scale temporal dynamics. Whether a deep learning architecture (such as the RNN) that allows the exploitation of the temporal dynamics of the EEG can be useful in such task, would require further research.

3.4.4 Summary

Overall, this study aimed to conduct a comparative analysis of the two most widely used deep learning architectures, rather than a hyperparameter optimization of the models, aiming to maximize their performance. The fact that cNNs were able to perform well given only 1 second of raw EEG data suggests that they could find utility both in electrophysiological investigation, but also in real-world applications, for assessment and monitoring of consciousness. Of course, apart from the choice of deep learning architecture and the non-necessity for feature engineering, the particular methodology of EEG preprocessing and cNN network design can still be a significant aspect of EEG representation, feature learning and knowledge discovery. In the next chapter, we investigate these questions, with the goal of creating a specialized model for EEG decoding.

Chapter 4 Convolutional Neural Networks and EEG Representation

4.1 Introduction

4.1.1 Overview

In this chapter, we focus on the development a novel convolutional neural network, based on the architectural design described in Chapter 3, which will be able to take full advantage of both the spatial and temporal structure of the EEG, irrespective of the system's configuration (no. of channels, channel locations, etc.). Specifically, we combine the advantages of the topomap projection and mesh representation, to create a 3D representation of the EEG epochs (scalp activity images vs time), and to allow 3D convolutional layers to explore possible spatio-temporal relationships in the data (in parallel with the work found in (Tan *et al.* 2017)). Another aim of this chapter, and in accordance with our goal to obtain a standard processing pipeline, is to investigate the effect of the EEG preprocessing parameters in a systematic way. Choices such as the number of channels and the spatial resolution of the EEG, the reference montage, the type of filtering applied, the manipulation of artifacts, as well as any type of epoch-wise sample normalization, are independent and specific to each work (while often without a given reasoning). As discussed in Chapter 2, the performance of machine learning algorithms is strongly dependent on the nature of the data, their representation and dimensionality, the number of available samples, and the level of noise. To understand the contribution of such factors, we perform a number of experiments and compare the performance of 2D and 3D cNN models, with respect to some of the most influential EEG parameters, namely: the montage (channel reference), the sample normalization method, the number of electrodes and the presence of high-frequency content, as well as the robustness of the models to EEG artifacts. The two model designs and the different EEG configurations are evaluated under the classification task of the three anesthetic states, as previously described. The strengths and weaknesses of each model, alongside the various preprocessing choices, are then discussed in detail. Finally, we derive a generic 3D convolutional neural network and a selected pipeline, which will help us proceed with our main research question.

4.1.2 Background

Our aim here is to explore the possibility of a convolutional neural network (cNN) with a unified architecture that will allow us to incorporate different EEG systems and datasets, having two goals in mind. The first goal relates to the consistent analysis of EEG data using a common network and overall methodology, which as discussed in the previous chapter, is missing and which in turn will support the integration and comparability of our findings. The second goal regards the transformation of any EEG channel configuration into a 2D structure, which will preserve the spatial properties of the signals, and exploit the strength of the cNN to extract potentially meaningful local representations, via the convolution operation.

The problem of creating a consistent EEG processing methodology starts with the lack of a standard preprocessing pipeline, that precedes the analysis of the signals by the neural network. To date, we are not aware of any systematic investigation on the effect of the EEG preprocessing methods with respect to any particular task or analysis, with the majority of researchers using their own techniques and parameters (e.g. regarding the channel selection, filtering parameters, or the manipulation of possible artifacts/noise). Beyond this pipeline, a feature extraction step is also commonly found among methods of analysis, especially within machine learning algorithms (EEG features are often fed into neural networks). Several signal processing techniques have been used for feature extraction and across a variety of EEG tasks, such as the power spectral density (PSD), wavelet transforms (WT), the Hilbert Transform (HT), the autoregressive model (AR), the filter bank common spatial patterns (FBCSP) algorithm, and others (often used alongside algorithms such as SVMs, k-nearest neighbours, linear discriminant analysis, and shallow NNs, with variable success (Wilaiprasitporn *et al.* 2020)). Nonetheless, human-engineered features based on prior assumptions (or expert knowledge) are not generalizable across studies or tasks, they can neglect information from the raw data, and they often lead to reproducibility problems (as these signals tend to be complex, with high inter-subject variability). To make matters worse, many of these techniques rely on a channel-wise analysis of the data, which potentially dismisses relevant information about the spatial dynamics of the EEG.

When it comes to the EEG channel configuration, and particularly for anesthesia research, there is a variety of studies using a number of channels, ranging from one to few hundreds. A large portion of these studies have explored the electrophysiology of the brain under clinical settings, with a limited number of EEG electrodes, as entailed by the preparation time and cost restrictions (e.g. BIS-like systems use 1-3 frontal electrodes, without full-coverage of the head). Given that there is no standard methodology for monitoring the depth of anesthesia, and since the precise mechanisms of anesthetics are unknown, experimental studies using multi-channel EEG systems become more important for neurophysiological investigation (Dubost *et al.* 2019). In addition, the acquisition of EEG recordings with high-density systems, often found in research, alongside the recognition and willingness from the community to share large datasets, opens the possibility for exploring the role of spatial

dimension in decoding brain activity. Of course, training large-scale networks with millions of parameters, based on the extraction of multivariate spatio-temporal features, is an area where deep learning thrives.

Given all the above, studies using deep learning have not fully explored the capabilities of the networks for EEG-based feature extraction, nor their relation to EEG preprocessing pipelines. Convolutional neural networks (cNNs) in particular, have shown their strength in revealing the spatial structure of the data. Conventionally, cNNs have been used with EEG input as a 2D matrix, where rows correspond to channels, and columns to time samples ("amplitude vs time" representation, as used in Chapter 3). However, there are two problems related to this form: 1) the number of channels might be different for different systems, resulting in variable input-matrix shapes, and thus, network parameters, and 2) the channel relationship across one dimension cannot be consistent or coherent, as electrodes exist in a 3D space, and feature learning is undertaken by kernels with shared parameters. Of course, whether cNNs can extract robust spatial EEG features, given the underlying assumptions and constraints of cNN kernels (e.g. with respect to spatial homogeneity), is still unclear from the literature.

4.1.3 Related Work

During the past few years, a variety of studies have used different EEG representations and designs of cNNs, partially investigating the above research question. While most of the efforts have been focused around EEG features and cNN input representations (rather than EEG preprocessing), there are several findings of interest to consider.

(Lai *et al.* 2019) made a comprehensive analysis on various arrangements of the EEG input, to examine their suitability towards a classification task. Six types of input arrangements were tested, namely: amplitude vs time, energy vs time (Hilbert Transform), amplitude vs time with channel rearrangement (based on Pearson correlation), and three corresponding arrangements, with the matrices transformed into images. Based on classification accuracies, they found that amplitude vs time performed best (intra-subject validation). Rearranged channels based on Pearson correlation showed significant decrease in performance, potentially due to the inconsistency of representation, as the arrangement would be subject to the given data, and not suitable for cross-subject generalization (moreover, highly correlated features are often avoided in several machine learning algorithms). Image representations performed worse on average, possibly due to the information loss from the compression of the images, as also recognized by the authors. Nevertheless, the above experimental designs did not take advantage of the known electrode locations.

(Alkanhal, Kumar and Savvides 2019) used spectral features over 30-sec windows as a series of images, by using a topomap projection (projection from the 3D to 2D space of channel locations). Although the window of analysis was large for our own task, the study claimed robust feature learning. (Yao, Plested and Gedeon 2018) focused on learning discriminative features from short-time EEG signals, using a channel-wise and Image-wise (topomap)

autoencoder approach. Again, spectral features (3 independent bands) were used as color channels, with Image-wise autoencoder performing better under a cross-subject validation. (Wilaiprasitporn *et al.* 2020) used multiple band-pass filtered EEG by mapping the electrodes into a 2D mesh, which preserved the spatial relationships of the sensors similar to the topomap projection image, achieving good classification performance (no cross-subject validation). These studies with topomap images/mesh used 2D convolutional layers, with kernels extracting only spatial information though (shared, or time-distributed 2D-cNNs).

(Tan *et al.* 2017) created a 3D architecture with multiple band-pass filtered EEG inputs, as video information (topomap image vs time), along with the optical flow information of this "EEG video", which used 3D kernels to extract spatio-temporal features. Results showed state-of-the-art performance over 1D cNNs, but without cross-subject validation.

Finally, (Bashivan *et al.* 2015) used spectral topomap images and compared cNN and mixed RcNN (recurrent convolutional neural networks) architectures, by including an LSTM layer before the classification layer, as a way to extract the temporal structure of the EEG. A similar approach was found in (Zhang *et al.* 2019). Although architectures could not be directly compared (e.g. due to differences in number and types of layers), resulting performances did not show any change of improvement for the mixed models. Nevertheless, the inclusion of the recurrent network significantly increased the number of trainable parameters (by 2-3 times). In addition, RNNs tended to have less stable training (due to the vanishing and exploding gradient problem (Pascanu, Mikolov and Bengio 2012)), and could potentially increase the complexity and interpretability of the trained model.

Of course, an accurate evaluation of all the above network architectures is hard, without comparing the classification problems, datasets, validation techniques, and overall model parameters.

4.2 Methods

4.2.1 Dataset Collection

For this analysis, we used the propofol anesthesia dataset (Murphy *et al.* 2011), described in detail in 3.2.1. Briefly, fifteen-minute spontaneous high-density EEG (EGI Hydrocel GSN, 257 channels) was recorded from 9 healthy participants during propofol anesthesia, at three different levels of consciousness, defined by the behavioral response of the participants (Ramsay score). The experimental design is depicted in Fig. 4.1. As shown in the previous chapter, the classification of three distinct anesthetic states (*Wakefulness*, *Sedation* and *Loss of Consciousness* (*LOC*)) provides a robust ground truth, yet a non-trivial problem, for testing the two cNN architectures and EEG preprocessing methods.

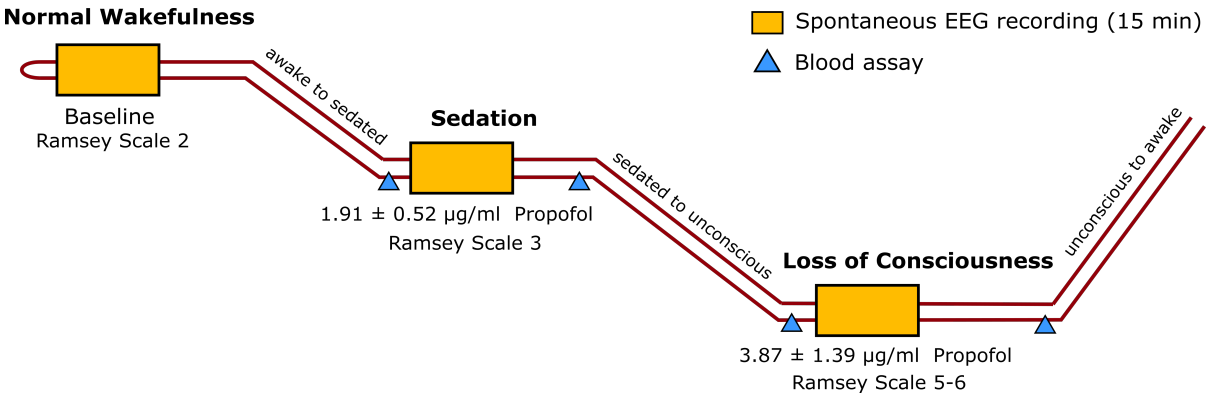

Fig. 4.1. Experimental design of the propofol anesthesia study. Nine participants underwent anesthetic induction into progressively deeper states of unconsciousness measured by behavior (Ramsay score)

4.2.2 EEG Pre-processing

The preprocessing methods are discussed here in detail. The specific choices for each parameter were selected during the course of the experiments, based on the performance of the models, or other discussed justifications, in cases where no statistically significant change was observed. The sequence of the pre-processing pipeline was executed, as presented below. All pre-processing steps were implemented using the *MNE-python* library with default settings (*ver. 0.18*) (Gramfort *et al.* 2014), unless specified otherwise.

Channel Selection:

- 10-20 System (10-20, 20 channels)
- High Density (HD, 173 channels)
- High Density + Peripheral Channels (full HD, 257 channels)

The effect of the spatial density of the EEG has not been studied extensively, with a variety of systems used in the literature, which give rise to comparability problems. A significant body of results has been reported using a limited number of electrodes (particularly in clinical settings) and the international 10-20 system, which usually comprises of a few dozens of electrodes (reduced or extended by utilizing intermediate locations). Nevertheless, research with high-density systems (hundreds of electrodes, with complete head coverage) allows for directly comparing the information gain among different density configurations, by selecting subsets of electrodes. Several studies have shown the capability of high-density configurations to detect even subcortical neuronal activity from deep structures (in contrast to conventional presumptions), by showing statistically significant correlations of the source reconstructed signals from the scalp EEG, with intracranial electrode recordings at the identified source locations (Krishnaswamy *et al.* 2017; Seeber *et al.* 2019). Overall, a limited

number of electrodes is not always sufficient for clinically significant EEG signatures, either due to sparsity of the head coverage, or due to the lack of electrodes at the relevant locations. Given all the above, three different configurations were chosen for the investigation of the spatial density of the EEG:

1) 10-20 System: The 10-20 channel selection comprises of 20 channels located as per the standard international 10-20 system, namely: F7, T3, T5, Fp1, F3, C3, P3, O1, Fz, Cz, Pz, Oz, Fp2, F4, C4, P4, O2, F8, T4, and T6. The given order was preserved in all experiments with the 2D representation, for representation consistency.

2) HD set: The high density set (HD) comprises of 173 channels from the EGI system (Hydrocel GSN) which cover the scalp densely, yet excluding the underside coverage of the head (i.e. below the ears, including the face)

3) Full HD set: The full high density set (full HD) comprises of all 257 channels of the EGI system, including the underside peripheral electrodes on the inferior surface of the head. These electrodes have also been shown to detect neuronal activity, particularly from certain brain regions (e.g. inferior temporal lobes or the ventral aspects of the frontal lobe (Luu *et al.* 2001))

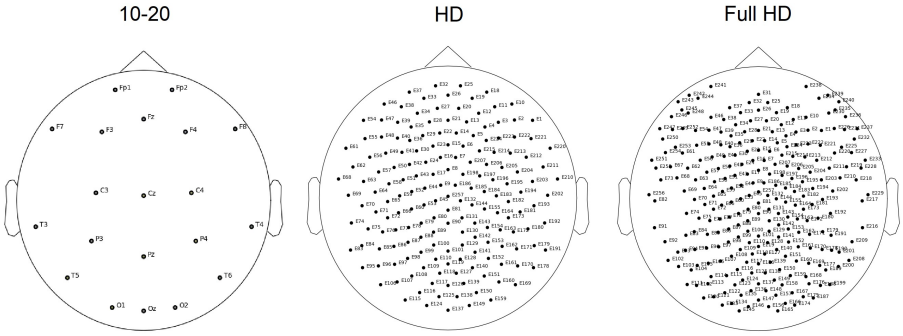

Fig. 4.2. EEG channel configurations: 10-20 (left), HD (center), Full HD (right)

Filtering:

- 0.5 – 40 Hz Band-Pass Filter, 50 Hz Notch Filter
- 0.5 – 100 Hz Band-Pass Filter, 50 Hz Notch Filter

The presence of high frequency information in EEG is another debatable aspect, with the scalp acting as a low-pass filter, and the high frequency content often attributed as noise (e.g. powerline noise, or artifacts from muscle activity, which overlap in the high beta and gamma bands >30Hz). In general, the majority of EEG studies using deep learning report the use of a band-pass filter, with a low-cut value of 0.1-0.5 Hz and a high-cut value of 30-50 Hz.

While using a low-pass filter can be beneficial for increasing the SNR trade-off overall, several studies have indicated the possibility of EEG capturing relevant neuronal oscillations at frequencies up to 100 Hz (Muthukumaraswamy 2013; Gotman 2013), and in particular within consciousness studies (e.g. high gamma synchrony has been correlated with consciousness (Koch *et al.* 2016)). For this reason, we tested 2 filtering settings:

1) 0.5 – 40 Hz: This is a typical range for EEG analysis (also used in the original study (Murphy *et al.* 2011). The low-cut frequency was set at 0.5 Hz, given that the window of analysis is 1 sec. A 50 Hz notch filter was also used, to remove the power-line noise.

2) 0.5 – 100 Hz: In order to test the contribution of high-frequency activity, we increased the high-cut frequency to 100 Hz. Again, a notch filter was used to remove power-line noise, at 50 Hz (including harmonics)

In both settings, other filter parameters were kept same (FIR, zero-phase, Hamming window, filter length of 6.6 sec, lower transition bandwidth: 0.5 Hz, -6 dB at 0.25 Hz, upper transition bandwidth: 10 Hz, -6 dB at 45/105 Hz, 53 dB stopband attenuation).

Resampling and Epoching:

The sampling rate of the data was set at 100 Hz, unless the experiment included a configuration with the high-frequency range (0.5 – 100 Hz), in which case the sampling rate was set at 200 Hz. Epoching was performed after resampling, to avoid edge artifacts. In all cases, data were segmented into 1 second non-overlapping epochs.

Artifact Cleaning:

- No Cleaning
- Automatic Cleaning
- Manual Cleaning

As already discussed in Chapter 2, one of the most critical limitations of EEG is artifact contamination, the detection and cleaning of which can be an important stage of preprocessing, for all kinds of subsequent analyses. Although there are several techniques that have been developed for such task (e.g. FASTER (Nolan, Whelan and Reilly 2010), PREP (Bigdely-Shamlo *et al.* 2015), Riemannian Potato (Barachant, Andreev and Congedo 2013), SNS (de Cheveigné and Simon 2008) and others), there is no consensus within the community on how to address the problem in a unified way, despite the need for a transparent, automatic and common standard. This is due to the fact that the recognition and classification of the different types of artifacts is subject to individual expertise and experience, while the intensity and nature

of artifacts can change across different EEG systems and participants. Moreover, in terms of the available techniques for artifact cleaning using signal reconstruction, there are often distortions that may or may not affect the underlying signals of interest (as for example in independent component analysis - ICA).

In general, data are annotated as 'bad' at the level of sensor/channel, or within a specific segment (epoch/trial) in time, that may contain an artifact (normally by visual inspection). A common strategy for the detection of bad segments relies on a simple peak-to-peak threshold metric (amplitude differences are compared to a manually set threshold value; if peak-to-peak amplitude exceeds a certain threshold, the segment is considered 'bad'), which can also be used to identify bad channels. While rejection of bad segments is often employed, given the problems or reconstructing the data (due to physical assumptions, time consumption, etc.), fixing whole bad channels is more practical (for retaining information).

Within the deep learning literature, there are several works which have used the raw EEG signals to train the networks (Schirrmeister *et al.* 2017), as well as works that employ some type of automatic or semi-automatic artifact cleaning method (Stober *et al.* 2015; Stober, Cameron and Grahn 2014; Van Putten, Olbrich and Arns 2018; Heilmeyer *et al.* 2019; Wilaiprasitporn *et al.* 2020). Given that we do not have any indications for their effect on the networks, we investigated three popular preprocessing approaches, in order to test the models' robustness to EEG artifacts:

1) No Cleaning: In this case, no artifact rejection or correction was employed. This is the simplest preprocessing pipeline, with minimal steps.

2) Automatic Cleaning: For this approach, three successive steps were implemented: 1) a peak-to-peak threshold was set, that determined the epoch rejection criterion, 2) channels that were flat or had >20% of epochs exceeding the peak-to-peak threshold, were considered bad, and were replaced by spatial interpolation (spherical splines, MNE), and finally 3) all epochs with a channel exceeding the peak-to-peak threshold, were dropped and excluded from further analysis. As there is no threshold value which is globally optimal for all EEG data, a large value of 800 μV was chosen as a conservative estimate ((Jas, Engemann, *et al.* 2017) has shown that thresholds can vary across different channels or subjects, ranging from 150 to 700 μV). This ensured that the rejected signals could not be produced by brain activity (the brain under anesthesia produces the strongest EEG signals, which go up to 100 μV)

3) Manual Cleaning: In this case, manual artifact rejection with visual inspection of the data was employed by an expert. This is a reliable approach for identifying and cleaning EEG artifacts, but nevertheless subject to fluctuation, bias and other factors related to experience and training.

Reference Channel:

- Average Reference
- Cz
- Frontal Reference (virtual electrode)

We chose to test three of the most common reference montages that have been used in anesthesia and EEG research in general, and which highlight different features of the data, namely: 1) the average reference, 2) the Cz electrode at the midline position, and 3) a frontal reference, chosen as the average of the channels Fp1, F3, Fz, F4 and Fp2.

The advantages and disadvantages of each representation have been discussed in detail in Chapter 2. Briefly, a common-electrode reference is good for broadly distributed abnormalities, particularly if the reference location is distant from the area of interest (e.g. the frontal reference would highlight the posterior areas). Cz is a common reference choice, and the pre-defined reference channel on the EGI system, which was used during the dataset collection. The frontal reference was chosen as a virtual average of five electrodes, in order to have a symmetric position with respect to the midline, but also to reduce potential facial artifact contamination emerging from a single frontal channel. Finally, the average reference is a versatile montage, in both capturing focal and broadly distributed abnormalities, although susceptible to bias with large amounts of noise or a low number of channels.

Normalization:

- Standardization
- Robust Standardization
- L2 Normalization
- Exponential Moving Standardization (EMS)

Normalization techniques are often implicitly required in many machine learning algorithms, to ensure the stability and predictability of several numerical and optimization conditions. For our case study, as the values of a typical EEG recording are in the range of 10-100 μV, the training of the model could be decelerated or even diverge, given that the activation functions, weights, and other parameters of a neural network presume values centered around 0, with a standard deviation close to 1 (saturation of activation functions, small gradients and an asymmetrical cost function can all affect gradient descent and model convergence). Parallel to deep learning normalization methods, input data can also be normalized in different ways (e.g. by features or by samples), with varying effectiveness on the training process, given that these methods can distort or discard information which can be relevant for the given task. For these reasons, we tested four different normalization methods of our EEG epoch instances, and their effect on model performance:

1) Standardization: Standardization is one of the most used methods, which centers the data to a mean of 0 and scales them to a standard deviation of 1. While input features can have different scales in general, the homogeneous nature of the EEG samples allow us to calculate the statistical properties of the signals throughout time and space. By concatenating all channels and samples, standardization can be computed for each instance (epoch) independently (in parallel to layer-wise normalization, in deep learning), offering a simple physical interpretation. Of course, by standardizing the scalp activity epoch-wise, large-scale temporal information could be discarded (mostly related to energy, as EEG signals are low-cut filtered and thus centered around 0)

2) Robust Standardization: This is a similar to the above technique, where statistical properties (mean and std) are computed according to an interquartile range. This is useful when data contain outliers, as in the presence of EEG artifacts, which can result in values that deviate orders of magnitude. For this method, the *RobustScaler* (*Scikit-learn*) was used (Pedregosa *et al.* 2012), with the default quantile range (0.25, 0.75).

3) L2 Normalization: L2 Normalization is another often used technique which imposes normalization of energy across the EEG samples, by dividing each sample with the L2 norm. This method is more susceptible to distortions from EEG outliers.

4) Exponential Moving Standardization (EMS): Exponential moving standardization (EMS) is a technique found in (Schirrmeister *et al.* 2017). The idea here is that the statistical properties of the signals are calculated and updated recursively throughout time, by including past information with a decay factor, instead of relying only on epoch-wise estimations. In this case, the calculations are performed channel-wise, using the following formulas:

$$x'_t = (x_t - \mu_t) / \sqrt{\sigma_t^2}$$

$$\mu_t = alpha\ x_t + (1 - alpha)\mu_{t-1}$$

$$\sigma_t^2 = alpha\ (x_t - \mu_t)^2 + (1 - alpha)\ \sigma_{t-1}^2$$

where x_t and x'_t are the original and the standardized signals for an electrode at time t, provided the exponential moving mean μ_t and exponential moving variance σ_t^2 at time t. Instead of using a fixed decay factor, as per the original study, we chose the term 1 – alpha (where alpha = 0.25 / sampling frequency), in order for the coefficients to contribute a weight which is independent of the sampling frequency. The formula was made to align with the decay factor in the original study, and which was empirically shown to provide a reasonable decay for statistical estimation, without strong dependence from potential noise (Fig. 4.3). The calculation of the initial values of μ_t and σ_t^2 are based upon the first four seconds of the data, as per the original study.

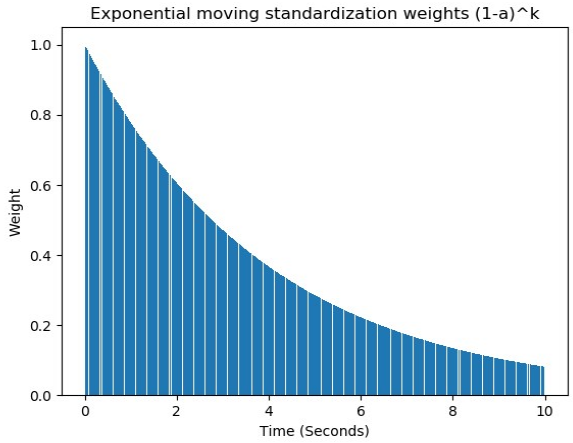

Fig. 4.3. EMS weight coefficients. The decay factor ensures a robust estimation of the statistical properties of EEG within a time-window of 10 seconds.

All of the above-mentioned normalization methods provide the advantage of an automatic preprocessing that is subject-independent, and can be used in a real-time scenario, using only past or current (online) information.

4.2.3 Convolutional Neural Network Architectures

Two convolutional neural network designs were used in the course of the experiments, in order to compare the effects of the 2D and 3D representations of the EEG, and the respective feature extraction within the spatial dimension. The architecture for both designs was based on the model described in 3.2.3. This 2D cNN model was also adjusted to support the input of the high-density EEG sets, with minimal changes. For the 3D convolutional neural network, a novel architecture was developed by incorporating a transformation of the channel representation to a topomap projection, which creates 2D images of the scalp activity for each time point. These 2D images, along with the 3rd dimension of time, result in a 3D representation of the EEG epochs. We consider this structure to be important for capturing the complex interactions of the spatio-temporal dynamics, contrary to a time-distributed 2D cNN found in previous studies. The 3D cNN model was designed to support a variety of EEG systems, irrespective of the number of channels or their locations.

In both network designs, the input representation reflects the amplitude of the signals vs time. As we showed and discussed in the previous chapter, a spectral representation does not offer any advantage over the raw EEG, and thus was not investigated (the limitations of Fourier analysis were discussed in detail in section 3.4.2). Moreover, a division of the EEG signals into multiple frequency bands (employed in several studies of section 4.1.3) was also avoided, to allow the networks to explore multivariate and non-linear representations across the whole signals' spectrum (research on neural oscillations and cross-frequency coupling has

also suggested the benefits of wide-band analysis (Cole and Voytek 2017)). In general, the design of the networks was prioritized to minimize – to the degree it is possible – the imposed computational assumptions/constraints on feature extraction.

2D Input Representation. This refers to the 2D matrix form of channels vs time samples, with the order of the channels being kept consistent for all channel selections (as mentioned in section 4.2.2 for 10-20, and as enumerated by the EGI system for the high-density sets). While this approach could be used in general for a specific set of channels (e.g. 10-20 system), different systems would have different sets of electrodes.

Topomap Image Extraction. For creating the 3D representation of the EEG, we implemented a topomap image extraction method for every time sample. During this procedure, the 3D coordinates of the channel locations are fitted within a unit sphere, after which they are projected to a 2D plane, or sensor map (by azimuthal equidistant projection, which preserves the relative distance between neighboring electrodes). As different EEG systems have different number of electrodes and locations, we center and scale the 2D coordinates based on the x_{min}, x_{max}, y_{min}, y_{max} values of the 10-20 system channels, which are included in most EEG systems ($center_{x,y}$: $0.5*(max_{x,y} + min_{x,y})$, scaling: $0.75/(max_{x,y} - min_{x,y})$). This way we make sure that the scaling of the 2D topomap image is consistent across different systems, and that the corresponding image locations align with same areas on the scalp. Over a 2D mesh, the values in-between the electrodes are interpolated with the *CloughTocher* scheme (cubic interpolation) (Renka and Cline 1984), and extrapolated up to the edges of the head circle (by using a mask), in order to avoid outlier values at its outlines (masked values are set to 0). After visual inspection of the resulting images, we fixed the image resolution to 30x30 pixels, which ensured that all local dynamics were essentially captured, even for the highest density sets. An example of an image extraction for a given time point can be seen in Fig. 4.4.

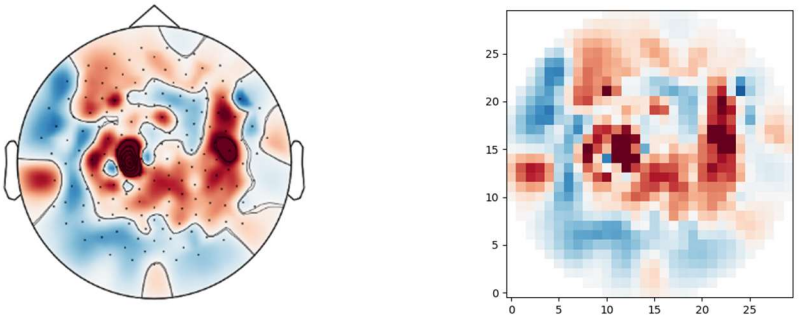

Fig. 4.4. Topomap Image Extraction. The topomap projection of the sensor activity, interpolated and extrapolated up to head circle (left), and the respective down-sampled 2D image representation (right).

2D Convolutional Neural Network for 10-20 system (cNN). The cNN model for the 2D representation of the 10-20 system is the reference size architecture described and depicted in 3.2.3. Briefly, it comprises of a sequential model with 2 functional layers: a feature extraction layer (2 convolutional layers with 32 and 64 feature maps, followed by max-pooling and dropout), and a classification layer (fully connected layer of 128 neurons, dropout, and 3 softmax units).

2D Convolutional Neural Network for High-Density sets (cNN – HD). The adjusted cNN model for the high-density sets was designed to share most of the parameters with original 2D cNN architecture. Again, two functional layers were used as described above. The convolution window of the 1st layer was changed to a kernel of 8x5 with 8x1 strides. Although the temporal windows remained the same throughout the network, we adjusted the channel dimensionality to compensate for the increased spatial density of electrodes (by ~8 times). Given these values, the resulted output tensors for both 10-20 and the HD sets were similar, and thus the rest of the architecture. The patch window for the max-pooling layer was 1x2, resulting in a compression of only the temporal dimension. This was selected for the network to be comparable to our 3D cNN architecture. All other parameters were kept constant to the original design. The adjusted architecture is depicted in Fig. 4.5.

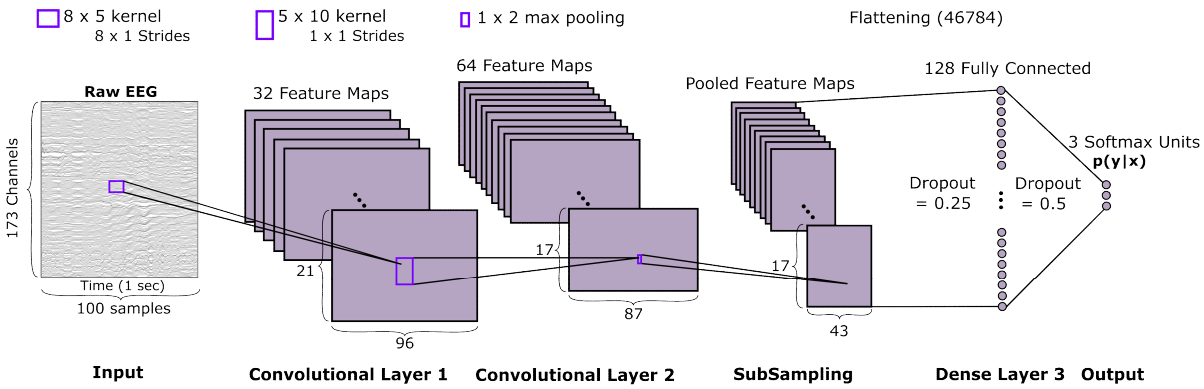

Fig. 4.5. The adjusted 2D convolutional neural network architecture for the HD set (2D cNN – HD).

3D Topomap Convolutional Neural Network (Topomap cNN). The 3D Topomap cNN is a sequential model based on the original 2D design. In this case, the topomap images extracted over the period of an epoch, create the 3D input samples to the network. The feature extraction layer consists again of two 3D convolutional layers (32, 64 feature maps), a max-pooling layer (3D) and a dropout layer. The classification layer remains as per the original design (128 neurons, 3 softmax units). The convolution windows were also kept constant within the temporal dimension. For the 1st convolutional layer, a kernel of 6x6x5 was used (6x6x1 strides), extracting information from a similar number of channels (based on the scalp region of focus) and which resulted in output tensors with similar dimensionality to the 2D design. For the 2nd

convolutional layer, a kernel of 2x2x10 (1x1x1 strides) was used, matching the dimensionality of the 2D cNN. The patch window for max pooling was 1x1x2, resulting in only temporal compression. All other parameters were kept constant as previously. The 3D Topomap cNN is depicted in Fig. 4.6.

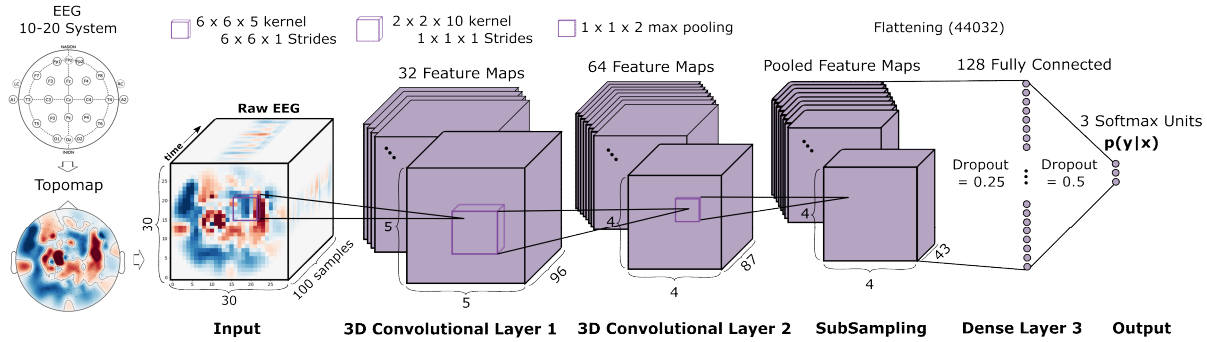

Fig. 4.6. The 3D Topomap Convolutional Neural Network. Topomap projection creates 2D images of scalp activity for each time point, which alongside the temporal dimension, results in a 3D representation of the EEG input samples.

4.2.4 Model Training and Evaluation

Model training and evaluation was based on the methodology presented in Chapter 3, and was kept consistent for all subsequent sets of experiments. Similarly to our previous analyses, our aim here was not to optimize each network for each experiment, but rather to compare the EEG configurations fairly, in order to reveal any computational advantages, under the classification task of the 3 anesthetic states - *Wakefulness*, *Sedation* and *LOC*.

Briefly, the EEG data were divided into training and testing sets, using a leave-one-participant-out cross validation paradigm, which ensures generalization performance and robust feature learning (9 participants, 3 states, 15 x 60 1-sec epochs ≈ 24,300 instances in total). One-hot encoding was used for the class target vectors. The models were trained using the categorical cross-entropy loss function with the Adadelta optimizer, and evaluated by their test accuracy. Initialization of network weights was done using the Xavier uniform initializer. A batch size of 100 was used, and for 10 training epochs (runs). All experiments were implemented in Python 3 using the *Keras*/*Tensorflow* libraries and a CUDA NVIDIA GPU (Tesla P100).

4.3 Experiment 1 – Reference Montage

In this set of experiments, we investigate the effect of the reference montage on the performance of the 2D cNN design, which has already been tested and provided us with a

baseline. Two models were used to check consistency across the 10-20 system channel selection, and the High Density set, in order to confirm that the observed effects were independent from the spatial resolution of the EEG (2 x 3 factorial design, Table 4.1). Presumably, the spatial resolution is one of the most influential factors of the EEG representation, with regards to the presence of information and noise, as well as the input and model parameters dimensionality. For the HD dataset, we used the adjusted model (cNN - HD), which ensured that the kernel sizes corrected for the number of incorporated channels. This increased the total number of trainable parameters from 2,921,219 to 23,013,219 (almost by 8 times).

Table 4.1. Experimental design for Reference Montage Comparison

Spatial Resolution		**Reference Montage**
10-20	**X**	Average Reference
		Cz
HD		Frontal Reference

4.3.1 Results

10-20. The convergence of all models was stable for the 10 runs, as indicated by the average categorical cross-entropy loss in Fig. 4.7. *Average Reference* (*AR*) performed better in all subjects, except two (S7, S8), with an average accuracy of 87.08%, although without a significant difference to other montages (Fig. 4.8). This small increase in performance can be found within the *Sedation* class, as depicted in the confusion matrices (Fig. 4.9), which has already been shown to be the most challenging state. Also, *AR* appeared to have the smallest variance across subjects.

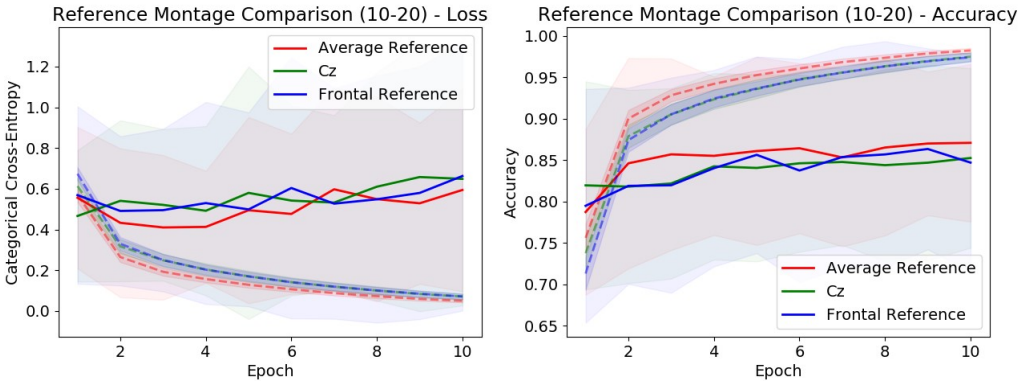

Fig. 4.7. Average categorical cross-entropy loss and accuracy history, for the three reference montages. Shaded areas correspond to std. Dashed curves correspond to training loss/accuracy. (10-20)

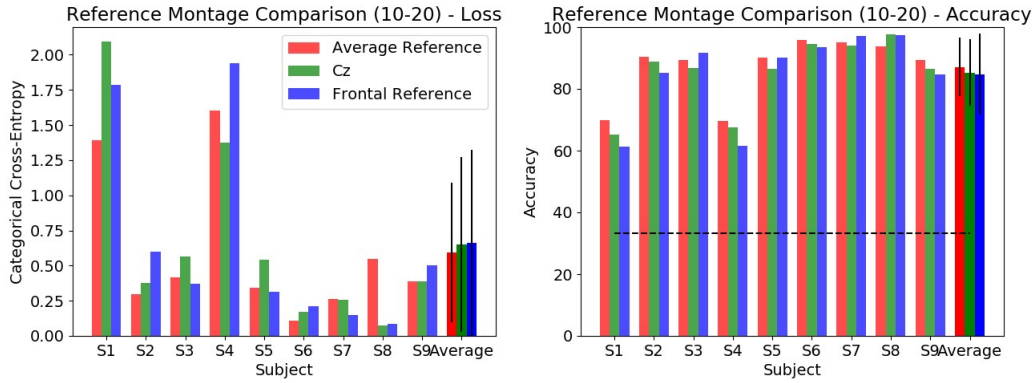

Fig. 4.8. Categorical cross-entropy loss and accuracy, for the three reference montages (10-20)

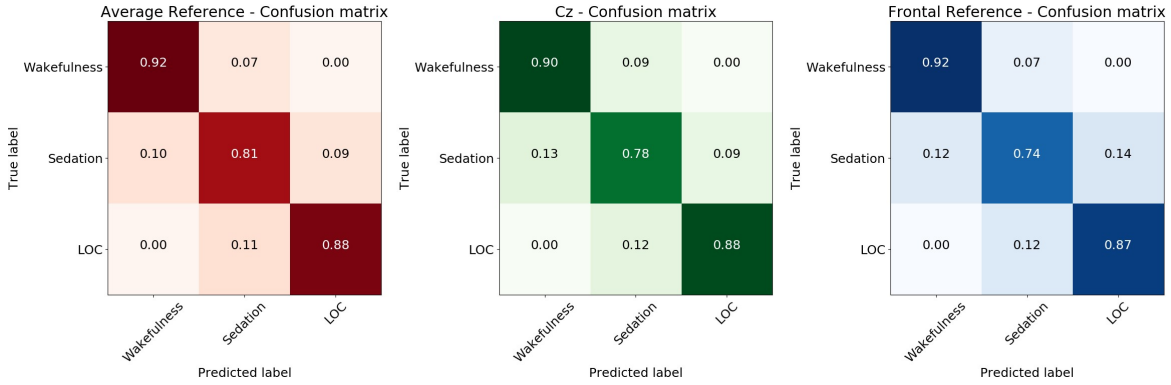

Fig. 4.9. Confusion matrices (normalized) for the three reference montages (10-20)

HD. The results were similar for the model trained with the HD dataset. Model performance was worse overall in comparison to 10-20 (79.86% accuracy for *AR*), but consistent with respect to the relative increase of the *average reference*, which was even more prominent for HD. Also, *AR* had again the smallest variance. This outcome highlights the potential strength of the *average reference* in high density settings, in agreement to the existing literature.

4.3.2 Discussion and Statistical Analysis

Overall, given that our EEG data are not significantly noisy (as it happens with recordings during general anesthesia), *average reference* seems to have the optimal performance (Fig. 4.10). It is also the most versatile and used option, as well as it is practical since it can be applied to any EEG system and channel configuration. Finally, the *average reference* is preferred in many other computational models (e.g. in source reconstruction models, due to averaging the distribution errors).

A statistical analysis using repeated-measures ANOVA (*statsmodels*) can been seen in table 4.2. The statistical significance of the effect from the reference montage was confirmed

with p = 0.014, as well as the difference between the two models using the 10-20 and HD configurations (p = 0.031).

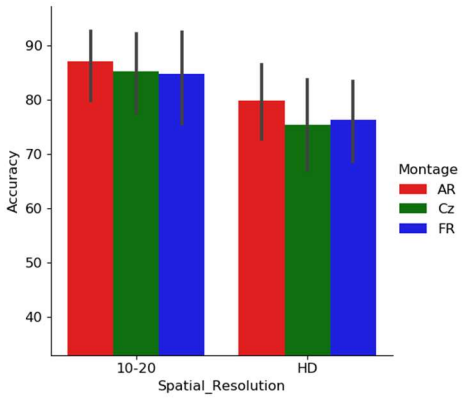

Fig. 4.10. Performance comparison for the 'reference montage x spatial resolution' experimental design. Error bars indicate the 95% confidence interval.

Table 4.2. Repeated-measures ANOVA for Reference Montage Comparison

	F Value	Num DF	Den DF	Pr > F
Montage	5.555422	2	16	**0.014717**
Spatial_Resolution	6.793104	1	8	**0.031307**
Montage:Spatial_Resolution	0.383291	2	16	0.687707

Based on our findings here, the *average reference* (*AR*) was kept fixed in all subsequent experiments, as the selection of the reference montage.

4.4 Experiment 2 - Normalization Methods

In this set of experiments, we investigate the effect of the EEG epoch normalization methods, using the 2D cNN design. Again, two models were used to check the consistency of findings for the 10-20 and HD configurations, resulting in a 2 x 4 factorial design (Table 4.3). For the HD dataset, the adjusted model was used (cNN - HD), which ensured that the kernel sizes corrected for the number of incorporated channels. This increased the total number of trainable parameters from 2,921,219 to 23,013,219 (almost 8 times).

Table 4.3. Experimental design for Normalization Method Comparison

Spatial Resolution		**Normalization Methods**
10-20	**X**	Standardization
		Robust Standardization
HD		L2 Normalization
		EMS

4.4.1 Results

10-20. The convergence of all models was stable for the 10 runs, as indicated by the average categorical cross-entropy loss (Fig. 4.11). There was no prevailing accuracy across subjects, with a similar average performance for all normalization methods, ranging from ~86% to ~88% (Fig. 4.12), albeit *Robust Standardization* and *L2 norm* had the smallest standard deviation (std=0.1 and std=0.09, respectively). Confusion matrices revealed *Robust Standardization* as the one with the most balanced within-state performance (Fig. 4.13).

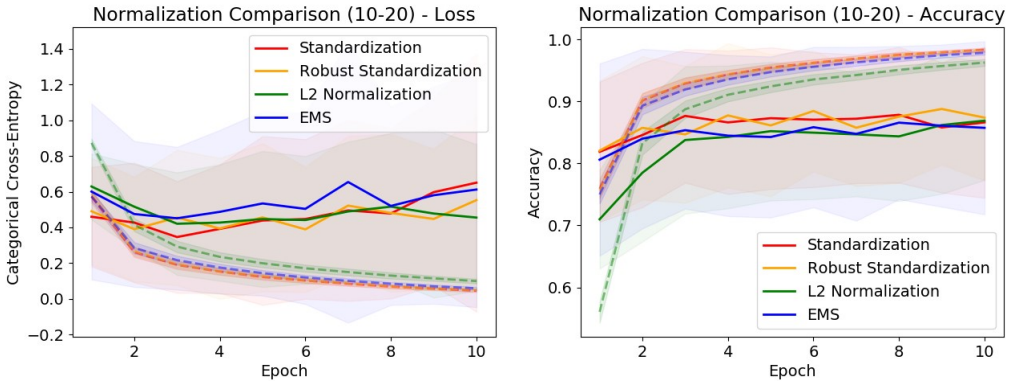

Fig. 4.11. Average categorical cross-entropy loss and accuracy history, for the four normalization methods. Shaded areas correspond to std. Dashed curves correspond to training loss/accuracy. (10-20)

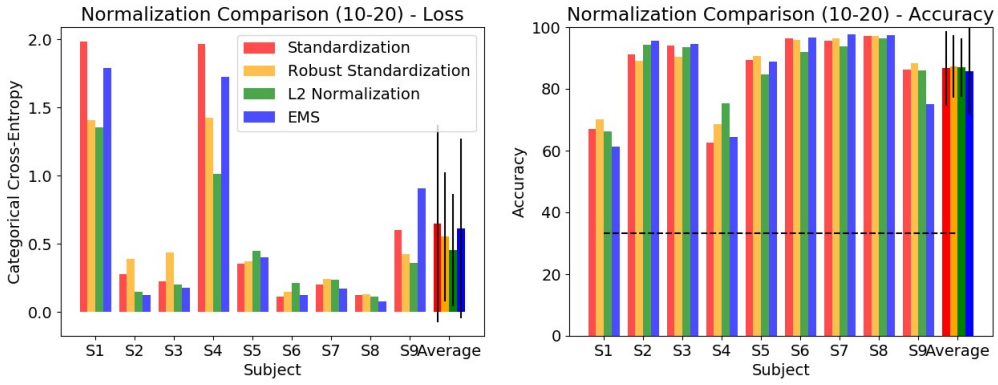

Fig. 4.12. Categorical cross-entropy loss and accuracy, for the four normalization methods (10-20).

HD. For the HD model, results showed a different perspective. Again, there was no prevailing accuracy across subjects, but *Standardization* performed significantly worse than the other three methods (69.3% for *Standardization*, 75-76% for *Robust Standardization* and *L2 Normalization*, and 78.18% for *EMS*, as seen in Fig 4.14). *Standardization* had also the largest variance across subjects (std=0.19). Confusion matrices revealed *Robust Standardization* and *EMS* as the methods with the most balanced within-state performance (Fig. 4.15).

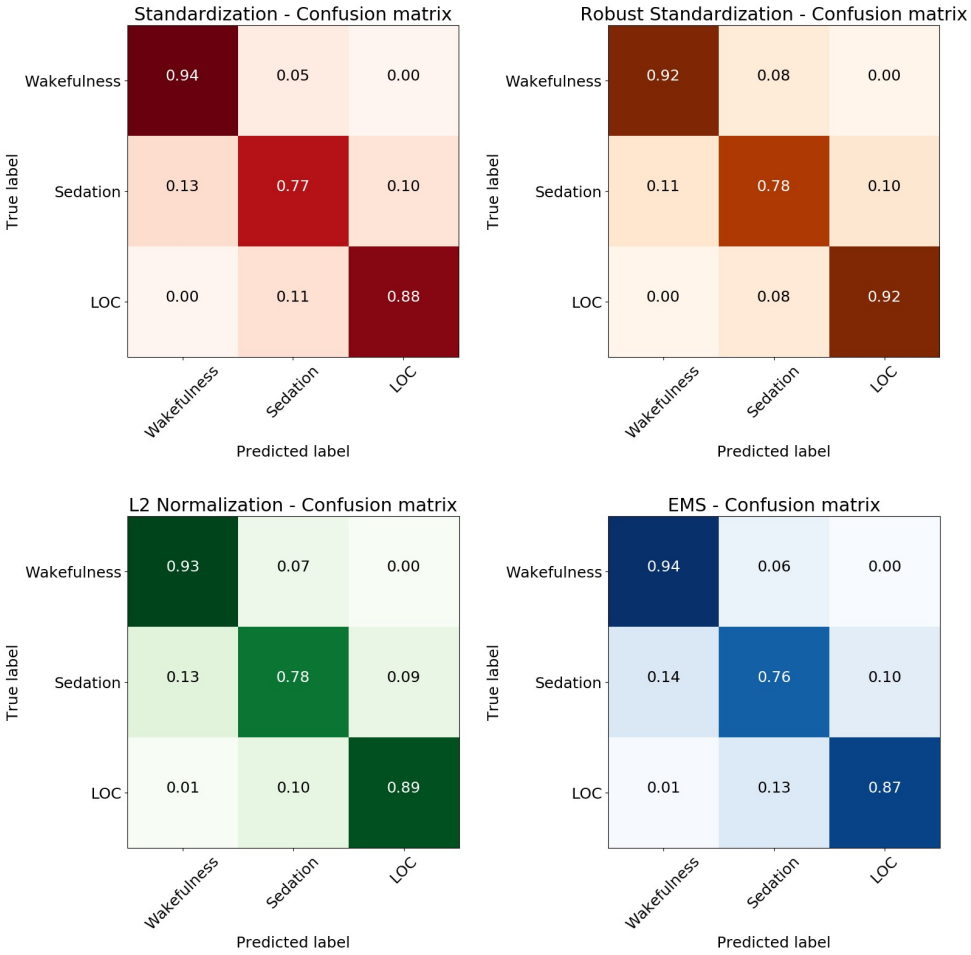

Fig. 4.13. Confusion matrices (normalized) for the four normalization methods (10-20)

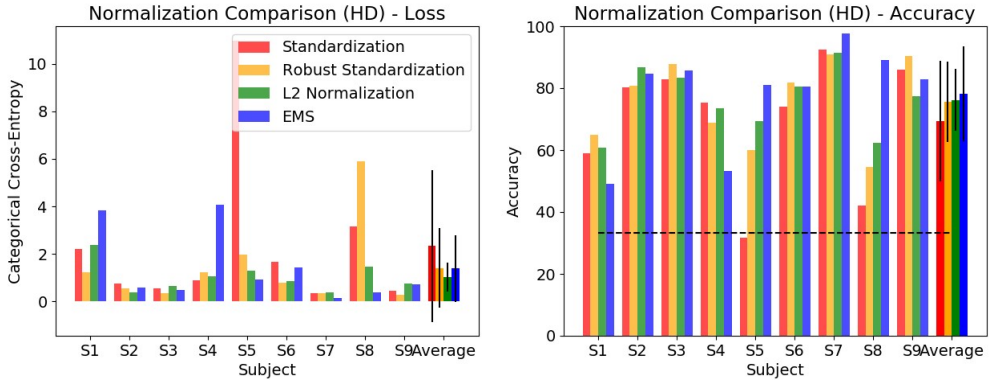

Fig. 4.14. Categorical cross-entropy loss and accuracy, for the four normalization methods (HD).

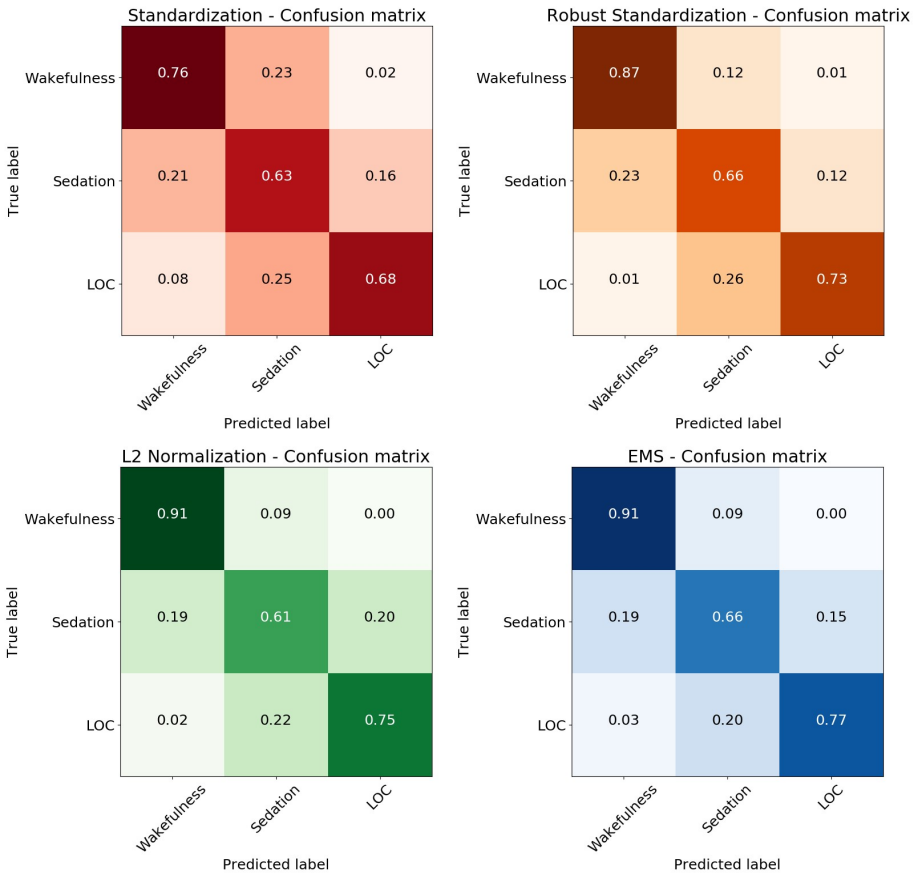

Fig. 4.15. Confusion matrices (Normalized) for the four normalization methods (HD)

4.4.2 Discussion and Statistical Analysis

Overall, the performance of the models did not reveal any general trend (Fig. 4.16). *Standardization* seems to be the most unstable method with the lowest performance, in contrast to *Robust Standardization*, which has the most robust performance (*L2 Normalization* following closely). For the HD dataset, *EMS* had a significant advantage over the other three normalization techniques.

A statistical analysis using repeated-measures ANOVA (*statsmodels*) can been seen in Table 4.4. Our results here have not indicated any significant effect for the normalization methods. The effect of the spatial resolution of the EEG (10-20 vs HD) was confirmed again, with $p = 0.015$. A small interaction between the two factors ($p = 0.17$) was driven from the results using EMS.

For the subsequent experiments, all normalization methods except *Standardization* were employed, in order to test how the various spatial resolution configurations, or the presence of high frequency content, might affect the performance of these methods, under the given classification task.

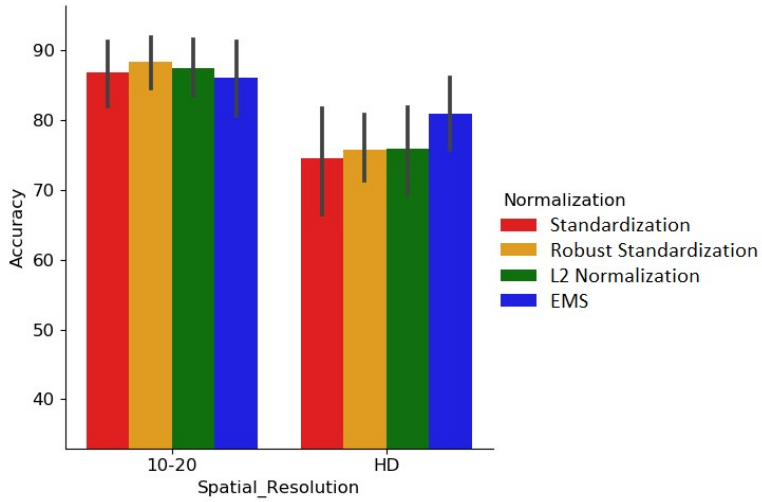

Fig. 4.16. Performance comparison for the 'normalization method x spatial resolution' experimental design. Error bars indicate 95% confidence interval.

Table 4.4. Repeated-measures ANOVA for Normalization Method Comparison

	F Value	Num DF	Den DF	Pr > F
Normalization	0.444714	3	24	0.723242
Spatial_Resolution	9.303622	1	8	**0.015817**
Normalization:Spatial_Resolution	1.782355	3	24	0.177409

4.5 Experiment 3 - Spatial Resolution and High Frequency Content

In this set of experiments, we investigate the effect of the spatial resolution of the EEG and the presence of high frequency content, using both 2D and 3D cNN architectures, in order to test and compare the respective representations and feature extraction approaches. Four different configurations were chosen for the channel selection and filtering parameters, namely: 1) the *10-20 system* (*10-20*), 2) the *HD set* (*HD*), 3) the *HD set including high-frequency content* (*HD + HF*) and 4) the *Full HD set*. As in the previous experiments, the original cNN architecture was used for the 2D representation of the 10-20 system, and the adjusted model (cNN – HD) for the 2D representation of the HD configurations. For the 3D topomap representations, the same 3D cNN design was used in all four configurations.

In both designs, we ensured that the cNN kernels incorporated a matching number of channels, so that the resulting models were all comparable with respect to trainable parameters. Moreover, the temporal dimensions of the convolution kernels were doubled in all models, to compensate for the doubling of the sampling rate (200 Hz), which allowed the inclusion of high-frequency content up to 100 Hz. The pool size of the max-pooling layer was also doubled

in the temporal dimension (pool_size = 4), as a way to acquire an equivalently compressed output-vector size. The total number of trainable parameters were 5,841,827, 6,196,323 and 10,071,139 for the 2D cNN models (for 10-20, HD and full HD, respectively). For the 3D Topomap model, the total number of trainable parameters were 5,812,067. Three normalization methods, the Robust Standardization, the L2 Normalization, and EMS, were also factors of experimentation. This resulted in a 2 x 3 x 4 factorial design (Model design x Normalization Method x Spatial/HF Configuration) (Table 4.5).

Table 4.5. Experimental design for Spatial Resolution and High-Frequency Content

Model Design		Normalization Method		Spatial Resolution and High-Frequency Content
2D cNN	X	Robust Standardization	X	10-20
				HD
3D Topomap cNN		L2 Normalization		HD + HF
		EMS		Full HD

4.5.1 Results

2D cNN – Robust Standardization. The average categorical cross-entropy loss showed no significant change over the 10 runs, indicating a stable convergence for all models (Fig. 4.17). The performance of the *10-20* configuration was significantly better in all subjects except one (S4 had a small decrease in comparison to *HD + HF*), with an average accuracy of 88.4%, against the high-density configurations (average accuracy ~76%) (Fig. 4.18). *10-20* also showed the smallest variance across subjects, along with the *full HD* (std = 0.09). Confusion matrices indicated that the performance decrease in the high density configurations could be found mostly within the *Sedation* class, and with a smaller effect, in *LOC* (Fig. 4.19).

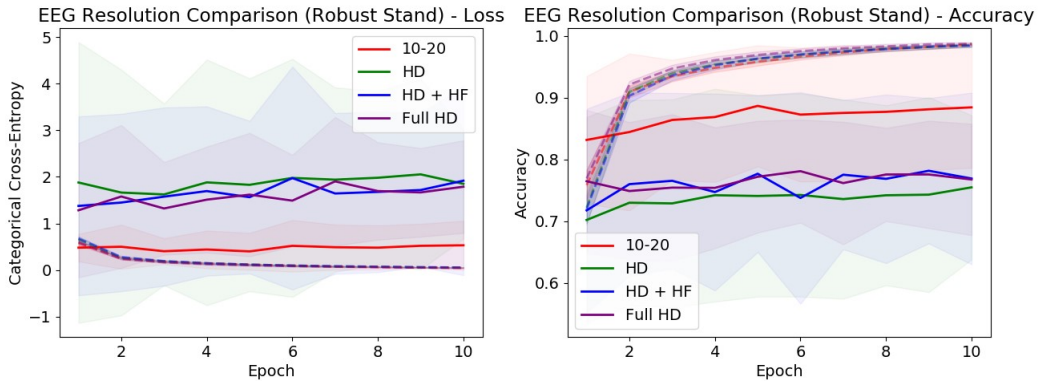

Fig. 4.17. Average categorical cross-entropy loss and accuracy history, for the four EEG resolution configurations. Shaded areas correspond to std. Dashed curves correspond to training loss/accuracy. (2D – Robust Standardization)

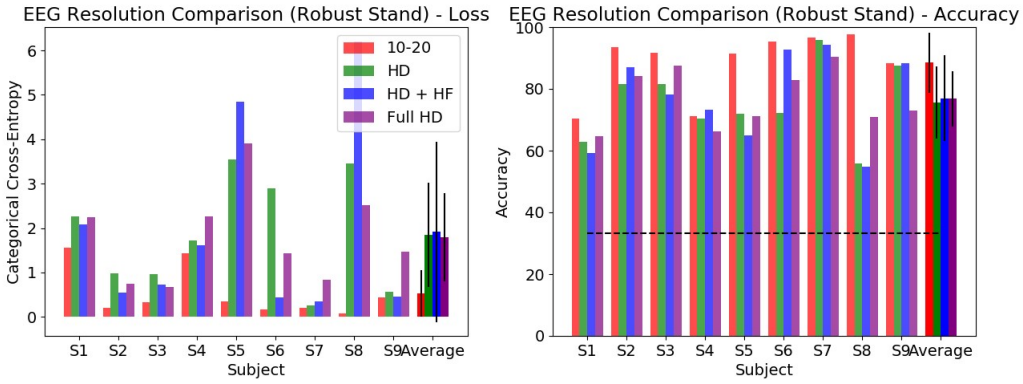

Fig. 4.18. Categorical cross-entropy loss and accuracy, for the four EEG resolution configurations (2D – Robust Standardization)

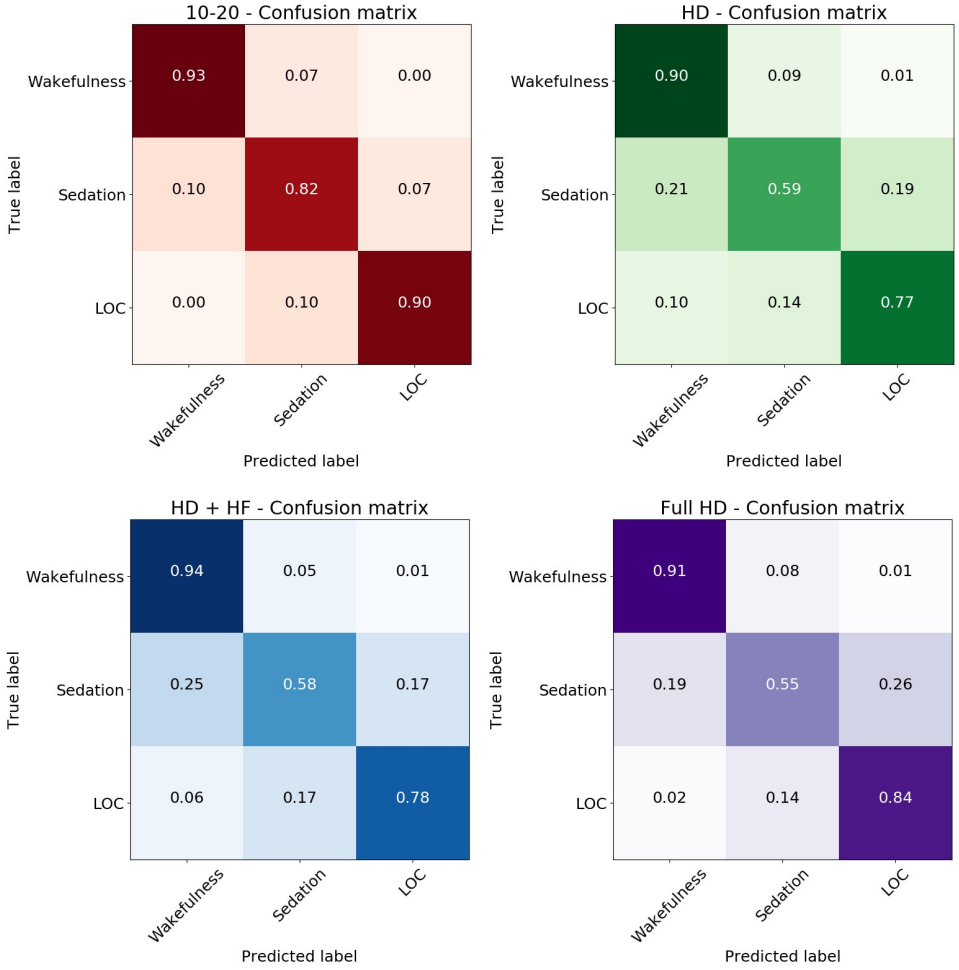

Fig. 4.19. Confusion matrices (normalized) for the four EEG resolution configurations (2D – Robust Standardization)

2D cNN – L2 Normalization. The results for L2 Normalization were similar. The *10-20* configuration had the highest accuracy in all subjects, with an average of 87.93%, against the *HD* configurations with an average of ~75% (again, *10-20* and *full HD* had the smallest variance across subjects). Confusion matrices indicated *10-20* to have the most balanced performance across the three classes.

2D cNN – EMS. The results using the EMS method were slightly different. As with Robust Standardization and L2 Normalization, the *10-20* configuration had the highest accuracy in almost all subjects (average accuracy of ~86%), albeit with a significant variance across subjects (std = 0.14) (Fig. 4.20, 4.21). Nevertheless, in contrast to the other methods, the high-density configurations (*HD*, *HD + HF*, *full HD*) alongside EMS achieved an increased average of ~82%. This increase was also evident from the confusion matrices, and particularly from the significant increase in performance within the *Sedation* class (Fig. 4.22).

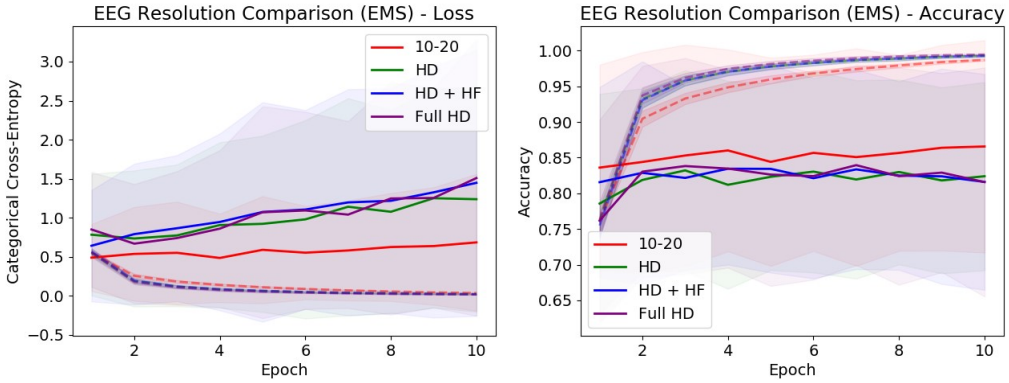

Fig. 4.20. Average categorical cross-entropy loss and accuracy history, for the four EEG resolution configurations. Shaded areas correspond to std. Dashed curves correspond to training loss/accuracy. (2D – EMS)

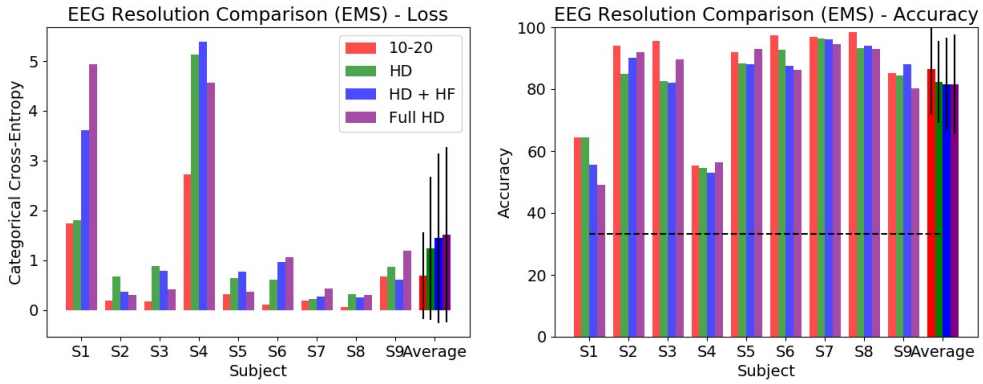

Fig. 4.21. Categorical cross-entropy loss and accuracy, for the four EEG resolution configurations (2D – EMS)

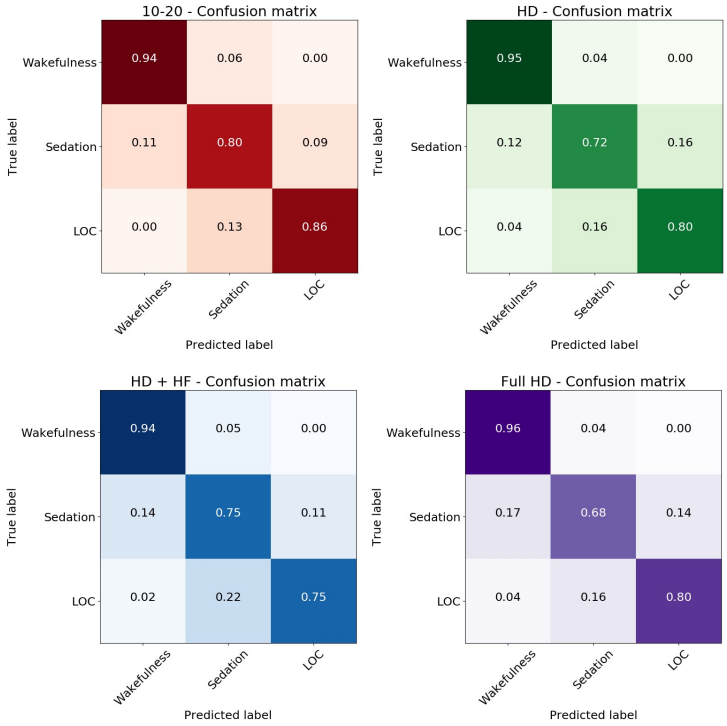

Fig. 4.22. Confusion matrices (normalized) for the four EEG resolution configurations (2D – EMS)

3D Topomap cNN – Robust Standardization. For the 3D topomap network design, results were moderately improved in Robust Standardization, when compared to the 2D model. The performance of the *10-20* configuration was the prevailing in the majority of subjects, reaching an average of 87% (std = 0.1), with *full HD* following with 81.96% (std = 0.12), and *HD*/*HD+HF* with an average of ~77% (Fig. 4.24). Confusion matrices revealed a more balanced performance across classes for the high-density settings, compared to the respective 2D model (Fig. 4.25).

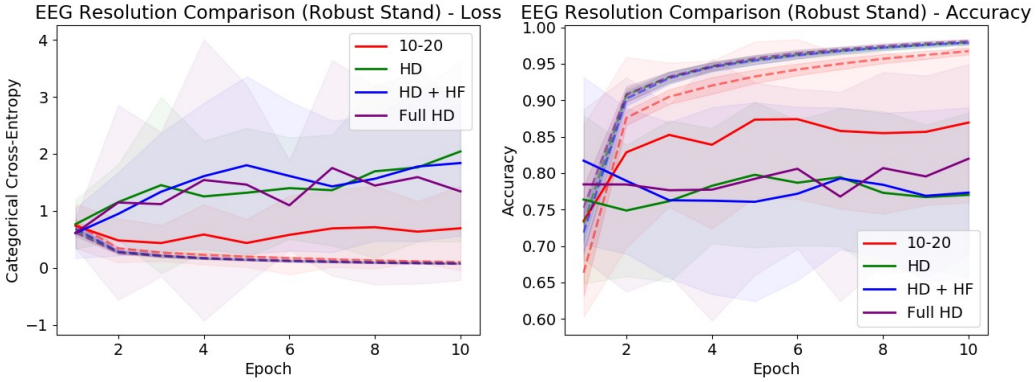

Fig. 4.23. Average categorical cross-entropy loss and accuracy history, for the four EEG resolution configurations. Shaded areas correspond to std. Dashed curves correspond to training loss/accuracy (3D Topomap – Robust Standardization).

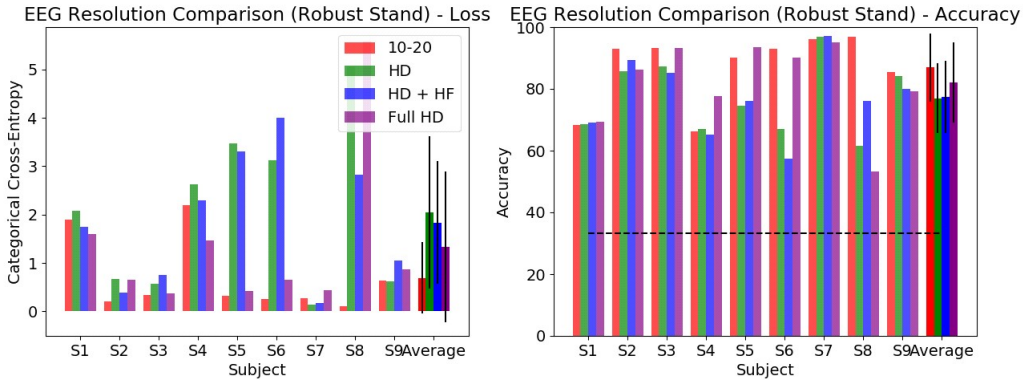

Fig. 4.24. Categorical cross-entropy loss and accuracy, for the four EEG resolution configurations (3D Topomap – Robust Standardization).

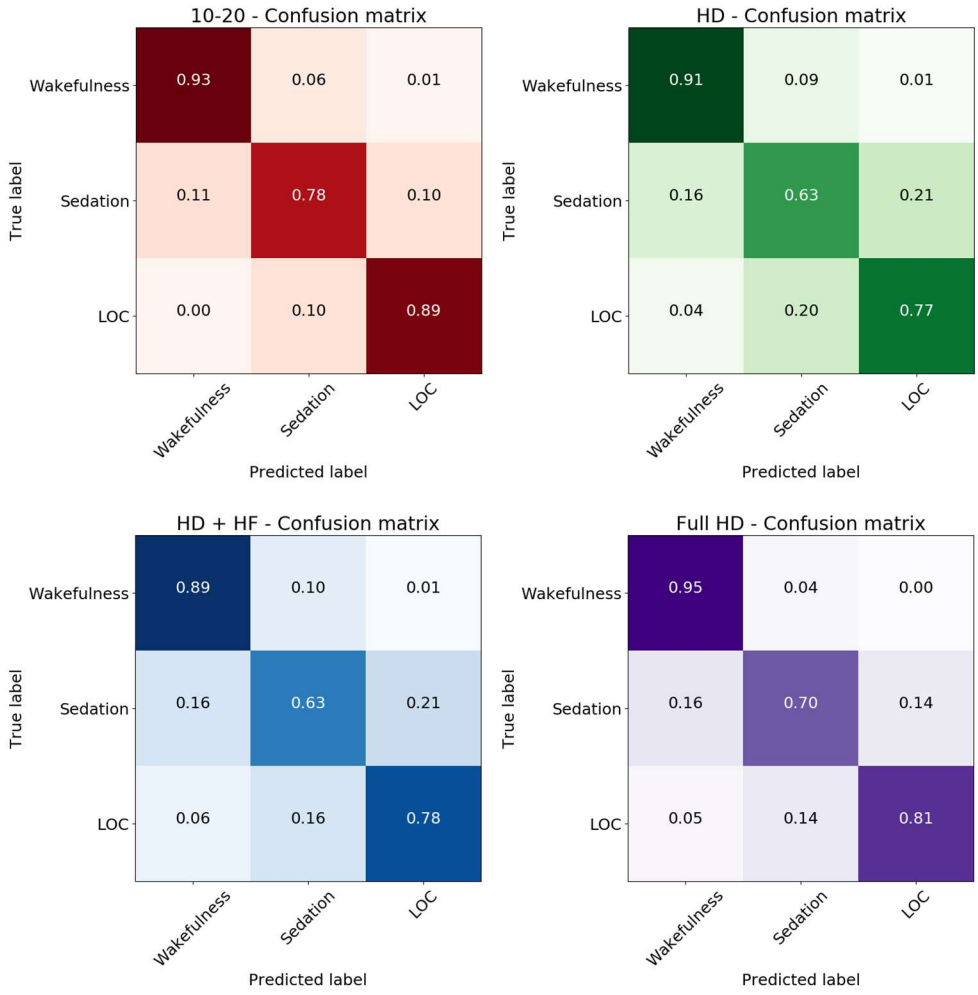

Fig. 4.25. Confusion matrices (normalized) for the four EEG resolution configurations (3D Topomap – Robust Standardization).

3D Topomap cNN – L2 Normalization. The 3D topomap model with the L2 normalization performed significantly worse than any other configuration. Categorical cross-entropy loss showed that the 10 runs of training did not necessarily reach a stable convergence (Fig. 4.26). Fig 4.27 clearly depicts this unstable behavior of the design pair. The per-subject accuracies showed no prevailing configuration, while for the *10-20* channels the model performed at chance level accuracy (33%) for all subjects, except S8. The *HD* and *full HD* configurations reached an average accuracy of ~71%, with *HD + HF* following with 64%. The confusion matrix of the *10-20* model showed a classification bias towards the *Wakefulness* state, revealing the failure of model's training (Fig. 4.28). For the high-density settings, confusion matrices revealed a minor instability within the *Sedation* state.

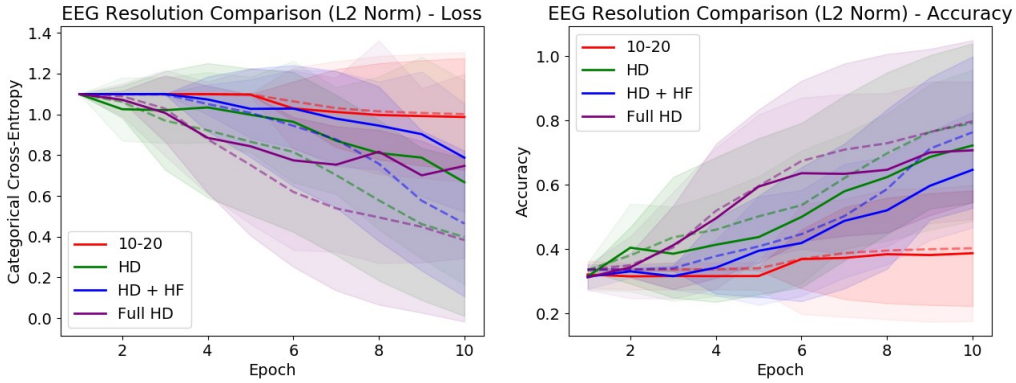

Fig. 4.26. Average categorical cross-entropy loss and accuracy history, for the four EEG resolution configurations. Shaded areas correspond to std. Dashed curves correspond to training loss/accuracy. (3D Topomap – L2 Normalization)

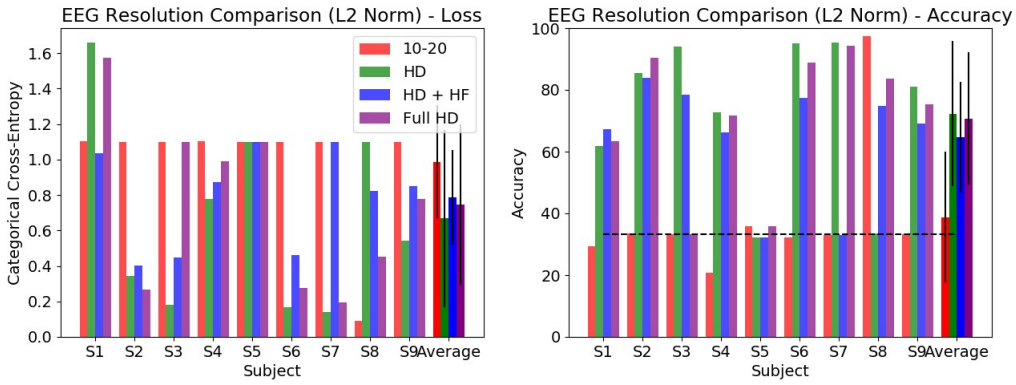

Fig. 4.27. Categorical cross-entropy loss and accuracy, for the four EEG resolution configurations (3D Topomap – L2 Normalization)

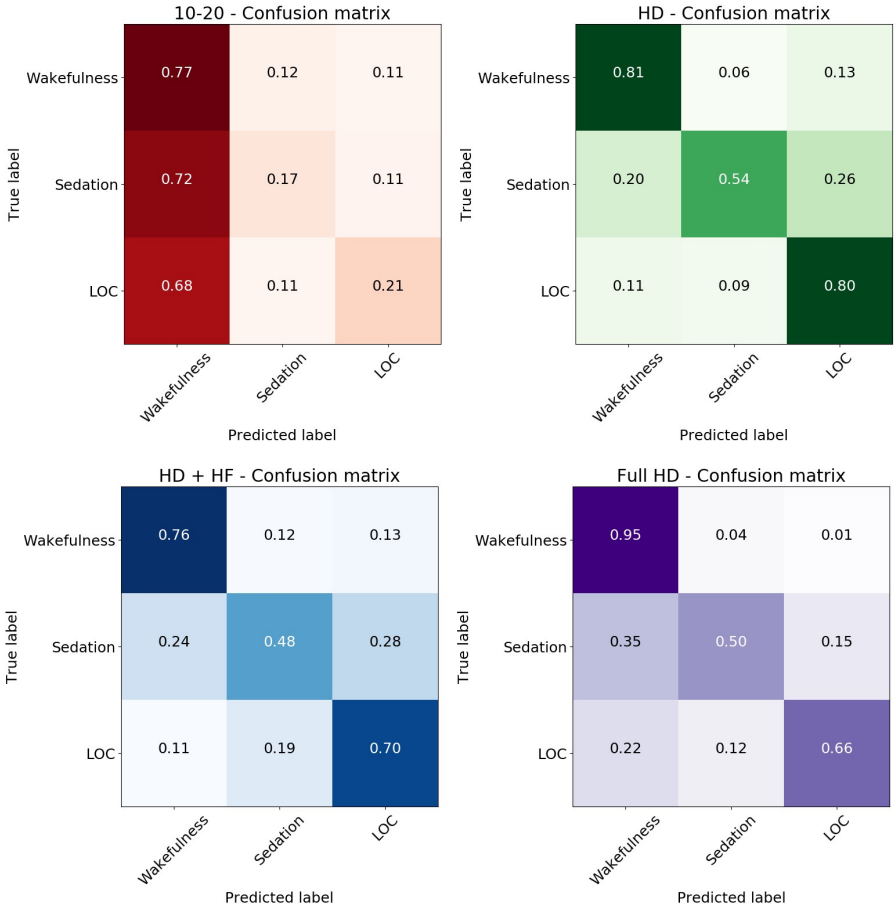

Fig. 4.28. Confusion matrices (normalized) for the four EEG resolution configurations (3D Topomap – L2 Normalization)

3D Topomap cNN – EMS. The results using the EMS were similar to the results obtained from the respective 2D representation model. All models had a stable convergence, based on the categorical cross-entropy curves. The *10-20* configuration achieved an average accuracy of 86%, with high-density configurations following with an average of ~83%, which was considerably higher than the one obtained from Robust Standardization and L2 Normalization.

4.5.2 Discussion and Statistical Analysis

Overall, the *10-20* channel configuration had the best performance against all other high-density configurations, irrespective of the model design and normalization method, reaching a peak accuracy of 88.4% with Robust Standardization (Fig. 4.29). The 3D Topomap network design showed some minor improvements over the high-density configurations, albeit with an unstable behavior for the 10-20 channels and L2 Normalization. Moreover, high-density configurations were further improved by using the EMS normalization method, as

already found from the previous experiment. The inclusion of high-frequency content (*HD + HF*), or the inclusion of the underside peripheral electrodes (*full HD*), did not seem to affect the performance of the *HD* dataset.

These findings were supported by the statistical analysis of the repeated-measures ANOVA, as seen in Table 4.6. The results using the L2 Normalization were excluded from the data as outliers that mislead the analysis, due to the instability of the 3D Topomap model and the chance level accuracy of the *10-20* system. The prevalence of the *10-20* configuration was confirmed by the significant effect of the spatial resolution, with p = 0.001. A significant interaction between the Normalization method and the spatial resolution was also present (p = 0.04), which was driven by the increased performance of EMS in high-density configurations.

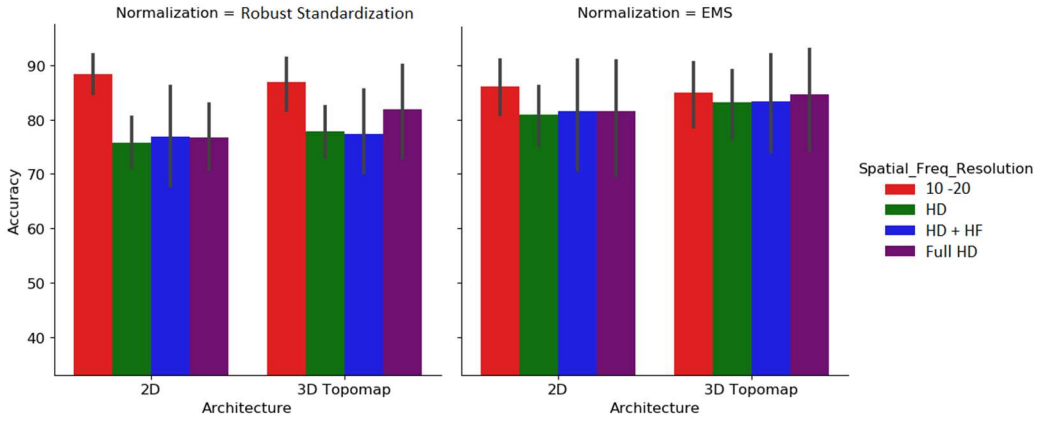

Fig. 4.29. Performance comparison for the 'model design x normalization method x spatial-frequency resolution' experimental design. Error bars indicate the 95% confidence interval.

Table 4.6. Repeated-measures ANOVA for Spatial Resolution and High-Frequency Content

	F Value	Num DF	Den DF	Pr > F
Normalization	0.786068	1	8	0.401158
Spatial_Freq_Resolution	7.063343	3	24	**0.001448**
Architecture	3.235074	1	8	0.109782
Normalization:Spatial_Freq_Resolution	3.034506	3	24	**0.048736**
Normalization:Architecture	3.16E-05	1	8	0.99565
Spatial_Freq_Resolution:Architecture	1.43632	3	24	0.256857
Normalization:Spatial_Freq_Resolution:Architecture	0.232566	3	24	0.872781

In the subsequent experiments, the comparison between the *10-20* and the *HD* channel configurations, along with the normalization methods of Robust Standardization and EMS, were kept under investigation, considering the interactions found here. L2 normalization was dismissed due to its instability with the 3D topomap network. The spatial resolution factor is particularly relevant for the investigation of the models' robustness to EEG artifacts, the levels of which can highly depend on the EEG sensor density.

4.6 Experiment 4 - Robustness to EEG Artifacts

In this set of experiments, we investigate the models' robustness to EEG artifacts, using the three artifact cleaning approaches described in 4.2.2. Both architecture designs of 2D and 3D representations were tested, along with the comparison between the two normalization methods (Robust Standardization and EMS), and the 10-20 and HD channel configurations, given the interactions found in the previous experiments. The original model designs and sampling frequency of EEG (100 Hz) were used here, as we excluded the high-frequency content. For the high density (HD) set with the 2D representation, the adjusted model (cNN – HD) was used. The total number of trainable parameters were 5,739,267 and 6,092,643 for the 2D representation models (cNN and cNN-HD, respectively), and 5,624,387 for the 3D Topomap. All experimental setups were tested under the three preprocessing approaches, resulting in a 2 x 2 x 2 x 3 factorial design (Model Design x Normalization Method x Spatial Resolution x Cleaning Approach) (Table 4.7).

Table 4.7. Experimental design for Model Robustness to EEG Artifacts

Model Design		Normalization Method		**Artifact Cleaning**
2D cNN	X	Robust Standardization	X	No Cleaning (10-20)
				Automatic Cleaning (10-20)
				Manual Cleaning (10-20)
3D Topomap cNN		EMS		No Cleaning (HD)
				Automatic Cleaning (HD)
				Manual Cleaning (HD)

4.6.1 Results

During the *automatic cleaning* approach, a number of epochs were dropped from the dataset, based on the peak-to-peak threshold of 800 μV, and after the interpolation of bad channels. Bad channel interpolation was limited, and mainly within the high-density datasets. Regarding epoch rejection, a small percentage was dropped that varied across participants (as expected, given that the peak-to-peak threshold was set to a conservative value), with a minor increase for the HD sets. Specifically, the average rejection rate was 0.81% for the 10-20 system, and 2.98% for the HD set. Given that there were ~24,300 epochs instances in total for all participants and anesthetic states, this resulted in a loss of ~196 samples for the 10-20 system, and ~724 samples for the HD set, which is acceptable for training and testing. Importantly, this approach ensured that there were no significant outliers during training.

2D cNN – Robust Standardization. For the 2D architecture with Robust Standardization, average categorical cross-entropy showed that all models had a stable convergence within the 10 runs (Fig. 4.30). The performance of the 10-20 channel selection was similar for all three artifact cleaning approaches, with an average of ~88% (Fig. 4.31). In contrast, HD datasets showed differences in performance, with *manual cleaning* having the highest accuracy at 84.07% (std = 0.11), following with *automatic cleaning* at 81.55% (std = 0.12) and *no cleaning* at 76.26% (std = 0.12). Confusion matrices further showed the gradual increase and decrease in performance for the corresponding 10-20 and HD models (Fig. 4.32).

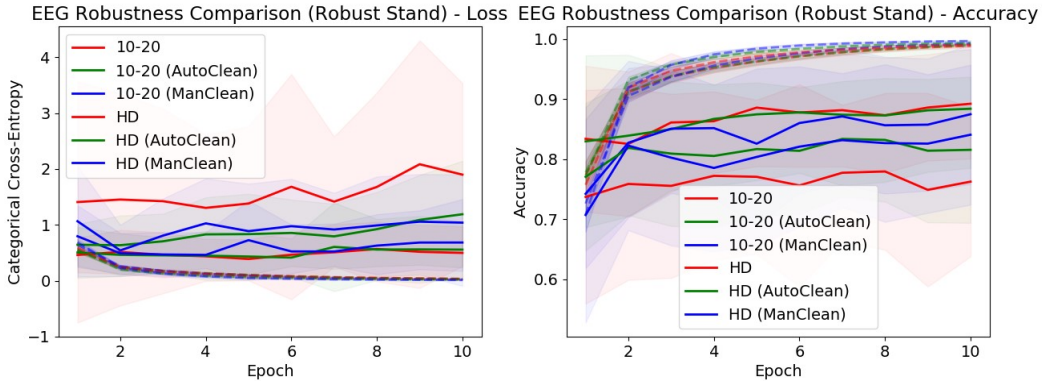

Fig. 4.30. Average categorical cross-entropy loss and accuracy history, for the 2 x 3 configurations of spatial resolution and artifact cleaning. Shaded areas correspond to std. Dashed curves correspond to training loss/accuracy. (2D – Robust Standardization)

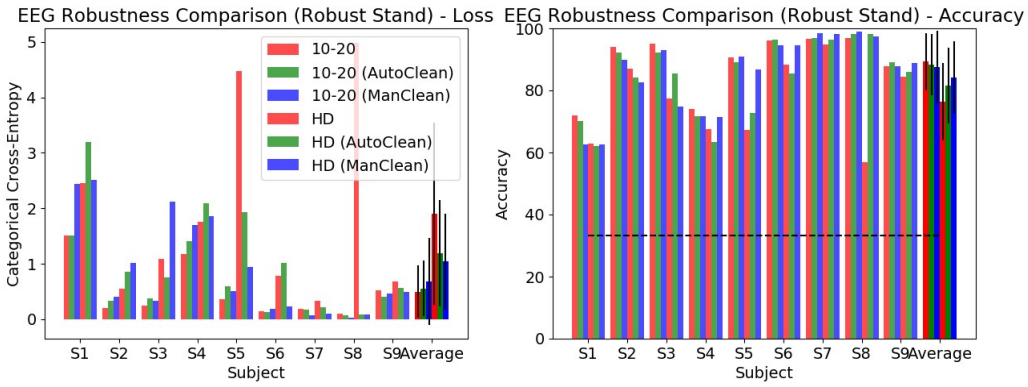

Fig. 4.31. Categorical cross-entropy loss and accuracy, for the 2 x 3 configurations of spatial resolution and artifact cleaning (2D – Robust Standardization)

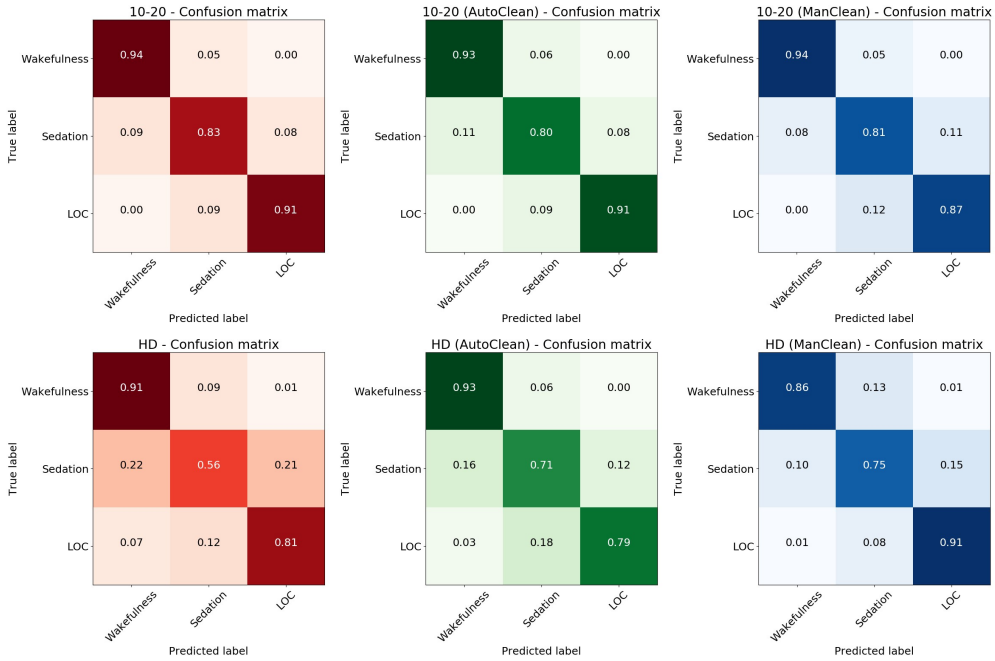

Fig. 4.32. Confusion matrices (normalized) for the 2 x 3 configurations of spatial resolution and artifact cleaning (2D – Robust Standardization)

2D cNN – EMS. When using EMS as the normalization method, results showed a similar picture. Again, all models converged for the 10 runs, as shown by the categorical cross-entropy loss (Fig 4.33). The three artifact cleaning approaches did not present significant changes in performance, albeit the overall increase for HD models (~85% accuracy for 10-20, ~83% accuracy for HD, on average, Fig. 4.34). Although EMS did not reach the peak accuracy achieved in the previous experiments, it further showed consistency across the different configurations, as well as a balanced performance, indicated by the confusion matrices (Fig. 7.34).

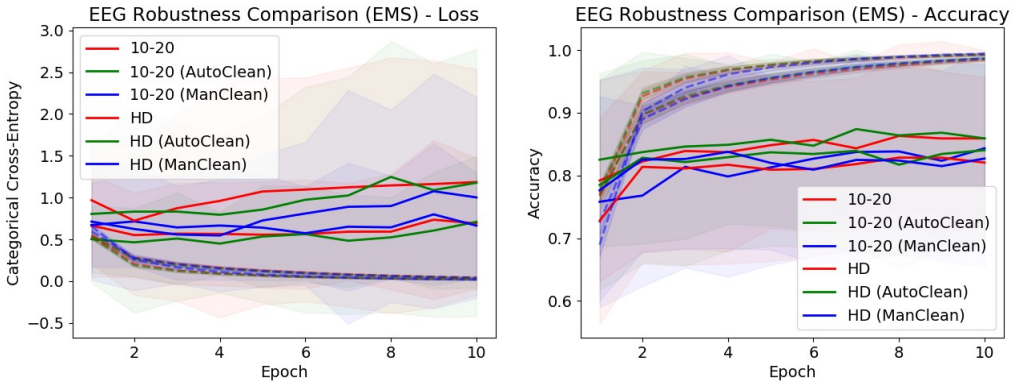

Fig. 4.33. Average categorical cross-entropy loss and accuracy history, for the 2 x 3 configurations of spatial resolution and artifact cleaning. Shaded areas correspond to std. Dashed curves correspond to training loss/accuracy. (2D – EMS)

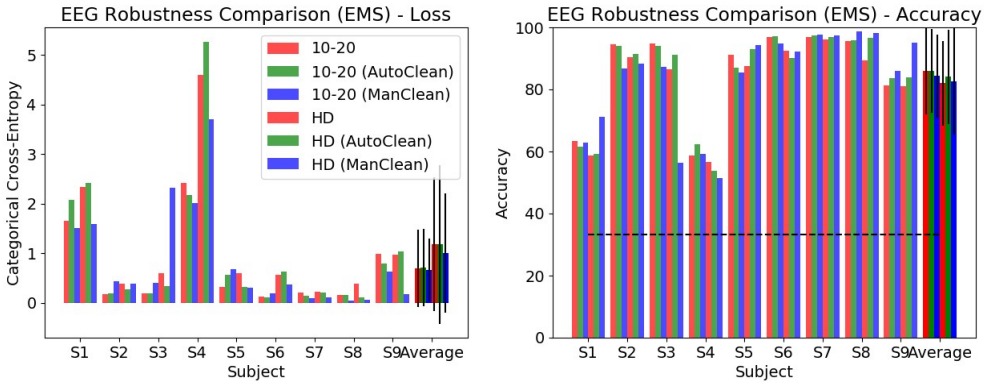

Fig. 4.34. Categorical cross-entropy loss and accuracy, for the 2 x 3 configurations of spatial resolution and artifact cleaning (2D – EMS)

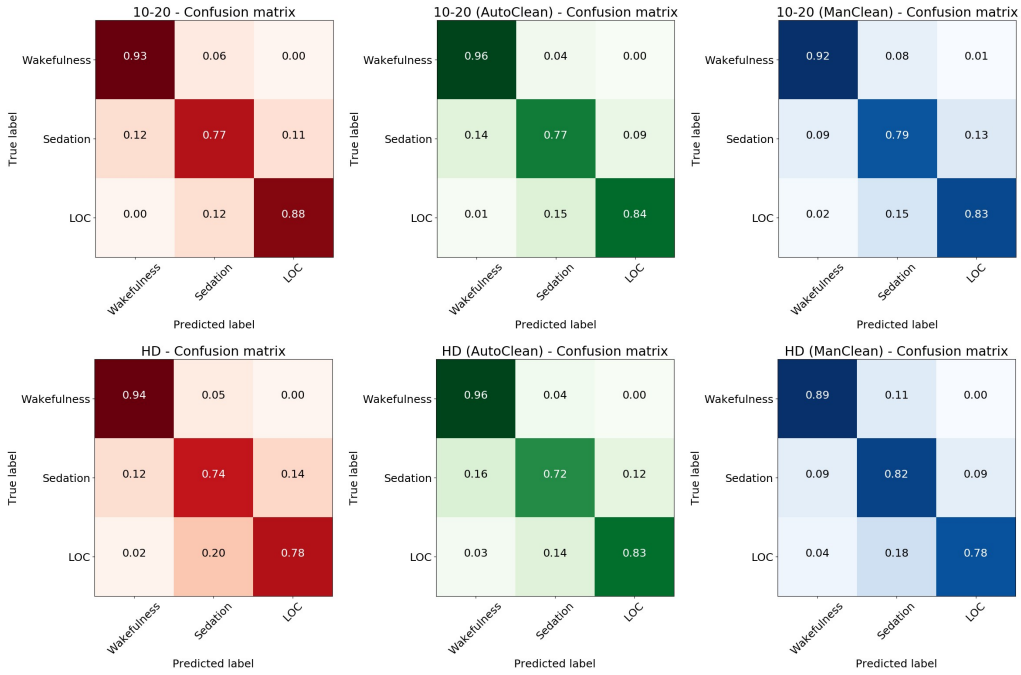

Fig. 4.35. Confusion matrices (normalized) for the 2 x 3 configurations of spatial resolution and artifact cleaning (2D – EMS).

3D Topomap – Robust Standardization. For the 3D Topomap architecture, the results showed a similar trend to the respective 2D model with Robust Standardization. All models had a stable convergence, based on the categorical cross entropy loss. For the 10-20 system, no artifact cleaning method had any prevailing performance (accuracy of ~86%). In contrast, *automatic cleaning* had the highest performance for the HD datasets (84.76%), following with the *manual cleaning* (82.13%), and finally the *no cleaning* approach (78.53%).

3D Topomap – EMS. The 3D Topomap architecture in combination with EMS had also similar results to the 2D model design. All models had a stable convergence, based on the categorical cross entropy loss. Average accuracy scores did not have any significant differences across the three cleaning approaches, or within the spatial resolution factor.

4.6.2 Discussion and Statistical Analysis

Overall, the three preprocessing approaches for artifact cleaning performed similarly for the 10-20 configuration (Fig. 4.36). On the contrary, high density configurations revealed more clearly the differences in performance and the strengths of each approach. As already discussed, EMS and the 3D Topomap representation, are two factors that can improve model performance, under the presence of high-density EEG (which is typically more noisy).

A statistical analysis using repeated-measures ANOVA (*statsmodels*) can been seen in Table 4.8. The interaction among the artifact cleaning approach, the normalization method and the 3D Topomap design, driven by the HD sets, can be seen from the results ($p = 0.06$).

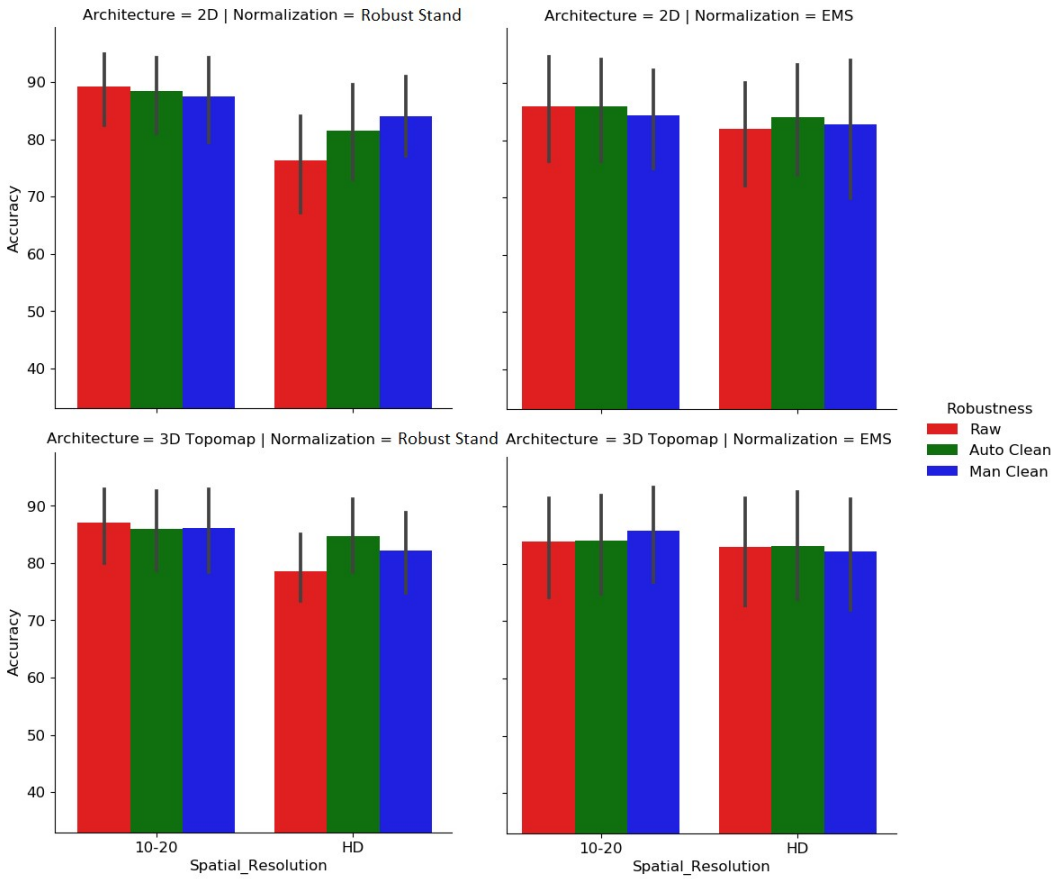

Fig. 4.36. Performance comparison for the 'model design x normalization method x spatial resolution x artifact cleaning' experimental design. Error bars indicate the 95% confidence interval.

Table 4.8. Repeated-measures ANOVA for Model Robustness to EEG Artifacts

	F Value	Num DF	Den DF	Pr > F
Normalization	0.04382	1	8	0.839422
Spatial_Resolution	12.85404	1	8	**0.007134**
Artifact_Cleaning	0.547377	2	16	0.588921
Architecture	0.792568	1	8	0.399305
Normalization: Spatial_Resolution	4.894954	1	8	**0.057859**
Normalization: Artifact_Cleaning	0.622216	2	16	0.549248
Spatial_Resolution: Artifact_Cleaning	1.724536	2	16	0.209782
Normalization: Architecture	0.00193	1	8	0.966032
Spatial_Resolution: Architecture	3.573121	1	8	0.095386
Artifact_Cleaning: Architecture	0.034293	2	16	0.96636
Normalization: Spatial_Resolution: Artifact_Cleaning	3.345339	2	16	**0.06112**
Normalization: Spatial_Resolution: Architecture	1.510882	1	8	0.25393
Normalization: Artifact_Cleaning: Architecture	1.469079	2	16	0.259572
Spatial_Resolution: Artifact_Cleaning: Architecture	3.175516	2	16	**0.068957**
Normalization: Spatial_Resolution: Artifact_Cleaning: Architecture	0.429352	2	16	0.658213

4.7 Discussion

Given our initial goals, we have shown the possibility of creating a novel 3D convolutional neural network design that is able to incorporate different EEG systems, and exploit the spatial structure of the EEG in a unified way. We have also worked towards the acquisition of a standard and automated pre-processing pipeline, by investigating the effects of several preprocessing parameters, the nature of which we consider crucial for EEG representation. To our current knowledge, we are not aware of any study that has tried to systematically investigate the above research questions. Nevertheless, there is a limited number of reviews that compare existing deep learning architectures for EEG analysis (Schirrmeister *et al.* 2017; Heilmeyer *et al.* 2019; Roy *et al.* 2019), as well as few studies on EEG representation comparison, albeit outside the context of machine learning (Yao *et al.* 2019; Lei and Liao 2017; Muthukumaraswamy 2013; Jiang, Bian and Tian 2019).

As already known and expected, the model, the nature of the data, their dimensionality, the presence of noise, and other variables, can be significant factors of machine learning performance. Even though the statistical comparisons were limited and restricted to our selected course of experiments, we acquired stronger or weaker indications for the advantages and disadvantages of the different methods. It is important to notice here that the repeated use of our validation data throughout our experiments can act as a ‘weak training set’, which might lead to information leakage in the selected processing pipelines. Of course, the selected classification task and model configurations used overall, might not be adequate to reveal some of the effects under study, in which case we would expect to find subtle differences. For all these reasons, we proceed with skepticism over the importance of the obtained results within

our data-limited analysis, alongside our understanding of the underlying processes and theoretical presumptions.

4.7.1 Reference Montage and Normalization Method

In Experiment 1, we showed that the average reference can be an optimal choice as the reference montage, given its high performance in both 10-20 and high density channel settings, its versatility in capturing focal and broadly distributed information, and its usability with a variety of EEG systems and methods. As our research focuses on anesthesia datasets with full head coverage and minimal noise from EEG artifacts, the potential drawbacks of this montage are significantly limited.

Moreover, during Experiment 2 we showed that robust standardization can be a reliable normalization method for the input epoch of our cNN model, providing consistent and balanced performance across subjects, whilst also being simple and robust to noise. On the other hand, the advantage of EMS in high-density settings can most likely be attributed to the use of a channel-wise standardization. Empirically, we know that high-density systems often have a number of unstable channels (cases of sensors with local or global high-amplitude noise), which can distort the information of the remaining ones, under an epoch-wise standardization approach. Nevertheless, as our experiments showed no requirement for using high-density sets, and a channel-wise standardization introduces significant spatial distortions, robust standardization remains a preferable option overall (the minor performance decrease of the 10-20 configuration could be possibly explained by the spatial information loss of EMS). Of course, a combination of the two methods is possible, by applying EMS with estimations calculated across all EEG channels (although still vulnerable to channel outliers, this would preserve a better statistical estimation of mean and std throughout time).

4.7.2 Spatial Resolution and High-Frequency Content

In Experiment 3, we investigated the effect of the spatial resolution and high-frequency content of the EEG, alongside the comparison between the 2D and 3D representation designs of the cNN. The acquisition of an HD dataset allowed us to do a rudimentary analysis on the information across different spatial densities of the EEG under anesthesia, which has not been researched in this context. The 10-20 system showed significantly better performance than any other HD configuration, giving us an optimal setup which is practical both from a research perspective, but also from a clinical one, as the use of HD systems is expensive and time consuming for hospitals. Notably, (Schirrmeister *et al.* 2017) reported similar findings for deep learning-based EEG analyses, with HD configurations resulting in decreased performance under a variety of classification tasks. However, the potential and theoretical information gain obtained with HD sets cannot be disregarded, based solely on the performance of the models.

Apart from the overall increase of noise in HD sets, our results could be partly explained by the considerable increase of input parameters in the EEG samples, which can be significant in the context of machine learning algorithms (20 x 100 = 2,000 input features for 10-20 in 2D, 173 x 100 = 13,300 for HD in 2D, and 30 x 30 x 100 = 90,000 for the 3D architecture). Given that the 3D cNN model was common in all four configurations, we can infer that input parameters, rather than model parameters, were the most determinant factor. The 10-20 system was able to achieve high performance with the 3D cNN despite the input dimensionality increase (as the information content remained low), yet still relatively worse compared to the 2D network, which preserved the original low dimensionality. In addition, while model parameters increased for the HD sets in 2D design (almost doubled for Full HD), performance patterns did not change with respect to the common 3D network. Of course, the exact relationship among the number of input features, training samples and model parameters is unclear, without a scientific or theoretical basis.

With regards to model parameters and the ability to learn from high-dimensional data, contemporary deep learning networks successfully incorporate several mechanisms, to avoid overfitting (e.g. dropout, L1/L2 regularization). On the other hand, the number of input features can be directly associated to the number of training samples (epochs), as finding patterns in high-dimensional spaces requires a sufficient number of examples (an empirical rule suggests training instances to be at least ~10 times the number of input parameters (Miotto *et al.* 2018); this is the case for the 10-20 system in 2D networks – 2,000 dimensions, ~21,600 training samples). Of course, such problem could be eventually solved with the inclusion of more data, or possibly by using a data augmentation technique (e.g. using an EEG-GAN, as found in (Hartmann, Schirrmeister and Ball 2018)), given the scarcity of EEG datasets in GA research.

Despite the amount of training data and the respective dimensionality problem, other theoretical reasons could be considered regarding the use of HD sets and the high-frequency content of the EEG. For example, deep structures (such as the thalamus) have be shown to be relevant in capturing the neural correlates of consciousness, making HD sets more viable for investigation. Moreover, high-frequency content has been shown to be relevant to perceptual abnormalities (Tekell *et al.* 2005), and in particular anesthetic agents (e.g. in ketamine (Maksimow *et al.* 2006)), which might not be evident from our results with propofol anesthesia. Nonetheless, our current findings suggest that the use of HD-EEG and high-frequency content decrease our model performance, most likely due to the increase of noise in HD settings.

4.7.3 Robustness to EEG Artifacts

Finally, in Experiment 4 we showed that deep learning models can be robust to EEG artifacts, under certain desired conditions. While different studies have used different approaches for training deep learning-based EEG models, with or without artifact cleaning, this is a particularly important aspect of methodology. The non-requirement for artifact cleaning is an important asset for deep learning, given the significant noise found in EEG

signals, as well as the time and expertise needed for manual intervention and data curation. This is in contrast to the majority of the conventional EEG methodologies, or other machine learning models, which require techniques such as feature selection, feature engineering, outlier detection, or other ways to improve data representation and performance.

More specifically, as we saw in the case of the 10-20 system, the detection and cleaning of EEG artifacts, either by an automated algorithm or by a human expert, did not affect model performance, which was robust across subjects already by using the raw data. On the other hand, artifact cleaning appeared to be more beneficial for HD configurations, which as already discussed, are more prone to artifact contamination. The accuracy gap between the 10-20 and HD configurations is further understood here, with HD performance significantly increasing (and in some cases approaching 10-20) under the suppression of noise, either by artifact cleaning or EMS (and to a lesser extent, by the 3D network design). Although we did not observe any correlation between subject performance and the rate of artifact detection/rejection, the HD performance increase is attributed to particular subjects that appeared to benefit from artifact cleaning (for automatic cleaning, an average of 3% epoch rejection resulted in ~5% accuracy increase). Based on these results, and given the dimensionality of HD configurations, we can infer that significant outliers – even within a small percentage of the data – can affect the training and testing of the models.

Furthermore, our results indicated that automatic cleaning can be as reliable as manual cleaning, which gives us the option to apply it globally without significant cost, and particularly in cases where high-density channel settings, increased dimensionality and a limitation to training samples, constrain model performance. Our approach also allows the preprocessing pipeline to remain automated, fast and reproducible. By observing the output of the algorithm in detail and by visual inspection of the dropped epochs, a peak-to-peak threshold of 800μV ensured that we did not dismiss signals of interest. However, the algorithm appeared sensitive in cases where the rejection of an epoch was based on few or even one bad channel (more evidently in HD). As data retention is generally important for EEG, and models appear robust to EEG artifacts, the epoch rejection requirement could be ideally relaxed (e.g. rejecting only epochs with >20% of channels exceeding peak-to-peak threshold). In any case, as our cNN model is time-invariant with predictions based solely on the epoch's time window, the preservation of temporal information is not of concern here. Such preprocessing pipeline could be used in future work for robust training, while allowing the test data to remain unprocessed for real-time prediction and evaluation.

Alternatively to this pipeline, other algorithms could be used for artifact cleaning, most notably based on deep learning techniques, as found in recent studies (e.g. with the employment of an independent deep autoencoder trained to denoise EEG signals (Yang *et al.* 2018)). Of course, such approach would require well-structured and standardized datasets for the development of systems for EEG denoising, which are currently missing from the research community (Zhang *et al.* 2020). Fine-tuned cleaning of large datasets requires substantial time of human expertise annotating artifacts (or artifact types), alongside any ground-truth

ambiguities. As the development of such dataset is hard and out of the scope of this work, our devised algorithm is a sensible method for implementation.

4.7.4 2D vs 3D Convolutional Neural Network Design

Summarizing our findings, we can infer that the 3D cNN architecture is an overall improvement over the 2D cNN design, both for the theoretical reasons discussed, but also for its computational efficiency and performance observed in Experiments 3 and 4. The 3D architecture showed a moderate performance increase over the 2D architecture, as revealed under the use of high-density EEG, where artifact contamination made the extraction of reliable features even more challenging. The homogeneous nature of the 3D kernels can be more effective in artifact detection and suppression, as noise tends to spread locally into neighboring sensors, and the 3D network preserves the topological properties of the signals. Moreover, the preservation of the spatial structure of the EEG, and the common network parameters (in contrast to the inconsistent representation of the 2D design), allowed for the extraction of spatio-temporal features that can be generalizable across different subjects and acquisition systems. Both of these reasons can account for the observed performance variation, based on the differences between the two architecture designs.

In spite of these improvements, the 3D network showed a minor disadvantage compared to the 2D network, when employed with the 10-20 system. Although performance changes here fall outside statistical significance, we suspect that the dimensionality increase, along with the highly correlated features of EEG epoch images, are likely responsive for this effect. While spatial filtering could in theory improve SNR, as shown in (Higashi, Tanaka and Tanaka 2014), the 90,000 features produced by the 30 x 30 topomap image resolution are significantly higher than needed for the 10-20 system. Therefore, the topomap image resolution can be decreased depending on the channel density required, or alternatively, replaced by a direct mesh representation of channel activity, similar to the approach found in (Wilaiprasitporn *et al.* 2020).

4.8 3D Convolutional Neural Network

In this section, we derive a generic 3D convolutional neural network and a standard pre-processing pipeline, based on the findings of this chapter. As initially discussed, these will allow us to fully exploit the spatio-temporal dynamics of the EEG, as well as to integrate a variety of EEG systems and datasets consistently, under a common methodology.

4.8.1 Input Pre-processing

The preferred pre-processing pipeline for the EEG input is specified here, following our previous discussion. Automatic artifact cleaning (step 5) is applied only on training data, as a way to ensure robust learning. The condition for epoch rejection is relaxed, with the requirement for more than 20% of channels exceeding the peak-to-peak threshold (as explained in section 4.7.3). Testing data remain unprocessed with regards to artifacts, as we want to evaluate the models over the continuous raw EEG signals, similarly to a real-time condition found in clinical settings.

EEG Pre-processing:

1. 10-20 System Channel Selection
2. Band-Pass Filtering (0.5 – 40 Hz, 50/60 Hz Notch Filter)
3. Resampling at 100 Hz
4. Epoching (1 sec, non-overlapping)
5. Automatic Artifact Cleaning (Training Data)
    a. Bad Channel Interpolation ('bad' if channel is flat or has >20% epochs exceeding peak-to-peak threshold of 800 μV)
    b. Epoch Rejection (if >20% channels exceeding peak-to-peak threshold of 800 μV)
6. Re-referencing to Average
7. Epoch-wise Robust Standardization (quantile range: 0.25 – 0.75)

4.8.2 Mesh Representation Design

Focusing on the advantages of the 3D network design, and specifically on the use of the 10-20 system channel configuration (revealed as the optimal setup), we further investigated several parameters of input representation. One of the main drawbacks of our 3D topomap model, related to the significant increase of the input dimensionality, as a result of the high-resolution topomap images. Hence, we naturally tested the use of lower-resolution images (6 x 6 images), the effect of spatial filtering (10 x 10 images with Average/Max pooling), as well as the use of a mesh representation (5 x 5), under our classification task (the mesh representation re-arranges the locations of the 10-20 channels into a 2D grid, by largely preserving their spatial relationships, as seen in Fig. 4.37). Given the selection of the 10-20 system, which comprises of 20 electrodes, we ensured that all of the above parameters were sufficient to capture the spatial dynamics of the EEG.

By incorporating a pooling layer for the spatial filtering of the 10 x 10 images, we were able to keep the dimensionality of the network's layers comparable, as a way to fairly assess the different input configurations. The results showed a significant improvement of the 2D mesh representation over the 6 x 6 topomap images, and to a lesser extent, over the 10 x 10 images with average pooling (not statistically significant). The difference between the mesh and 6 x 6 topomap images could be explained due to interpolation distortions that can occur in low resolution settings (in contrast to the mesh, which preserves the original signals).

Moreover, we tested a variant of the 2D mesh representation, by excluding epoch standardization centering (in order to avoid introducing noise in the 'empty' – zero valued – cells of the grid), which showed no statistically significant difference in performance. Therefore, we retained the original architecture, by replacing the topomap image extraction with the mesh representation, as depicted in Fig. 4.37.

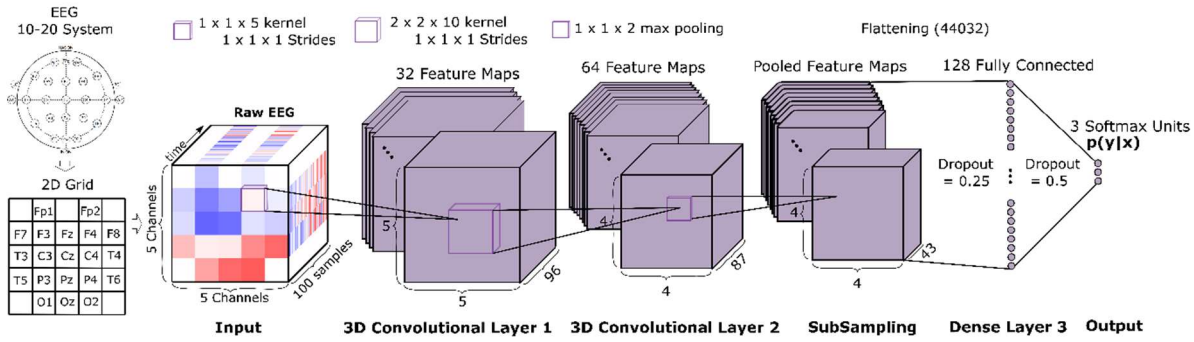

Fig. 4.37. The 3D Convolutional Neural Network architecture. The 10-20 system channels are arranged into a 2D mesh/grid, the time courses of which create the 3D representation of the EEG input.

This 3D architecture preserves all the advantages found in our analysis, and offers a computationally efficient solution for the 10-20 system channel selection (total number of trainable parameters: 5,718,529). A minimal number of cells in the grid are 'empty' (in this case five), whilst any channel can be replaced, depending on the spatial needs of the given EEG system (e.g. in several systems alternative electrodes can be found within the 10-20 system). While this model is used throughout our subsequent analyses, the 3D topomap cNN design remains an alternative robust design, in cases where high-density configurations are needed.

4.8.3 Summary

Overall, we have developed a novel cNN architecture and an automated pre-processing pipeline that allow us to process different EEG systems and datasets, in a unified and consistent way. We have also explored the effect of several parameters of EEG representation, and the capacity of 3D convolutions to extract robust spatio-temporal features from the EEG. By deriving an optimal processing pipeline, we were able to acquire a 2% performance increase within our classification task (up to 88% accuracy). Of course, besides the representation parameters focused here (which we consider to be most impactful), other parameters in deeper layers of the network can also affect the performance of the model. Setting aside such hyper-parameter optimization, the nature of the task under training and the respective ground-truth are among the most critical aspects of feature learning and performance. Given this premise, in the next chapter we focus on a variety of learning tasks, with an aim to better understand the EEG under GA, the respective clinical ground-truths, and the limitations of deep learning.

Chapter 5 Predictive Analysis of Behaviorally, Pharmacologically, and Psychometrically defined Anesthetic States

5.1 Introduction

5.1.1 Overview

In this chapter, we explore the capabilities of deep learning for electrophysiological investigation and estimation of anesthetic-induced states of unconsciousness. Specifically, we exploit the acquisition of data from experimental designs that control for several clinical variables, such as the anesthetic agent and the administration mode, which are used to target a particular behavioral, pharmacological, or psychometrical response. As none of these responses can provide us with an infallible ground-truth for consciousness in the brain, it is important to test the consistency of our models' predictions, and compare them to the respective EEG signatures. Moreover, we investigate the effect of the learning algorithm on the dynamics of our 3D cNN model (derived in the previous chapter), by conducting classification and regression experiments under a targeted ground-truth measure. In Experiment 1, we explore the behavior and performance of the model in a propofol study, when trained and tested under behaviorally-defined anesthetic states, characterized by the Ramsay scale. In Experiment 2, we further explore pharmacologically-defined states of propofol, by incorporating two ground-truth measures – the targeted plasma concentrations (as estimated by the Marsh PK model) and plasma concentrations measured arterially from blood samples. In Experiment 3, we explore psychometrically-defined anesthetic states from a ketamine study, which incorporates self-report measures from an altered-states-of-consciousness (ASC) questionnaire. Our results highlight the features and limitations of the models, in relation to each learning task and clinical ground-truth. This allows us to make inferences about the nature of our EEG data, as well as to derive an optimal training strategy. We further discuss about the nature of recovery, the relation of the models to the large-scale temporal dynamics of EEG and the depth of anesthesia, and we compare the utility of behavioral, pharmacological, and psychometrical measures. Finally, we compare the performance of our model to results from other related studies and methods of contemporary clinical practice.

5.1.2 Background

Our aim here is to focus on the nature of the optimization task undertaken by our deep learning model, by exploring different learning objectives and anesthetic definitions, based on behavioral, pharmacological, and psychometrical evidence for consciousness. One of the goals of this analysis is to investigate and understand the EEG data, in relation to the administered agent, the various anesthetic depths, and other clinical variables (i.e. pharmacological, behavioral and psychometrical measures) that are currently used as ground-truth information within GA practice. Another goal of this chapter is to test the predictive power of deep learning under different learning tasks, and reveal an optimal training strategy that will enable our model to impartially capture the electrophysiological features reflecting states and levels of unconsciousness.

Investigating the neurophysiological changes under anesthesia is a challenging problem, given our rudimentary understanding of the anesthetics' action, and the limitations of our tools to explore the underlying brain mechanisms. Specifically with regards to EEG, the richness of the electrophysiological signals and the various sources of noise have created the need for multivariate pattern analysis techniques, which allow for the decoding of more complex (and possibly hidden) brain states. However, while deep learning has shown its strength for data-driven analyses, its vulnerabilities to high-dimensional patterns make the selection of the objective function and ground-truth, particularly important. In our own study, ground-truth is in one sense defined by medical standards, as the information collected and used by clinicians to characterize the different anesthetic states. Although we theoretically understand the causal chain from the administration of a drug's dose, to changes in brain activity, and eventually to changes in EEG and clinical outcomes (e.g. behavioral unresponsiveness), the exact interactions remain unknown (due to the biological and pharmacological complexities of pharmacokinetics/pharmacodynamics). Nevertheless, by acquiring indirect (to consciousness) pharmacological, behavioral, and psychometrical measures, and by observing the model's behavior under their instantiation, we can better assess the nature of the respective EEG signatures.

At the same time, the employment of different clinical variables as the target of an EEG-based predictive task, can reveal the learning capacity of our model in relation to these complex interactions. Supervised learning is particularly efficient in discovering features directly relevant to a given task, while specifically for EEG decoding it has been shown that different algorithms can affect the outcomes of feature learning (Stober *et al.* 2015). Meanwhile, literature has shown a variety of studies using classification and regression methodologies in similar research tasks, without a systematic investigation on their effectiveness, or their theoretical and practical implications (e.g. in relation to definitions of consciousness, or the large-scale temporal dynamics of EEG – defined here as the dynamics beyond the model's window of analysis). In this sense, an optimal training strategy can be derived, with respect to a supervised learning algorithm and ground-truth encoding, whilst evaluated upon the model's

ability to make predictions that are accurate and generalizable (or interpretable, based on known facts).

Both above-described goals are connected to several theoretical considerations and unsolved problems, recognized within GA research (Bonhomme *et al.* 2019), which can advance clinical practice and consciousness research in general. One of the main unsolved problems, regards our ability to sensitively and specifically distinguish across different consciousness states that emerge during anesthesia. Notably, a state of disconnected consciousness often occurs during GA, which can be assessed by retrospective self-reports. Different anesthetic agents and doses can produce a variety of states, including connected consciousness (oriented consciousness, with awareness of the environment), disconnected consciousness (a dream-like state, without perception of the environment) and unconsciousness (no subjective experience) (followed by explicit or implicit memories). Understanding their neurophysiological and phenomenological properties can contribute not only to the identification of the full NCC, but also to the development of better clinical indices for tracking the DoA (dreaming and connectedness occur frequently during surgeries, with incidences reported up to 5%). Especially with regards to disconnected consciousness, a significant distinction between loss of behavioral response (LOBR) and loss of consciousness (LOC) must be made here, as unresponsiveness is not equated to unconsciousness (particularly during light anesthesia, where a transition from LOBR to LOC has been hypothesized to take place (Sanders *et al.* 2012)).

Finally, our investigation connects to other theoretical considerations and unstudied phenomena recognized in (Bonhomme *et al.* 2019). Notably, the between-studies variations in experimental designs, anesthetic agents, doses, and administration modes have led to many difficulties in the comparison of findings and reproducibility overall. The majority of research studies have focused on specific agents, and mostly during the induction and maintenance phases of anesthesia. Nevertheless, questions about the dose-response relation (and the inter-individual response to a drug), the transitional phases of anesthesia, the direction of transitions, and other asymmetries (e.g. recovery compared to induction), remain open. In the next sections, we describe how our approach tackles some of these questions by creating predictive models of behavior, drug concentrations and psychometrics.

5.1.3 Related Work

Literature review of the past few years has shown an increasing number of studies using learning-based methods, and especially deep learning methods, for the analysis of EEG under anesthesia (Lalitha and Eswaran 2007; Huang *et al.* 2013; Liu *et al.* 2015; Jiang *et al.* 2015; Sun *et al.* 2019a; Saadeh, Khan and Altaf 2019; Gu, Liang and Hagihira 2019; AlMeer and Abbod 2019; Liu *et al.* 2019; Dubost *et al.* 2019). Almost all of the studies we are aware have focused on a methodological investigation of the models, as an EEG-based tool for monitoring the depth of anesthesia (DoA) (rather than an explorative tool), whilst evaluated on a behavioral

ground truth. Despite the differences in methodology, several observations can be drawn from these works, mostly in relation to the selected ground-truths. However, while clinical scales are used on the basis of behavior as a reliable marker for consciousness, pharmacological or psychometrical variables remain unexplored in the context of machine learning. In addition, the majority of the studies have analyzed patient data under clinical settings (with small number of electrodes), which although increase the clinical relevance of the findings, are harder to control (than research environments) in terms of investigating specific agents and anesthetic depths (typically, multiple medications are co-administered and different depths are targeted, depending on the patient/operation needs).

With regards to learning tasks, there are several studies on classification analyses of anesthetic states, focusing on 2 to 3-state problems (Lalitha and Eswaran 2007; Saadeh, Khan and Altaf 2019; Dubost *et al.* 2019; Liu *et al.* 2019; AlMeer and Abbod 2019; Gu, Liang and Hagihira 2019), and a few studies on regression analyses, incorporating either discrete behavioral assessments (e.g. RASS scores (Sun *et al.* 2019a)) or continuous representations of consciousness levels (e.g. expert assessment curves, found in (Liu *et al.* 2015; Jiang *et al.* 2015)). Notably, (Liu *et al.* 2019) and (Sun *et al.* 2019a) achieved high performances using cNN models under a 3-state classification task (reaching 93.5%), and a regression-over-RASS-scores task (reaching MAE of ~1), respectively. However, given the significant variations found across tasks, ground-truths and depths of anesthesia, a comparative evaluation cannot be made (notably, only 66% of the studies used a cross-subject validation approach). Hence, a systematic investigation on the selected algorithm and learning objective can be important, in order to understand the behavior of the models, before any other methodological considerations (e.g. with respect to model architecture or EEG features).

When it comes to the characterization of altered states of consciousness (such as disconnected consciousness), or subtle changes in anesthetic depth and quality, behavioral scales exhibit significant limitations. For example, the assessment of disconnected consciousness is often made by retrospective reports, or by controlled experiments of intermittent sedation cycles (as found in (Radek *et al.* 2018)), which are not found in clinical datasets. Moreover, the assessment of connected consciousness can also be restricted by the administration of neuromuscular blocking agents (typically used in surgeries), unless a communication method is established (e.g. using the isolated forearm technique, as in (Tacke *et al.* 2020)). Beyond connected and disconnected states, more subtle changes in anesthetic depth and the transitional phases of anesthesia are difficult to assess, unless there is a systematic control of drug administration. In this respect, few studies have explicitly analyzed states of intermediate sedation (sedation before LOBR or LOC) (Lalitha and Eswaran 2007; Saadeh, Khan and Altaf 2019; Gu, Liang and Hagihira 2019), or incorporated data from all transitional phases of anesthesia (Liu *et al.* 2015; Sun *et al.* 2019a; Dubost *et al.* 2019). Most importantly though, the majority of the works have shown the existence of large-scale temporal dynamics throughout the transitions of anesthesia (in both classification and regression analyses), with a small subset of the models exhibiting dynamics correlated with the anesthetic depth.

Given all the above, questions related to specific agents, depths, pharmacological and psychometrical measures, as well as the algorithmic approach and its effect on the uncharacterized dynamics of anesthesia, constitute the focus of our analysis.

5.2 Methods

5.2.1 Datasets Collection

The data used throughout the experiments of this chapter were acquired from three independent studies, found in (Murphy *et al.* 2011), (Chennu *et al.* 2016), and (Vlisides *et al.* 2017; Vlisides *et al.* 2018), with consent given from the corresponding authors.

Liege Anesthesia Dataset. The dataset acquired in (Murphy *et al.* 2011) has been described in detail in section 3.2.1. For this work, we expanded the dataset to include the 4th state of *Recovery*, which followed the transition from loss of consciousness to wakefulness, as denoted by the two consecutive behavioral assessments (clear and strong response to command – Ramsay 2). Importantly, all states were determined based upon reaching and sustaining the desired Ramsay score, with recordings following a 5-minute equilibration period, after reaching the appropriate effect-site concentration (steady-state recordings were ensured by adjusting a constant-rate infusion of propofol, alongside the effect-site estimations from the Marsh model – Alaris TIVA/TCI mode). Moreover, for 3 out of 9 participants, we were able to acquire the EEG recordings during the transitional states, namely: from wakefulness to sedation (*WS*), from sedation to loss of consciousness (*SL*) and from loss of consciousness to recovery (*LR*). During the transitional states, propofol infusion rates were increased or decreased, according to the desired target. The experimental design of the full dataset is shown in Fig. 5.1.

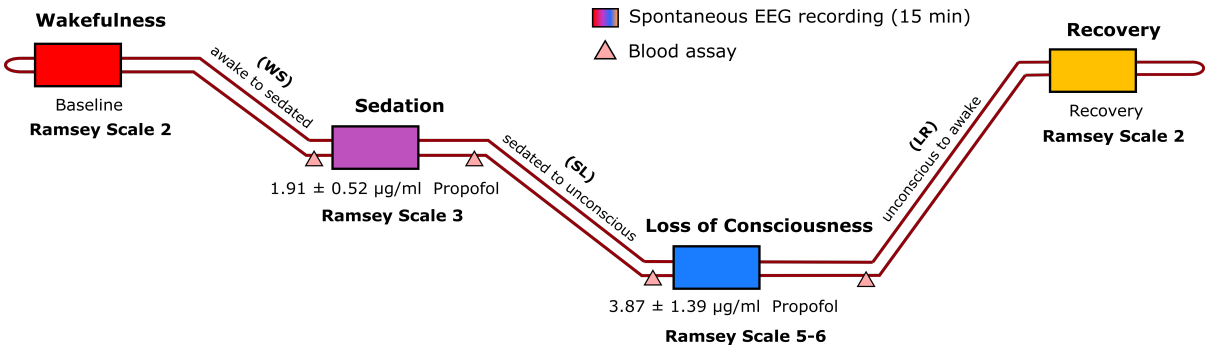

Fig. 5.1. Experimental design of the *Liege Anesthesia Study*. Anesthetic induction was guided by behaviorally-steady states of progressively deeper levels of unconsciousness, defined by the Ramsay scale.

Cambridge Anesthesia Dataset. The dataset acquired in (Chennu *et al.* 2016) is also based on a propofol study, in which the experimental design is described in detail. Briefly, the study was approved by the Cambridgeshire 2 Regional Ethics Committee and in accordance with the Declaration of Helsinki. All participants were neurologically healthy and gave written informed consent.

Approximately, seven minutes of spontaneous high-density electroencephalography (hd-EEG, 128 channel EGI Hydrocel GSN) was recorded from 20 participants (mean age = 30.85, SD = 10.98; 9 male, 11 female) during propofol sedation at four different states: *Baseline Wakefulness*, *Mild Sedation*, *Moderate Sedation* and *Recovery* (eyes-closed resting-state). Each state was determined by a desired ("target") plasma concentration, controlled by a computerized syringe driver that achieved and maintained the required propofol infusion rate for that plasma target (Alaris TCI mode, using the Marsh model). The targeted blood-plasma levels were 0.6 µg/ml for *Mild Sedation* (a relaxed but still behaviourally responsive state) and 1.2 µg/ml for *Moderate Sedation*, while a 10-minute equilibration period was allowed before recordings, to attain pharmacologically-steady states. For *Recovery*, EEG was recorded 20 minutes after the cessation of infusion, to ensure that propofol concentrations would approach zero (based on pharmacokinetic simulation). Blood samples were also taken between the anesthetic states, in order to characterize the inter-individual variability and to confirm similarity to target concentrations. The experimental design of the study is shown in Fig. 5.2.

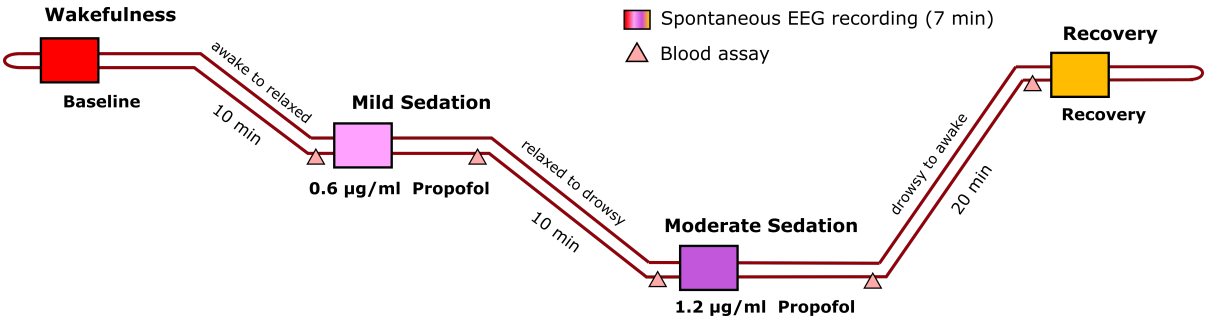

Fig. 5.2. Experimental design of the *Cambridge Anesthesia Study*. Anesthetic induction was guided by pharmacologically-steady states of progressively deeper levels of sedation, defined by propofol plasma concentrations.

Michigan Anesthesia Dataset. The dataset acquired in (Vlisides *et al.* 2017) is based on a ketamine study, in which the experimental design is described in detail. Briefly, the study was approved by the University of Michigan Medical School Institutional Review Board, and written informed consent was given by volunteers. Physical examination and medical history were obtained from all participants, to avoid exclusion criteria based on the American Society of Anesthesiologists physical status class I (e.g. cardiovascular or respiratory problems, hypertension, apnea, asthma, neurologic disorders, psychiatric disorders, pregnancy, and others.)

Spontaneous high-density electroencephalography (hd-EEG, 128 channel EGI Hydrocel GSN) was recorded from 15 healthy participants (age 20-40; 7 male, 8 female) during ketamine anesthesia at four different states of consciousness: *Wakefulness*, *Sub-anesthetic Sedation*, *Loss of Behavioral Response* (*LOBR*) and *Recovery*. After the initial 10-minute baseline period (*Wakefulness*), a continuous intravenous infusion of 0.5 mg/kg racemic ketamine was administered over a 40-min period (*Sub-anesthetic Sedation*, eyes closed), followed by a 1.5 mg/kg anesthetic bolus dose that led to loss of behavioral response (*LOBR*), and then a 10-min *Recovery* period. Before, during and after the anesthetic bolus, participants were instructed to squeeze an object on either the right or their left hand, based on a randomized auditory loop (one command every 30 sec). *LOBR* was denoted when participants ceased to respond to two consecutive commands. Additional medications were also administered when needed, for nausea and vomiting prophylaxis (ondansetron and scopolamine patch).

In contrast to the previous studies, drug concentrations here were not guided by target-controlled infusions (effect-site or plasma targeting), but rather on dosing strategies directly relevant to clinical care for depression (sub-anesthetic dose) or anesthetic induction (bolus dose). For the purpose of our analysis, two steady-states were extracted from the sub-anesthetic period and during *LOBR*, in accordance with the analysis found in the original studies (Vlisides *et al.* 2017; Vlisides *et al.* 2018). The *Sedation* steady-state consists of the final block of the sub-anesthetic period (25-40 min), which was assumed to have reached a pharmacologically-steady state (plasma concentration was approximately 180 ng/ml at the end of this phase). *LOBR* consists of a 5 min block after the moment of cessation of response to commands (3.8 min for one participant). Moreover, for 14 out of 15 participants, we were able to acquire all in-between transitional state recordings, namely: from wakefulness to steady-sedation (*WS*), from steady-sedation to the end of the sub-anesthetic state (*PS*), from the end of the sub-anesthetic state to *LOBR* (*SL*), and from *LOBR* to recovery (*LR*). The experimental design of the study is shown in detail in Fig. 5.3.

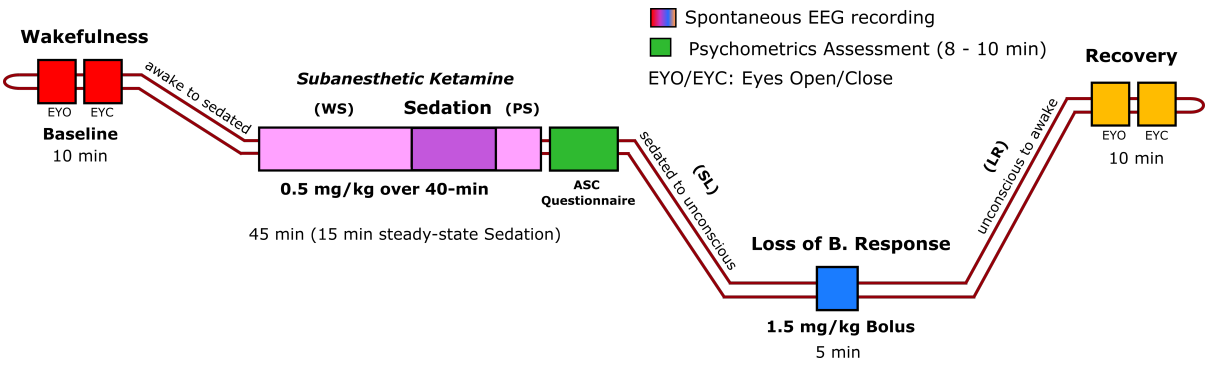

Fig. 5.3. Experimental design of the *Michigan Anesthesia Study*. The anesthetic states were guided by dosing strategies directly relevant to clinical care for either depression (sub-anesthetic dose) or anesthetic induction (bolus dose).

Apart from the behavioral paradigm denoting *LOBR*, self-report measures of altered states of consciousness (ASC) were also recorded during and after the study, using a validated questionnaire (IRB approved). Besides its anesthetic effects, ketamine is known to have psychoactive properties that include perceptual distortions, cognitive impairment, and feelings of disconnection from the body and environment. In contrast to other agents (such as propofol), states of disconnected consciousness (dream-like states) can be induced by ketamine, even under profound behavioral unresponsiveness (as revealed by a significant body of work (Bonhomme *et al.* 2019)). The questionnaire used here consisted of 83 items related to 13 subscales (perceptual dimensions), namely: *experiences of unity*, *spiritual experience*, *blissful state*, *insightfulness*, *disembodiment*, *impaired control and cognition*, *anxiety*, *complex imagery*, *elementary imagery*, *audio-visual synaesthesia*, *changed meaning of percepts*, *transcendence of time and space*, and *ineffability* (further definitions can be found in the original study and in (Studerus, Gamma and Vollenweider 2010; MacLean *et al.* 2012)). The response for all items was from 0 (no, not more than usual) to 10 (yes, very much more than usual), with the subscale score being the average of all items within that scale. These psychometrics were assessed twice, once after the sub-anesthetic period and before *LOBR* (*study score*), and once using an online questionnaire within 48 hours of the study (*lifetime history score*), which reflected the rating of these experiences throughout the whole lifetime of the participant. By looking at subscales with good intra-scale reliability (Cronbach's alpha > 0.7) and the highest deviation between study and lifetime scores, three subscales were identified in the original study as most relevant to ketamine sedation: *disembodiment*, *transcendence of time and space*, and *complex imagery*.

5.2.2 EEG Pre-processing

All EEG datasets were commonly pre-processed using the pipeline derived in the previous chapter (section 4.8.1), which has been shown to be generally optimal. Briefly, 20 electrodes are selected based on the 10-20 system, which are conventionally present in full-coverage EEG acquisition systems, namely: Fp1, Fp2, F7, F3, Fz, F4, F8, T3, C3, Cz, C4, T4, T5, P3, Pz, P4, T6, O1, Oz, and O2. Band-pass filtering is then applied between 0.5 and 40 Hz, with an additional notch filter at 50 or 60 Hz, depending on the frequency of the respective power line noise. After filtering, the data are resampled at 100 Hz and segmented into 1 sec non-overlapping epochs. For training data, an automatic artifact cleaning procedure is performed, for bad channel interpolation and bad epoch rejection, based on the 800 μV peak-to-peak threshold. Finally, the data are re-referenced to the average and epochs are normalized with Robust Standardization (*scikit-learn*), using the 0.25-0.75 quantile range. All of the above steps are executed automatically using the *mne* library and can be applied online.

5.2.3 Deep Learning Model

The deep learning model used in all subsequent experiments was the 3D convolutional neural network, described in section 4.8. This architecture is able to explore the spatio-temporal structure of the EEG, without strong and explicit priors on the electrophysiological nature of the signals and features of interest. Moreover, it allows the integration of our selected datasets and the possibility to jointly train the network in a consistent manner, if required by our research question. By using a fixed methodology on EEG processing overall, we are able to compare our findings irrespective of model design, with a focus on the task under optimization.

Briefly, the 3D cNN is a sequential model that uses a 3D representation of the EEG epoch inputs, by arranging the 10-20 system channels into a 2D mesh structure (time courses being the 3rd dimension). The first functional layer comprises of two 3D convolutional layers (32 and 64 feature maps, respectively), which perform temporal-only and spatio-temporal nonlinear filtering, followed by a MaxPooling and a Dropout layer. The second functional layer comprises of a flattened fully-connected layer (128 units), followed by a Dropout layer, and finally an output layer. All activation functions in convolutional and dense layers are ReLU units, with the exception of the output layer. The output layer consists of either N softmax units for N-class classification tasks, or one linear unit for regression tasks. Kernel sizes, strides, pooling and dropout rates remain as depicted. All other hyperparameters are set to default values in keras, unless specified otherwise. The 3D cNN model design is summarized in Fig. 5.5.

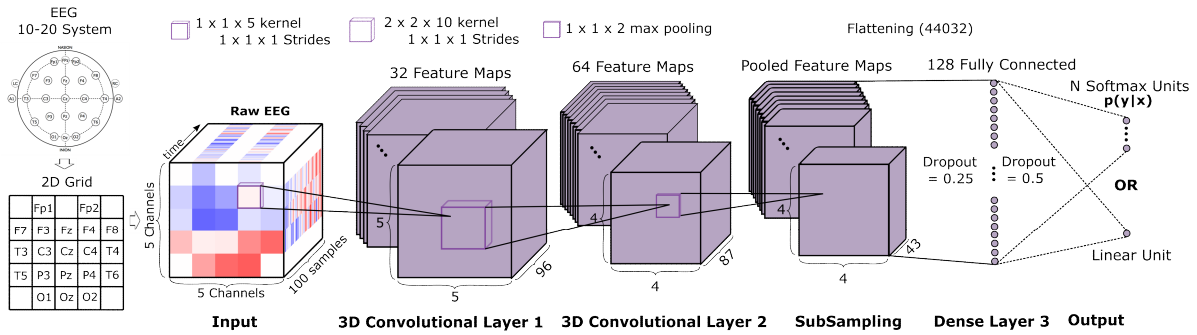

Fig. 5.5. The 3D Convolutional Neural Network model. The output layer is replaced by N softmax units for classification, or a linear unit for regression, depending on the nature of the learning task.

5.2.4 Model Training and Evaluation

Model training and evaluation was kept consistent with the previous analyses in Chapters 3 and 4, with several additional parameters that allowed us to include different datasets and learning tasks (i.e. regression). For the purpose of predictive analysis of the variously defined anesthetic states, we are interested in observing the behavior of the network

under different classification and regression tasks, while the main methodology remains fixed. The idea here is that any observable differences can be attributed to factors such as the nature of the data (e.g. anesthetic agent), the dataset size, the number of states under training, and of course, the given ground-truth, alongside its underlying assumptions (the discrete or continuous nature of consciousness represented in the various anesthetic states, defined by behavior, drug concentration or psychometrics). While these factors cannot be studied completely independently, as they heavily depend on the respective dataset and experimental design, they determine the optimization process and can significantly affect model training, representations and performance.

Specifically, a leave-one-participant-out cross validation paradigm was used for splitting the data intro training and testing sets, for all experiments, ensuring cross-subject generalization. For classification tasks, one-hot encoding was used for the target vectors, with the categorical cross-entropy (CCE) as the loss function and accuracy as the performance metric. For regression tasks, the behavioral (Ramsay scores), pharmacological (plasma concentrations) or psychometrical (ASC scores) measures were used as regressands, with the mean-squared-error (MSE) as the loss function, and mean-absolute-error (MAE) as the performance metric. All models were trained using the Adadelta optimizer, with a batch size of 100 and for 10 training epochs. Initialization of network weights was done with the Xavier uniform initializer. Model creation, training and evaluation were implemented in Python 3 using the *Tensorflow*/*Keras* library and a CUDA NVIDIA GPU (Tesla P100).

Classification vs Regression. Both main types of supervised learning algorithms were investigated throughout the experiments, given that each one reflects the discrete and continuous aspects of the different anesthetic states and levels of unconsciousness. As mentioned in section 5.1.3., literature review has revealed a number of works using either approach, albeit classification was the most prominent. While we briefly discussed some theoretical concerns regarding the short-scale analysis window of our input in Chapter 3, large-scale temporal dynamics of the brain's activity and its corresponding EEG patterns can be directly relevant to the performance of each approach (despite the fact that the cNN model is time-invariant across epochs and does not consider large-scale temporal relations). From an engineering point of view, this can be due to the underlying assumptions of the respective learned function and the imposed restrictions on optimization (either from the loss function or the architecture of the model), which ultimately determine the feature learning and the predictive behavior of the model. From a theoretical point of view, the question of learning algorithm is associated to implicit assumptions about the nature of consciousness, which nevertheless remain unclear (e.g. the notion of consciousness levels, EEG signatures, and their interaction, as continuous or discrete phase transitions). Despite this fact, we make presumptions of steady and transitional states within each anesthetic paradigm, given the selected ground-truth measures taken during the experiments.

Sample Weighting. In order to account for dataset differences found in the selected experimental setups, a sample weighting method was developed and used, particularly having in mind cases of unbalanced datasets (e.g. with respect to the size of the states and the availability of ground-truth, as prominently appears in the *Michigan Anesthesia Dataset*), or cases of training with multiple heterogeneous datasets, aiming at a cross-study and cross-drug analysis approach.

In general, sample weighting and class weighting are used in machine learning for adjusting the cost function, as a way to apply different weights to the independent sample or class losses that might be over- or under-represented (and in which case, contribute to a greater or lesser extent on model training and optimization). More specifically, a weight vector can be applied as coefficients to the per-sample losses during the calculation of the cost function J:

$$J(X, Y; W, b) = \sum_{i=1}^{m} w_i \, L(\hat{y}^{(i)}, y^{(i)})$$

where w_i is a sample weight for the i^{th} sample, and m is the number of samples within the current training batch.

For our own case study, we devised a formula towards an overall unbiased model, that takes into consideration several weighting factors, such as the number of epochs per state (w_{1i}), the number of states per subject (w_{2i}), the number of subjects per dataset (w_{3i}), the number of datasets per agent (w_{4i}), and the number of agents (w_{5i}), based on the identity of the i^{th} sample. We also implemented a weighting factor that reflects the number of instances per target (w_{6i}), irrespective of other identifiers, for each i^{th} sample belonging to that target (target weighting). By multiplying all weighting factors, we calculate sample weights as follows:

$$w_{sample_i} = \left(\frac{1}{w_{1_i} \, w_{2_i} \, w_{3_i} \, w_{4_i} \, w_{5_i}} \right) \left(\frac{1}{w_{6_i}} \right)$$

All weighting factors were applied throughout the experiments, unless specified otherwise (target weighting was dismissed during training with unique regressands). Importantly, these weights are calculated based on the training data and are used only during training. Before feeding the weights to the loss function, we normalized the values by $\frac{N_{training}}{\sum w_{sampl\ i}}$, which ensures that the sum of all factors is always $N_{training}$ (and thus, the overall cost is comparable to the default summation of losses).

Overall, although it has been suggested that sample weighting is mostly impactful during the early stages of training (Byrd and Lipton 2018), we empirically found that the inclusion of weighting either improved, or did not affect model performance, and thus was retained in our analysis.

5.3 Experiment 1 – Behaviorally-defined Anesthetic States

In this set of experiments, we investigate the effect of learning task and ground-truth in the predictive behavior and performance of our model, when trained and tested under behaviorally-defined anesthetic states. Specifically, we explore what the model reveals about the EEG signatures of steady and transitional behavioral states, and we test the ability of deep learning to reasonably generalize across states and participants (which is currently missing).

Aside from any theoretical understanding, such findings could lead to novel electrophysiological markers, which are potentially valuable in clinical practice, either for the detection of the moment of *LOC*, or for the production of a continuous index measuring the depth of anesthesia (similarly to many commercial DoA monitors). The assessment of levels of unconsciousness in many diagnostic contexts relies heavily on behavioral measures of responsiveness, which nevertheless have been shown to be generally unreliable (discussed in detail in Chapter 2). On the other hand, and particularly for measuring the DoA, current methodological approaches (such as the widely used BIS index) have shown weaknesses in application across agents and anesthetic states, when evaluated on patients' unresponsiveness (Kreuzer 2017). Therefore, it is important to evaluate our ground-truth, and assess whether EEG provides further information beyond the standard clinical measures, coming either from behavior or from current methodological techniques of EEG analysis.

For this investigation, we used the *Liege Anesthesia Dataset* that allows us to analyze behaviorally-steady states, as well as in-between transitional states, whilst providing us with a robust ground-truth and adequate data for training. As mentioned in section 5.2.1, the four main states were recorded upon reaching and sustaining the desired Ramsay score for each state (Ramsay 2, 3, 5-6, and 2, for *Wakefulness*, *Sedation*, *LOC* and *Recovery*, respectively). The appropriate effect-site concentration that corresponded to the desired behavioral state was equilibrated and adjusted independently for each subject by the TCI (target-controlled infusion) device, allowing us to consider the recorded states as behaviorally-steady.

Through our 3D cNN model, we conducted classification and regression experiments with differing number of states – starting with the two prominent states of *Wakefulness* and *LOC*, followed by the inclusion of *Sedation*, and finally *Recovery*. This progressive increase in the number of states allowed us to introduce more and more subtle changes in levels of sedation, during the training and testing of our model. For classification, we used one-hot encoding based on the number of trained states, as our target vectors. In case of regression, the Ramsay score of each state was used as the trained regressands. The model was then tested on unseen steady and transitional states, to reveal the interplay of the various anesthetic states.

5.3.1 Classification Results

The results of the classification analysis are summarized in Table 5.1. The loss and accuracy obtained by the model is shown for each of the three experiments, which progressively introduce intermediate levels of sedation. All models had a stable convergence during the 10 training epochs, indicated by the average categorical cross-entropy loss and accuracy curves (Fig. 5.6).

Table 5.1. Behaviorally-steady state classification results

States under Training	Loss (CCE)	Accuracy		
		Chance	***Per State***	**Total**
Wakefulness, LOC	0.08	*50%*	*(98%, 99%)*	98.5%
Wakefulness, Sedation, LOC	0.66	*33%*	*(94%, 79%, 90%)*	87.7%
Wakefulness, Sedation, LOC, Recovery	1.24	*25%*	*(74%, 73%, 91%, 61%)*	74.8%

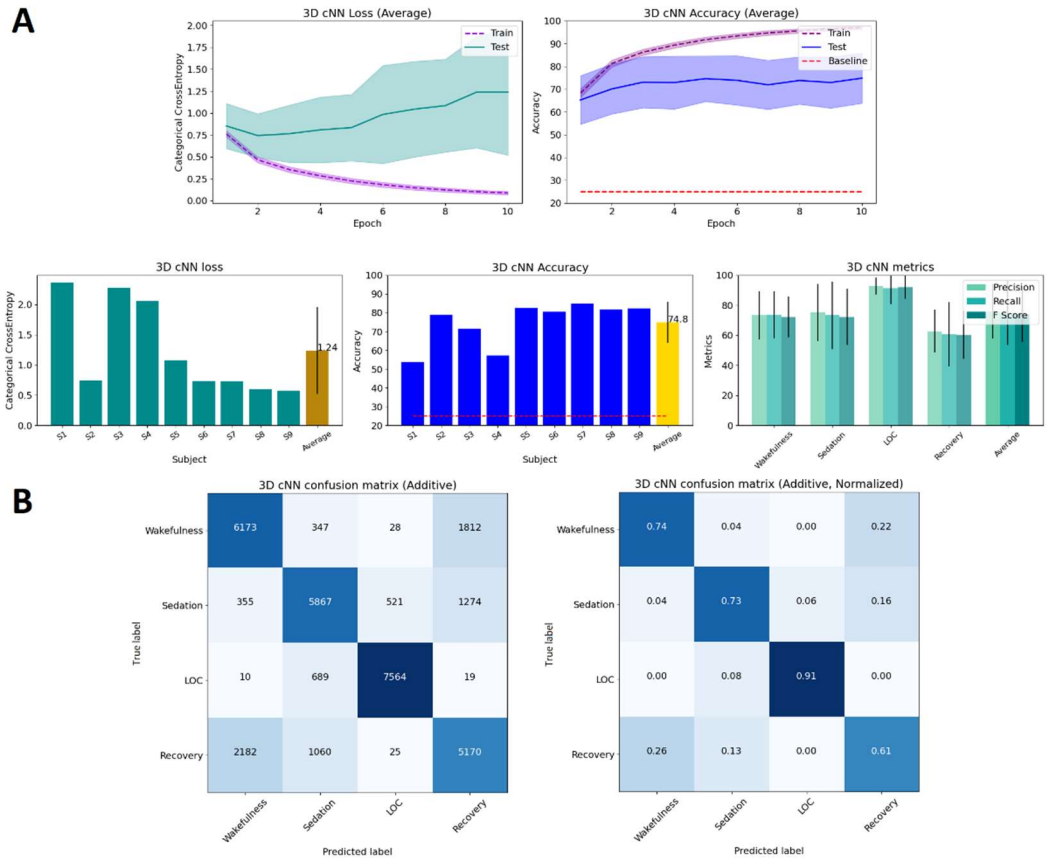

Fig. 5.6. Classification results of the *Liege Anesthesia Dataset*, for the 4 behaviorally-steady states. A) Average categorical cross-entropy loss and accuracy curves (top). Subject-wise losses, accuracies and other metrics (bottom). B) Additive confusion matrix (left), and normalized additive confusion matrix (right).

As already seen in other studies (section 3.1.3), the classification task of the two most behaviorally-distinct states (*Wakefulness* and *LOC*) is a problem solved easily by contemporary algorithms, even under a cross-subject validation approach (in our case with 98.5% accuracy), but without offering us significant depth. The 3-state task is a more balanced problem, which we have used as a baseline, and which we have analyzed and discussed in detail in Chapter 3. Briefly, an accuracy of 87.7% is obtained, essentially due to the difficulty in predicting the intermediate state of *Sedation*, and particularly for 4 out of 9 subjects (S1, S4, S5 and S9). Most interesting here though are the results regarding the 4-state classification task, which includes the state of *Recovery*. The detailed cross-validation accuracies, losses, confusion matrices, and other class metrics are depicted in Fig. 5.6.

In this case, an average accuracy of 74.8% is achieved, with three participants showing a low accuracy in the range of 50-70% (S1, S3 and S4), and the remaining five showing an accuracy close to 80%. Again, subjects 'S1' and 'S4' were the most difficult test cases, as observed similarly in the 3-state task. Notably, all subjects in all experiments showed a performance that was significantly higher than chance levels (50%, 33% and 25% accuracy, respectively), which indicates the learning capacity of deep learning, even for subtle changes in levels of sedation.

To better understand the behavior of the model, and the resulting decrease in performances, we observe the confusion matrices in Fig. 5.6. The normalized confusion matrix (additive across all subjects) shows a significant proportion of the EEG misclassified between *Wakefulness* and *Recovery*, and to a lesser extent, between *Recovery* and *Sedation*. While the model can clearly identify *Recovery* as a distinct state from *Wakefulness*, with higher than chance accuracy, it also reveals that the EEG signature of *Recovery* shares common characteristics with *Wakefulness* and *Sedation* (possibly indicating a state of an intermediate level of sedation). This distinction between *Recovery* and *Wakefulness* was also evident by the individual class-wise accuracies, with few participants having *Recovery* over-classified as *Wakefulness*/*Sedation*. Such findings are contrary to our behavioral ground-truth, where *Wakefulness* is indistinguishable from *Recovery*.

Another way to understand these predictions is by visualizing the model's output over time (one prediction per 1-sec epoch). The average predictions of the softmax output, which show the probabilities for each of the four states, can be seen in detail in Fig. 5.7. We calculate the average values across subjects, by taking the minimum number of epochs per state, and aligning them at the beginning of each state. Probabilities here reveal again the shared characteristics of *Recovery* with *Wakefulness* and *Sedation*, observed previously in the confusion matrices. Moreover, the prominent values of each state show an overall trend of decreasing probability for *Wakefulness* at the end of *Wakefulness* state, and an increasing probability for *LOC* at the beginning of *LOC* state, respectively. This observation confirms the existence of large-scale temporal dynamics, which have already been hypothesized, as well as the transitional nature of EEG, within presumably steady-states.

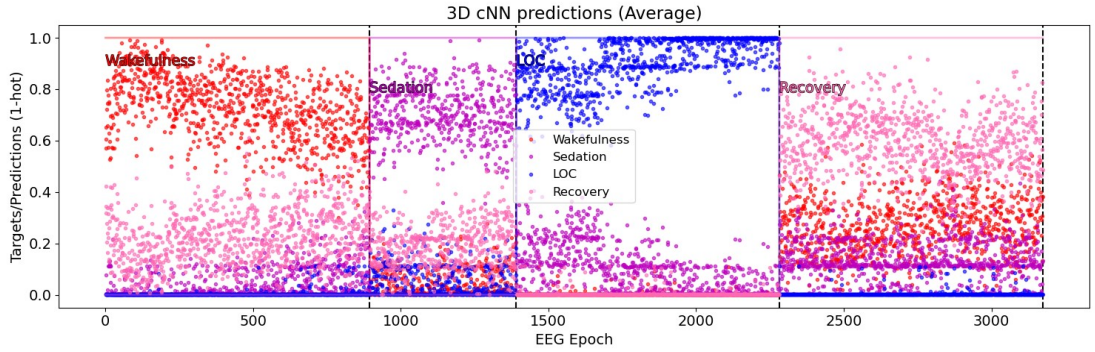

Fig. 5.7. Softmax predictions for the unseen test subjects (average), showing the probabilities of the 4 trained classes over time (1-sec epochs), for the 4 behaviorally-steady states.

Overall, the 4-state classification task is a challenging problem, due to the similarities across *Wakefulness* and *Recovery*.

Testing on Transitional States. In order to further test the above observations, we used the model trained in all steady-states (4-state classification task) and visualized its predictions over a recording period that included both behaviorally-steady and in-between transitional states (*Wakefulness to Sedation*: *WS*, *Sedation to LOC*: *SL*, *LOC to Recovery*: *LR*), given the availability of intermediate recordings for 3 (out of 9) participants. As these recordings were almost consecutive in time (following the anesthetic paradigm in Fig. 5.1), we concatenated the states and averaged across the 3 test participants, as previously described (Fig. 5.8).

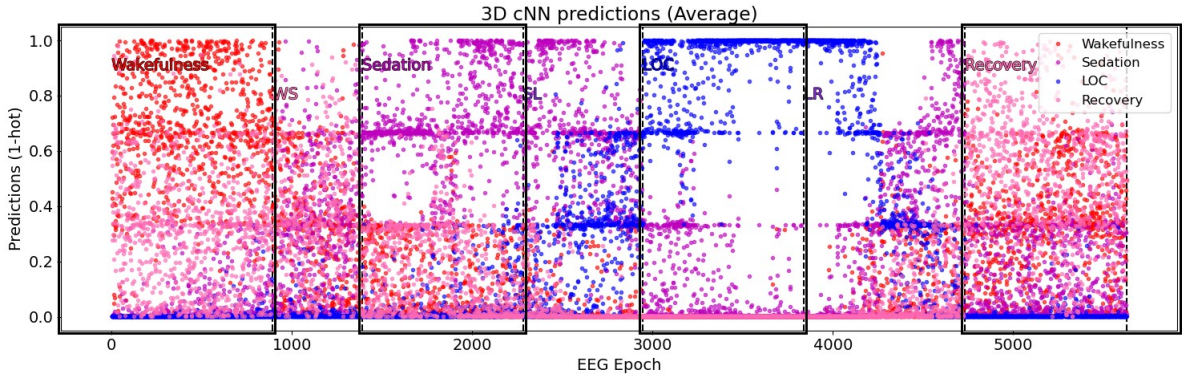

Fig. 5.8. Softmax predictions for the unseen test subjects (average), showing the probabilities of the 4 trained classes over time (1-sec epochs), for all states (behaviorally-steady and transitional states). States under training are highlighted with black boxes.

Fig. 5.8 shows a similar trend for the prominent probabilities, although with a much noisier pattern that can be attributed to the limited number of subjects, and other possible inter-individual differences reflected in the EEG (e.g. due to variations in the length and timing of the events, the individualized response to anesthetic induction and recovery, etc.). In general,

we can track the anesthetic paradigm and the participants' transition from one state to the next, both within steady and transitional states, albeit appearing as step-like probability increments (mostly within the transitional states). As discussed in section 5.2.4, this shows the limitations of classification (and to an extent, of behavioral measures) with respect to the transitional dynamics of the EEG, which nevertheless seem to play a significant role (especially during the states' transitions, which theoretically and clinically are the most crucial to understand and capture). Additionally, high-probability predictions of *Recovery* during *WS* verify again the idea of *Recovery* as a mildly-sedated state, given the initiation of propofol infusion during *WS* (evidence of residual propofol during *Recovery* can be found in (Chennu *et al.* 2016)).

5.3.2 Regression-to-Ramsay-Score Results

The results of the regression analysis are summarized in Table 5.2. The loss and mean-absolute-error (MAE) obtained by the model is shown for each of the three experiments, which introduce intermediate levels of sedation. As already pointed out, regression enables the representation of transitional dynamics, which we have shown to be present and relevant in our analysis. Also, in contrast to classification, the regression task allows us to reasonably test the model in all states (whether included in training or used for testing), and consistently compare the experiments. Overall, all models had a stable convergence during the 10 training epochs, indicated by the average mean-squared-error loss and mean-absolute-error curves (Fig. 5.9).

Table 5.2. Regression-to-Ramsay-Score results

States under Training	Loss (MSE)	Mean Absolute Error (MAE)	
		Per State	**Total**
Wakefulness, LOC	0.58	*(0.3, 0.84, 0.69, 0.36)*	0.54
Wakefulness, Sedation, LOC	**0.50**	***(0.29, 0.54, 0.76, 0.41)***	**0.50**
Wakefulness, Sedation, LOC, Recovery	0.51	*(0.31, 0.63, 0.73, 0.41)*	0.52

Unseen test states are underlined. Optimal model is highlighted in bold.

The results obtained here give us some previously observed, and some novel insights, regarding the EEG under general anesthesia, and particularly with respect to the *Liege Anesthesia Dataset*. In terms of the difficulty of each experiment, we observe an analogous trend of decreasing performance, as we include more states, similarly to the classification analysis. However, Table 5.2 reports the model's evaluation in all four states, given the shared ground-truth and uniformity of the behavioral scale. This provides us with a direct comparison on the effect of the chosen states under training, in the performance of our model (Total MAE refers to the MAE of all 4 states. Underlined values refer to unseen test states).

Training by regression-to-Ramsay-scores revealed a predictive bias, which was affected by several factors, such as the number of samples in each state, the given training states, and the assumption of the Ramsay scale to represent changes in levels of consciousness linearly. In this case, sample weighting had an observable effect on the results, most evidently in the 3rd experiment with the inclusion of *Recovery* under training, due to the increment of

instances with common targets (Ramsay 2). Specifically, *Sedation* was significantly overestimated by 1.12 MAE in the 2-state experiment, while *LOC* was significantly underestimated by 1.07 MAE in the 4-state experiment. As some of these predictive biases were corrected by sample/target weighting, any differences observed here can be attributed to the effects of the training states and the Ramsay scale. The detailed cross-validated performances and the visualization of the model's predictions can be seen for the 3-state experiment, which had the best overall performance (MAE: 0.5, within Ramsay scale), in Fig. 5.9 and Fig. 5.10, respectively.

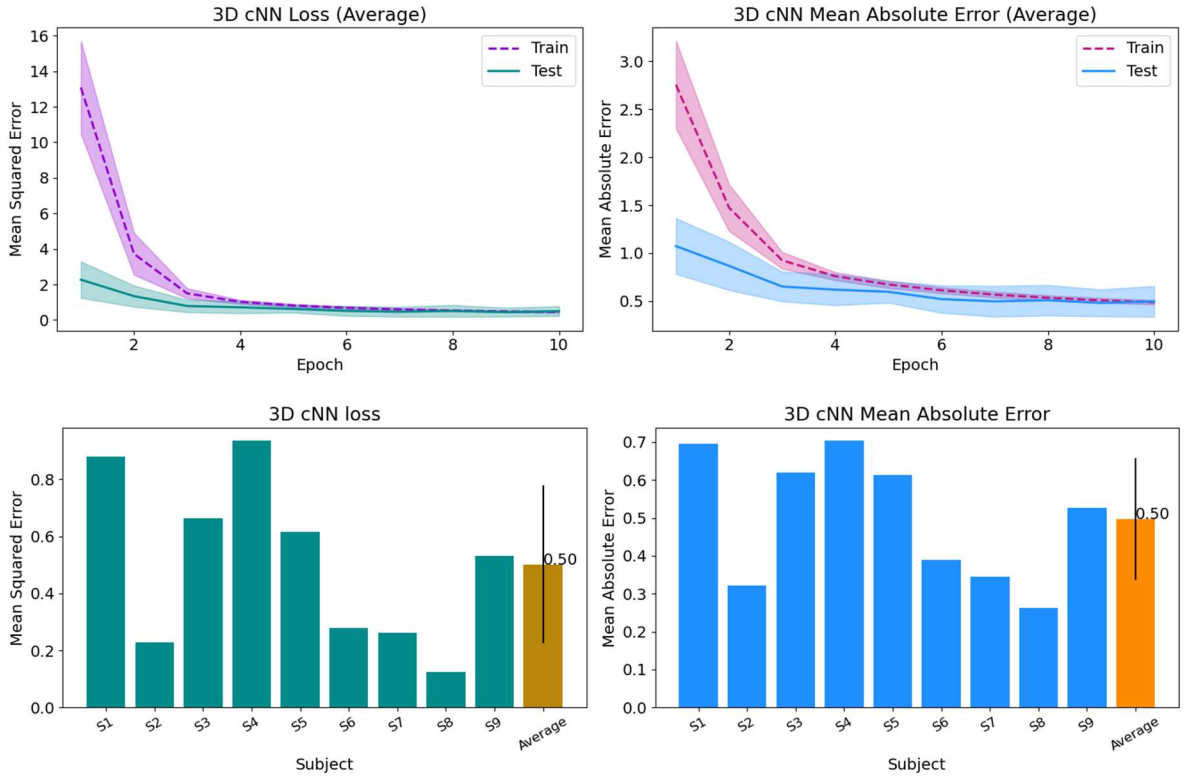

Fig. 5.9. Regression-to-Ramsay-score results of the *Liege Anesthesia Dataset*, for the model trained under the 3-state task (*Wakefulness*, *Sedation*, *LOC*). Average mean-squared-error loss and MAE curves (top). Subject-wise losses and mean-absolute-error values (bottom).

Fig. 5.9 shows that the model achieved the lowest MAE at 0.5, as well as the most balanced per-state MAE, when trained under *Wakefulness*, *Sedation* and *LOC*. Specifically, four participants (S1, S3, S4, S5) obtained an MAE between 0.6 – 0.7, with the rest five participants having an MAE lower than 0.5 (0.2 – 0.5). To better understand these results, we observe the average-across-participants predictions for each of the 3 experiments (calculated by the minimum number of epochs per state, across subjects, as shown in Fig. 5.10). As previously described, *Sedation* and *Recovery* were the hardest to predict, which is revealed by several observations here. The inclusion of *Sedation* during training significantly lowered the MAE of the state and its separation from *Wakefulness*, which shows its importance in training

and the (non-linear) complexity of the electrophysiological transition from *Wakefulness* to *LOC*. In contrary, the inclusion of *Recovery* did not seem to affect its minor over-estimation by the model (MAE of 0.41), in agreement to our previous evidence of *Recovery* as a mildly-sedated state. Finally, the MAE range of *LOC*, although distinguishable from the other states, could be partly explained by the linearity assumption of the Ramsay scale.

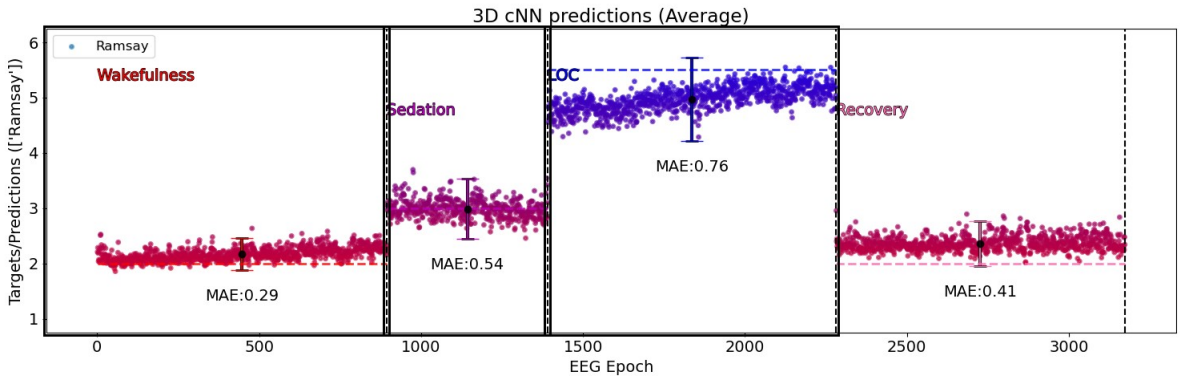

Fig. 5.10. Ramsay score predictions for the four anesthetic states of the unseen test subjects (average). Horizontal dashed lines indicate the Ramsay score ground-truth. States under training are highlighted in black boxes.

Testing on Transitional States. Similarly to classification, in order to further test the model's behavior under the regression task, we visualized the Ramsay score predictions in both behaviorally-steady and transitional states, for the three available participants (S1, S6, S8) from which we had access to all intermediate recordings (Fig. 5.11).

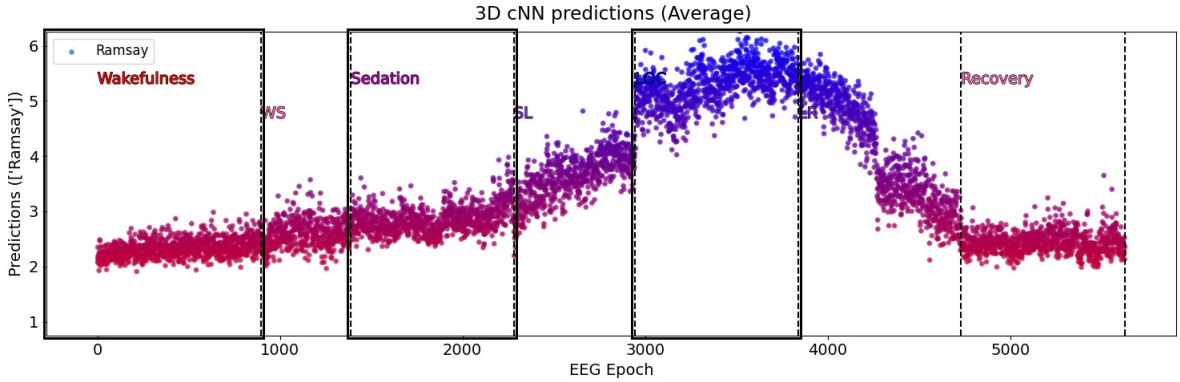

Fig. 5.11. Ramsay score predictions for the steady and in-between transitional anesthetic states of the three available test subjects (average). States under training are highlighted in black boxes.

Due to the continuity of the recordings, the concatenation of the states allow us to track the anesthetic paradigm depicted in Fig. 5.1. In contrast to the probabilistic approach of the softmax output, the learning task here enables the model to represent the various levels of

sedation in a progressive manner, which seems to be more appropriate based on our empirical observations, but also in terms of our research goals. As seen from Fig. 5.11, the Ramsay predictions show an increasing trend that continues up to the end phase of *LOC*, followed by a more rapid decrease during *LR* and up to the beginning of *Recovery*, re-confirming the transitional nature of EEG and its large-scale temporal dynamics.

This trajectory can also be tracked in individual participants, which is one of the most important factors of evaluation for our model. In Fig. 5.12, we show the predictions of a test subject (S6) among the three with all available transitional states (median performance - MAE of 0.38), along with the time-frequency representation of each state (spectrogram), which is representative both in terms of EEG signatures and model performance. Overall, predictions appear robust and in agreement with the expectations of GA practice (pharmacologically), which validates our model's objective and performance.

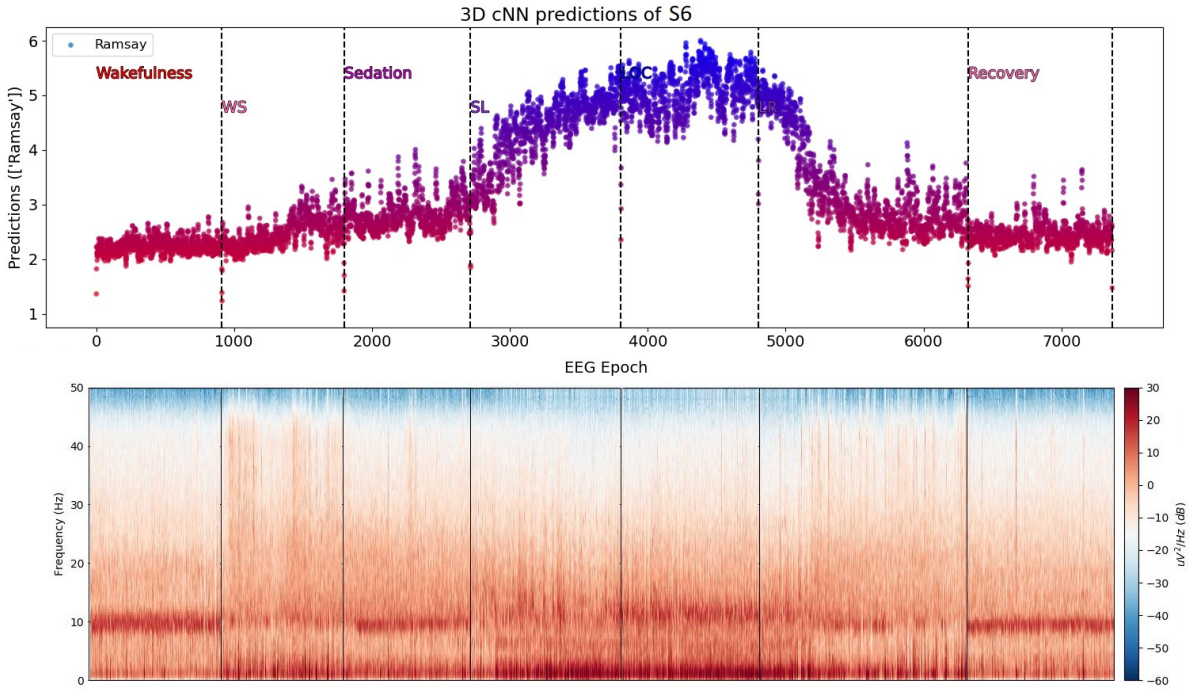

Fig. 5.12. (Top) Ramsay score predictions for all anesthetic states (steady and transitional) of a median-performance subject (Moving-average filter applied, kernel_size=5). (Bottom) PSD of the corresponding states (Method=Welch, channel_aggregation=mean, window=1 sec, n_fft=256)

Visual inspection of the Ramsay predictions and the mean spectrogram of the corresponding epochs does not reveal any clear association between the band-power dynamics of the EEG and the depth of anesthesia. As briefly discussed in Chapter 3, the spectrogram shows a decrease of alpha (8-12 Hz) and increase of spindle and beta activity (12-25 Hz) during *Sedation* and transition to *Recovery* (Ramsay 3), while *LOC* (Ramsay 5-6) seems to be accompanied by significant increase in delta (0-4 Hz) and moderate increase in high-alpha and theta (4-8 Hz) activity. Moreover, gamma activity (25-40 Hz), which has been reasonably

dissociated from possible muscle activity, appears occasionally during *Sedation* and transition to *Recovery* (Murphy *et al.* 2011). While these findings have been reported in the original study and have been replicated several times in the literature, the nature and interaction of these signatures appear to be more complicated to understand and interpret, both in terms of the underlying neurophysiology of the brain, but also with regards to model prediction here.

Post-hoc Analysis of 'Bad' Subjects. As a final step to this analysis, we performed a post-hoc investigation for the three subjects with the lowest performance (S1, S3 and S4 had, in average, the lowest performances, in both classification and regression analyses), hoping to better understand the data and the trained model. In all three participants, we observed a significant underestimation of *LOC* (MAE > 1), particularly at the beginning phase of the state, that strongly affected the resulting performance. Besides the concerns we postulated regarding the appropriateness of the Ramsay scale, these findings reflect the inter-individual differences in EEG signatures, which naturally appear either inherently (e.g. due to age), or in relation to the anesthetic procedure (e.g. due to infusion rate differences and PD impact). By visual inspection of the individual spectrograms we observed a strong positive correlation between Ramsay predictions, at the levels around *LOC* (Ramsay > 3), and the power of delta activity, which appeared to be present in all participants (see Fig. 5.12 for reference. This hypothesis has been studied in detail in (Mhuircheartaigh *et al.* 2013b)). Of course, whether these predictions (and the relative decrease of delta power) show a peculiarity of the model, or reveal possible ground-truth errors, is unclear from the data.

Additionally, we observed a peculiar pattern during the *Sedation* state of S4, which had the worst performance in the dataset (Fig. 5.13). By further investigating the procedural details during the experiment's recordings, we acquired evidence (in the form of handwritten notes by the clinician at site) of reversed order for the steady-state recordings of *Sedation* and *LOC*, in 4 out of 9 participants (S2, S4, S7, and S9), which were also confirmed by their file timestamps. This is explained by the inter-subject variability of the pharmacodynamic impact of propofol, which drove several participants directly into unconsciousness, before reaching and sustaining a state of *Sedation* (by lowering the drug's infusion rate). The pattern observed in Fig. 5.13 can now be better understood by reversing the two states, where a more natural transition for the levels of sedation is revealed (and in agreement with the infusion rates). While this experimental diversion did not have an impact for 3 out of 4 subjects, possibly due to the 5-minute equilibration period before the steady-state recordings, it highlights several points on model evaluation. Most importantly though, it provides novel evidence for the potential of deep learning to decode and predict the transitional signatures of individuals.

Overall, understanding the EEG patterns that emerge during anesthesia, and the corresponding predictions from our deep learning model, is a task that critically depends on the evaluation of several factors that strongly contribute to the nature of the data, and thus in model training. Considering factors related to the anesthetic procedure (e.g. the rate of drug administration, the phenomena of neural inertia and hysteresis, the induction and emergence

asymmetry, etc.), and other ground-truth considerations, such as the distinction between *LOC* and *LOBR*, will allow us to better estimate the role of our learning objective and its limitations in model optimization.

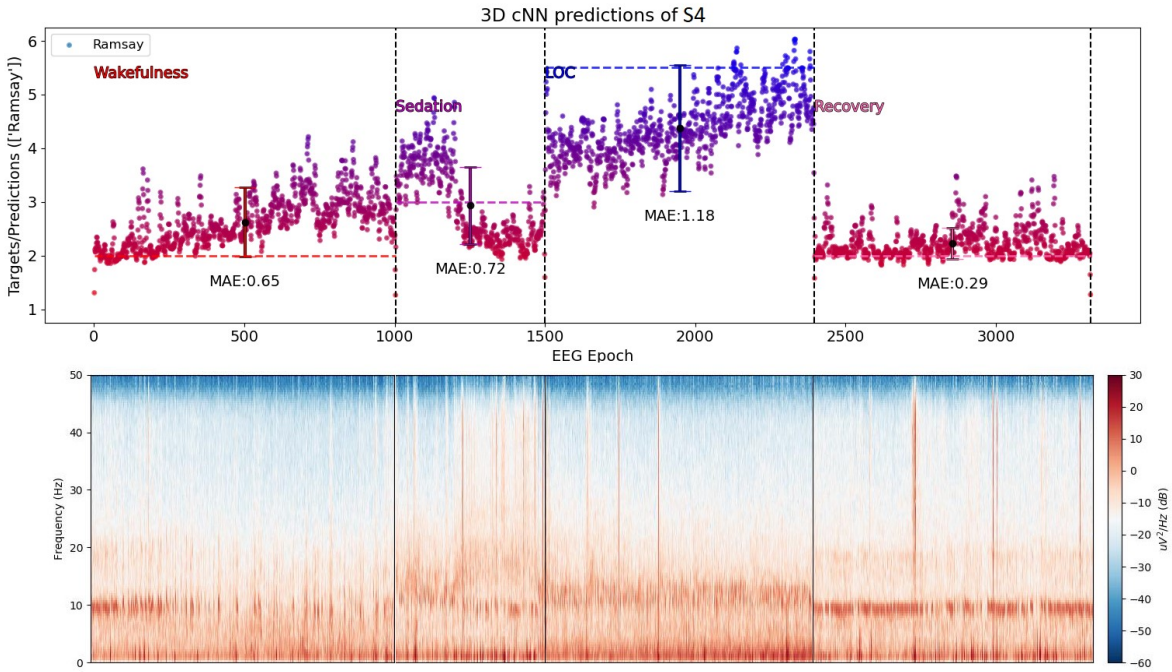

Fig. 5.13. (Top) Ramsay score predictions for the 4 anesthetic states of the worst-performing subject (Moving-average filter applied, kernel_size=5). (Bottom) PSD of the corresponding states (Method=Welch, channel_aggregation=mean, window=1 sec, n_fft=256)

5.4 Experiment 2 – Pharmacologically-defined Anesthetic States

In this set of experiments, we investigate the effect of learning task and ground-truth in the predictive behavior and performance of our model, when trained and tested under pharmacologically-defined anesthetic states. Similarly to our previous experiment, we explore what the model reveals about the EEG signatures of pharmacological steady-states, and we test the ability of deep learning to reasonably generalize across states and participants.

From a clinical perspective, understanding the relationship between drug doses and patient response requires the consideration of complex interactions between the administered doses and the plasma (or effect-site) concentrations of the drug (pharmacokinetic-PK phase), the effect-site concentrations and the clinical effect (pharmacodynamic-PD phase), as well as their coupling. The delivery of the drug and its effects over the different tissues in the body is determined by a variety of phenomena (such as the observed hysteresis, caused by the temporal delay in drug equilibration), which are mathematically described in PK/PD models. Nevertheless, the available PK/PD models implemented in current TCI (target-controlled

infusion) devices have significant limitations in accurately estimating the required drug concentrations, for the specific needs of each patient, due to a variety of pharmacological and biological factors. Hence, it is important to evaluate the pharmacological ground-truth used in contemporary anesthesia practice, and the possibility for EEG signatures to contain information (in this case, related to propofol concentrations), which can potentially help us better understand this complex interaction.

For this investigation, we used the *Cambridge Anesthesia Dataset* (section 5.2.1) that enables us to analyze pharmacologically-defined anesthetic states, whilst providing us with two different ground-truth measures. The first ground-truth relates to the per-state plasma concentrations which were targeted by the TCI device, and were common across participants. In contrast to *Liege Anesthesia Dataset*, where states of consciousness were defined based on behavioral evidence for consciousness, the ground-truth here was based on states determined upon reaching specific drug concentrations (plasma targeting). The 10-minute equilibration period before the recording of mild and moderate sedation states is considered adequate to avoid any strong pharmacological instabilities (due to infusion rate changes or PK errors from the syringe pump), thus allowing us to define them as pharmacologically steady. The second ground-truth relates to the plasma concentrations measured from the blood samples (taken between states), which can reflect some of the PK properties and clinical outcome of the individual participants.

We used our 3D cNN model to conduct classification and regression experiments with differing number of states – starting with the two prominent states of *Baseline* (*Wakefulness*) and *Moderate Sedation*, followed by the inclusion of *Mild Sedation*, and finally *Recovery*. This progressive increase in the number of states allowed us to introduce more and more subtle changes in levels of sedation, during the training and testing of our model. For the classification task, one-hot encoding was used based on the number of the trained states, as our target vectors. In case of regression, the propofol concentrations (ng/ml), as targeted by the TCI device or measured by the blood samples, were used as the trained regressands. In the case of blood sample measures, the average value of the two samples (taken at the beginning and ending phase of the recording) were used to represent the states of *Mild Sedation* and *Moderate Sedation*, with *Recovery* measured once at the beginning. This was in accordance with the original analysis found in (Chennu *et al.* 2016), which confirmed that the sample measures were similar at the beginning and ending phases.

5.4.1 Classification Results

The results of the classification analysis are summarized in Table 5.3. The loss and accuracy obtained by the model is shown for each of the three experiments, which progressively introduce intermediate levels of sedation during training and prediction. All models had a stable convergence during the 10 training epochs, indicated by the average categorical cross-entropy loss and accuracy curves (Fig. 5.14).

Table 5.3. Pharmacological steady-state classification results

States under Training	Loss (CCE)	Accuracy		
		Chance	***Per State***	**Total**
Baseline, Moderate Sedation	0.90	*50%*	*(84%, 83%)*	83%
Baseline, Mild Sedation, Moderate Sedation	2.49	*33%*	*(68%, 42%, 64%)*	57.6%
Baseline,Mild Sedation,Moderate Sedation,Recovery	2.79	*25%*	*(57%, 30%, 65%, 38%)*	47%

As seen from the performance of each experiment, the 2-state problem is a balanced problem which can be reasonably solved by the model (83% accuracy), in contrast to the 3-state and 4-state classification tasks, in which accuracy drops significantly due to the inclusion of *Mild Sedation* and *Recovery* (57.6% and 47% accuracy, respectively). Although the anesthetic levels attained in this experimental design were lower in comparison to the *Liege Anesthesia Dataset* analysis in section 8.3 (0.4 and 0.9 μg/ml average arterial concentrations in *Mild*/*Moderate Sedation*, over 1.9 μg/ml of the *Sedation* state in *Liege Anesthesia Dataset*), we similarly observe the difficulty in differentiating mildly-sedated states (i.e. *Mild Sedation* and *Recovery*, with *Recovery* already indicated as a mildly-sedated state in Experiment 1). These observations reveal the distinctive limitations of the model in classifying anesthetic states with small differences in levels of drug concentration. By looking at the subject-wise accuracies of each experiment, we recognized two participants (S8 and S9) that systematically showed low performance. Overall, the pattern of classification accuracies is consistent across experiments, with the model trained under all states being the most informative. The detailed cross-validation accuracies, losses, confusion matrices, and other class metrics are depicted for the 4-state model in Fig. 5.14.

The results obtained here give us a complementary picture regarding the electrophysiological nature of anesthetic states, observed previously in our behavioral classification analysis (section 5.3.1). The normalized confusion matrix (Fig. 5.14, B) shows a significant proportion of EEG epochs misclassified between *Mild Sedation* and *Moderate Sedation* or *Recovery*, as well as between *Recovery* and *Baseline Wakefulness*. This reconfirms our hypothesis that *Recovery* is similar to a mildly-sedated state, at least from an electrophysiological perspective. These class-wise accuracies also reveal the model's inability in distinguishing mild to moderate changes in drug concentrations, as appear here for pharmacologically-neighboring states. An interesting note here is that these classification errors appear more prominently in particular subjects, as we can see from the large variance across participants' performances (Fig. 5.14, A). In general, predictions tend to concentrate within the most robust classes of *Baseline Wakefulness* and *Moderate Sedation*, with the electrophysiological effects of *Mild Sedation* and *Recovery* becoming more subject-specific. While we observe that the model can learn some of the group statistics for the four anesthetic states, it is unable to identify all the individualized effects from the drug given the targeted concentration changes (0.6 μg/ml increase steps). Of course, this limitation could also reflect the lack of strong individual EEG signatures to be picked up by a model trained on targeted (rather than actual – effect-site) drug concentrations.

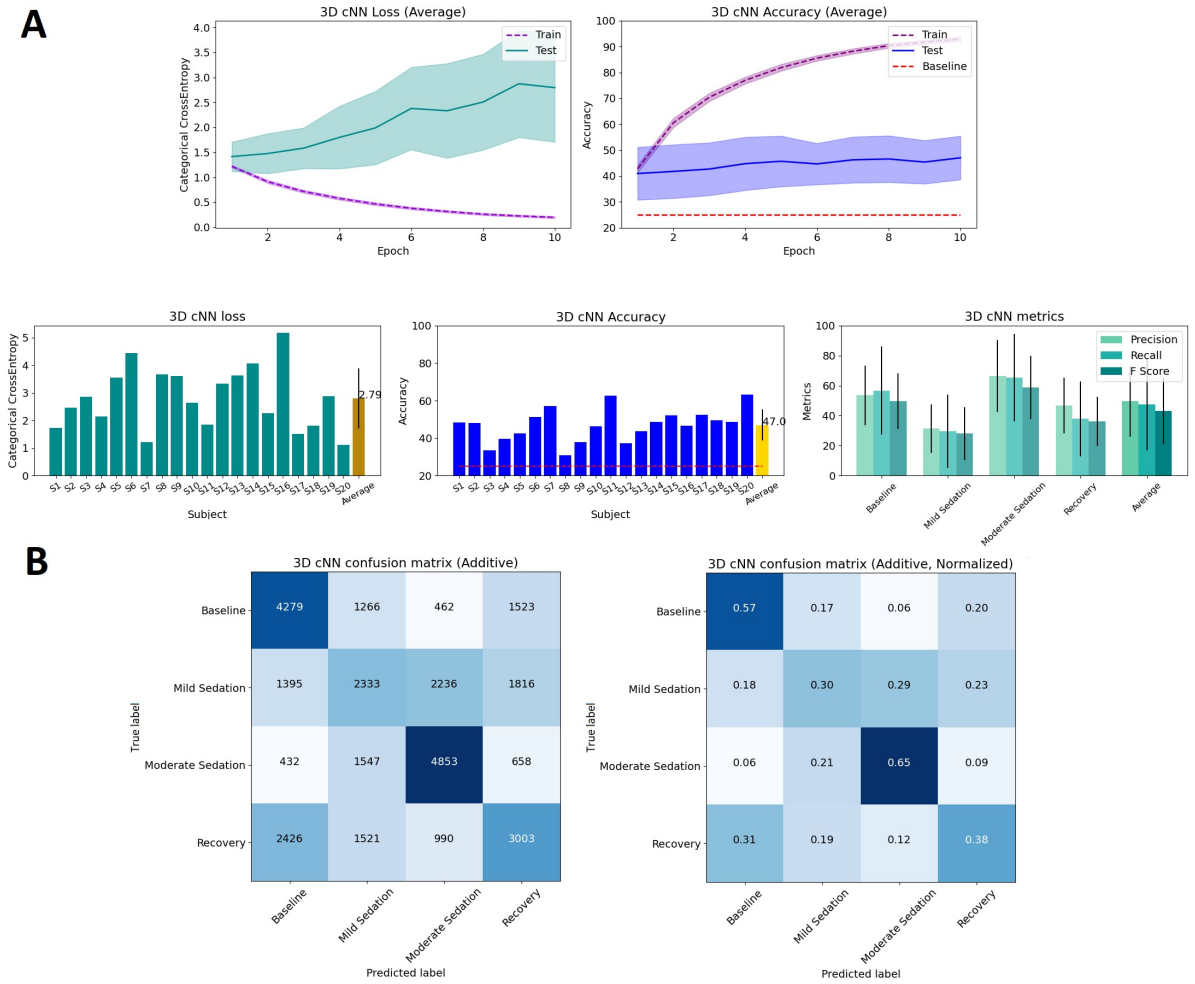

Fig. 5.14. Classification results of the *Cambridge Anesthesia Dataset*, for the four pharmacologically-steady states. A) Average categorical cross-entropy loss and accuracy curves (top). Subject-wise losses, accuracies, and other metrics (bottom). B) Additive confusion matrix (left), and normalized additive confusion matrix (right).

Some of these observations can also be seen with the visualization of the 4-class model's output over time (one prediction per 1-sec epoch). The average predictions of the softmax output, which show the probabilities for each of the four states, can be seen in detail in Fig. 5.15. We calculate the average values across subjects, by taking the minimum number of epochs per state, and aligning them at the beginning of each state.

The probabilities here reveal again the shared characteristics of *Mild Sedation* and *Recovery*, as well as the inability of the model to correctly classify mild changes in drug concentration. In addition, by looking at the temporal trajectory of the prominent probabilities, we observe relatively stable and consistent large-scale dynamics, with minor deviations, in comparison to the strong dynamics observed in our behavioral analysis (Experiment 1). These deviations are likely to reflect PK-modelling limitations in achieving real pharmacologically-

steady states (effect-site concentrations might increase over the period of *Moderate Sedation*, due to the delayed accumulation of the drug).

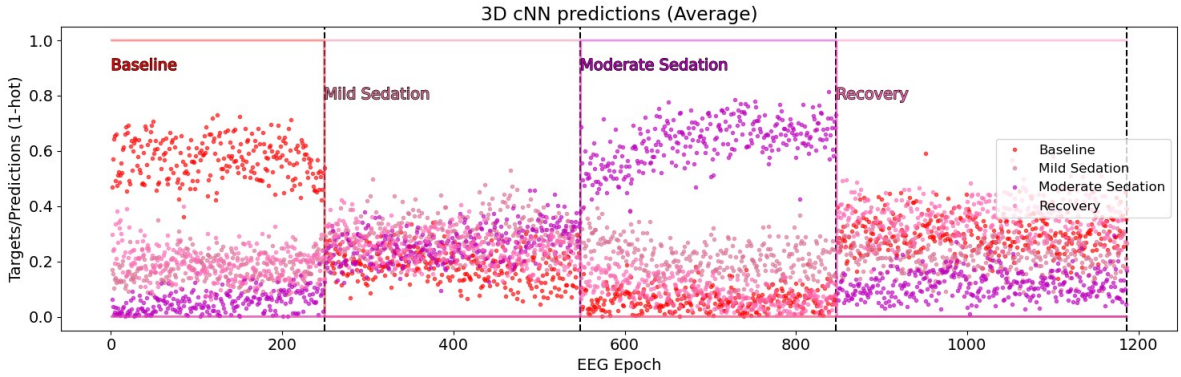

Fig. 5.15. Softmax predictions for the unseen test subjects (average), showing the probabilities of the 4 trained classes over time (1-sec epochs), for the 4 pharmacological steady-states.

5.4.2 Regression-to-Target-Concentrations Results

A complementary analysis based on the shared pharmacological ground truth was performed, by training a regression model to predict the targeted plasma concentrations (estimated by the Marsh PK model). Specifically, the concentrations used for the states of *Baseline*, *Mild Sedation*, *Moderate Sedation*, and *Recovery* were 0, 0.6, 1.2 and 0 μg/ml, respectively.

The results of the regression analysis are summarized in Table 5.4. The loss and MAE obtained by the model is shown for each of the three experiments, which introduce intermediate levels of sedation. Importantly, regression enables a continuous representation of the output (which is naturally imposed here, as drug concentrations are continuously adapted), allowing us to test the model in all states (whether included in training or used for testing), and consistently compare the experiments. Overall, all models had a stable convergence during the 10 training epochs, indicated by the MSE loss and MAE curves (Fig. 5.16).

Table 5.4. Regression-to-Target-Concentrations results

States under Training	Loss (MSE)	Mean Absolute Error (MAE)	
		Per State	**Total**
Baseline, Moderate Sedation	236324	*(295, 366, 407, 440)*	378.55
Baseline, Mild Sedation, Moderate Sedation	**204087**	***(290, 297, 389, 440)***	**355.09**
Baseline, Mild Sedation, Moderate Sedation, Recovery	210013	*(313, 328, 427, 407)*	367.45

Unseen test states are underlined. Optimal model is highlighted in bold.

Overall, the experiments showed results consistent with the previous classification analysis, while offering us more depth in terms of the model's predictive behavior with respect to propofol concentrations. Table 5.4 indicates the model trained under the 3-state task as

optimal (lowest MAE), similarly to our behavioral regression analysis. As already shown in section 5.3.2, the inclusion of *Recovery* appears to mislead model training and decreases the total performance, due to the ambiguous nature of *Recovery* and the incorrect assumption of our ground-truth. The detailed cross-validated performances and the visualization of the model's predictions can be seen for the 3-state experiment, in Fig. 5.16 and Fig. 5.17.

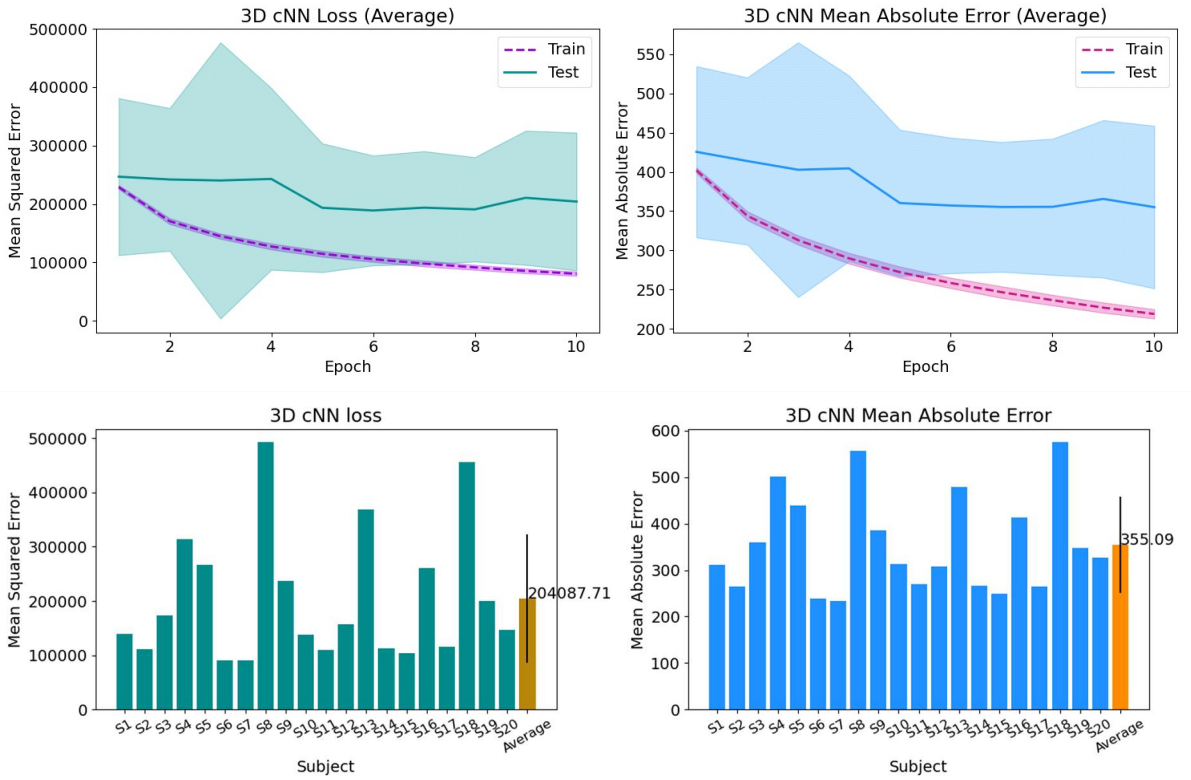

Fig. 5.16. Regression-to-Target-Concentrations results of the *Cambridge Anesthesia Dataset*, for the model trained under the 3-state task (*Baseline*, *Mild Sedation*, *Moderate Sedation*). Average mean-squared-error loss and MAE curves (top). Subject-wise losses and mean-absolute-error values (bottom).

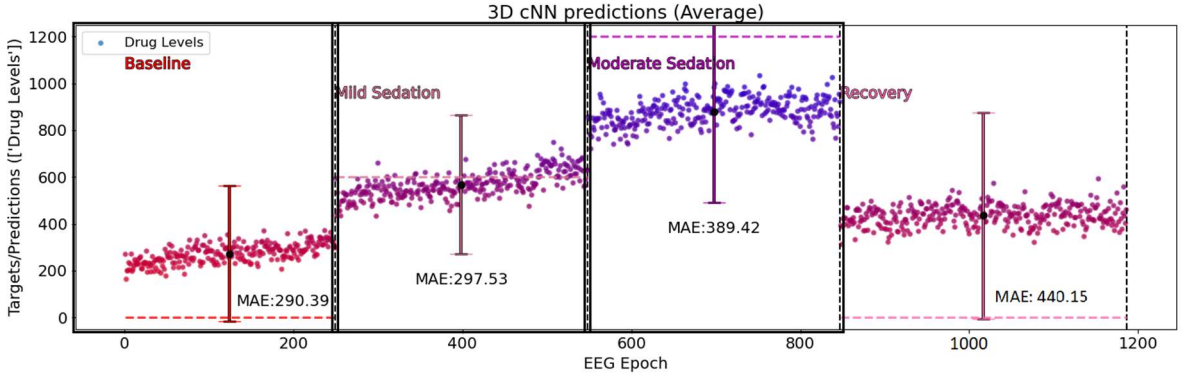

Fig. 5.17. Target concentration predictions for the four anesthetic states of the unseen test subjects (average). Horizontal dashed lines indicate the TCI device's ground-truth. States under training are highlighted in black boxes.

As we see from Fig. 5.16, a large variability of performances across participants remains here, with two subjects (S4 and S8) acquiring consistently high MAE (S4, S8 and S18 had an MAE > 500). The average-across-participants predictions (calculated by the minimum number of epochs per state) reveal an estimation bias for 3 out of 4 states (Fig. 5.17), which is independent of the state length and target (due to sample weighting), and reflects the inability of the model to strongly differentiate the selected concentration changes. Once again, we see *Recovery* predictions fitting between the levels of *Baseline* (*Wakefulness*) and *Mild Sedation* states, pointing to its shared electrophysiological characteristics. This is also confirmed by the presence of residual propofol, as measured from the blood samples at the beginning of *Recovery* (290 ng/ml, average), and which contrasts our presumed PK ground-truth (concentration should approach 0, 20 minutes after the cessation of infusion).

While predictions appear more cohesive in comparison to our classification task (both across states and in terms of temporal trajectories), individual and group level statistics remain noisy. This is evident from the per-state MAEs (Fig. 5.17), which show the significant variance in individual responses. Interestingly, and despite this variation, model predictions reveal a shorter range of concentrations in comparison to our current ground-truth (target concentrations), which shows a limitation coming either from the data or/and the model. In this regard, a ground-truth limitation can be consistent with these results, given the differences between the targeted concentrations and the average arterial concentrations, measured from the blood samples (see Table 5.5). Notably, while the Marsh model is optimally used in plasma targeting mode (Absalom *et al.* 2009), literature has suggested a high predictive PK error to be introduced in comparison to other models (Glen and White 2014). These observations lead us naturally to the analysis of the next section, where we trained our model to predict actual concentrations of propofol measured in the blood.

Table 5.5. Targeted Concentrations vs Blood Sample Concentrations

State	Target Concentrations	Blood Sample Concentrations (Average)
Baseline	0 ng/ml	0 ng/ml
Mild Sedation	600 ng/ml	446 ng/ml
Moderate Sedation	1200 ng/ml	900 ng/ml
Recovery	0 ng/ml	290 ng/ml

5.4.3 Regression-to-Blood-Sample-Concentrations Results

In the previous section, we showed that the trained model had significant variation and noise in its predictions, with respect to our 1st ground truth (Marsh model estimations of the targeted plasma concentrations). In this section, we investigate our 2nd pharmacological ground truth, consisting of propofol concentrations measured from the blood samples of the participants (see Appendix A1). Although plasma targeting may not be proportional to changes in EEG and clinical effect, measures from blood samples can reflect some of the PK interplay between the drug and the effect-site, which is the brain (currently, there are no methods to

directly measure the effect-site concentrations of intravenous anesthetics in the brain (Absalom *et al.* 2009)). However, while we know from the original study that there was significant divergence in blood sample concentrations across participants, and that some participants were more susceptible to the anesthetic action than others, the two did not seem to be connected (Chennu *et al.* 2016). In this experiment, we investigate whether EEG contains information that reflects this PK/PD divergence, by training a regression model to predict the blood sample concentration measures (as a surrogate for effect-site concentrations).

The results of the regression analysis are summarized in Table 5.6. The loss and MAE obtained by the model is shown for each of the three experiments, which introduce intermediate levels of sedation. Again, we were able to test the model in all states (trained and unseen) and consistently compare the experiments. Overall, all models had a stable convergence during the 10 training epochs, indicated by the MSE loss and MAE curves (Fig. 5.18). Target weighting was dismissed, as we had discrete targets for each state and each participant.

Table 5.6. Regression-to-Blood-Sample-Concentrations Results

States under Training	Loss (MSE)	Mean Absolute Error (MAE)	
		Per State	**Total**
Baseline, Moderate Sedation	102635	*(207, 243, 360, 183)*	247.87
Baseline, Mild Sedation, Moderate Sedation	82229	*(227, 199, 317, 148)*	222.92
Baseline, Mild Sedation, Moderate Sedation, Recovery	**76956**	***(212, 187, 319, 120)***	**209.83**

Unseen test states are underlined. Optimal model is highlighted in bold.

As we see from Table 5.6, the total MAE is considerably lower in all experiments in comparison to the MAEs obtained in the previous section. The results are also different from our two previous regression analyses, with the 4-state task having the best performance here (total MAE of 209). In this case, the ground truth of *Recovery* was represented more accurately by the inclusion of the blood sample measure, which positively contributed to model training and overall performance (in contrast to the PK estimation or the behavioral assessment seen in Experiment 1). A large variability across individual performances is still present, with one participant showing consistently high error (S8). The detailed cross-validated performances and the visualization of the model's predictions can be seen for the 4-state task, in Fig. 5.18 and Fig. 5.19, respectively.

By comparing our findings with the 1st pharmacological analysis of the previous section, we observe similar patterns, with the individual and group levels statistics improved here (resulting in lower MAEs). This can be partly explained by the decrease in the average range of drug concentrations, but also by the fact that the EEG better reflects the inter-individual differences in effect-site concentrations (and thus to a degree, the blood sample measures). In Fig. 5.19, we see predictions for *Recovery* having the lowest MAE, with the average concentration levels fitting between *Baseline* and *Mild Sedation* (290 ng/ml). On the contrary, *Moderate Sedation* had the highest MAE, as drug concentrations had the highest variance in that state, due to the individualized PK/PD impact of propofol (reported in detail in (Chennu *et al.* 2016)).

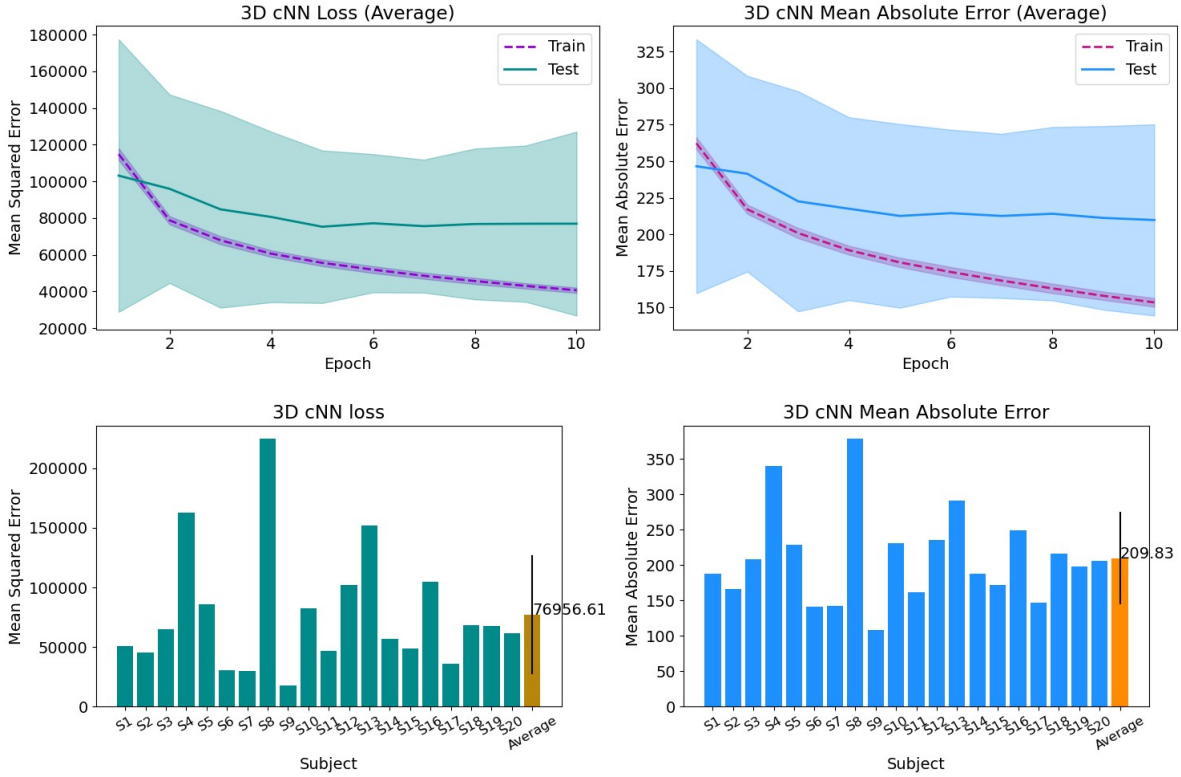

Fig. 5.18. Regression-to-Blood-Sample-Concentrations results of the *Cambridge Anesthesia Dataset*, for the model trained under the 4-state task (*Baseline*, *Mild Sedation*, *Moderate Sedation, Recovery*). Average mean-squared-error loss and MAE curves (top). Subject-wise losses and MAE values (bottom).

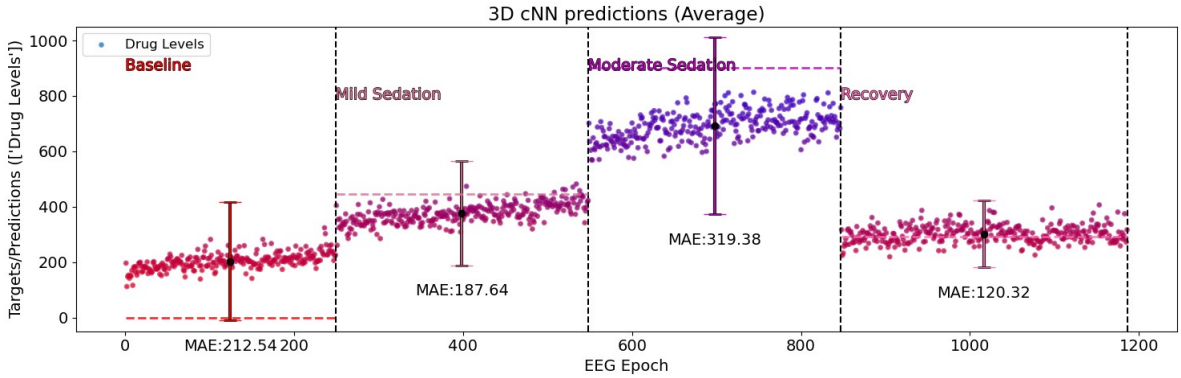

Fig. 5.19. Blood-sample-concentration predictions for the four anesthetic states of the unseen test subjects (average). Horizontal dashed lines indicate the average blood sample concentrations.

Given that these metrics reflect the average-across-participants performance, and we want to understand the predictive ability of our model in individual cases, we compare our predictions with the respective (unique) blood sample concentrations, for each state and participant independently. As previously observed, the variance of the individual predictions constrains an objective assessment on the error of concentration changes across states.

However, we can observe the relative-to-our-ground-truth estimations within each state of sedation (*Mild Sedation*, *Moderate Sedation* and *Recovery*), by computing an average-concentration prediction. Fig. 5.20 shows the blood sample concentrations and average-concentration predictions, for each subject and state. Based on the coefficient of determination (R^2), we can infer the limited capacity of the model for individual predictability (as values close to zero indicate an average-based performance).

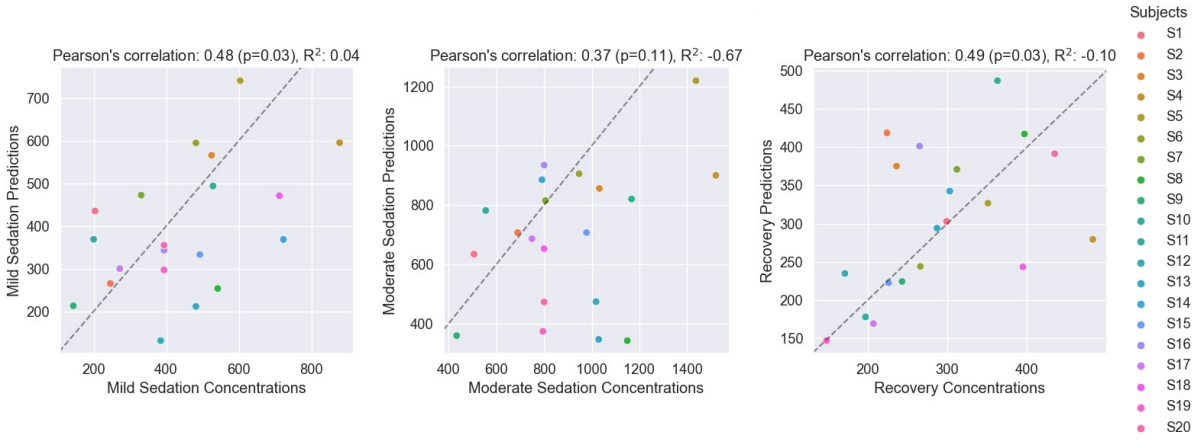

Fig. 5.20. Blood sample concentrations (ground truth) and model predictions (average), for each subject and state of sedation (*Mild Sedation*, *Moderate Sedation* and *Recovery*)

Overall, the assessment of pharmacologically-defined predictive models, as the ones we developed throughout this experiment, requires the consideration of many independent factors (e.g. related to the anesthetic procedure, such as the anesthetic doses used and the population statistics), especially if we want to compare them with other current models of clinical practice (the standardized PK/PD models). Nevertheless, our results still offer us significant insights on the understanding of the electrophysiology of anesthesia, by incorporating all the parameters and observations from our behavioral and pharmacological experimental designs.

5.5 Experiment 3 – Psychometrically-defined Anesthetic States

In this set of experiments, we investigate the effect of learning task and ground-truth in the predictive behavior and performance of our model, when trained and tested under psychometrically-defined anesthetic states. Specifically, we explore what the model reveals about the EEG signatures of states characterized by levels of disconnected consciousness, and we test the ability of deep learning to adequately generalize across participants.

As we have mentioned in the Introduction, general anesthesia (GA) is more complex than simply a state of unconsciousness, with different brain mechanisms and elements of consciousness retained or suppressed, depending on the anesthetic agent and its concentration levels. One of the main states produced under GA, is a state of disconnected consciousness (a

dream-like state, without awareness of the environment), which is typically assessed by explicit reports after recovery from anesthesia. Understanding the electrophysiological changes during disconnected consciousness can have significant implications for both research and clinical applications. Theoretically, the detection and contrast of altered states of consciousness with distinct phenomenological features, coming from different agents and study paradigms, can contribute to the identification of the full NCC (as the functional elements that characterize these states are unknown). From a clinical perspective, detecting covert awareness in unresponsive patients is crucial, given the recent reports of higher incidences of connected and disconnected consciousness during surgical procedures (Mashour and Avidan 2015). As previously pointed out, contemporary monitors like the BIS are not sensitive enough to distinguish among these different states. Hence, it is important to evaluate whether EEG has any bearing in decoding disconnected consciousness online.

For this investigation, we used the *Michigan Anesthesia Dataset* that enables us to analyze psychometrically-defined anesthetic states, based on the self-report measures from the altered states of consciousness (ASC) questionnaires (during and after the study). Contrary to our two previous datasets, the anesthetic states here were not determined upon reaching and sustaining specific drug concentrations (based on behavioral assessment or plasma targeting, as we saw in Experiments 1 and 2), but rather guided by different dosing strategies (continuous infusion during *Sedation*, and bolus dose during *LOBR* – see section 5.2.1). Nevertheless, here we focus on the characterization of the altered nature of consciousness states under ketamine, which was evident in all participants by the psychometrics of the study (ketamine is one of the most used agents that typically produce feelings of disconnection from the body and the environment). Of course, we cannot presume any state as psychometrically-steady, given that a retrospective assessment does not allow for the characterization of specific temporal events.

Using our 3D cNN model, we conducted classification and regression experiments with different sets of anesthetic states, depending on the nature of the task. Initially, we performed classification with an increasing number of steady-states (defined in section 5.2.1), and we tested the model on transitional states, in order to explore the electrophysiological distinction across the dose-dependent ketamine-induced brain changes (similarly to our previous analyses). In case of regression, we used the steady-states of *Wakefulness* and *Sedation*, along with the psychometrics recorded during and after the study, in order to create measures of disconnected consciousness (DC) that were used as the trained regressands. The choice of the *Sedation* state as the main focus of our analysis was based on the fact that the ASC questionnaires took place immediately after the sub-anesthetic period (which was also the longest). Hence, a measure of disconnected consciousness was associated to the pharmacologically-steady period that preceded these reports (defined here as *Sedation*). Specifically, we used the average scores of the three subscales recognized as most reliable and relevant (see section 5.2.1), to create three different DC measures for each participant, namely: the *study scores* (assigned to *Sedation*), the *lifetime + study scores* (assigned to *Wakefulness* and *Sedation*, respectively) and the *relative change scores* (assigned to *Sedation*).

5.5.1 Classification Results

The results of the classification analysis are summarized in Table 5.7. The loss and accuracy obtained by the model is shown for each of the three experiments, which progressively introduce intermediate levels of sedation. All models had a stable convergence during the 10 training epochs, indicated by the average categorical cross-entropy and accuracy curves (Fig. 5.21).

Table 5.7. Ketamine steady-state classification results

States under Training	Loss (CCE)	Accuracy		
		Chance	***Per State***	**Total**
Wakefulness, LOBR	0.30	*50%*	*(97%, 90%)*	94.6%
Wakefulness, Sedation, LOBR	1.06	*33%*	*(86%, 69%, 83%)*	77.6%
Wakefulness, Sedation, LOBR, Recovery	1.89	*25%*	*(74%, 54%, 72%, 64%)*	64.2%

As shown in Table 5.7, the results obtained here closely resemble what we have already seen in propofol anesthesia (section 5.3.1), albeit there is a significant decrease in performance for this dataset. Similarly to propofol, the task of discriminating the two most behaviorally-distinct states of *Wakefulness* and *LOBR* is an easily solved problem, with the total accuracy reaching 94.6% (cross-subject validation). The 3-state task is a more balanced problem, due to the difficulty in differentiating the intermediate state of *Sedation* (particularly for S11 and S12), resulting in a drop of accuracy at 77.6%. Finally, the 4-state task is the most challenging due to the inclusion of *Recovery* (64.2% accuracy), which again shows characteristics of both *Wakefulness* and *Sedation*. Importantly, all subjects in all experiments showed a performance significantly higher than chance levels. In Fig. 5.21, we depict the detailed cross-validation accuracies, losses, confusion matrices, and other class metrics, for the model trained under the 4-state task, which is the most informative.

The detailed cross-validated accuracies show a large variance across participants' performance (Fig. 5.21, A), with 3 out of 15 subjects acquiring a low accuracy in the range of 40-50% (S4, S5, and S15). This performance decrease (and the performance of the 3-state task, respectively) can be better understood by looking at the confusion matrices (Fig 5.21, B). Specifically, the normalized confusion matrix shows a proportion of the EEG during *Sedation* misclassified as *Wakefulness*, as well as a proportion of EEG misclassified between *Recovery* and *Sedation* or (to a lesser extent) *Wakefulness*. While these biases are more evident in particular subjects, they reveal the electrophysiological similarities found across wakefulness and low-dose ketamine states (state lengths or target imbalances do not explain these biases, which have been corrected by sample/target weighting). Considering *Recovery*, we have already recognized its shared characteristics with mildly-sedated states during propofol anesthesia (in Experiments 1 and 2), with similar findings observed here for ketamine.

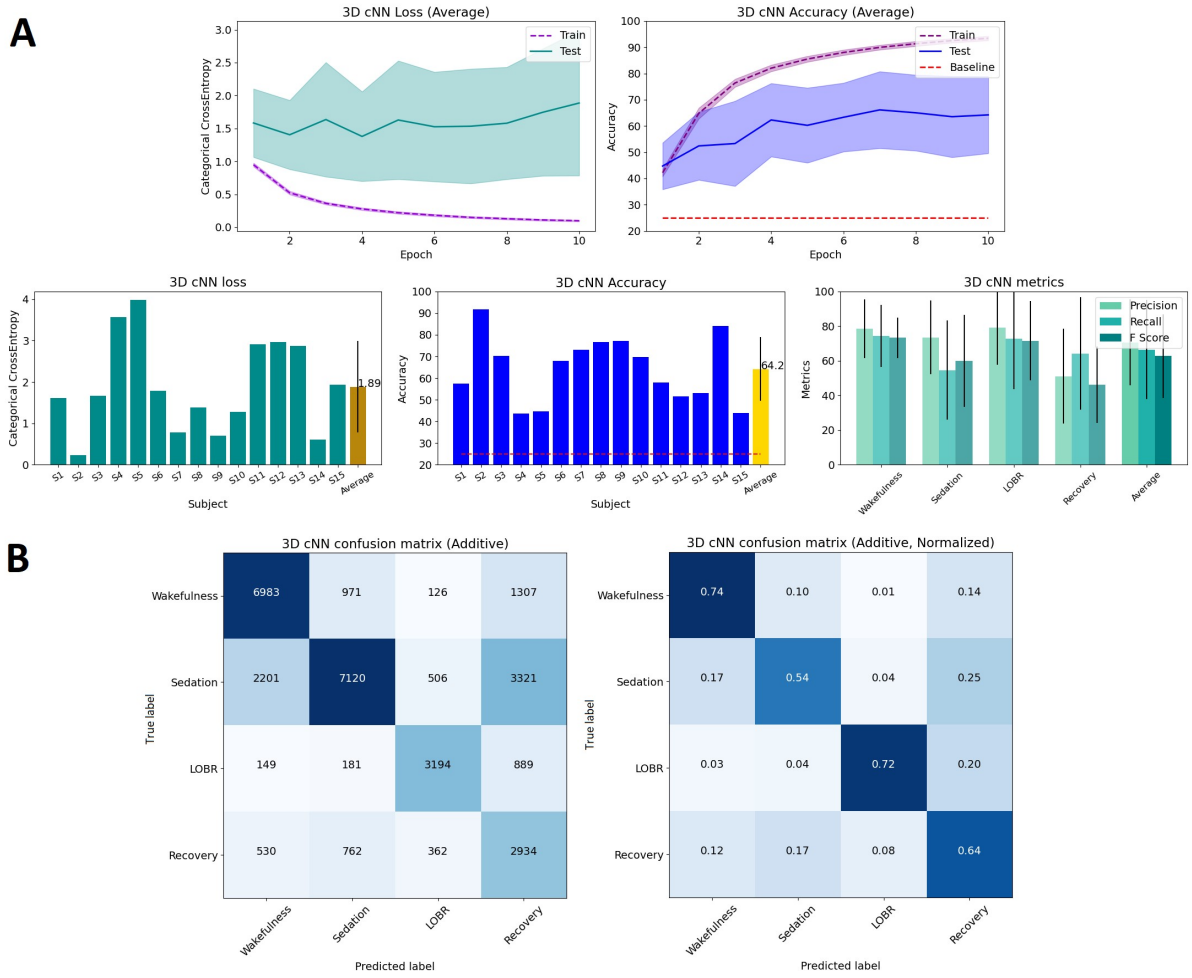

Fig. 5.21. Classification results of the *Michigan Anesthesia Dataset*, for the four steady-state task. A) Average categorical cross-entropy loss and accuracy curves (top). Subject-wise losses, accuracies, and other metrics (bottom). B) Additive confusion matrix (left), and normalized additive confusion matric (right).

To better understand the temporal dynamics of the model, we visualized the model's output over time (one prediction per 1-sec epoch). The average predictions of the softmax output, which show the probabilities for each of the four states, can be seen in detail in Fig. 5.22 (average values are calculated across participants, by taking the minimum number of epochs per state, and aligning them at the beginning of each state). By focusing on the prominent probabilities of each state, we see that the temporal dynamics of the model are overall consistent, with a weak indication of a rapid rise and fall during *LOBR* (due to the bolus dosing strategy).

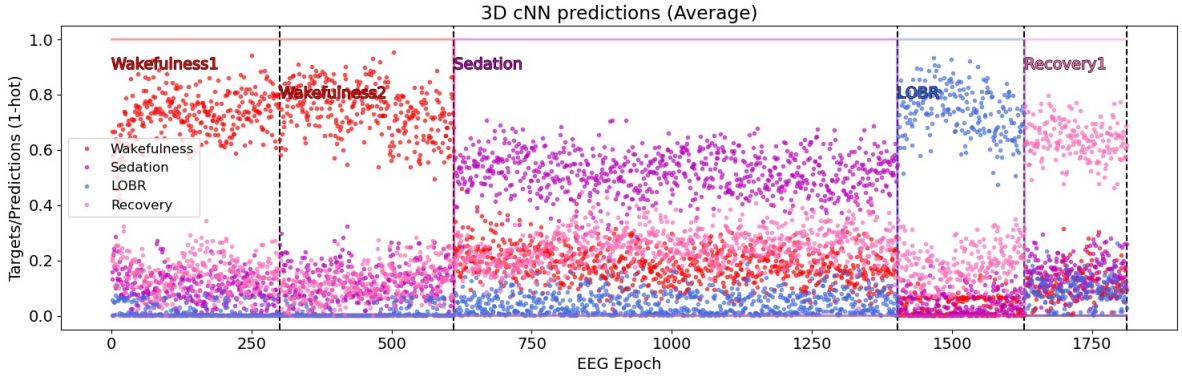

Fig. 5.22. Softmax predictions for the unseen test subjects (average), showing the probabilities of the four trained classes over time (1-sec epoch), for the four ketamine steady-states.

Testing on Transitional States. In order to further test the above observations, we used the model trained in all steady-states (4-state classification task) and visualized its predictions over the recording period that included both steady and transitional states. This was possible for 14 out of 15 participants, for which we had access to all intermediate EEG recordings (*Wakefulness* to *Sedation – WS*, *Post Sedation – PS*, *Sedation* to *LOBR – SL* and *LOBR* to *Recovery – LR*). Given the continuity of the recordings (following the anesthetic paradigm in Fig. 5.3), we concatenated the states and performed averaging, as previously described (Fig. 5.23).

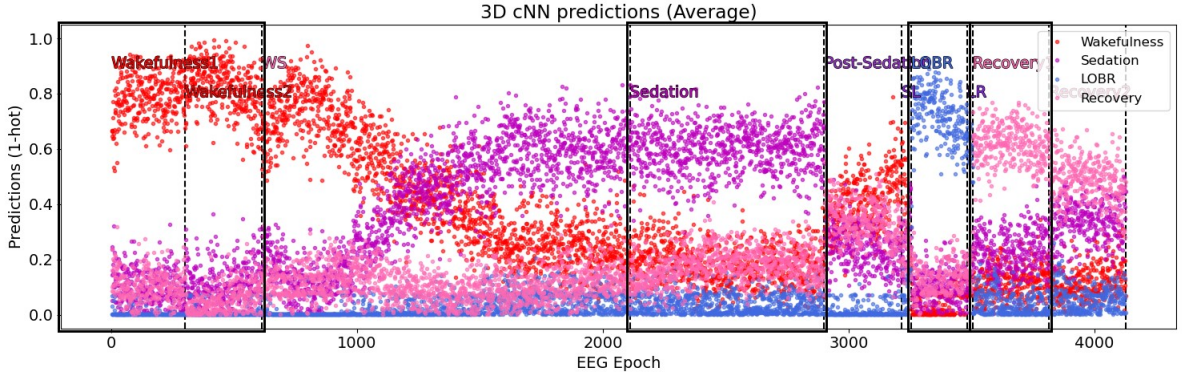

Fig. 5.23. Softmax predictions for the unseen test subjects (average), showing the probabilities of the four trained classes over time (1-sec epochs), for all ketamine states (steady and transitional states).

Fig. 5.23 shows the general trend of transitions from *Wakefulness* to *Sedation*, and from *Post-Sedation* recovery to *LOBR* and back, which consistently tracks the anesthetic paradigm of the study. The sub-anesthetic period (*WS* + *Sedation)* was long enough (~40 minutes) for the large-scale dynamics to be captured by the EEG and our model, as depicted by the decreasing and increasing probabilities of the *Wakefulness* and *Sedation* classes. Even though only the last block of the sub-anesthetic period was considered as pharmacologically steady (a

rather conservative estimate of the original study, based on which we defined the steady-state *Sedation*), the model showed a predictive convergence several minutes before reaching *Sedation* (ketamine concentration was approximately 180 ng/ml at the end of the state). In contrast, the transitions to, during, and from *LOBR* were too short for the detection of transitional dynamics with high granularity. Overall, these data confirm the existence of large-scale temporal dynamics over our ketamine dataset, with the implications and advantages of regression, discussed previously in Experiment 1.

As a final step to this analysis, we visualize the predictions and the corresponding time-frequency representation (spectrogram) of the epochs over all anesthetic states (steady and transitional), for a single test subject (S6) which was representative both in terms of EEG signatures and model performance (median performance – 67% accuracy). Fig. 5.24 shows that the prominent probabilities of each state follow the general trajectory of the softmax predictions observed previously for the group (Fig. 5.23), with specific peculiarities that are subject-specific, and which are described below.

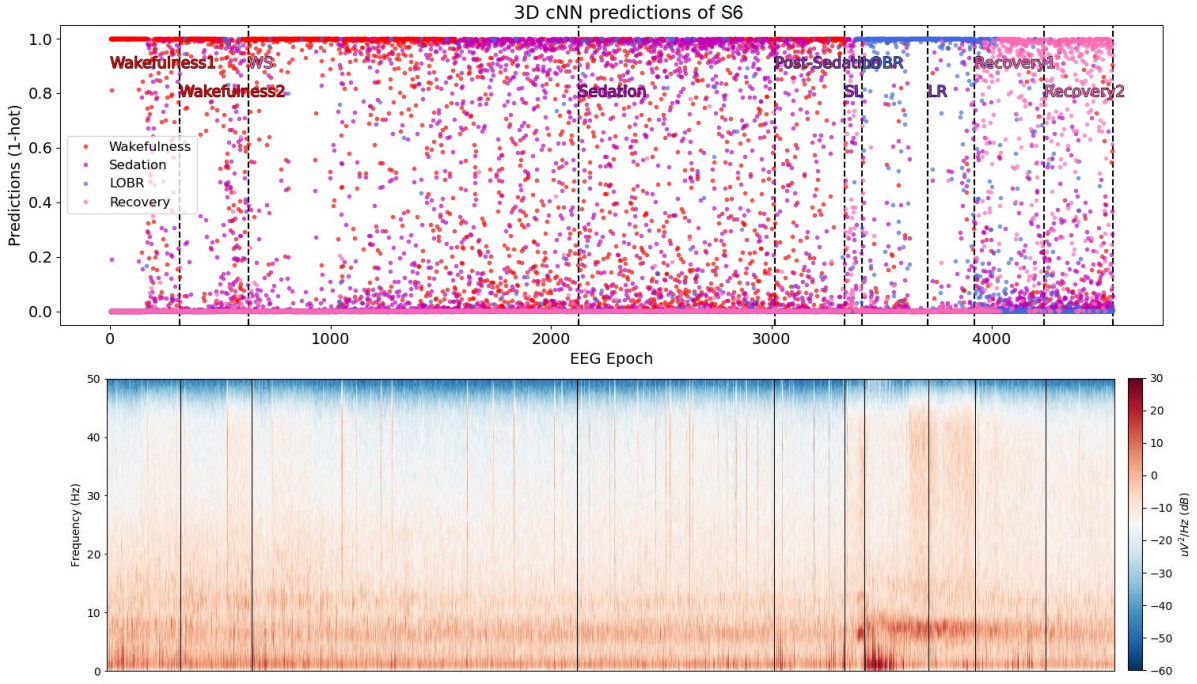

Fig 5.24. (Top) Softmax predictions over all anesthetic states (steady and transitional) for a median-performance subject. (Bottom) PSD of the corresponding states (Method=Welch, channel_aggregation=mean, window=1 sec, n_fft=256).

At first sight, visual inspection of the mean spectrogram does not reveal any clear associations between the band-power dynamics of the EEG and the dose-dependent ketamine states. In the broader context, ketamine is an agent with unique features across all levels of analysis (molecular, neural, and behavioral), which is also reflected in its distinct

electrophysiological characteristics from other intravenous or inhaled anesthetics. Specifically, several studies have reported changes in the band-power dynamics, which generally include a decrease of alpha activity during the sub-anesthetic state of *Sedation*, an increase in delta and gamma power during *LOBR*, as well as a general theta power increase throughout ketamine anesthesia (most evidently during *LOBR*) (Vlisides *et al.* 2017). Some of these changes appear to be shared with propofol, as previously observed in section 5.3.2 (e.g. the decrease of alpha activity during *Sedation*, and the increase of delta during *LOBR*). Regarding the presence of disconnected consciousness, alpha activity has previously been correlated inversely with the experience of disembodiment (and has been suggested to play a role in conscious orientation to time and space (Vlisides *et al.* 2018)). Nevertheless, only the delta and gamma signatures can be seen in S6, and not consistently throughout the *LOBR* state. Moreover, while we have associated the experience of DC to the steady-state of *Sedation* (which preceded the reports), a significant decrease of alpha activity is not evident from the figure.

One of the subject-specific peculiarities we observe in Fig. 5.24, is the intermittent prediction of *LOBR* during *Sedation* (with high confidence/probability ~1), which also seems to correlate with the short-lived periods of delta, beta, and gamma activity observed in the spectrogram. This could indicate that loss of responsiveness, or even unconsciousness, may happen periodically for short periods during ketamine, without our ability to assess these phenomenological dynamics. Of course, to better understand these signatures and their relation to phenomenology, would require further analysis and ground truth testing. Overall, the distinction between states of connected and disconnected consciousness is difficult to make, due to their electrophysiological similarities, already demonstrated in the literature (discussed in detail section 5.6).

5.5.2 Regression-to-Psychometric-Score Results

In this section, we focus on the estimation of specific levels of disconnected consciousness (DC), by regression to measures created from the average scores (levels) of the three subscales (*disembodiment*, *transcendence of time and space*, and *complex imagery*), identified most consistently within and across participants. These scores were based on the *study* and *lifetime* questionnaires (see section 5.2.1), and represented individual levels of DC for each participant (overall, study scores were higher than lifetime history scores; see Appendix A2). As briefly mentioned in the experiment description, we created three different ground-truth measures, namely: 1) the *study score*s, 2) the *lifetime* + *study scores*, and 3) the *relative change scores*.

Given that the *study scores* were recorded after the subanesthetic period of the experiment, and the *lifetime scores* were recorded within 48 hours after the study, we associated each group of scores to the steady states of *Sedation* and *Wakefulness*, respectively. The state of *Sedation* was chosen as the pharmacologically-steady period that preceded participants' self-reports during the study. *Wakefulness* was associated to *lifetime scores*, as these scores

represented an individual baseline for disconnected consciousness that each participant reported to have in their daily life. Based on this association, the 1st ground-truth measure (*study scores*) used only the study scores as a DC measure during *Sedation* (*0* was assigned to *Wakefulness*), the 2nd ground-truth measure (*lifetime + study scores*) used both *lifetime scores* and *study scores* for *Wakefulness* and *Sedation*, and the 3rd ground-truth measure (*relative-change scores*) used a relative change score (*study scores – lifetime scores*) for *Sedation* (*0* was assigned to *Wakefulness*). The relative change score was created to assess the subjective change of the individuals' DC scores, as a way to normalise inter-individual variability within the participants' self-reports. Given the difficulty in assessing a measure of disconnected consciousness in the brain, we trained separate models for all three ground-truth measures, and explored whether the EEG signatures predicted any of these, using a cross-subject generalization approach.

The results of the regression analysis are summarized in Table 5.8. The loss and MAE obtained by the 3D cNN is shown for each of the three experiments, which introduce a given ground-truth for levels of DC. Overall, all models had a stable convergence during the 10 training epochs, indicated by the MSE loss and MAE curves (Fig. 5.25). Target weighting was dismissed, as we had discrete targets (scores) for each participant.

Table 5.8. Regression-to-Psychometric-Score results

Psychometric Ground-Truth	Loss (MSE)	Mean Absolute Error (MAE)	
		Per State	**Total**
Wakefulness (0), Sedation (Study Score)	0.12	*(0.14, 0.36)*	0.27
Wakefulness (Lifetime Score), Sedation (Study Score)	0.09	*(0.15, 0.30)*	0.24
Wakefulness (0), Sedation (Relative-Change Score)	**0.08**	***(0.13, 0.26)***	**0.21**

Model with lowest MAE is highlighted in bold.

Table 5.8 shows a total MAE ranging from 0.2 to 0.3 (within the scoring scale), which is hard to evaluate and compare across the experiments, due to the variable nature of the ground-truth measures. By comparing the MAE values between the lowest-error model in the 3rd experiment, and the models in the 1st and 2nd experiments, we did not find any statistically significant difference in performance (with $p=0.07$ and $p=0.27$, based on *two-sided t-tests*). Focusing on the per-state MAEs, we observe that the errors stem mainly from the state of *Sedation*, where DC levels had the largest variation in the group. This variation is also likely responsible for the large variability across individual performances. Specifically, for 3 out of 15 subjects (S9, S11, and S12), we consistently acquired a low performance with MAE > 0.3. Given the similarity of results across the various ground-truth measures, we depict the task trained under the *relative-change scores*, which had the lowest MAE. The detailed cross-validated performances and the visualization of the model's predictions can be seen in Fig. 5.25 and Fig. 5.26, respectively.

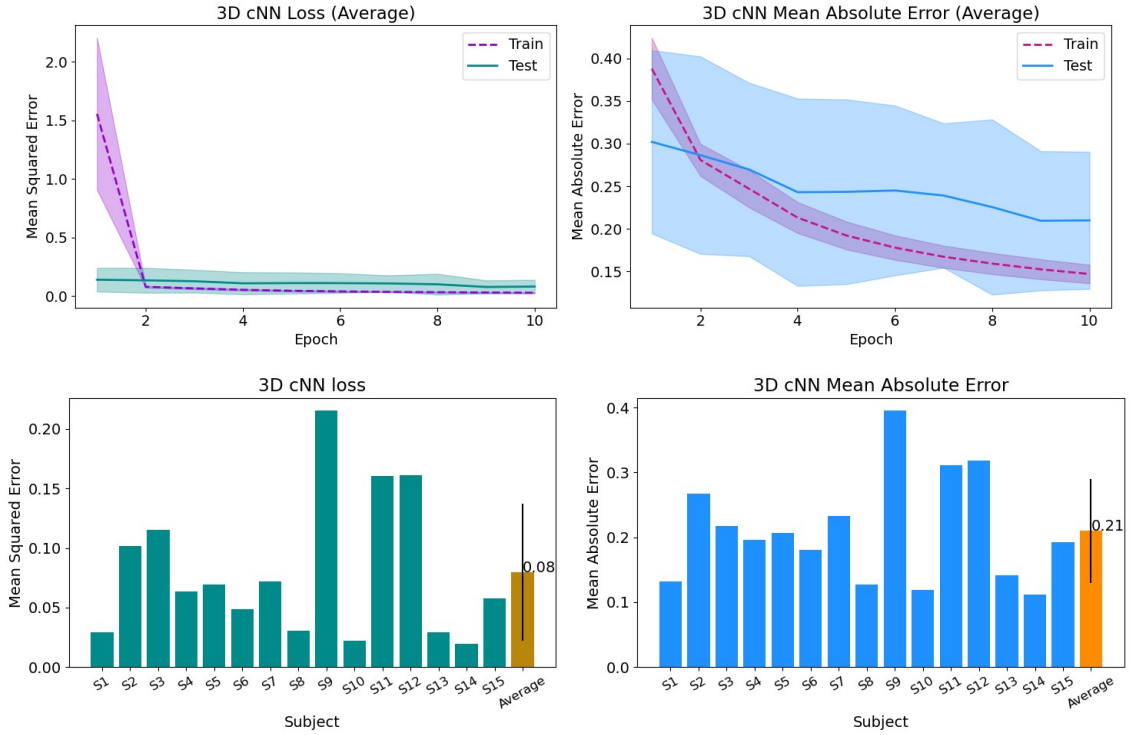

Fig. 5.25. Regression-to-Psychometric-Score results of the *Michigan Anesthesia Dataset*, for the model trained under the *relative-change score* ground-truth. Average mean-squared-error loss and MAE curves (top). Subject-wise losses and MAE values (bottom).

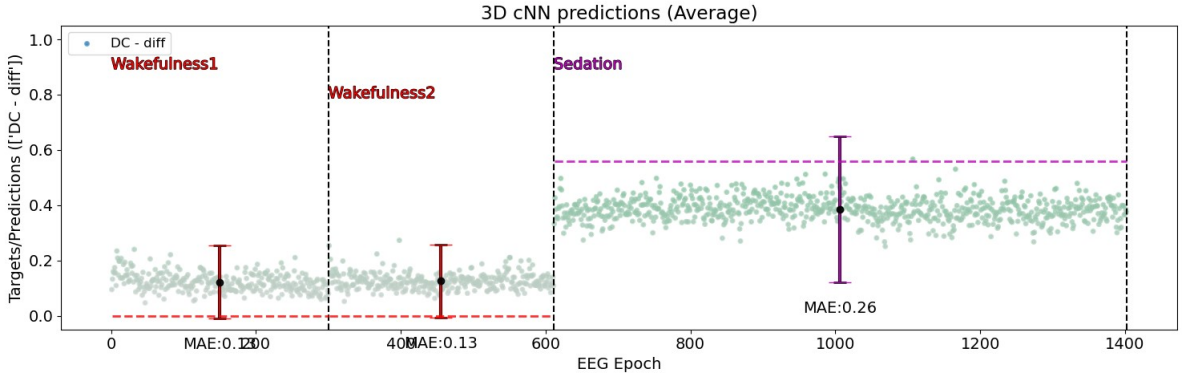

Fig. 5.26. Psychometric-score (DC) predictions for the states of *Wakefulness* and *Sedation* of the unseen test subjects (average). Dashed lines indicate the average score of each state (*relative-change scores*)

Due to the significant variation and noise of the model within *Sedation* (Fig. 5.26), and in order to further understand its behavior in individual cases (rather than statistically across the group), we compared model predictions with the respective (unique) DC scores, for each participant independently. As previously performed (section 5.4.3), we can observe the relative-to-our-ground-truth estimations, by computing an average-score prediction within the state of *Sedation*. Fig. 5.27 shows the models' predictions and the DC values for each subject,

under all three psychometric ground-truths. The relational plots show no correlation between the two values, which indicates the inability of our model to make individualized predictions of psychometrically-defined levels of DC, based on the EEG.

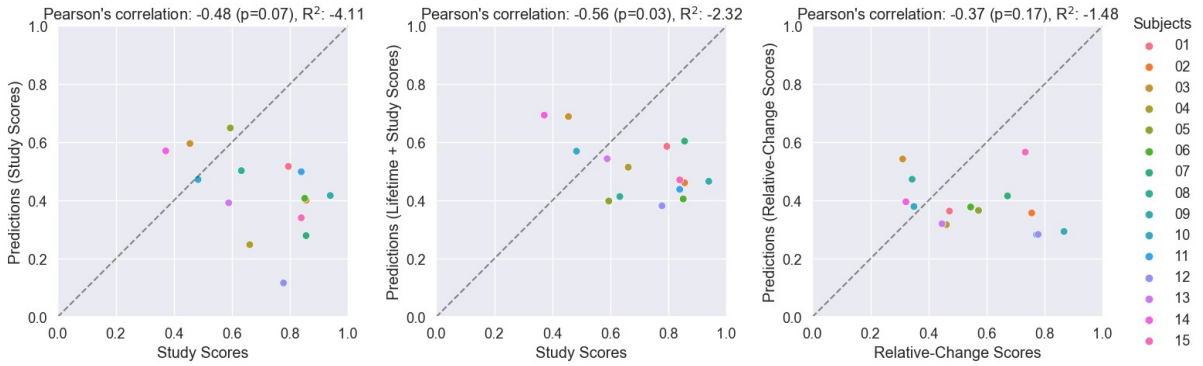

Fig. 5.27. Disconnected consciousness measures (ground truth) and model predictions (average), for each subject and experiment (predictions are based on the *Sedation* state).

Overall, the task of investigating and estimating states characterized by disconnected consciousness is a hard problem, which already appears with the definition and formation of a ground-truth for altered states of consciousness (especially for measuring specific levels of DC, as attempted here). Of course, such task reflects a number of limitations exhibited by our current tools of investigation, which mainly reflect the assessment of covert awareness through post-anesthetic explicit reports (which on the one hand require recall, and on the other cannot indicate the timing of specific events), as well as the comparison of (subjective) phenomenological experiences across participants, using questionnaires. In the next section, we discuss in detail all the findings and insights, gathered from our previous experiments.

5.6 Discussion

Our experiments focused on the importance of the optimization task undertaken by our deep learning model, for the purpose of predictive analysis of various anesthetic states. The first aspect of optimization regards the employment of a selected ground-truth, defined by a clinical standard. Behavioral, pharmacological, and psychometrical evidence are all used in medicine for the assessment of states of consciousness, as we also find during the different stages of anesthesia (e.g. behavioral measures are used at the induction and recovery of anesthesia, while pharmacological measures are mostly used during its maintenance). While any anesthetic agent and dose can be characterized by these measures, our analysis relied on the data and tasks allowed by the selected studies/experimental designs, in order to make

inferences and compare findings. The choice of propofol (Experiments 1 and 2) and ketamine (Experiment 3) can be representative for our research goal, given the distinct properties of the two drugs (pharmacologically, electrophysiologically, and phenomenologically), which synergistically contribute to the identification of the full NCC. Moreover, both anesthetic agents are often used in GA, which enhances the clinical relevance of our findings.

The second aspect of optimization, which relates to the learning algorithm and ground-truth encoding, has been shown to strongly affect the performance of machine learning models (Korotcov *et al.* 2017), and particularly in EEG analysis has been shown to determine the outcomes of feature learning (Stober *et al.* 2015). Of course, while the question of features and learning objective can be guided by the posed research problem, we have shown throughout the experiments that the trained models can enrich or dismiss information relevant to our investigation, depending on the optimization strategy. We have also shown that the existence of large-scale temporal dynamics (i.e. changes in EEG signatures over the period of several epochs) affect the model's predictive behavior under classification and regression tasks, as hypothesized, with findings that sometimes agree with our clinical assumptions, and sometimes disagree, are novel, and up to interpretation (e.g. indications for continuous or discrete transitions of levels of consciousness).

Given that our model training and evaluation methodology was kept consistent across experiments (without effects from other processing hyperparameters), we were also able to observe differences related to the nature of the data, such as the drug under analysis, the dataset sizes, and the various anesthetic states and depths. While some of these factors differed significantly among the datasets used (e.g. state lengths varied within and across studies, most evidently in the case of *Michigan Anesthesia Dataset*), sample and target weighting managed to correct a number of biases observed in our experimental results. Besides these factors, all findings were evaluated upon 1) the cross-subject generalizability of the model, due to the susceptibilities and possible biases that arise from EEG and neural networks (discussed further in section 3.4.1), and 2) the interpretability of model predictions, or their generalization to known facts. Finally, we focused on some of the model's failures (statistically or in specific subjects), rather than the actual performance metrics, which are more important to understand the behavior of the models. Below we discuss some of the most relevant findings.

5.6.1 Recovery as a State of Mild Sedation

One of the observations that were repeatedly made throughout the experiments, was the electrophysiological similarities of *Recovery* with states of mild sedation, typically appearing during the induction to anesthesia. With respect to the literature, we have not found any similar studies on EEG analysis under GA, that explicitly investigate *Recovery* as a distinct state. Several studies include post-anesthesia recordings, which use an identical ground-truth to pre-anesthetic recordings (*Wakefulness*), either by direct labeling or by some behavioral assessment (Liu *et al.* 2015; Sun *et al.* 2019a; Dubost *et al.* 2019), and few have included states of

intermediate sedation (*Mild/Moderate Sedation*) (Lalitha and Eswaran 2007; Saadeh, Khan and Altaf 2019; Gu, Liang and Hagihira 2019). Nevertheless, transitional states of sedation can be important to understand, as the crucial events towards *LOC* and *Recovery* are of significant relevance to both research and clinical applications.

The evidence for this phenomenon were found in both classification and regression analyses. In Experiments 1 and 3, we observed a similar pattern of *Recovery* sharing signatures to both *Wakefulness* and *Sedation* during classification (as depicted by the confusion matrices and model output), for both propofol and ketamine, and an almost identical signature to the transition from wakefulness to sedation (*WS*) in *Liege Anesthesia Study* (regression-to-Ramsay-scores model also estimated the state with an intermediate Ramsay score). Comparable results were also observed during the analysis of pharmacologically-defined states in Experiment 2, by the similarities of *Recovery* to *Mild Sedation* (in both classification and regression predictions). Also, the blood sample measures of the study showed residual propofol during the beginning of *Recovery* state, contrary to PK estimations (~0.3 μg/ml). As previously discussed, blood sample measures are the best surrogate for effect-site concentrations, which give us a more reliable indication for electrophysiological change in the brain.

Most importantly, the model was able to make a distinction between *Wakefulness* and *Recovery*, which although subtle, contrasts with our assumed behavioral and pharmacological ground-truths (both states were originally determined by Ramsay 2 or an absence of drug concentrations). The performance of the models also increased or decreased accordingly, with the inclusion or exclusion of *Recovery* during training (except in the case of regression-to-blood-sample-measures, where a more reliable ground-truth was presented).

5.6.2 Large-scale Temporal Dynamics of EEG Appear Consistent with the Depth of Anesthesia

In resting-state experiments, a steady state of consciousness is commonly supposed, which is also reflected on the ground-truth and evaluation of deep learning studies. However, recent research has shown that the resting state is a rather dynamic state, and particularly in terms of changes in vigilance, which appear most prominently (Zacharias *et al.* 2020). In one resting-state EEG/fMRI study, the authors found that one third of the subjects exhibited unstable wakefulness and loss of wakefulness within 3 minutes. Of course, such changes have also been observed under various anesthetics, including propofol and ketamine, which most evidently affect levels of vigilance (the effect of sub-anesthetic doses of ketamine have also been correlated to light sleep, which in turn has been characterized by unstable dynamics and cyclic alternating patterns (CAPs) (Musso *et al.* 2011)).

Our observations based on the models' predictions over time revealed the existence of large-scale temporal dynamics - over several seconds or minutes within presumable steady states - in agreement with the recent literature. We also showed the transitional nature of EEG

under anesthesia in both classification and regression tasks, along with its connection to the anesthetic depth, throughout our experiments (most evidently in Experiments 1 and 3). For both propofol and ketamine anesthesia, the model was able to track the anesthetic paradigm found in *Liege Anesthesia Study* and *Michigan Anesthesia Study*, during steady and transitional states (especially when using regression – see Fig. 5.11). This was confirmed by two independent observations. The first one relates to the changes in drug concentrations that were known to take place (in both studies, by independent dosing strategies), and which matched model's predictions. The second observation is based on Experiment 2, where the pharmacologically-steady states showed the most consistent model predictions (a reasonable expectation, given that the EEG strongly reflects changes in effect-site concentrations). While EEG patterns can show significant variability due to various biological factors (e.g. age), the drug administration – characterized by the concentration levels and the rate of titration – showed the most direct effect on the model's predictive behavior.

A number of relevant studies, estimating states and levels of unconsciousness using machine learning, have also revealed the existence of temporal dynamics through a variety of models, using windows of analysis that span 1 to 10 seconds (Kangas *et al.* 1997; Lalitha and Eswaran 2007; Jiang *et al.* 2015; Sun *et al.* 2019a; Saadeh, Khan and Altaf 2019). However, the relation of the models' dynamics to the anesthetic depth is not always clear, with predictions that are either unstable/noisy, or lack a sensible interpretation on the basis of our current understanding (e.g. transitions from wakefulness to deep anesthesia, and vice versa, as found in (Saadeh, Khan and Altaf 2019)). In one of the most successful works, (Sun *et al.* 2019a) used a cNN model which showed large-scale temporal dynamics (over several hours) that naturally tracked the depth of anesthesia, as denoted by the Richmond Agitation-Sedation Scale (RASS). Although the model used 4-sec epoch windows, it exhibited a delay in response to the known ground-truth, with a lag of 0.6 – 6 minutes (the authors attributed this lag to the use of ordinal regression, and the addition of an LSTM network, which learned to smoothen the model's output).

In our own analysis, when looking at individual subjects, changes in model predictions appear to take place over several minutes, with transitions that sometimes span seconds. Nonetheless, the evaluation of these fine-grained dynamics can be challenging without a respective behavioral or pharmacological ground-truth. An interesting finding based on the model trained in the 1st experiment, was the predicted changes in anesthetic depth over steady-defined states, despite the use of effect-site targeting that aimed to sustain a behavioral level of consciousness. For example, the general increase of depth observed during *LOC* (Fig. 5.10) is consistent with reports of increasing depths of unresponsiveness after the moment of LOBR ((Radek *et al.* 2018) showed different degrees of arousability and probabilities of content of experiences, for participants under increasing doses of propofol, within the range of 2-3 μg/ml). Another finding which points to the validity of our model, was reflected in the decrease of depth during the *Sedation* state of our worst-performing subject (Fig. 5.13), which revealed the

order of transition from *Wakefulness* to *LOC* and back to *Sedation* (despite the 5-min equilibration period that preceded the recordings).

Overall, and based on these findings, it is important to note that besides any ground-truth and PK limitations, model's behavior can be affected by changes in brain activity that are not directly caused by concentration changes, but emerge from phenomena such as neural inertia (the process where the brain resists behavioral state transitions, as in the shift from conscious to unconscious states) and hysteresis (the asymmetry of induction and emergence from anesthesia, due to the lagging of the brain's response to changes). For example, we know that the anesthetic concentration at which consciousness is lost is higher than the concentration at which consciousness is regained, due to hysteresis (Tarnal, Vlisides and Mashour 2016). Nevertheless, we do not currently have the data to compare and evaluate such observations. It may be critical though to obtain such evidence from further experimental designs, by combining multiple targeted measures, considering the limitations of implicit and indirect evidence of individual approaches (which lack the qualitative and quantitative properties exhibited by the brain).

5.6.3 Regression vs Classification for Tracking States and Levels of Consciousness

As discussed in section 5.2.4, we investigated both main types of supervised learning algorithms throughout our experiments, given that each one reflects the discrete and continuous aspects of the variously defined anesthetic states and levels of unconsciousness. Besides any technical differences, and with respect to our research investigations, two theoretical approaches can be considered and posed as a learning objective, which ultimately determines the algorithm of choice. The first one is based on the notion of consciousness states as discrete phase transitions, which can be classified as separate neurophysiological states. The second approach, regards the consideration of discrete or continuous levels of consciousness that may share neurophysiological transitions, and which can be exploited by regression. Given that we do not have sufficient evidence for either notion (we recognize both distinct and transitional electrophysiological signatures throughout the different depths of anesthesia (Patrick L. Purdon *et al.* 2015)), we tested both approaches, allowing the model to interpret the raw data. Literature review shows a limited number of studies using regression methodologies (Jiang *et al.* 2015; Sun *et al.* 2019a), with the majority of studies using classification, on the basis of this first implicit assumption (Lalitha and Eswaran 2007; AlMeer and Abbod 2019; Liu *et al.* 2015; Dubost *et al.* 2019). In one study we are aware, (Ferreira *et al.* 2021) used classification for the problem of the detection of the exact moment of LOC.

In this work, we argue that a continuous representation by regression is more appropriate for our research investigation, which is based on clinical definitions for levels of unconsciousness (mainly estimated by behavioral or pharmacological measures). Model

predictions of individual subjects under classification did not show strong temporal consistency, and were not informative with respect to specific levels of unconsciousness. This was observed especially during the intermediate levels of sedation, where the critical events of transitioning to and from LOC take place (as we observed in section 5.3.1, by the noisy and step-like traces of softmax probabilities). On the contrary, regression was able to capture the large-scale temporal dynamics of EEG and track the anesthetic trajectory of individuals, as discussed in detail in the previous section. As initially hypothesized, the existence of such dynamics contributed to the behavior and performance of the respective algorithms. Moreover, this learning objective is in alignment with current DoA monitors, which use continuous indices to denote the various anesthetic depths (Patrick L. Purdon *et al.* 2015). Finally, regression has the advantage of a trained model that can be consistently tested in novel and unseen anesthetic states, unlike classification models.

With regards to our initial theoretical notions, the model interpretation of the data during our behavioral analysis (section 5.3) revealed specific anesthetic depths across different electrophysiological classes, but also changes in depths within a particular class (if we consider the electrophysiological classes as the prominent changes in band-power dynamics, that have been systematically reported in the literature - see Fig. 5.12). Of course, further understanding of these findings would require further interpretation of the model. Overall, our results showed the possibility of deep learning in investigating and estimating levels of consciousness, by continuous measures of research and clinical value.

5.6.4 Behavioral Measures are more Reliable than Pharmacological Measures as a Ground-truth for Consciousness

One of the main investigations across the experiments of this chapter, and the 2nd aspect of optimization and ground-truth, was the comparison among the different definitions of anesthetic states based on behavioral, pharmacological, and psychometrical evidence for consciousness. So far, all the studies that we have referenced in our discussion, and which relate to our research question, have used behavioral measures as a ground-truth for consciousness (using scales such as the Ramsay or the RASS). Even though this ground-truth is often implicitly accepted as accurate and valid by clinical standards (in case of standard clinical scales), we need to critically assess our findings, under the limitations of behavioral evidence for understanding consciousness (discussed in detail in Chapter 2). Besides behavioral or other physiological and clinical signs that have been typically used by anesthesiologists to assess the anesthetic depth, pharmacological measures are rarely taken into consideration (and thus, the pharmacological ground-truth is almost never evaluated). While we are aware of few machine learning studies that have incorporated other physiological signals alongside EEG to study anesthesia (mostly as input variables for the prediction of DoA, as in (Subramanian *et al.* 2020)), we are not aware of any study focusing on the prediction of

states characterized by drug concentrations (either TCI-estimated or arterially measured). Nevertheless, here we make several observations, particularly with respect to propofol, where we had independent access to both behavioral and pharmacological data.

Based on the analysis of Experiments 1 and 2, we can infer that behavioral measures are more reliable than pharmacological measures, as a ground-truth for consciousness. First of all, the classification analysis of the experiments showed that behaviorally-steady states had higher electrophysiological similarity across participants than pharmacologically-steady states, as defined by plasma targeting (using the Marsh model). This was evident from the performance of the respective models, and particularly on the states of *Sedation* and *Moderate Sedation* of the corresponding studies (91% over 83% accuracy), which were closest by anesthetic depth (based on arterial concentrations). Moreover, regression-to-Ramsay-scores showed the ability of the model to represent levels of unconsciousness (almost) linearly using the Ramsay scale, and allowed us to reasonably interpret the temporal dynamics of EEG (as argued in the previous sections). In contrast, regression to target or blood sample concentrations showed significant variation and noise within both group and individual level predictions, with blood sample measures slightly improving the results. All these outcomes can be explained by the complexity of the inter-individual response to the drug, which can be captured by the EEG signatures and partially, by behavior (or any other clinical effect).

These findings are also consistent with empirical clinical practice and the existing literature. From a clinical perspective, the differences between the targeted plasma concentrations and the actual effect-site concentrations in the brain relate to a number of limitations that arise from contemporary TCI (target-control infusion) devices (Absalom *et al.* 2009). These limitations arise from either the pump's volumetric inaccuracies (mostly relevant for short half-life drugs and high dose rates), or from PK modelling deviations, as compartment modelling, rate constants and effect-sites are all determined by various pharmacological and biological variables. Current PK models (including the Marsh) have been developed with a limited number of patients, whose physical characteristics (e.g. age, weight, height, etc.) are not always representative, and thus there is no generally superior model. This is also reflected in clinical practice, where anesthesiologists often perform TIVA (total intravenous anesthesia) using manual adjustments of infusion rates (unlike TCI modes), with doses that are empirically adapted to patient characteristics, the intensity of noxious stimulation and as a function of other co-administered medications.

With respect to existing literature, (Chennu *et al.* 2016) showed in the original study that our 2nd experiment was based on, that besides the inter-subject variability of drug concentrations (as shown from the blood sample measures), dose-depended signatures and the overall anesthetic trajectory varied across participants (including the loss and recovery of consciousness). The authors also showed that arterial concentrations were not proportional to clinical effect, as measured by an auditory discrimination task, revealing the individual susceptibility to anesthetic dosage (specifically, subjects with weaker alpha-band networks were more likely to lose responsiveness). Of course, even by acquiring an accurate estimation

of drug concentrations, understanding the relation between drug doses and clinical effect is not a trivial problem, as the available methods for measuring effect-site concentrations and clinical effects are both indirect, and are often used circularly during PD modelling. In this respect, EEG is a strong candidate for capturing neurophysiological changes beyond the effect-site concentrations, and which are not directly evident from clinical effects (e.g. behaviorally), but ideally reflect states and levels of unconsciousness.

Overall, and despite the advantages of behavioral measures over pharmacological, there are still several limitations associated with contemporary clinical scales. One of such limitations is reflected on the temporal granularity in which responses can be recorded, and which does not allow us to assess the transitional nature of the observed signatures/predictions. Moreover, the clinical assessment is limited both qualitatively regarding contents of experience, but also quantitatively in terms of levels of unconsciousness, as it relies on the grading of specific non-scalable responses (usually verbal, ocular or motor responses). This was mostly evident during the state of *Recovery*, where Ramsay scores were unable to differentiate the subtle changes in levels of sedation. Last but not least, behavioral measures are (by definition) unable to detect covert awareness, which is particularly important for assessing the critical transition from LOBR to LOC. Therefore, the issue of ground-truth remains a challenge for the research field, and particularly for deep learning during training and evaluation.

5.6.5 Limitations of Psychometrical Measures for Investigation and Estimation of Altered States of Consciousness

As we mentioned in the previous section, one of the main investigations of the chapter regarded the evaluation of psychometrical evidence for consciousness, which are rarely explored in similar studies (especially within the machine learning literature). Given that subjects under GA can be disconnected from the environment and still having conscious experiences, the assessment of the electrophysiological changes during disconnected consciousness becomes important for both research and clinical implications. Ketamine is a dissociative anesthetic (with various psychoactive effects) that allows us to study such altered states, as it induces profound unresponsiveness, but also vivid dreams that subjects report upon emergence. However, when it comes to psychometrical measures, and based on the analysis of Experiment 3, there are still significant constrains that arise from the timing and nature of the subjective reports.

Starting by a comparison to propofol anesthesia, the electrophysiological distinction of the dose-dependent ketamine-induced changes appeared to be a more challenging task, already due to the agent itself. This was evident from the significant decrease in performance overall (and the larger variance in performances across participants), when comparing the corresponding n-state classification tasks of the propofol and ketamine studies (sections 5.3.1

and 5.5.1). Particularly for the state of *Sedation* under ketamine, confusion matrices showed a bias towards the prediction of *Wakefulness*, which could be partly explained by the milder sub-anesthetic dosing strategy of the *Michigan Anesthesia Study* (which has been associated to Ramsay 2-3 in (Bonhomme *et al.* 2016)), but also to several ground-truth considerations. Specifically, the resemblance of ketamine anesthesia to states of wakefulness has also been demonstrated in other studies, both phenomenologically (by reports of dreaming), but also electrophysiologically (by measuring the complexity of the cortical responses to TMS stimulation, as in (Sarasso *et al.* 2015)). Additionally, in contrast to Experiment 1, the characterization of ketamine *Sedation* as a unitary behaviorally-defined or psychometrically-defined state across time and across participants is questionable, given the retrospective reports and the subjective nature of the psychometric scoring.

These limitations were also evident in our regression analysis, where we tried to predict specific levels of disconnected consciousness (DC), from the EEG signatures of *Sedation*. Using three different ground-truth measures for DC levels (*study scores*, *lifetime + study scores*, and *relative-change scores*), our model showed a large predictive variation within *Sedation*, with no statistically significant differences across measures, and no ability to predict the subject-individual DC scores (section 5.5.2). In a similar study based on the *Michigan Anesthesia Dataset*, (Vlisides *et al.* 2018) made an exploratory analysis that aimed to uncover possible relationships between the psychometrics and the EEG. Using Spearman's correlation between the study scores of each subscale in the ASC questionnaires (11 subscales in total), and in relation to basic channel and source metrics (spectral power features computed from the *Sedation* state), the authors found no statistically significant correlations, after correction for multiple comparisons (Bonferroni's method). Of course, while deep learning can be used for more complex non-linear multivariate pattern analysis, we selectively focused on the average score of the three most relevant scales (identified in (Vlisides *et al.* 2018)), as our best prior hypothesis. As a final point to this analysis, the state of *LOBR* was left unexplored in relation to disconnected consciousness, due to the lack of subsequent records of the ASC psychometrics (albeit it has previously been associated to reports of DC, for all participants found in (Sarasso *et al.* 2015)).

In general, there are several other theoretical and technical limitations that affect our current tools of investigation and the study of altered states of consciousness. For example, although the systematic collection of retrospective reports is the only available method to detect covert awareness, we need to critically assess them, as participants may forget or confabulate their experiences upon awakening (although in case of ketamine it is unlikely, given that reports tend to be highly structured, explicitly narrative, emotionally rich, and extended in time, similarly to sleep dreams). Also, it becomes important to continuously refine the ASC dimensions in terms of psychometric validity and reliability, temporal effects and inter-individual differences. Finally, other technical considerations regarding specifically the *Michigan Anesthesia Study*, could relate to the small sample size of participants (which might not reached statistical significance), the relatively low dosing strategy, potential cognitive

impairment during the assessment (or misunderstandings), environmental factors (e.g. a bright lighting and the lack of privacy could lead to weaker effects) and even a lack of placebo condition, all of which could affect the credibility of our investigation.

5.6.6 Deep Learning-based EEG and Comparison to Clinical Practice

Throughout our experiments, we saw a number of deep learning models attaining different levels of performance on the basis of certain factors, such as the learning algorithm, the ground-truth measure, the anesthetic agent, and others. While we compared and discussed several of these factors within our own results, the heterogeneity across similar studies makes a comparative evaluation challenging, due to the significant differences in anesthetic and methodological protocols. Most notably, a large number of studies found in the related work (section 5.1.3) have analyzed patient data, which although create more clinically relevant findings, are not as suitable for a basic analysis of specific agents and dosages, as experimental setups found in research (clinical environments are not as controlled, with several medications co-administered depending on patients' needs). Nevertheless, it is important to review the current state-of-the-art models for DoA prediction, and discuss their relation to contemporary clinical practice.

Based on our behavioral analysis, we obtained the best performing model using a regression-to-Ramsay-scores optimization strategy, as revealed by the average MAE and the observed model dynamics. Using the EEG pre-processing pipeline and the 3D cNN derived in Chapter 7, and under the 3-state regression task (which included the most reliably encoded states), we were able to achieve a total MAE of 0.5 (within the Ramsay scale). (Sun *et al.* 2019a), which had the most successful results upon the same criteria, achieved a similar performance using a cNN-based EEG model trained under the Richmond Agitation-Sedation Scale (RASS), reporting a total MAE of 1. While the range of sedation is similar for the two scales (RASS used 2 additional levels of sedation, in comparison to Ramsay), the analysis of ICU patients, the variety of anesthetics, and the use of 2 frontal EEG channels makes the performance of the study mostly of clinical significance. This was evident by the predictive accuracy of the model (allowing one RASS level deviation) that surpassed the accuracy of a median technician-nurse agreement (70% over 59%). (Jiang *et al.* 2015) has also showed that neural networks can outperform contemporary DoA monitors like the BIS, as evaluated by anesthetic depth curves created by expert doctors (correlation coefficient of 0.73 over 0.62). Besides this, BIS like systems cannot perform with all anesthetics (e.g. ketamine or dexmedetomidine), and have been shown to exhibit a delay in response ((Zanner, Pilge, E.F. Kochs, *et al.* 2009) found 0.4 – 2 min delay before a new state was detected, which might not be acceptable in certain operations). On the other hand, in our own work we focused on experimental setups using full-head coverage EEG systems (10-20), which managed to obtain more stable performance across participants, and shorter delay responses from our model (in comparison to (Sun *et al.* 2019a) or BIS (Zanner, Pilge, E.F. Kochs, *et al.* 2009)).

Regarding our pharmacological analysis and the problem of estimating drug concentrations, we obtained moderate results by employing a regression-to-blood-sample-concentrations optimization strategy. When trained under TCI-estimated concentrations, we showed that our 3D cNN distinguished plasma concentrations of ~1 μg/ml on average (with an accuracy of 83%), which was constrained by the inter-subject variability in response to the drug (performance increased by regression to arterial concentrations). Although we showed a limitation in predicting drug concentrations for individuals, we need to consider clinical factors that contribute to the training behavior and performance of the model (e.g. patient population characteristics or range of concentrations). Specifically, while our model trained on data within the range of 0 – 2 μg/ml, propofol concentrations in clinical practice can range from 0 to 4 mg/ml, depending on the operation. Based on the median absolute performance error (MDAPE) used in (Glen and White 2014), and defined as the percentage of error between the concentrations measured and predicted (an established methodology found in (Varvel, Donoho and Shafer 1992)), we can assess the inaccuracy of our model with an IQR median performance of 37%. Nonetheless, such performance error is not clinically acceptable, as it has been suggested that MDAPE should be in the region of 20-30% (albeit the Marsh model itself does not pass these criteria) (Glen and White 2014).

5.6.7 Summary

Overall, we focused on the investigation of our EEG data, through different learning tasks and ground-truth measures, in order to uncover the relationships across the pharmacological, electrophysiological, behavioral, and psychometrical variables. We discussed the appropriateness of regression analysis with respect to theoretical considerations and the temporal dynamics of anesthesia, as well as the advantages of behavioral measures against pharmacological or psychometrical evidence for measuring levels of consciousness. Despite the limitations that arise from current tools of investigation, our findings highlight the ability of the 3D cNN model to detect anesthetic levels with higher accuracy than any other contemporary machine learning model, or method of clinical practice, under an optimal optimization strategy. Moreover, we showed the adequacy of deep learning-based EEG to refine and enrich (interpolate) its predictions, beyond any noise arising from the EEG or the ground-truth. In the next chapter, we test the predictive behavior and generalizability of our model under novel setups, by incorporating unseen experimental paradigms and anesthetic agents, in order to further investigate its robustness and learning capacity.

Chapter 6 Cross-study and Cross-drug Generalization of Anesthetic-Induced Unconsciousness

6.1 Introduction

6.1.1 Overview

In this chapter we test the predictive power of our deep learning model under a cross-study and cross-drug generalization task, as a way to explore its performance robustness and the validity of our findings thus far. Models within the EEG and deep learning literature have shown weaknesses in replication of their results, due to a number of factors, such as the lack of consistent methodologies, proper prior hypotheses, or validation frameworks with independent datasets. Especially within the context of deep learning-based EEG for tracking the depth of anesthesia, there is an absence of reproducibility of the models, which is partly driven by the difficulties in acquiring or publicly sharing anesthesia datasets. With this goal in mind, we exploit the acquisition of four anesthesia datasets and perform a series of experiments, where we evaluate our model across studies and anesthetic agents, using predefined behavioral ground-truths. We also explore the possibility of improving model generalization using a training strategy that incorporates multiple anesthetics, whilst assuming a common electrophysiological learning target. Specifically, in Experiment 1 we focus on cross-study generalization using two propofol studies with distinct experimental designs and paradigms for measuring behavioral unresponsiveness. In Experiment 2, we focus on cross-drug generalization using a model trained on propofol and tested on two unseen anesthetics – ketamine and xenon. In Experiment 3, we further explore cross-drug generalization by creating a mixed trained model using the agents of propofol and ketamine, which exhibit distinct electrophysiological and phenomenological characteristics. Our results demonstrate the robustness of the model across different experimental designs, and its capacity for learning generalized cross-drug features of unconsciousness, given an appropriate setup. We finally discuss some of the differences across anesthetics, our limitations in characterizing altered states and depths of anesthesia, and their implications on model training and performance.

6.1.2 Background

Our aim here is to focus on the problem of generalization and reproducibility within the framework of deep learning-based EEG, in order to assess our model's strength in impartially estimating levels of anesthetic-induced unconsciousness. Specifically, we test our predictive model across different studies and anesthetic agents, and we explore a training strategy that could allow us to capture potential common features of unconsciousness. The existence of such features, besides aiming at an improved cross-study and cross-drug generalization, would also have theoretical implications for general anesthesia research.

The problem of generalization and reproducibility, while fundamental for science and engineering, is particularly evident throughout both EEG and deep learning literature, as indicated by recent studies. With regards to EEG research, the neuroscience community is showing growing awareness about the limited evidence of many findings, and has only recently begun to appreciate the importance of large-scale studies, in an effort to improve the replicability and statistical power of experiments (Pavlov *et al.* 2021a). Meanwhile, generalization is one of the fundamental issues and goals of machine learning. Specifically for deep learning, as neural networks can be trained to fit anything, the generalization hypothesis lies on their ability to discover appropriate low-dimensional latent spaces (or "manifolds"), where the data can be encoded and interpolated (generalization can be thought as interpolation upon this manifold). Even though the success behind this idea is guided by different methodological concerns compared to EEG research, both fields share several key insights; the statistical power of the models rely significantly on the size and quality of the data (representability), the appropriate formulation of the research question or task, and the given methodological choices (e.g. the processing pipelines or the architectural priors of neural networks). In Chapter 4, we have already discussed the lack of standardized EEG methodologies as part of the reproducibility problem in EEG research, and our concomitant efforts to acquire a consistent deep learning-based pipeline, that allowed the integration and comparability of our findings (generalization testing). Here, we focus on two notions of predictive generalization, in terms of extrapolating to new datasets defined by novel experimental designs and anesthetic agents.

The first notion of generalization which is important to evaluate, concerns the cross-study generalization. As we have previously mentioned, studies with EEG under GA can vary across a number of variables, such as the experimental or clinical setting, the EEG device, the recording environment, the anesthetic type and administration protocol, and others. In addition, EEG experiments often suffer from small sample sizes (due to cost and complexity), noisy datasets, computational requirements (data curation), and experimenter degrees of freedom (e.g. in statistical tests or regions of analysis). All of the above can result in analytic flexibility, which creates the need for improved model robustness. Meanwhile, deep learning studies have shown much evidence of faulty performance, with the networks detecting spurious correlations in all kinds of biomedical engineering tasks (e.g. related to imaging equipment noise or task-

specific confounds) (Fellner *et al.* 2016; DeGrave, Janizek and Lee 2021; Mahmood *et al.* 2021). Specifically within our own analysis, EEG artifacts and mislabeled epochs are prospective pitfalls that could lead to model overfitting. Therefore, it is important to extent our datasets, pre-define our ground-truth hypotheses, and appropriately assign our training/testing sets based on our learning goal.

The second notion of generalization, which is more relevant to our research inquiry, concerns the cross-drug generalization. One of the theoretical considerations and unsolved questions recognized in GA research, regards the discovery of the common mechanisms of anesthetic-induced unconsciousness, or even the possibility of finding a unitary signature for measuring levels of unconsciousness (although less commonly presumed, given that different altered states seem to occur through different pathways in the brain) (Bonhomme *et al.* 2019). However, while there has been progress in uncovering many electrophysiological markers for a variety of anesthetics, findings are often limited to specific agents, specific mechanisms and even specific levels of sedation, with no clear continuity or distinction across states and agents (Patrick L. Purdon *et al.* 2015). Hence, we are interested in exploring the possibility of deep learning to discover common cross-drug features, and the potential to improve its performance by training under multiple anesthetic agents (or, more "representative" datasets). Such potential would also have clinical implications, given the ineffectiveness of contemporary DoA monitors to perform under different anesthetics with independent action (Hans *et al.* 2005).

6.1.3 Related Work

In the previous chapter, we mentioned a number of works using learning-based methods for the analysis of EEG and the estimation of levels of consciousness under GA (section 5.1.3). While several of these studies share similar learning objectives and anesthetic definitions with our work here, a comparative evaluation of the models is still limited by the variation found in methodological choices, validation frameworks, and most importantly, the datasets under analysis. As we have emphasized, the particular selection of the data under training and testing has a profound effect on the resulting model and the respective performance. Most importantly though, none of the deep learning studies we are aware has further validated the acquired model under a cross-study or cross-drug generalization task.

Despite this gap, there are several works on theoretically or empirically-based EEG metrics for discriminating states of consciousness, which have been shown to generalize across a variety of conditions (including anesthetic states). Typically, these metrics can be categorized to methods that quantify the information or spectral content of the EEG signals (e.g. approximate entropy (Bruhn, Röpcke and Hoeft 2000), spectral entropy (Klockars *et al.* 2012), bispectral index (Ellerkmann *et al.* 2010)), methods that evaluate the spatial extent or synchronization of brain activity (e.g. measures of functional/effective connectivity, DCM (Boly *et al.* 2012; Muthukumaraswamy *et al.* 2015), granger causality (Engel and Singer 2001)) or a mixture of the two (e.g. PCI (Casali *et al.* 2013), graph theoretical measures (Chennu *et*

al. 2014; Chennu *et al.* 2016)). Here, we focus on some of the techniques that can be used to progressively track the anesthetic depth, and which have been tested across a variety of datasets and anesthetic agents.

One of the most successful metrics developed over the past years is the perturbational complexity index (PCI), which is based on the estimation of information differentiation and integration of cortical activity. In one study, (Casali *et al.* 2013) showed in 18 healthy volunteers undergoing propofol, xenon or midazolam anesthesia (N=6 for each agent) that the PCI index was able to discriminate states of consciousness (wakefulness) and unconsciousness, as determined by the MOAAS (Modified Observer's Assessment of Alertness and Sedation) scale. (Bai *et al.* 2015) investigated the permutation Lempel-Ziv complexity (PLZC) index, which, similarly to PCI, measures the complexity of the EEG timeseries by quantifying the distinct patterns of activity. Specifically, the authors incorporated two studies with 10 healthy volunteers under propofol and 19 patients under sevoflurane anesthesia, and showed that the PLZC values correlated highly to effect-site concentrations in both agents, as estimated by PK/PD modelling (using standard sigmoid models). Another set of widely used techniques, are based on entropy measures. (Liang *et al.* 2015), made a systematic comparison of 12 entropy indices on 48 patients under sevoflurane and isoflurane anesthesia (N=19 and 29, respectively), upon their ability to predict the effect-site concentrations of the drugs (using PK/PD modelling similar to (Bai *et al.* 2015)). In general, permutation entropy indices performed adequately well, with Renyi PE having the best overall performance. (Mhuircheartaigh *et al.* 2013b) and (Sleigh *et al.* 2019) explored the appearance of slow wave activity (SWA) and its saturation point as potential indicators for the transition to LOBR and LOC, respectively. Specifically, the two studies used healthy participants under propofol (N=16) and ketamine (N=15) and showed a sharp increase of SWA after LOBR for both agents, with a weaker correlation of SWA to effect-site concentrations and the point of recovery, for ketamine. Finally, (Colombo *et al.* 2019) investigated the spectral exponent (decay-rate) of the EEG's PSD as a metric of unconsciousness, given the association of anesthetics to EEG slowing and redistribution of spectral power. By incorporating 15 healthy subjects under propofol, ketamine and xenon anesthesia (N=5 for each agent), the authors showed that the spectral exponent reliably separated conditions of consciousness and unconsciousness, as determined by Ramsay scores.

In spite of some partial success with the above techniques, each one shows specific strengths and limitations, when it comes to generalization and reproducibility. For example, the PCI and the spectral exponent have been shown to generalize only within the 2-class problem of discriminating wakefulness from states of unconsciousness (a relatively easy problem, as seen in section 5.3.1), without being able to progressively track levels of anesthetic depth (for propofol, PCI has been shown to produce intermediate values during sedation). On the contrary, PLZC, entropy measures and SWA have been tested as continuous indices against the effect-site concentrations, showing correlations with the different anesthetic depths. Nevertheless, complexity measures cannot differentiate robustly states with distinct electrophysiological and phenomenological properties (e.g. from anesthetic induction to LOC,

or from deep anesthesia to recovery). Moreover, they have not shown generalization across all tested GABAergic anesthetics. This could relate to the fact that the PK/PD modelling used for evaluation is determined by EEG parameters, which likely vary across anesthetics (hence, creating a problem of a circular ground-truth). SWA has been strongly correlated with the moment of LOBR for both propofol and ketamine, but cannot account for pharmacological variations or the moment of recovery (SWA is not a sufficient condition for consciousness). Finally, none of the above methods can discriminate wakefulness from states of disconnected consciousness, which have been shown to appear during LOBR under ketamine.

Overall, considering the lack of reproducibility within the deep learning-based EEG literature, and given our reasoning for the problems that can emerge within EEG and deep learning studies, it is important to focus on the problems of cross-study and cross-drug generalization. Besides generalization and reproducibility, a comparison of our deep learning approach with these non-learning-based methods can also be made, given the variety of methodological differences and limitations, which we discuss in detail in the next chapter.

6.2 Methods

6.2.1 Datasets Collection

For this investigation, we expanded our data from the previous chapter by including a novel dataset acquired in (Sarasso *et al.* 2015), which can serve as an additional test set for our cross-study and cross-drug generalization analysis. We also elaborate on our existing datasets by incorporating information related to a behavioral ground-truth for assessing levels of consciousness, for the states found in *Cambridge Anesthesia Study* and *Michigan Anesthesia Study* (previously characterized and used for pharmacological a psychometrical investigation).

Liege Anesthesia Dataset 1. The dataset acquired in (Murphy *et al.* 2011) has been described in detail in sections 3.2.1 and 5.2.1. Briefly, fifteen-minute spontaneous high-density EEG was recorded from 10 healthy participants during propofol anesthesia at four different states of consciousness: *Wakefulness*, *Sedation*, *LOC*, and *Recovery*. Each state was determined based upon reaching and sustaining an individualized effect-site concentration that corresponded to a desired behavioral response (Ramsay score), namely: Ramsay 2 (fully awake) for *Wakefulness* and *Recovery*, Ramsay 3 (slow response to command) for *Sedation*, and Ramsay 5-6 (no response) for *LOC*. For 3 out of 10 participants, the EEG recordings during all transitional states (wakefulness to sedation – *WS*, sedation to loss of consciousness – *SL*, and loss of consciousness to recovery – *LR*) were also available for analysis. During the transitional states, propofol infusion rates were increased or decreased, according to the desired target.The experimental design of the dataset is depicted in Fig. 6.1.

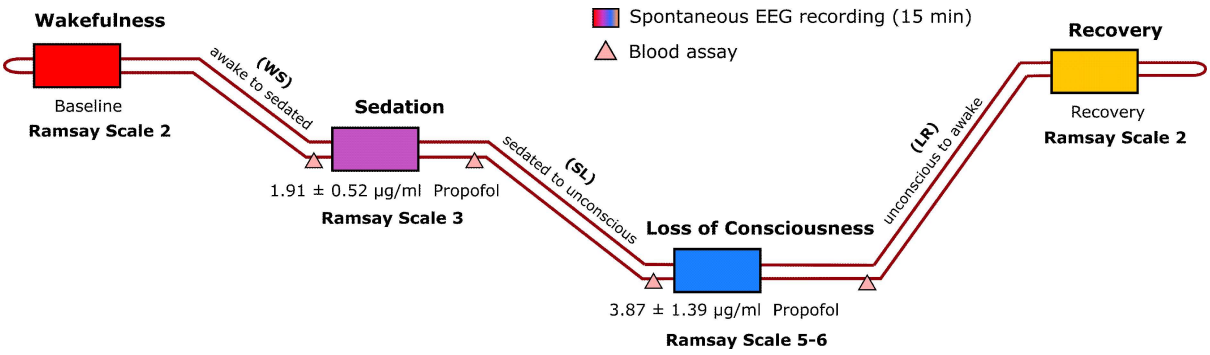

Fig. 6.1. Experimental design of the *Liege Anesthesia Dataset 1*. Four different levels of consciousness were attained based on the behavioral response of the participants, assessed by the Ramsay scale.

Cambridge Anesthesia Dataset. The dataset acquired in (Chennu *et al.* 2016) has been described in detail in section 5.2.1. Briefly, seven-minute spontaneous high-density EEG was recorded from 20 participants during propofol anesthesia at four different states of consciousness, namely: *Baseline Wakefulness*, *Mild Sedation*, *Moderate Sedation* and *Recovery*. Each state was determined by a desired ("target") plasma concentration, controlled by a computerized syringe driver (Alaris TCI mode, using the Marsh model) that achieved and maintained the required propofol infusion rate (0.6 µg/ml for *Mild Sedation* and 1.2 µg/ml for *Moderate Sedation*. *Recovery* was associated to 0 µg/ml, 20 minutes after cessation of infusion).

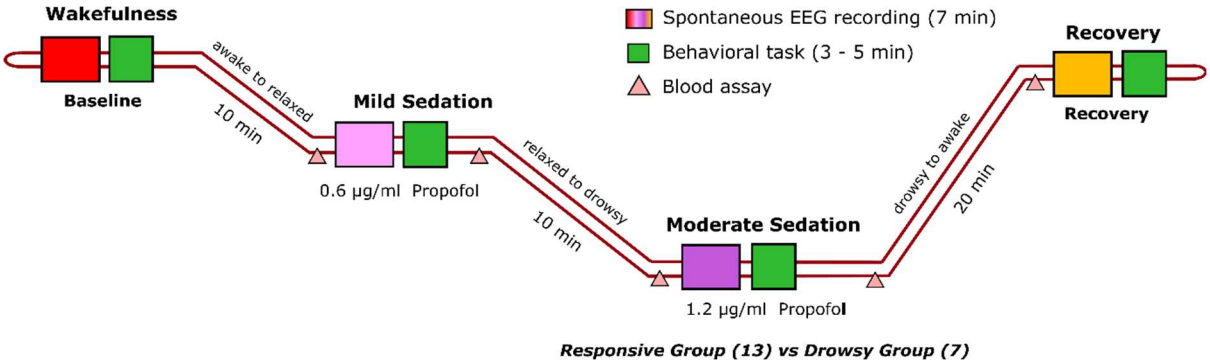

Fig. 6.2. Experimental design of the *Cambridge Anesthesia Dataset*. Two different participant subgroups were identified during *Moderate Sedation* based on their behavioral responsiveness (*Responsive* and *Drowsy*), assessed by the 'hit rates' of the auditory discrimination task.

At each of the four states, and after the resting-state period, a simple behavioral task was performed that involved a fast discrimination (button press) between two possible binaurally auditory stimuli (buzz or noise). An assessment of levels of consciousness was then performed by measuring the hit rates (percentage of correct responses) and reaction times of the participants' responses (the delay between an auditory onset and the button press). During *Moderate Sedation*, and based on binomial modelling of the hit rates, the authors identified two

subgroups of 13 and 7 participants, characterized as *Responsive* and *Drowsy*, which reflected the inter-individual variability of the pharmacodynamic impact of propofol. The experimental design of the dataset is depicted in Fig. 6.2.

Michigan Anesthesia Dataset. The dataset acquired in (Vlisides *et al.* 2017) has been described in detail in section 5.2.1. Briefly, spontaneous high-density EEG was recorded from 15 healthy participants during ketamine anesthesia at 4 different states of consciousness: *Wakefulness*, *Sub-anesthetic Sedation*, *Loss of Behavioral Response* (*LOBR*) and *Recovery*. The two anesthetic states were characterized by different dosing strategies, namely: a continuous intravenous infusion of 0.5 mg/kg racemic ketamine over 40 minutes during *Sub-anesthetic Sedation*, and a 1.5 mg/kg anesthetic bolus dose in *LOBR*. For each anesthetic state, a steady-state was extracted in accordance with the original analysis, based on the pharmacologically-steady period of the sub-anesthetic block (*Sedation*), and the 5-min period after cessation of response to commands (*LOBR*), respectively. For 14 out of 15 participants, the EEG recordings during all transitional phases were also available for analysis.

In order to behaviorally characterize the steady-state of *Sedation*, we draw evidence from two independent studies (Bonhomme *et al.* 2016; Salah and Alansary 2019), which have used the Ramsay scale to assess the responses of participants under sub-anesthetic doses of ketamine. Based on the experiment in (Vlisides *et al.* 2017) and their work, the corresponding Ramsay scores for the four steady states are: Ramsay 2 for *Wakefulness* and *Recovery*, Ramsay 2-3 for *Sedation* and Ramsay 5-6 for *LOBR*. The experimental design of the dataset is depicted in Fig. 6.3.

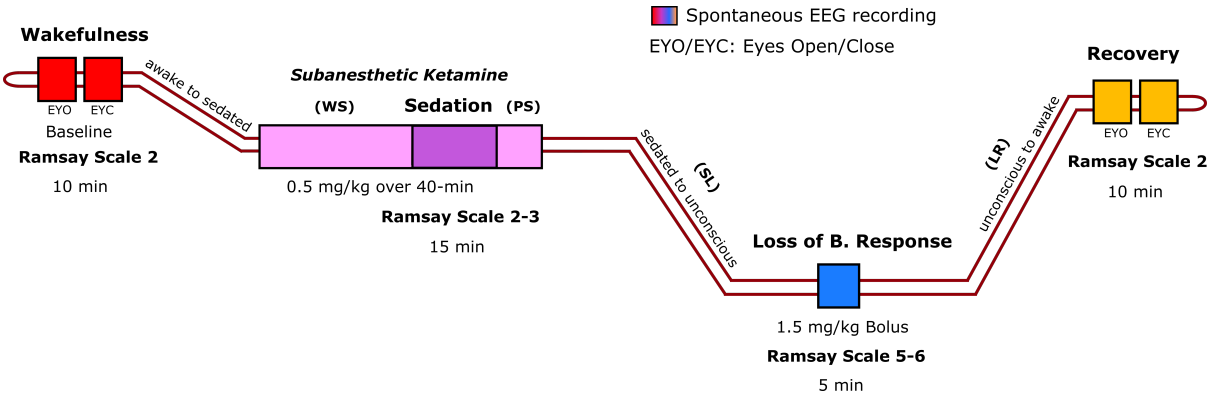

Fig. 6.3. Experimental design of the *Michigan Anesthesia Dataset*. Four different levels of consciousness were attained, which corresponded to different behavioral responses within the Ramsay scale.

Liege Anesthesia Dataset 2. The dataset acquired in (Sarasso *et al.* 2015) is based on a TMS-EEG anesthesia study, in which the experimental design is described in detail. Briefly, the study was approved by the local ethical committee of the University of Liege and all participants gave written informed consent. Physical examination and medical history were

obtained from all participants, to exclude conditions incompatible with the experimental procedure (anesthesia or TMS related).

High-density electroencephalography (TMS-compatible hd-EEG, 60 channel Nexstim eXimia system) was recorded from 10 participants (age 18-28) during xenon and ketamine anesthesia, at two different states: *Wakefulness* and *Loss of Consciousness* (*LOC*)/*Loss of Behavioral Response* (*LOBR*). Only one type of anesthetic was administered to a given participant (randomly assigned, with N=5 for xenon and ketamine, respectively), in order to independently study their effects on the brain. At baseline (*Wakefulness)*, spontaneous EEG was recorded before the TMS-EEG session, for approximately 5-6 minutes. After anesthetic induction and during *LOC*/*LOBR*, spontaneous EEG was recorded for several minutes before and after the TMS-EEG session (5-6 minutes). Xenon anesthesia was maintained using a Dräger PhysioFlex closed circuit ventilator (62.5 ± 2.5% in oxygen), with a total amount ranging from 24 to 32 liters. For Ketamine anesthesia, a 2 mg/kg intravenous infusion (diluted in 10mL of 0.9% normal saline) of racemic ketamine over 2 minutes was used and maintained at 0.05 mg/kg/min over the experimental procedure. During all procedures, participants were monitored (electrocardiogram, BP, etc.) and received metoclopramide (2mg) for nausea and vomiting complications by the anesthetics. As in this work we focus on resting-state EEG measures, we accumulate only the periods of EEG without TMS, which allow us to investigate the behaviorally-steady states of the two conditions.

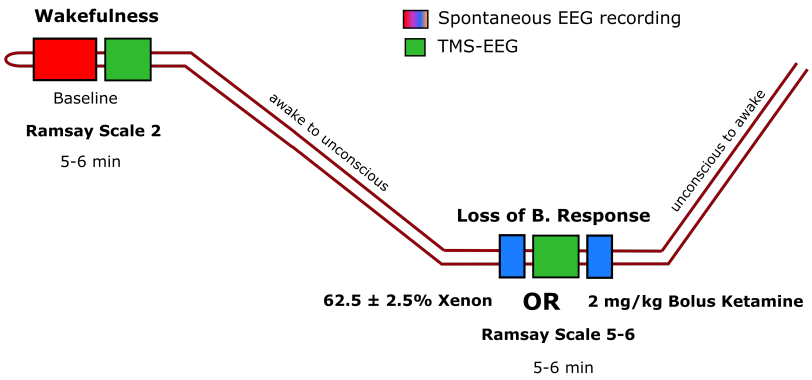

Fig. 6.4. Experimental design of the *Liege Anesthesia Dataset 2*. Two different levels of consciousness were attained for both agents (ketamine/xenon) based on the behavioral response of the participants, assessed by the Ramsay scale.

The levels of consciousness were commonly assessed for both agents with the Ramsay scale, similarly to the *Liege Anesthesia Study 1* (a verbal command to squeeze the hand of the investigator, every 30 sec). *LOC*/*LOBR* was denoted upon reaching Ramsay score 5-6 (no response to external stimuli) and after three consecutive assessments (Ramsay 2 – clear response – was assigned during *Wakefulness*). Moreover, self-report measures were collected after participants recovered from deep anesthesia (and asked to confirm their responses after

one hour), in order to investigate the possibility of covert experiences during the unresponsive period. Conscious experiences were considered present when participants were able to describe their content (any kind of activity, such as thoughts, images or emotions). In case of ketamine, all participants reported having experiences unrelated to the external environment during *LOBR*, which were vivid, visually rich and extended in time (in agreement with the *Michigan Anesthesia Study*. A more detailed description of a ketamine report can be found in the original study). In case of xenon, participants had no explicit recall of any experience (*LOC*). The experimental design of the dataset can be seen in Fig. 6.4.

6.2.2 EEG Pre-processing and Deep Learning Model

The EEG pre-processing pipeline and the deep learning model used throughout the subsequent analyses were kept common for all datasets and experiments (and in alignment with the previous chapter). As we have discussed, the selected pre-processing pipeline and the generic 3D cNN derived in Chapter 4, allow us to automatically and consistently integrate different EEG devices and channel configurations, while introducing minimal prior computational assumptions (on the nature of EEG and the respective model's architecture design). The EEG decoding methodology is briefly depicted below (detailed description in section 4.8).

EEG Preprocessing. All preprocessing steps are executed sequentially using the *mne* and *scikit-learn* libraries.

1. 10-20 System Channel Selection
2. Band-Pass Filtering (0.5 – 40 Hz, 50/60 Hz Notch Filter)
3. Resampling at 100 Hz
4. Epoching (1 sec, non-overlapping)
5. Automatic Artifact Cleaning (Only for Training Data)
    a. Bad Channel Interpolation ('bad' if channel is flat or has >20% epochs exceeding peak-to-peak threshold of 800 μV)
    b. Epoch Rejection (if >20% channels exceeding peak-to-peak threshold of 800 μV)
6. Re-referencing to Average
7. Epoch-wise Robust Standardization (quantile range: 0.25 – 0.75)

3D Convolutional Neural Network. The architecture of the 3D cNN is shown in Fig. 6.5. All activation functions are ReLU units, with the exception of the output layer. Other non-specified hyperparameters are set as default in *Tensorflow* (*v2.6.0*).

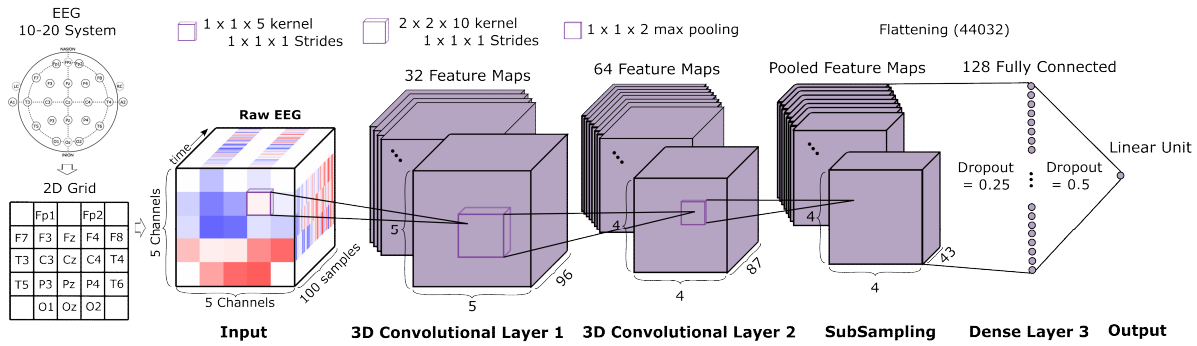

Fig. 6.5. The 3D Convolutional Neural Network model.

6.2.3 Model Training and Evaluation

In the previous chapter, we focused on the importance of the optimization task, with respect to the learning algorithm and the selection/encoding of a clinical ground-truth. Based on our investigation, we derived an optimal training strategy that relied on a regression analysis over behavioral measures of unconsciousness. This strategy was evaluated upon two different notions of generalization. The first one related to the adequacy of the model in refining and enriching its predictions over uncharacterized (or 'fuzzy') anesthetic states, most notably observed in the large-scale transitional dynamics of the EEG and the predicted anesthetic depths, which followed the anesthetic protocols. The second notion related to the cross-subject validation framework, which ensured the dismissal of features reflecting subject-dependent confounds or noise. In this chapter, we are interested in evaluating our model's robustness and learning capacity in extracting features beyond and unrelated to the experimental design or the particular anesthetic action, with an aim towards the acquisition of electrophysiological features that impartially reflect levels of unconsciousness.

More specifically, we employed a regression-to-behavioral-scores training, using either the Ramsay scale or the 'Behavioral-Responsiveness' (BR) scale (defined in Experiment 3), under the 3-state task that incorporates *Wakefulness*, *Sedation* and *LOC*/*LOBR* (*Recovery* was excluded from training given its unreliable assessment - discussed in section 5.6.1). In case of training under one study, the training/testing sets were split by individual datasets, with all participants of a given dataset used either for training or testing. In the case of cross-study training, a leave-one-participant-out cross-validation (LOPOCV) paradigm was used for the evaluation of participants within the training datasets. In both cases, all testing datasets consisted of participants from studies unseen during the training phase, which ensured a mismatched training/testing distribution. This is important to assess generalization across particular desired features, beyond study design and anesthetic type, as well as to estimate a cross-study or cross-drug generalization error.

The loss function and the performance metric for the regressands were the mean-squared-error (MSE) and the mean-absolute-error (MAE), respectively. Sample/target

weighting was applied to training data (described in section 5.2.4), given its advantage to correct for biases that stem from unbalanced or heterogeneous datasets. All models were trained using the Adadelta optimizer (*learning rate* = 1), with a batch size of 100 and for 10 training epochs. Initialization of network weights was done with the Xavier uniform initializer. Model creation, training and evaluation were implemented in Python 3 using the *Tensorflow*/*Keras* library and a CUDA NVIDIA GPU (Tesla P100).

6.3 Experiment 1 – Cross-study Generalization to Propofol Anesthesia

In this experiment, we investigate the performance of deep learning in a cross-study generalization task, by incorporating a test dataset with a novel experimental setup. Specifically, we use the *Liege Anesthesia Dataset 1* to train a regression-to-Ramsay-scores model under the 3-state task (see section 5.3.2), which has been studied, and given that it provides us with adequate data and a robust ground-truth for learning (as it includes reliably characterized states, ranging from *Wakefulness* to *LOC*). We then use the *Cambridge Anesthesia Dataset*, which is also based on a propofol anesthesia study, in order to test our predictive model in correctly estimating the unresponsiveness of participants, measured under an unseen behavioral paradigm (auditory task).

6.3.1 Results

The behavior of the trained model has been reported in detail in section 5.3.2. Briefly, a stable convergence is attained during the 10 training epochs, as indicated by the average mean-squared-error/mean-absolute-error curves (Fig. 5.9), with a cross-subject validation performance of 0.5 MAE.

In order to test our model, we first need to define a ground-truth on which we can evaluate the predictions obtained for the *Cambridge Anesthesia Dataset*. As the model is trained on Ramsay scores, and given that we do not have a ground-truth estimation of Ramsay scores for our test dataset, we can exploit the behavioral analysis employed in the original study for assessing our model's performance. Based on this analysis, two subgroups of participants were identified – a *Responsive* (Group 1 – 13 participants) and a *Drowsy* (Group 2 – 7 participants) – which reflected the inter-individual variability of the pharmacodynamic impact of propofol, leading to different levels of behavioral impairment.

Fig. 6.6 shows the output of the model over time (one prediction per 1-sec epoch) for each of the two groups, averaged across participants (calculated by taking the minimum number of epochs per state, and aligning them at the beginning of each state, across subjects). Predictions appear consistent overall for both groups, with a slow trend of increasing anesthetic

depth, which can be explained by the pharmacodynamic impact of the drug's accumulation over time (Fig. 6.6 A and B). The range of predictions fall within Ramsay 2-3 for the *Responsive* group (Group 1), and Ramsay 2-4 for the *Drowsy* group (Group 2), in agreement with the behavioral evidence found in (Chennu *et al.* 2016). Overall, the model is able to correctly estimate the behavioral responses of the participants, for both groups identified in the original study.

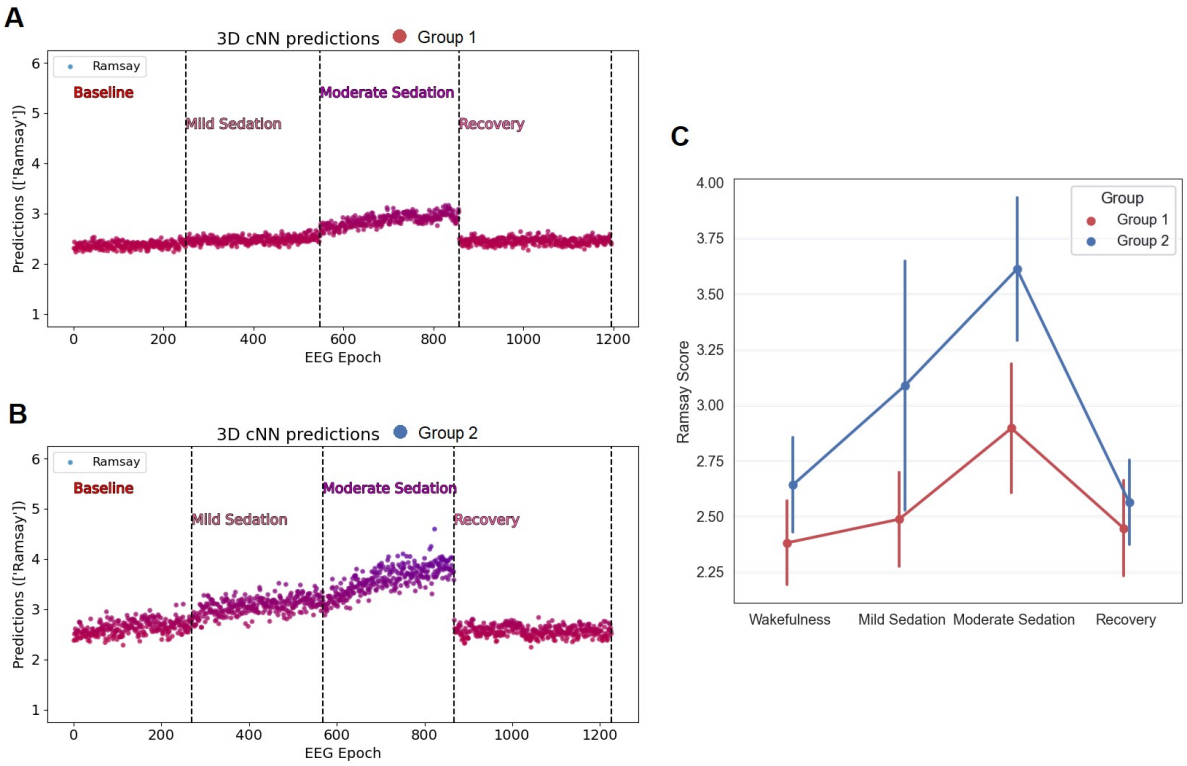

Fig. 6.6. Cross-study generalization results over the four anesthetic states of the test subjects in *Cambridge Anesthesia Dataset*. A) Ramsay score predictions of the *Responsive* group (Group 1, average). B) Ramsay score predictions of the *Drowsy* group (Group 2, average). C) Mean and standard deviation of Ramsay scores, per state and per group.

To confirm our results with this ground-truth, we performed a statistical test by entering the mean Ramsay scores for each participant and each state (Fig. 6.6 C), into a mixed ANOVA model (*fitrm*/*ranova*, *MATLAB*) with one non-repeated measure (participant group – *Responsive* or *Drowsy*) and one repeated measure (level of sedation – *Wakefulness*, *Mild Sedation*, *Moderate Sedation* and *Recovery*). Our results showed a significant interaction effect between participant group and sedation level ($F(3) = 6$, $p = 0.005$), which was mainly driven by the effect of the group during *Moderate Sedation* ($p = 0.0001$). We also observed smaller group effects during *Mild Sedation* ($p = 0.004$) and *Wakefulness* ($p = 0.01$), which indicate the

predisposition of the *Drowsy* group into increased unresponsiveness, both from the drug, but also from their prior resting-state EEG. These observations are also in agreement with the findings reported in (Chennu *et al.* 2016).

Given all the above, our findings highlight the model's ability for cross-study generalization over participants' levels of unconsciousness, defined by independent behavioral assessments and methodological analyses.

6.4 Experiment 2 – Cross-drug Generalization to Ketamine and Xenon Anesthesia

In this experiment, we investigate the performance of deep learning in a cross-drug (and cross-study) generalization task, by incorporating a test dataset with two novel anesthetic agents – ketamine and xenon. Specifically, we use the same model from the previous experiment (model trained under propofol, based on the *Liege Anesthesia Dataset 1*) and test its predictive behavior in *Liege Anesthesia Dataset 2*, which includes unseen anesthetic states produced by agents with distinct mechanisms and phenomenology. Due to the small size of *Liege Anesthesia Dataset 2* (which makes it unsuitable for training), we use it as test set for both cross-study and cross-drug analysis purposes, in order to acquire a generalization error. From the group receiving ketamine, we were able to acquire the data from only 4 out of 5 participants.

6.4.1 Results

The results of the experiment are shown in Fig. 6.7. The mean-squared-error (MSE) and mean-absolute-error (MAE) curves of our test set show that the 10 training epochs were adequate for convergence (Fig. 6.7 A). An average MAE of 2.09 was obtained for the ketamine participants and an MAE of 1.00 for the xenon participants, with the error stemming mainly from the anesthetic states during *LOBR*/*LOC* (Ramsay 5-6).

As we observe from Fig. 6.7 (B), all ketamine participants had a relatively high MAE, with xenon participants showing a larger variability of performances, ranging from 0.5 to 1.5 MAE (0.5 is the average performance of the trained propofol model using LOPOCV). Generally, predictions appear consistent during *Wakefulness* (MAE of ~0.45) and *LOBR* under ketamine (albeit with significant deviation from the ground-truth – MAE of 3.20), while a larger variability is present for the participants during *LOC* under xenon (MAE of 1.73) (Fig. 6.7, C). By comparing the generalization errors with the respective states during training (reported in section 5.3.2), we observe a minor error increase for *Wakefulness* (0.45 over 0.29 MAE), a moderate increase for *LOC* under xenon over propofol (1.73 over 0.76 MAE) and a significant increase for *LOBR* under ketamine (3.20 MAE). Although these results could be

partially explained by the between-study variations in experimental and methodological protocols, the significant error increase during the anesthetic states suggest a generalization barrier over the effects of ketamine and xenon.

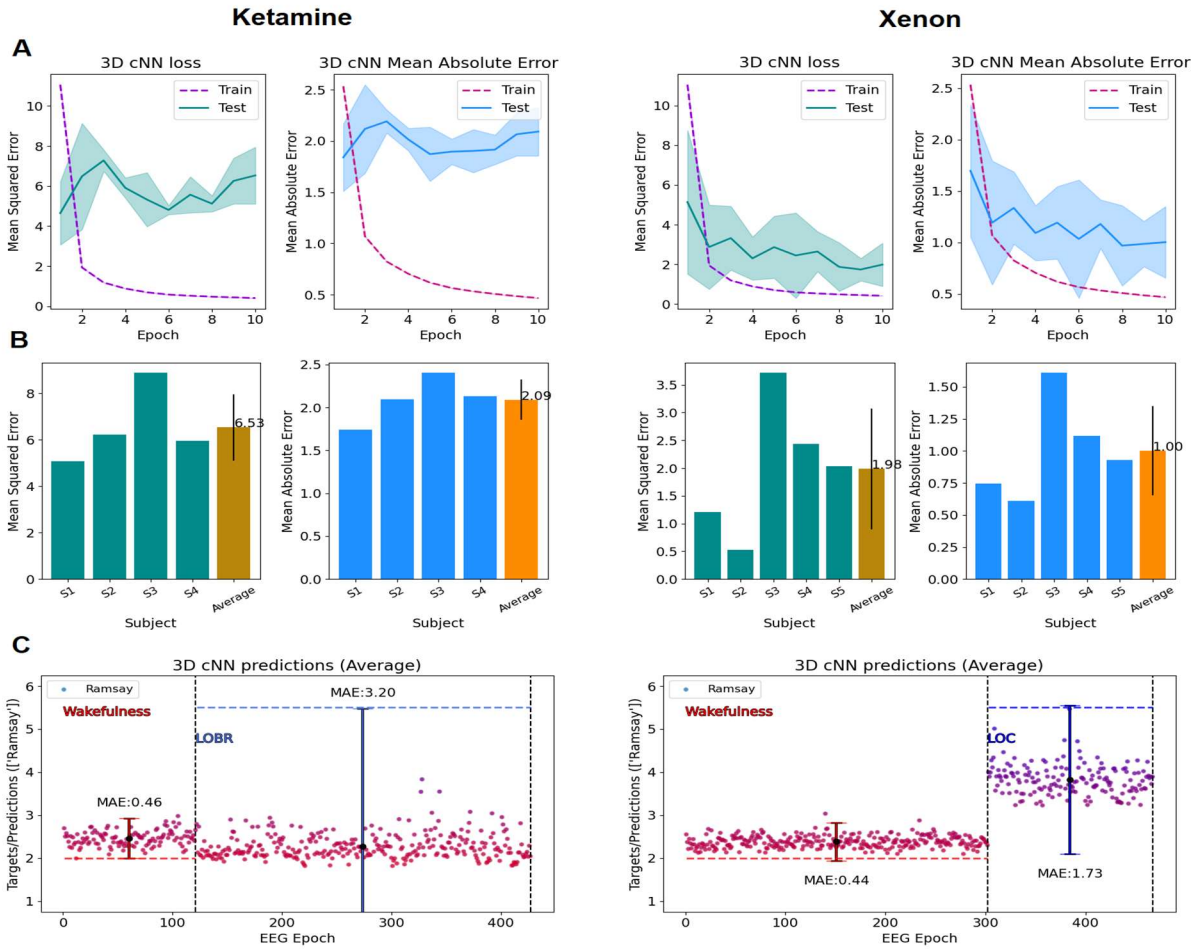

Fig. 6.7. Cross-drug generalization results of the test subjects in *Liege Anesthesia Dataset 2*, for ketamine and xenon respectively. A) Average mean-squared-error and mean-absolute-error curves. B) Subject-wise losses and MAEs. C) Ramsay score predictions for the 2 anesthetic states of the unseen test subjects (average, N=4 for ketamine, N=5 for xenon).

To better understand the behavior of the model and this barrier, we visualize the Ramsay predictions and the mean spectrogram of the corresponding epochs, for the best- and worst-performing subjects in the ketamine and xenon group, respectively (Fig 6.8). The two subjects were representative both in terms of the EEG signatures and the predictive behaviors observed.

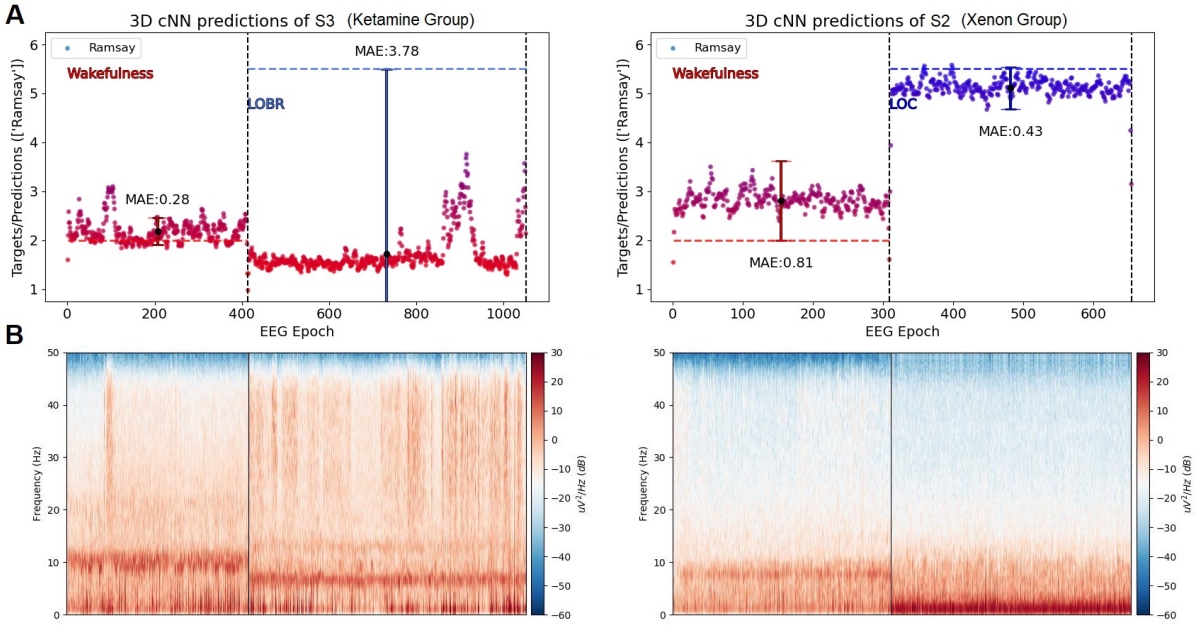

Fig. 6.8. Cross-drug generalization results of the best- and worst-performing subjects (S3 in ketamine group, S2 in xenon group). A) Ramsay score predictions for the two anesthetic states of the ketamine and xenon subjects (moving-average filter applied, kernel_size=5). B) PSD of the corresponding states and epochs (method=Welch, channel_aggregation=mean, window=1 sec, n_fft=256).

In the case of ketamine, the model predicted a Ramsay score of ~2 for both *Wakefulness* and *LOBR*, with occasional spikes of increases up to Ramsay 3-4, potentially associated to the increases observed in delta (0 – 4 Hz) and high gamma (25 – 45 Hz) activity. This behavior is consistent with the signatures and predictions reported in Chapter 5 (5.5.1), for *LOBR* under ketamine (*Michigan Anesthesia Study* analysis). In the case of xenon, the model predicted a Ramsay score of ~3 for *Wakefulness* and a score of ~5 for *LOC*, potentially associated to the observed weak alpha and strong delta activity, respectively. This behavior is consistent with the signatures and predictions of *Sedation* and *LOC* under propofol, reported in section 5.3.2 (*Liege Anesthesia Dataset 1* analysis). Together, these findings indicate the possible role of delta activity as a common marker of unconsciousness, and potentially, the model's inclination towards the acquisition of features reflecting *LOC* (*LOBR* states have not been included during training).

Overall, our results here show the limitations of the model to fully capture relevant electrophysiological features that impartially reflect depths of anesthesia, when trained under propofol and tested under novel anesthetics, such as ketamine and xenon. Of course, understanding the distinctive mechanisms and qualities (phenomenological properties) of unconsciousness across different agents is important to assess our model. In the next section, we explore the possibility of improved generalization using a cross-study and cross-drug training, by incorporating data from both propofol and ketamine studies.

6.5 Experiment 3 – Cross-study and Cross-drug Training on Propofol and Ketamine Anesthesia

6.5.1 Behavioral-Responsiveness (BR) Scale

So far, and throughout the experiments of the previous chapters, we have been training our deep learning model using a single-study/single-agent dataset approach, and mostly under a behavioral ground-truth for assessing levels of unconsciousness (such as the Ramsay scale, found in *Liege Anesthesia Dataset 1*). However, anesthetic states produced by different agents can show differences in their electrophysiological and phenomenological characteristics, despite any common behavioral evidence of unresponsiveness. Therefore, we need to consider the implicit information provided to the model, when moving towards a cross-study/cross-drug training approach.

For this reason, an alternative scale to Ramsay scores was devised and used here ('Behavioral Responsiveness' scale), in order to semantically differentiate the ground-truth, potential feature learning, and predictions of the model, when trained under both states of *LOC* and *LOBR*. As we discussed in the previous chapter, the distinction between *LOBR* and *LOC* forces us to differentiate the electrophysiological signatures of unconsciousness from signatures of disconnected consciousness, often emerging during *LOBR*. While we have strong evidence of unconsciousness during the state of *LOC* in *Liege Anesthesia Dataset 1* and *Liege Anesthesia Dataset 2* (for propofol and xenon, respectively), the anesthetic state of ketamine found in *Liege Anesthesia Dataset 2* (*LOBR*) has been associated with disconnected experiences for all participants (we suspect the same is true for *LOBR* in *Michigan Anesthesia Dataset*, given the comparable dosing, although we do not have direct evidence in the form of retrospective reports). Therefore, when training on both states of *LOC* and *LOBR*, we developed and used a behavioral responsiveness (BR) scale to reflect the *intersection* of the two signatures, which had the same behavioral profile of unresponsiveness, but not necessarily the same state of unconsciousness.

We devised the BR scale in alignment with other scales used in commercial DoA monitors, which index a patient's levels of unconsciousness within the range of 0 – 1 (or 100) (Zanner, Pilge, E. F. Kochs, *et al.* 2009). A direct mapping to Ramsay scores (Table 6.1) can be obtained by using the following formula:

$$BR = 1 - \frac{(Ramsay\ Score - 1)}{5}$$

Table 6.1. Ramsay Sedation Scale and Behavioral Responsiveness

Participant's Response	Ramsay Score	BR Score
Anxious and agitated, restless or both	1	1.0
Co-operative, oriented and tranquil	2	0.8
Responding to commands only	3	0.6
Brisk response to light glabellar tap or loud auditory stimulus	4	0.4
Sluggish response to light glabellar tap or loud auditory stimulus	5	0.2
No response to stimulus	6	0.0

6.5.2 Experiment

In this experiment, we investigate the performance of deep learning in a cross-study and cross-drug training task, by incorporating two datasets with distinct experimental setups and administered anesthetic agents – i.e. propofol and ketamine. Specifically, we use the *Liege Anesthesia Dataset 1* and *Michigan Anesthesia Dataset* to train a regression-to-BR-scores model under the 3-state task, which includes *Wakefulness*, *Sedation* and *LOC*/*LOBR*. The selection of these studies was based upon two criteria. Firstly, both datasets are adequately extensive with respect to anesthetic states and participants, which makes them suitable for training. Secondly, the inclusion of data from propofol and ketamine is appropriate for cross-drug feature learning, given their distinct pharmacological, electrophysiological and phenomenological properties (as we showed in the previous experiment, the model trained with propofol was not able to assess the depth of *LOBR* under ketamine). As previously mentioned, the creation of a representative training dataset is an important aspect for deep learning, which is susceptible to data biases.

The target values used during training were: 0.8 for *Wakefulness* (Ramsay 2), 0.7 for ketamine *Sedation* (Ramsay 2-3), 0.6 for propofol *Sedation* (Ramsay 3), and 0.1 for *LOC*/*LOBR* (Ramsay 5-6), based on the Ramsay scores reported in section 6.2.1. We initially evaluate our training participants using a LOPOCV approach (25 participants; ~ 54,000 epoch instances in total), and compare our results to those obtained by the respective single-study models. We then further test our model on unseen transitional anesthetic states, and on *Liege Anesthesia Dataset 2*, in order to obtain a cross-study and cross-drug generalization error (similarly to our previous experiment).

6.5.3 Results

The results of the cross-drug model are summarized in Fig. 6.9. The mean-squared-error (MSE) and mean-absolute-error (MAE) curves show that the model had a stable convergence during the 10 training epochs. The average cross-validated performance obtained was 0.09 MAE.

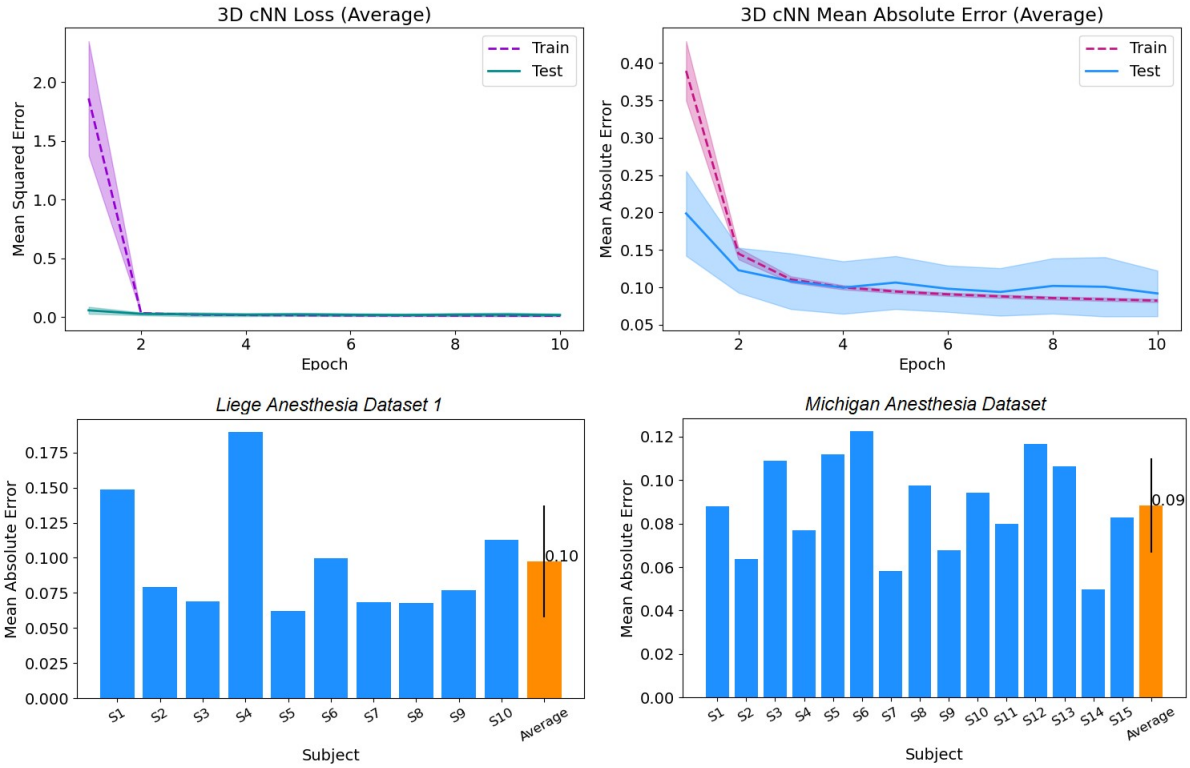

Fig. 6.9. Cross-study and cross-drug training results of the regression-to-BR-scores model under the 3-state task (*Wakefulness*, *Sedation*, *LOC*), for *Liege Anesthesia Dataset 1* and *Michigan Anesthesia Dataset*. Average MSE loss and MAE curves (top). Subject-wise MAE performances (bottom).

As we see from Fig. 6.9, the mean and variance across participants' performances were similar for both datasets, with MAE values ranging from 0.05 to 0.12, except for S1 and S4 in *Liege Anesthesia Dataset 1*, which had the worst performance (already recognized in section 5.3.2, where we postulated a post-hoc analysis explanation). By comparing the average MAEs obtained from the respective single-study trained models, we observed no statistically significant difference in the performance acquired by our mixed model here (0.09 over 0.1 for *Liege Anesthesia Dataset 1*, and 0.11 over 0.09 for *Michigan Anesthesia Dataset*). Moreover, training with the BR scale over Ramsay scores did not affect model performance (as expected, due to their linear mapping), which we can assess by normalizing the MAEs with respect to the target ranges (0.09 MAE_{BR} over 0.10 MAE_{Ramsay}, with $MAE_X = MAE / (X_{max} - X_{min})$).

To further assess the predictive behavior of the model, we visualize the average-across-participants predictions over time, for the two trained datasets (calculated by the minimum number of epochs per state, and aligning them at the beginning of each state) (Fig. 6.10).

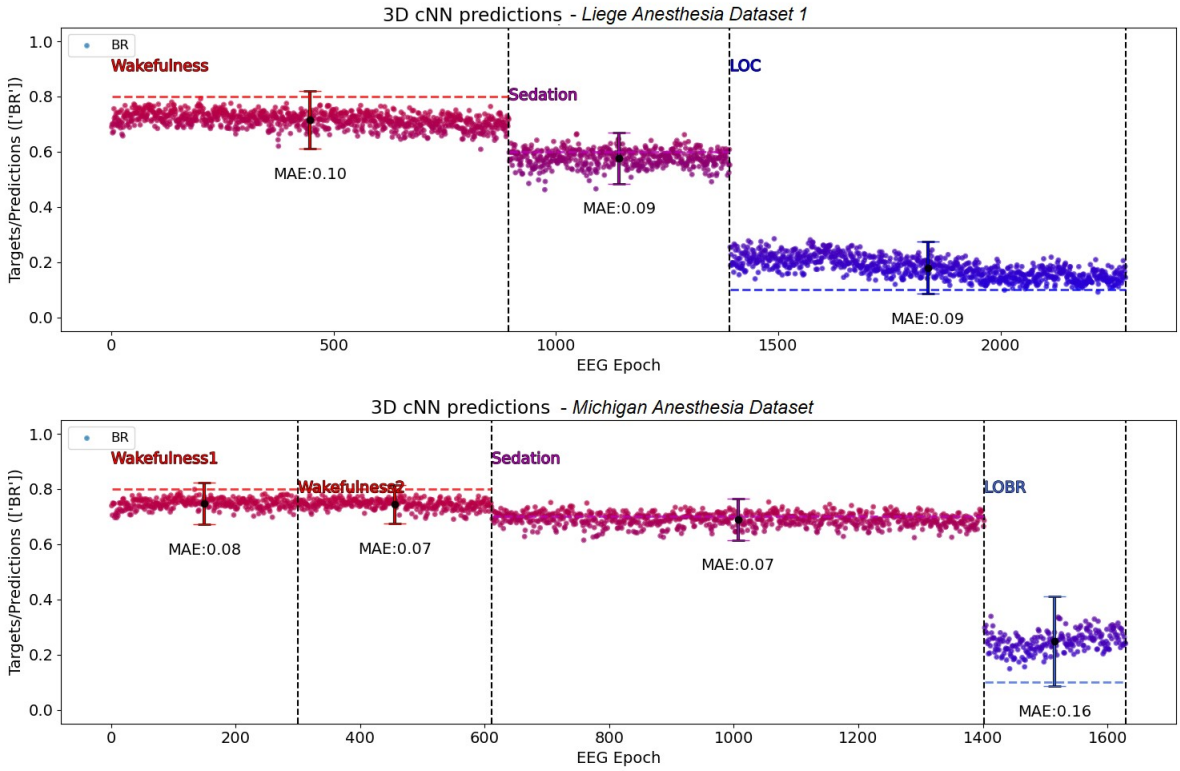

Fig. 6.10. Behavioral responsiveness (BR) predictions for the three anesthetic states of the unseen test subjects. Average predictions for the *Liege Anesthesia Dataset 1* participants (N=10, top). Average predictions for the *Michigan Anesthesia Dataset* participants (N=15, bottom). Horizontal dashed lines indicate the BR score ground-truth.

Fig. 6.10 shows that predictions were robust and consistent over time for both studies, with the largest MAE (0.16) obtained during *LOBR* under ketamine. Most notably though, the model was able to distinguish the states of *Wakefulness* and sub-anesthetic ketamine (*Sedation*) of the *Michigan Anesthesia Dataset*, despite their phenomenological similarities in participants' levels of behavioral responsiveness (sub-anesthetic ketamine can decrease vigilance and increase reaction times, as reported in (Maksimow *et al.* 2006) and (Micallef *et al.* 2002)). This was confirmed statistically by taking a mean prediction of each state per participant, and applying a paired t-test ($p = 0.0003$).

Testing on Unseen Anesthetic States. As a final step to this analysis, we tested our cross-drug model over unseen anesthetic states, which include the transitional states of the two training datasets, as well as the states from the unseen study and agent (xenon) of *Liege Anesthesia Dataset 2*. Fig. 6.11 depicts the average BR predictions of the test subjects which incorporated all anesthetic states (steady and transitional recordings, if available), for each of the three datasets.

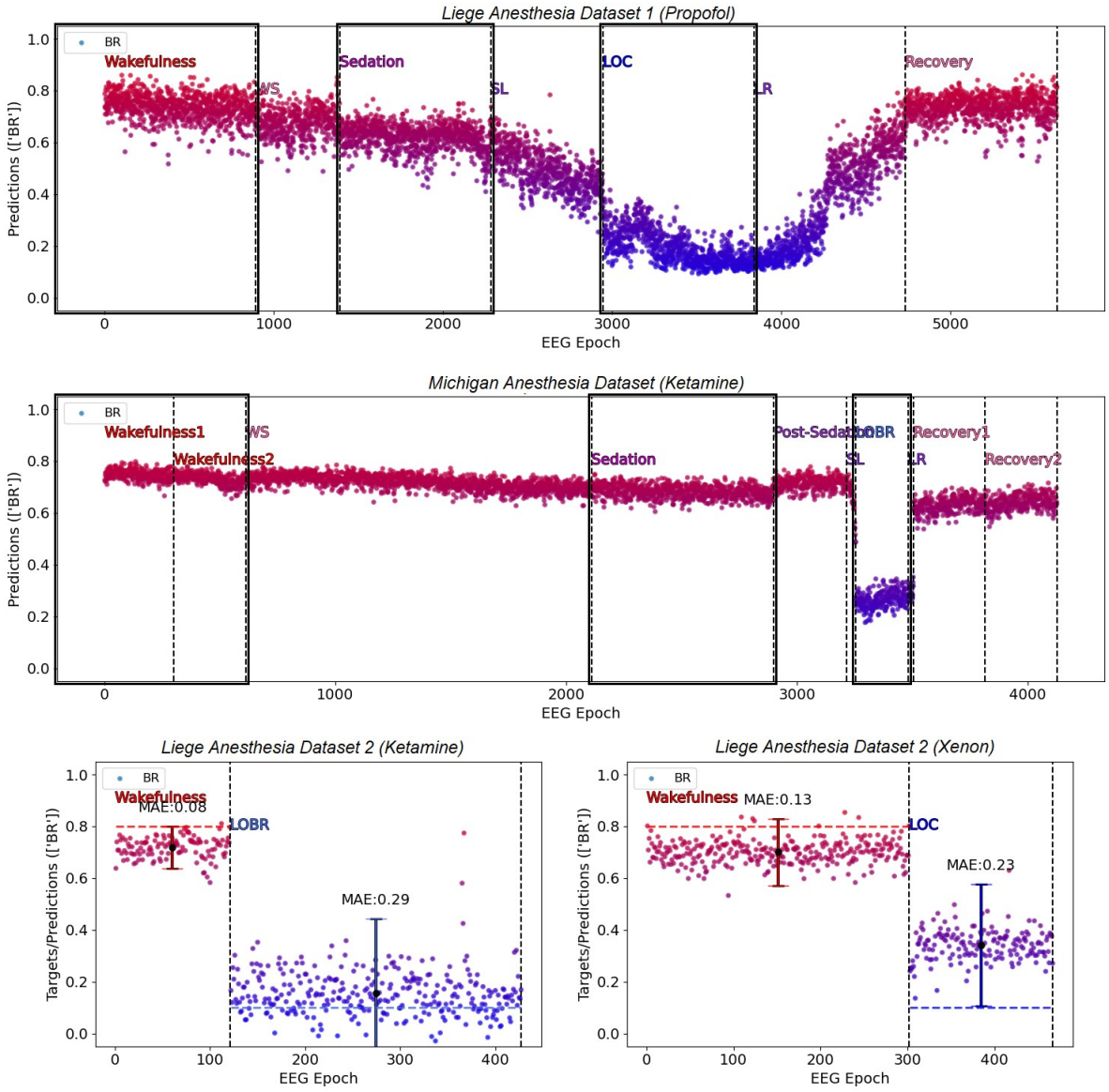

Fig. 6.11. Behavioral responsiveness (BR) predictions of the test participants in *Liege Anesthesia Dataset 1* (N = 3), *Michigan Anesthesia Dataset* (N = 14) and *Liege Anesthesia Dataset 2* (N = 4 for ketamine, N = 5 for xenon) (average). States under training are highlighted in black boxes.

While previously observed for *Liege Anesthesia Dataset 1*, our results here show the ability of the model to track the anesthetic paradigms of both training studies, using behavioral measures of unconsciousness. The performance of our regression model remains reliable, as discussed in Chapter 5, with levels of sedation progressively following the administration of the drugs, both statistically but also within the trajectories of individual participants (see section 5.3.2 for more details).

With regards to *Liege Anesthesia Dataset 2*, we observe a significant performance increase for the ketamine participants during *LOBR*, and a moderate performance increase for

the xenon participants during *LOC*, in comparison to the single-study model in Experiment 2 (0.29 MAE_{BR} over 0.64 MAE_{Ramsay}, and 0.23 MAE_{BR} over 0.34 MAE_{Ramsay}, normalized). In terms of ketamine, while such improvement is expected for the model trained under both propofol and ketamine, it reveals that the cross-study generalization error is quite small (the predicted BR score is even lower for the test state of *Liege Anesthesia Dataset 2*, compared to the trained *LOBR* of *Michigan Anesthesia Study*, which can be explained by the deeper anesthetic depth attained by the continuous infusion of ketamine). In terms of the cross-drug generalization error during *LOC* under xenon, the mixed model performed better than the propofol model in Experiment 2, albeit with some variation over specific subjects. This was evident by comparing the average predictions of the propofol model and the respective mixed model (by converting Ramsay scores to BR scores, or vice versa, using the formula in 6.5.1), for the 5 xenon participants ($p = 0.08$, paired t-test). This reveals the potential capacity of the cross-drug model in acquiring better electrophysiological features, that generalize over novel, unseen anesthetic agents.

Given all the above, we have shown that a cross-study and cross-drug training strategy can be used for improved generalization over novel experimental setups and anesthetic agents, towards an impartial predictive model of anesthetic depths.

6.6 Discussion

The challenge we have addressed in this chapter is relevant for both consciousness research and the investigation of the electrophysiological signatures, but also for clinical practice and the creation of automated systems for monitoring DoA. While there have been several techniques developed for EEG to tackle the above problems (e.g. PCI, Algorithmic complexity indices, SWAS, and others), studies on machine learning-based methods lack the reproducibility found in other works of the field. Given that EEG datasets include noisy and ambiguous data, and as we have already observed significant variations across participants' spectral signatures, the task of generalization to novel experimental designs and drugs can be considered a robust assessment for relevant feature learning. This is particularly important for deep learning models in general, which are over-parameterized, and thus tend to overfit in the presence of noisy samples. Of course, we need to consider our findings within the limits of our experiments and the selected available datasets.

6.6.1 Cross-study Generalization

Our results have highlighted the capacity of the model for cross-study generalization, throughout the experiments of this chapter. In our main cross-study analysis (Experiment 1), we showed that the model trained on *Liege Anesthesia Dataset 1* was able to correctly generalize over the novel experimental setup of *Cambridge Anesthesia Dataset*, by correctly

estimating the Ramsay scores of participants, across anesthetic states. The generalization here stems from the fact that the two studies had significant differences in their experimental designs, most evidently with respect to the anesthetic administration protocol and the assessment of behavioral unresponsiveness.

Regarding the anesthetic administration protocol, an Alaris TCI device (using the Marsh PK model) was employed in both studies to administer propofol to healthy participants, under different titration modes. In the trained study (*Liege Anesthesia Study 1*), an individualized effect-site concentration was targeted and maintained for each participant (TIVA mode with manual adjustments) upon reaching a desired Ramsay score. In the test study (*Cambridge Anesthesia Study*), the TCI device was set to obtain constant plasma concentrations of propofol (plasma-targeting mode), irrespective of the participant's behavioral response. Despite the differences in propofol concentrations across studies and participants (measured either by TCI estimations or blood sample measures), we hypothesize that the model was able to generalize from the EEG an estimate of the effect-site concentration levels and clinical outcome (as already discussed in section 5.6.4), which distinguished the two test groups (plasma concentrations were similar for the two groups). Moreover, the increasing trend of anesthetic depth observed in the predictions of both groups (Fig. 6.6) was in alignment with our observations in Chapter 5, and our hypothesis on the pharmacodynamics of hysteresis (section 5.6.2).

Most notably though, and regarding the behavioral assessment, the two studies had different methodological protocols for measuring levels of unresponsiveness. While the trained model used a ground-truth of Ramsay scores for its optimization (a response to a verbal command asking the subject to squeeze the hand of the investigator), the ground-truth of the test study was based on the *Responsive* and *Drowsy* groups, derived by the hit rate analysis of the auditory discrimination task (a prior independent analysis). Of course, while there is no way to know which features of the data represent the learned subspace from which the model generalizes (interpolates), the training ground-truth acts always as an estimate for the features shared across the analyzed states and participants. In view of this fact, the model was able to correctly estimate the Ramsay scores for each group of participants – Ramsay ~3 at the end of *Moderate Sedation* for the *Responsive* group, and Ramsay ~4 for the *Drowsy* group – in agreement with the increased reaction times and decreased hit rates, reported in (Chennu *et al.* 2016). In addition, the subtle (although statistically significant) group effects found during *Wakefulness* and *Mild Sedation* were in alignment with the findings of weaker alpha band networks at the baseline state of the *Drowsy* group (reported in the original study), which predicted the predisposition of the group towards unresponsiveness.

Further evidence of cross-study generalization was also found in Experiments 2 and 3, where we incorporated the *Liege Anesthesia Dataset 2* as a test set. A generalization error can be estimated by comparing the MAEs of test states with the respective trained states (evaluated under a LOPOCV paradigm), in cases where they were included during training (e.g. during *Wakefulness* in Experiment 2, with 0.45 over 0.29 MAE_{Ramsay}, and during *LOBR* under

ketamine in Experiment 3, with 0.29 over 0.16 MAE_{BR}, normalized). This generalization error could stem from differences in the experimental procedures of *Liege Anesthesia Study 2*, such as the difference in the EEG device (Nexstim eXimia, TMS compatible) and channel configuration (T7, T8, P7, and P8 channels were used instead of T3, T4, T5 and T6, found in the EGI Hydrocel system).

Finally, other cross-study differences affecting model performance (increase in generalization error) could relate to participant demographics, such as the age distributions, which we know to significantly contribute to the spectral characteristics of the EEG signatures (Purdon *et al.* 2015) (e.g. mean age = 22 for *Liege Anesthesia Dataset 1*, and 31 for *Cambridge Anesthesia Dataset*). Overall, our models' features were shown to generalize accurately to novel EEG configurations, and most importantly, to unseen anesthetic protocols and behavioral assessments.

6.6.2 Cross-drug Generalization

Beyond cross-study generalization and based on our findings in Experiments 2 and 3, our model has also shown the potential for cross-drug generalization. Although the single-agent model trained under propofol had limited predictive power on *Liege Anesthesia Dataset 2*, which introduced the agents of ketamine and xenon (Experiment 2), the cross-drug trained model showed the capacity for improved generalization, when tested on the unseen agent of xenon (Experiment 3).

Specifically, the predictive behavior and biases of the propofol model in Experiment 2 revealed limitations, as well as potentially learned features, that could help us better understand the cross-drug generalization problem. When tested on participants under ketamine, predictions remained consistently at a Ramsay score of ~2 during both *Wakefulness* and for most of the duration of *LOBR*, despite the obvious phenomenological differences in behavioral responsiveness. For participants under xenon, although the model could differentiate *Wakefulness* from *LOC* with a higher anesthetic depth (Ramsay ~4 on average), there was significant variability and noise within the group. These results could be viewed and interpreted in light of several known characteristics of propofol, ketamine and xenon. For example, the partial predictability of the model under xenon could be explained by the pharmacological, electrophysiological and phenomenological similarities of propofol and xenon, as both agents act as GABA agonists, they are accompanied by strong delta activity, and have been associated with profound unconsciousness during deep anesthesia. In contrast, ketamine is an NMDA antagonist with distinct features among anesthetics, accompanied by strong gamma activity, and associated with disconnected consciousness during deep anesthesia. While the model had been trained to predict the anesthetic depth of *LOC*, it had not encountered states of *LOBR*, as found in ketamine. Also, the electrophysiological resemblance of ketamine *LOBR* with states of *Wakefulness* has already been discussed in section 5.6.5, while for specific brain regions ketamine has been shown to increase metabolism, cortical connectivity and possibly, levels of

(covert) consciousness (Zacharias *et al.* 2020) (interestingly, the model associated *LOBR* with a slightly decreased anesthetic depth to *Wakefulness*, observed in Fig. 6.8). Meanwhile, the appearance of delta activity as a common feature of behavioral unresponsiveness across all three agents (and correlated to our model's prediction of increased anesthetic depth - albeit sporadically during ketamine *LOBR*), is in agreement with the findings in (Sarasso *et al.* 2015), and has been extensively studied in (Ní Mhuircheartaigh *et al.* 2013; Sleigh *et al.* 2019). Nonetheless, our results in Experiment 2 could indicate a model behavior consistent with adaptations to propofol signatures (such as the delta activity), and thus we cannot presume any generalized learned mechanism.

Following our results in Experiment 3, the cross-study and cross-drug training approach showed the capacity of the model to learn over multiple anesthetic agents, with a potential for improved cross-drug generalization. Initially, by comparing our mixed model with the respective single-study trained models (on *Liege Anesthesia Dataset 1* or *Michigan Anesthesia Dataset*), we ensured that there was no overfitting or degradation of performance within each study, when training with two distinct anesthetics under a common behavioral ground-truth. This is an open challenge in deep learning, as neural networks can be powerful enough to overfit, but often unable to generalize, given the increased complexity of the task. Moreover, we showed that the model was able to consistently track the anesthetic paradigms of both studies, based on the known titration trajectories. Most interestingly though, albeit without a strong statistical significance ($p = 0.08$) (likely due to the limited number of participants, $N = 5$), we found that the cross-drug model improved the generalization error over the unseen state of xenon *LOC* (*Liege Anesthesia Dataset 2*). Despite the inclusion of an agent (ketamine) with divergent properties from propofol and xenon, the model was able to acquire better performance compared to the propofol model found in Experiment 2. This shows the potential of deep learning to discover novel, cross-drug, and generalizable features – instead of simply utilizing multiple features – beyond any drug-specific mechanisms (the electrophysiological signatures of which have already been recognized in the literature).

Finally, these findings can be indicative of the similarities and differences among the various anesthetic agents, in terms of their electrophysiology, but also in terms of changes in levels of consciousness that are captured by the model, but may not be consistent with our behavioral ground-truth (as in the case of covert awareness during ketamine). While we do not have sufficient or definitive evidence of a unitary mechanism across the tested agents, deep learning has shown its effectiveness in utilizing multiple, and likely cross-drug anesthetic features, in order to create a unified predictive model of DoA. This is particularly important from a clinical perspective, due to the weakness of contemporary DoA monitors in performing across anesthetics with independent action (Hans *et al.* 2005).

6.6.3 Summary

Overall, we have shown the capacity of our 3D cNN model to accurately detect progressive, fine-grained changes in levels of unconsciousness, over unseen datasets that incorporate novel anesthetic protocols, EEG devices, behavioral assessments of unresponsiveness, and even anesthetic agents. These findings highlight the robustness and validity of the model to learn appropriate task-related features, but also the potential of deep learning to discover common signatures of unconsciousness. Of course, further testing our discussed hypotheses on the nature of the data and the respective behavior of the model would require extensive analysis on the network and its learned features. In general, understanding the data, the electrophysiological and phenomenological characteristics of each agent, the appropriate computational priors, and the limitations of clinical scales, is crucial for the development and engineering of a successful model of analysis. In the next chapter, we review our findings, we assess the advantages and limitations of our model, and we pose possible future research directions.

Chapter 7 General Discussion

7.1 Overview of Research

In this section, we summarize the research conducted throughout the previous chapters (Chapters 3 – 6), towards the engineering of a model for the investigation and estimation of anesthetic-induced states of unconsciousness. Specifically, we emphasize our initial goals and the respective research gaps, the experiments performed, as well as our main findings and contributions, within the fields of Deep Learning-based EEG (DL-EEG) and General Anesthesia.

7.1.1 Chapter 3 – Deep Learning for EEG Decoding of Anesthetic-Induced Unconsciousness

Our investigation began by exploring the effectiveness of deep learning in extracting relevant electrophysiological features, from the resting-state EEG of healthy participants under general anesthesia. While literature has shown a number of findings and EEG-based techniques for decoding states and levels of unconsciousness, robust neurophysiological markers are still missing. Meanwhile, the theoretical and mathematical assumptions of contemporary methods of analysis constrain a data-driven investigation of novel electrophysiological signatures. This is an area in which deep learning has shown great potential. Moreover, deep learning offers an end-to-end feature learning approach, which is suitable for the creation of automated systems for clinical use.

With these goals in mind, and given the lack of standard models within the recent field of DL-EEG, we compared two widely used deep learning architectures – cNNs and MLPs – alongside the effect of the models' input representation, on their performance under a classification task (with three anesthetic states characterized by decreasing levels of consciousness). Using leave-one-participant-out-cross-validation (LOPOCV), we showed that cNN architectures significantly outperformed MLPs, both in terms of resource utilization (they required a lower number of parameters and training time), but also in terms of their capacity to extract more effective spatio-temporal features. Furthermore, we showed that cNNs were able to achieve high performance using only one-second segments of the raw EEG signals (with minimal pre-processing), without the need for feature extraction, which typically constraints the data space, and thus, the discovery of novel signatures. Overall, our findings highlighted the potential of cNNs to discover and utilize generalizable, cross-subject features (often

overlooked in EEG studies), with implications for both electrophysiological investigation and real-world applications in DoA monitoring.

7.1.2 Chapter 4 – Convolutional Neural Networks and EEG Representation

Following the results from our previous analysis, we further focused on the development of a convolutional neural network with a unified architecture, that allowed us to incorporate different EEG systems and datasets, under a common processing methodology. There were two distinct reasons that led us to this investigation. The first one related to the lack of a standardized model for EEG analysis, which prevents a comparative evaluation of findings and the possibility of integrating heterogeneous datasets. The second reason related to the exploitation of the spatial structure of EEG, given that previous works have dismissed the spatial dynamics of the signals, and hence their role in decoding brain states (especially in anesthesia studies, which often employ a restricted number of electrodes).

After a number of experiments and theoretical considerations, we tested several pre-processing parameters of EEG representation, alongside 2D and 3D cNN architecture designs, on their performance under a common classification task (defined in the previous chapter). Our results showed that the average reference montage and an epoch-wise standardization of the input provides a simple, versatile format for capturing the EEG dynamics with minimal distortions. We also found that model performance significantly varied as a function of the spatial resolution of the EEG, given the increased parameters and noise of high-density configurations, with the 10-20 system achieving an optimal performance. Finally, we showed that the 3D cNN (a design based on the projection of the scalp activity into a 2D map) was robust to EEG artifacts, with artifact cleaning becoming more impactful for high-density systems (in our case, automatic cleaning performed equally to expert manual cleaning). We concluded our analysis by deriving an optimal pre-processing pipeline and a generic 3D cNN model, which we employed throughout our subsequent analyses.

7.1.3 Chapter 5 – Predictive Analysis of Behaviorally, Pharmacologically, and Psychometrically defined Anesthetic States

Having explored the basic parameters of EEG representation and network design, we moved our investigation to the nature of the optimization task undertaken by deep learning. The first goal of this analysis was to understand the EEG data in relation to particular clinical variables and anesthetic depths, defined by behavioral, pharmacological, or psychometrical evidence for consciousness. As none of these measures provide an infallible ground-truth for consciousness, we exploited the strength of deep learning to decode the complex brain states that emerge from the various pharmacokinetic and pharmacodynamic interactions. The second

goal was to test the predictive power of our cNN model under different learning tasks, and reveal an optimal training strategy that impartially captures the electrophysiological dynamics reflecting states and levels of unconsciousness (the optimization strategy has been shown to be one of the most impactful aspects of deep learning).

Using classification and regression algorithms, we performed a number of experiments with an increasing number of behaviorally, pharmacologically, or psychometrically-defined states, and observed the behavior of the model, under propofol and ketamine anesthesia. These observations connected to several theoretical considerations, which are open in GA research. Our results initially revealed the predictive biases of the model, and most evidently the electrophysiological nature of *Recovery* (as a state of mild sedation), which deviated from the assigned behavioral and pharmacological ground-truths. Observations of model predictions over time also revealed the existence of large-scale temporal dynamics, and the transitional nature of EEG, which were consistent with the depth of anesthesia. This behavior allowed us to track the anesthetic paradigm of the studies (in alignment with the drug titration changes), and the direction of transitions to various levels of unconsciousness. Based on these findings, we further reasoned that regression analysis was able to enrich the information of the model, in agreement with our clinical assumptions and research goals. Finally, we discussed the limitations of psychometrical measures for the investigation and estimation of disconnected consciousness. Overall, we showed the ability of our model to progressively detect anesthetic depths with higher granularity than any other contemporary clinical measure (behavioral or pharmacological), under an optimal regression-to-Ramsay-scores strategy (achieving an MAE of 0.5, within the Ramsay scale).

7.1.4 Chapter 6 – Cross-study and Cross-drug Generalization of Anesthetic-induced Unconsciousness

In our last research chapter, we tested the reproducibility of our findings and the robustness of our model in estimating levels of unconsciousness, by employing cross-study and cross-drug generalization tasks. These two notions of generalization were considered important for both technical and theoretical reasons. Models within the EEG and Deep Learning literature have shown replication weaknesses that generally stem from analytic flexibility (in terms of hypotheses or methods (Pavlov *et al.* 2021a)) and the lack of appropriate validation frameworks (e.g. due to non-representative datasets, or our inability to detect and control noise, confounds and spurious correlations (Fellner *et al.* 2016)). From a theoretical perspective, detecting generalized signatures across different anesthetic agents has many implications for research, as we are unaware of any possible common mechanisms and markers of anesthetic-induced unconsciousness.

For this analysis, we used our 3D cNN model and the regression task derived in the previous chapter, to conduct experiments with two test datasets that incorporated a novel

experimental setup, and an unseen anesthetic agent (xenon). Our results highlighted the capacity of our model to extract generalized, task-relevant features, given the differences exhibited by the studies under training and testing (propofol anesthesia). Specifically, we showed that the model was able to correctly estimate the anesthetic depth and clinical outcome of the participants, under an unseen titration protocol, and a prior independent behavioral assessment of unresponsiveness. Moreover, we highlighted the capacity of the model for improved, cross-drug generalization, using a cross-drug training strategy. Although the model trained under propofol showed limited predictive power when tested on ketamine and xenon (most likely due to the pharmacological, electrophysiological and phenomenological differences across the drugs), we found evidence of improved performance for xenon, using our mixed trained model with propofol and ketamine. The fact that such improvement was exhibited by learning from agents with divergent properties to the one tested, points us to the potential of deep learning for discovering common mechanisms of action.

7.2 Deep Learning-based EEG – Assessment and Limitations

In this section, we have a general discussion on the assessment and limitations of our model, in light of the recent Deep Learning-based literature review. Specifically, we evaluate our methodological framework in terms of datasets, EEG pre-processing, deep learning architectures and training methodology, validation and reproducibility schemes, and model interpretability.

7.2.1 Overview

As we have mentioned throughout this thesis, EEG is a complex signal that requires advanced signal processing techniques, feature extraction, and often several years of training, in order to be correctly interpreted. The challenges exhibited during EEG processing (related to its low SNR, non-stationarity, and inter-subject variability) have often led to domain-specific processing pipelines, which nonetheless have shown varying success (Lotte *et al.* 2018). Deep learning has been very successful in many different domains, and has shown great promise for EEG decoding over the past years, as it can simplify this pipeline by automatic end-to-end learning of pre-processing, feature extraction and task-decoding modules (whilst reducing the need for expert knowledge and manual curation).

Of course, the question of whether deep learning has significant advantages over traditional EEG processing approaches, remains open. On the one hand, deep learning can improve and extend the existing methods, by offering hierarchical feature extraction from the raw (or minimally-processed) EEG data, state-of-the-art performances, and the development of various learning tasks (e.g. predictive or generative), towards better generalization. On the other hand, the neuroscientific community has shown skepticism over its use for data analysis.

Notably, deep learning requires large datasets for training, with EEG datasets found in research being typically small compared to other fields (such as computer vision, or natural language processing), as they are more expensive to collect, and often inaccessible due to privacy concerns (especially clinical datasets, with some initiatives trying to tackle this problem).

Here, we evaluate and compare several aspects of model creation and validation, as they have been reviewed in (Roy *et al.* 2019) for 156 studies within the DL-EEG literature (~30% of the studies related to clinical domains, such as epilepsy detection and sleep staging).

7.2.2 Datasets

The availability of large datasets is often mentioned as one of the main enablers of deep learning. Looking at similar studies, the total amount of data used during the training, validation and testing of the models varied significantly across domains of application, both in terms of EEG recording times, but also in terms of the number of extracted samples (instances). Specifically, (Roy *et al.* 2019) found a median dataset duration of 360 minutes, with a median of 14,000 samples (depending on the epoching strategy), across all 156 studies. In our own work, datasets consisted of ~560-600 minutes recordings (with the exception of *Liege Anesthesia Dataset 2* used for testing, at ~100 minutes), with our cross-study/cross-drug analysis acquiring a total of ~1,860 minutes and ~111,600 samples (combining all four datasets). While the size of the epoch windows appears to be guided by the domain under study (with no standard guidelines), no correlation to model performance has been observed (for our research, a small window size was desirable considering the criticality of the temporal resolution in estimating the DoA, as discussed in section 3.4.3).

In terms of the number of subjects, there is also significant variability across studies, with ~50% of datasets containing less than 13 subjects. (Völker *et al.* 2017) tested the impact of the number of subjects under training (from 1 to 30), using a LOPOCV approach, and showed an increase in performance with diminishing returns above 15 subjects. We also observed a similar effect, with 9-10 subjects being adequate for a stable performance (the addition of multiple datasets in Chapter 6 did not reveal any statistical change). Nevertheless, in cases with limited numbers of subjects (as our work here), we empirically found that the quality and nature of the data for any particular subject, was more impactful during training and testing, than the actual number of subjects. Of course, the potential of DL-EEG models reside in the combination of data from multiple subjects and datasets, towards the acquisition of common, generalizable features.

Finally, several studies have explored various data augmentation techniques, in order to increase the dataset sizes. A majority of such studies have used overlapping windows for the extraction of epochs (samples), albeit without significant performance improvements (notably, (Schirrmeister *et al.* 2017) implemented a cropped-training strategy, using sample-wise sliding windows, with no significant effect). In Chapter 3, we also observed that a 50% window overlap – which doubled the number of samples – did not affect model performance. Other

simple techniques, such as adding Gaussian noise to the data, have been shown to be impactful in cases of insufficient data (Wang *et al.* 2018). More recently, few works have tried to use generative adversarial networks (GANs) to produce artificial EEG signals, with one study reporting a 3% accuracy improvement (Wang *et al.* 2018). Lastly, data augmentation can also be useful in cases of classes with limited number of samples (in our own analysis, we showed that sample/target weighting was an effective method for dealing with unbalanced datasets).

Overall, and despite the scarcity of anesthesia datasets, our experiments showed that the dataset sizes, the number of subjects, and our epoching strategy, were all adequate for successful training and testing of the models, throughout our work (in alignment with the current literature).

7.2.3 EEG Pre-Preprocessing

The problem of EEG pre-processing was investigated in detail in Chapter 4, as we were motivated by the lack of standardized processing pipelines, and the general dismissal of the spatial dynamics of EEG in respective models. During the past few years, the EEG community has started to recognize the need for standard pipelines (as with the initiative of *EEGManyPipelines.org* (Pavlov *et al.* 2021b)), and the effects of seemingly subtle differences across processing routines (Robbins *et al.* 2020). Based on our own findings, and as initially hypothesized, the nature of the data, the dimensionality, and the presence of noise, can be significant factors affecting model performance. Given this premise, it is surprising that only 63% of studies report their pre-processing pipelines (Pavlov *et al.* 2021b), with the results becoming incomparable and irreproducible. As we mentioned, one of the main reasons behind DL-EEG analysis, is its potential for automated feature learning, when artifact cleaning and feature engineering are demanding, time-consuming and require expertise. Therefore, it is critical to assess which levels of pre-processing are required for deep learning models. While our analysis tried to systematically investigate the above research questions, it is important to recognize the limitations of our work, based on the evaluation of a classification task under a single dataset (albeit we tried to incorporate theoretical justifications for our results).

Starting with the recording parameters of EEG, the acquisition system and the channel configuration is the first important aspect of processing. In general, there are no established standards for any domain of application, with a large variety of EEG devices used throughout the DL-EEG literature, and with various electrode densities (the number of electrodes varied from 1 to 256, with ~50% of works using between 8 and 62). Nevertheless, few studies have investigated the effect of the spatial density of EEG. Specifically, (Shah *et al.* 2017) and (Chambon *et al.* 2018) showed that increasing the number of channels from 4-6 up to 22 significantly improved their model performance. Moreover, (Schirrmeister *et al.* 2017) found that high-density configurations (128 channels) led to worse performances, compared to configurations with 20-40 channels. Both of these findings are consistent with our results, and the selection of 20 channels (10-20 system) as a preferable channel configuration.

When it comes to further pre-processing steps, different studies have relied more or less on a number of steps and algorithms, typically for EEG filtering, down-sampling, re-referencing and artifact handling. While there is no clear evidence on the specific effects of these methods, (Shah *et al.* 2017) and (Chambon *et al.* 2018) emphasized the role of the reference montage, as part of the comparative evaluation. Besides this, artifact handling is considered one of the most important aspects of pre-processing. (Roy *et al.* 2019) showed that only 1/3 of the studies employed an artifact handling technique, which reveals the capacity of deep learning to perform effectively using the raw data. In cases of artifact handling, some techniques relied more on expert human knowledge (e.g. on visual inspection of the signals, for high-variance segment identification, or amplitude thresholding), while others required prior knowledge for hyperparameter tuning of an algorithm (e.g. ICA). Overall, the non-requirement for artifact cleaning is an important asset for deep learning, given the expertise and time needed for manual curation. Even for cases that artifact cleaning is beneficial (as we showed with HD sets, which tend to be noisier), our results indicated that simple automatic procedures can perform as effectively as manual curation.

A final aspect of pre-processing, before feeding the data to the neural networks, can be a feature extraction step. A review across the DL-EEG literature revealed that almost half of the studies used the raw EEG signals as input to the models, with the remaining using frequency domain (36%) or other types of features (Roy *et al.* 2019). In Chapters 3 and 4, we already argued that the raw representation of the EEG outperformed a PSD representation for our task, whilst also being preferable for a data-driven investigation of novel electrophysiological features (non-constrained from a Fourier-type spectral decomposition). More recently, (Cho and Jang 2020) made an explicit comparison of the raw signals against Fourier-type representations under a seizure detection task, showing the use of cNNs with the raw EEG as the optimal configuration. Given that feature engineering is one of the most demanding steps of traditional EEG processing, and the majority of the models performed effectively without it, we have strong evidence for the capacity of deep learning for robust feature extraction.

7.2.4 Deep Learning Architecture and Training

One of the most crucial aspects in all deep learning studies, is the choice of the neural network architecture and its relation to the characteristics of the input data. Since our own analysis in Chapter 3, several studies have been published reporting similar findings, regarding the application of convolutional neural networks to raw EEG signals, as a successful architecture and input modality for EEG decoding (Heilmeyer *et al.* 2019; Roy *et al.* 2019; Cho and Jang 2020). These results are also evident by the significant growth of DL-EEG studies since 2015, with an increasing proportion employing cNNs (>50%). The success behind this architecture has been hypothesized to lie in their ability to effectively train feature extraction and task-decoding modules simultaneously, contrary to previously used models that were mostly effective under a two-step procedure (independent feature learning, alongside MLPs,

RBMs or DBNs). This is in alignment with our results in Chapter 3, where MLPs performed adequately given a PSD representation of the input. Despite the fact that cNNs do not explicitly take into account time dependencies (as RNNs), recent findings have shown their effectiveness in processing time series (Bai, Kolter and Koltun 2018).

This ability for simultaneous feature extraction is generally attributed to the imposed hierarchical processing of deep learning models. Although there is no absolute definition or rule regarding the number of layers within a cNN (or, how "deep" the networks need to be), the majority of studies utilized architectures with at most 10 layers (2-3 layers was the most common depth). While few of these studies tried to explicitly investigate the effect of the number of layers in the performance of the models, there was no consensus on the results (O'Shea *et al.* 2017; Kwak, Müller and Lee 2017) . In general, some of the most widely employed cNN models of the previous years have utilized 2-5 layers (Heilmeyer *et al.* 2019), suggesting that shallower models are more appropriate for EEG processing, compared to state-of-the-art models found in computer vision (currently utilizing 16-20 layers).

Besides the above-described aspects of architectural design, the training strategy and the optimization process is also of great impact to deep learning. As previously discussed, the vulnerabilities of neural networks to high-dimensional patterns and their susceptibility to overfitting, make the selection of the algorithms particularly important in determining the generalization performance. In Chapter 5, we explicitly investigated a number of learning tasks and ground-truths, in order to understand our data, and formulate an appropriate training strategy. However, there is only one study we know that has investigated the effect of training algorithms on the outcomes of EEG feature learning (Stober *et al.* 2015). The majority of DL-EEG studies have employed classification algorithms under a variety of tasks, depending on the domain of application (hence, it would be impossible to have a comparative evaluation). From an engineering perspective, almost all studies have incorporated at least one form of regularization to the networks, such as Dropout or L1/L2 weighting. Regarding optimization, the Adam and the Adadelta optimizers have most often been utilized, as generally effective algorithms, due to their adaptive learning-rate tuning, and their fast convergence (empirically, we found that Adadelta had a more stable behavior).

Finally, other methodological priors can play a role in model performance, which can be driven by EEG-specific characteristics. Deep learning models have shown great performance by mirroring the shape of information in the data, through their architectural design. In this work, we have only investigated the basic parameters of EEG representation, as we prioritized the minimization – to the degree it was possible – of the imposed computational assumptions (discussed in Chapter 4). Nevertheless, by imposing further mathematical assumptions on the nature of EEG and the respective feature extraction, we can significantly constrain the underlying learned function, towards a more appropriate (robust) solution (we have already set prior constrains on the compositionality of the electrophysiological features and their spatio-temporal invariance). For example, a common idea found across several studies was the separation of the temporal and spatial information processing, during the first

layers of the models (similar to our 3D cNN design) (Roy *et al.* 2019). Such behavior often resembles filter-bank or common-spatial-pattern processing stages, by incorporating layer regularization and constrains on convolutional kernels/activation functions.

Given all the above, the hyperparameter search space can be large and time-consuming for the optimization (tuning) of the models. Many of the DL-EEG studies under review have not reported such model tuning, with a large proportion discussing model performance after several trial-and-error attempts of alternative design hyperparameters (we explicitly avoided such hyperparameter optimization, in order to avoid information leakage from our dataset evaluation into model architecture). However, deep learning literature has shown that particular combinations of hyperparameters can perform better or worse, irrespective of their individual performance. For this reason, a very cautious analysis is required, in order to understand the effects of such design choices, and ideally in alignment with ideas found within contemporary EEG analysis, aiming at an improved model interpretability.

7.2.5 Validation and Reproducibility

One of the main questions initially posed, was whether deep learning-based EEG models have significant advantages over traditional EEG processing pipelines. While DL-EEG is a new field and has not established standard benchmark datasets or validation frameworks, and given that methodology varies significantly across studies (as we discussed in Chapter 3), we can estimate an answer based on the reported results of similar studies outside our domain of analysis. Many researchers have conducted experiments to compare the performance of deep learning models with established state-of-the-art EEG processing methods (often incorporating feature extraction and machine learning techniques), used to solve a particular task. Based on an analysis made in (Roy *et al.* 2019), the majority of the studies reported a performance increase when they employed deep learning techniques (>95%), with a median accuracy gain of 5.4% against the current state-of-the-art, across all kinds of EEG application domains (>100 studies, including clinical domains). This finding provides us with strong evidence over the capacity and flexibility of deep learning models, to achieve state-of-the-art performance in EEG decoding.

Of course, while within-study comparisons are sensible under a common framework of analysis (i.e. datasets, methods, and validation), large-scale cross-study comparisons are still limited by the variation in methodology (and hence, findings are limited by their statistical power). One of the main problems recognized and discussed in Chapter 3, was the variation in validation procedures and the respective performance metrics. Machine learning models are always evaluated on a measure of generalization performance (i.e. how well they perform on unseen data), typically by employing a cross-validation paradigm, the selection of which can have a profound effect on the results (especially in EEG, where signals can show significant variability across subjects and experiments). Given this fact, it is important to note that more

than 60% of studies have not reported the use of any form of cross-validation (Roy *et al.* 2019), and thus the evaluation of many findings remains highly restricted.

In general, EEG studies have focused on either intra-subject or inter-subject decoding tasks (depending on whether the model has been trained and tested on a particular subject, or across many subjects), which can significantly differ with respect to their expected performance. (Turner *et al.* 2017) and (Hajinoroozi *et al.* 2016) showed that intra-subject decoding always provided better performance than inter-subject decoding tasks, for several deep learning models and in spite of the significant decrease in training data. (Deiss *et al.* 2018) observed a decrease of accuracy from 75% to 38% when using cross-subject validation (inter-subject training, but testing solely on unseen subjects), under a seizure detection task. These results are in agreement with our review on previous DL-EEG studies on anesthesia (see section 3.1.3), with the 3-state classification task solved almost perfectly using an intra-subject validation approach, contrary to our performance (~87% accuracy). However, while intra-subject validation can be useful in certain tasks, cross-subject approaches are more applicable to real-life scenarios, and provide us with stronger evidence of generalization performance (as deep learning is susceptible in capturing subject-specific signatures, unrelated to the task). Over the past years, there is a general paradigm shift in machine learning for more systematic and robust methods of evaluation, which is also evident in the DL-EEG literature, as an increasing number of studies appear to adopt cross-subject validation frameworks (albeit it is still ~20%) (Roy *et al.* 2019). In our work, and for our selected tasks, we showed that deep learning models were capable of cross-subject generalization. For this reason, we proceeded to compare our findings only with anesthesia studies incorporating cross-subject validation approaches (Leave-N-Subjects-Out).

Finally, the validity of any model eventually rests on the reproducibility of its results, which is a cornerstone of science and fundamental for the field to move forward. To date, the reproducibility of the majority of DL-EEG studies is still limited by the lack of standardized methods, open datasets, and code accessibility. However, during the past few years, the recognition of the EEG community regarding the need for large-scale studies that enable the replicability of findings, has led to several collaboration efforts (e.g. with the initiatives of *EEGManyLabs* (Pavlov *et al.* 2021b), *OpenML* (Vanschoren *et al.* 2014), and *MOABB* (Jayaram and Barachant 2018)). Currently, about half of the DL-EEG studies have utilized public datasets, and especially within the clinical domains (such as sleep and epilepsy), where data are more difficult to acquire (Roy *et al.* 2019). Of course, the availability and sharing of anesthesia datasets, which tend to be more expensive to collect, is still a major restriction for our research. In this work, we explicitly moved towards the acquisition of a standard pipeline that allowed us to integrate various anesthesia studies, compare our findings, and reproduce our results (Chapters 4 and 6). Our analysis provided evidence for the ability of deep learning to perform effectively under various experimental setups, towards improved cross-study generalization.

7.2.6 Model Inspection and Interpretability

Another major way to validate the performance of deep learning models and compare findings – with respect to other DL-EEG models or traditional EEG processing pipelines – is by methods of model inspection and interpretability. From this perspective, our research has focused on investigating the behavior of the models under different learning tasks and anesthetic ground-truths (Chapter 5), mainly by inspecting their predictions over time, whilst assessed over particular clinical measures. Nevertheless, the validity of our observations can be limited by our current understanding of the clinical variables characterizing our EEG data, and several post-hoc interpretations, which can mislead our findings (the results of many studies within consciousness research have been generally guided by prior theoretical assumptions and post-hoc explanations, as indicated in (Yaron *et al.* 2022)).

Alternatively, models can be inspected and interpreted internally (in terms of the layers prior to the output), both for the discovery of novel features that have been extracted for a given task, but also for understanding how these features are utilized by the models. During the past few years, researchers have shown increased interest in the field of AI interpretability and explainability (XAI). This has led to the development of many techniques that aim to understand model behavior, either in terms of causal relationships (between inputs and outputs), or in terms of human interpretability and understanding (especially for deep learning models, which are considered as 'black boxes'). Specifically within the field of DL-EEG, several studies have tried to apply interpretability techniques, such as weight analysis, analysis of activations, input-perturbation network-prediction correlation maps, and others, with varying success (Roy *et al.* 2019). Interestingly, (Lawhern *et al.* 2018) was able to find spectral features, topomaps, and ERPs that corresponded to existing knowledge within the literature of the paradigm under study, using a combination of such methods.

Given that neural networks can easily diverge into learning spurious patterns, algorithmic robustness (e.g. via dataset curation, regularization, and cross-validation) and interpretable models can be crucial for the transparency and trustworthiness of deep learning. However, current approaches of interpretability present several limitations with respect to their explainability of various phenomena (Ghassemi, Oakden-Rayner and Beam 2021). This is particularly important within clinical domains, where humans and machines are competing or co-operating for the decision-making of patients' management. Explainable models are often required by medical professionals and patients in order to be trusted, considering the ethical and societal implications. Overall, further research is required in our work, in order to explore the effects of the layers, the hyperparameter design, and the extraction of features with potential neuroscientific or clinical interest.

7.3 EEG Methods for Analysis and Estimation of Anesthetic-Induced Unconsciousness

In this section, we discuss a number of EEG methods that have been developed over the past years, for the analysis and estimation of levels of consciousness and the depth of anesthesia. Specifically, we compare our model to several theoretically- and empirically-driven techniques, as well as to other deep learning models with the same objective, in terms of their advantages and limitations for research and clinical purposes.

7.3.1 Overview

As initially stated in the introduction of this thesis, our understanding of the neurophysiological signatures of the full neural correlates of consciousness (full NCC) remains incomplete (Koch *et al.* 2016). The lack of robust EEG markers specifically, has driven many researchers to investigate the electrophysiological correlates of states and levels of unconsciousness under general anesthesia, using a variety of methodologies. However, no method of analysis has been universally successful across all anesthetic conditions, so far. At the same time, medical practice has a fundamental shortcoming in reliably assessing levels of consciousness, using behavioral, pharmacological, and physiological measures. Despite the fact that GA mainly targets the brain, commercial EEG devices for DoA monitoring have shown faulty performance under particular anesthetic agents and depths (Barr *et al.* 1999; Kearse *et al.* 1994), as well as susceptibility to EEG artifacts and the co-administration of other drugs (Schuller *et al.* 2015). This has led to the more recent development of novel techniques for real-time monitoring of the DoA. Here, we highlight the differences across the most prominent non-learning-based and deep learning-based methods.

7.3.2 Non-Learning-Based Models

In Chapter 6 (section 6.1.1), we mentioned several theoretically- and empirically-based EEG metrics that have been widely investigated in similar studies, and which have shown adequate generalization across various anesthetic conditions. Notably, the perturbational complexity index (PCI) (Casali *et al.* 2013), the permutation Lempel-Ziv complexity (PLZC) (Bai *et al.* 2015), various entropy measures (Liang *et al.* 2015), the slow-wave activity saturation (SWAS) (Mhuircheartaigh *et al.* 2013b), and the spectral exponent (Colombo *et al.* 2019), have all been extensively tested, and can be in principle used to progressively track levels of unconsciousness or the DoA. For the purpose of this discussion, we further evaluate the characteristics of these techniques (beyond their generalizability), and other methods used

in anesthesia research, with respect to their underlying assumptions, their capacity for electrophysiological investigation, and their potential for improved brain monitoring.

From a theoretical perspective, these metrics generally rely on the quantification of the information or spectral content of EEG, the spatial extent and synchronization of brain activity, or a mixture of these two. However, the exact approach of each technique is based on particular mathematical and theoretical assumptions on the nature of EEG and/or the nature of consciousness. For example, PCI presumes that consciousness requires the differentiation and integration of cortical activity (as per the integrated information theory (Oizumi, Albantakis and Tononi 2014)), and that a measure of these two can be mathematically approximated by the compressibility of the EEG response to a TMS perturbation. Similarly, Lempel-Ziv complexity and entropy measures presume that consciousness require a level of complexity, which can be estimated by algorithmic complexity or information-theoretic measures in EEG signals. SWAS relies on the observation that many anesthetic agents favour the hyperpolarization of cortical neurons (oscillating at low frequencies, ~1 Hz), a behavior associated to the thalamocortical and cortico-cortical communication barrier hypothesis. Finally, the spectral exponent exploits the spectral signatures of the signals and the association of unconsciousness to EEG slowing. Contrary to these techniques, deep learning models can extract a variety of task-related features, within the constraints imposed by the architectural design of the networks. Specifically for our own 3D cNN design, while we make no electrophysiological assumptions on the nature of consciousness, our computational priors presume a compositionality of EEG features, characterized by local invariance across space and time (this assumption can be weaker for the spatial dynamics of EEG).

When it comes to investigating the electrophysiological correlates of anesthetic states, theoretically- and empirically-driven techniques have several advantages and limitations, compared to learning-based methods. For example, the results obtained by PCI and SWAS have been further supported by independent findings of studies on the respective theories and hypotheses (e.g. studies on the IIT theory, or the thalamocortical disruption hypothesis). In general, the lower analytic flexibility in hypotheses and methods of these techniques is a significant aspect for the validity of the results, compared to deep learning, which as discussed is vulnerable to false pattern recognition (and thus requires a careful understanding of the data and task under analysis). On the other hand, many of the above-mentioned techniques are limited by their interpretability (e.g. the complexity and entropy measures), and have been evaluated by post-hoc analysis approaches (e.g. studies on Lempel-Ziv complexity and SWAS have evaluated the metrics against PK/PD models which incorporate parameters of EEG response). Most importantly though, non-learning-based metrics do not allow for a data-driven discovery of novel electrophysiological signatures. This is an area where deep learning has shown great success in the past years (methods of model inspection have been discussed in section 7.2.6).

From a practical or clinical perspective, deep learning has shown the strongest advantages for tracking levels of consciousness and the creation of DoA monitoring systems.

As we discussed in section 6.1.1, several of the previously mentioned metrics have shown their ability to discriminate states of consciousness from unconsciousness, across agents like propofol, xenon and ketamine. Nevertheless, there are significant limitations when it comes to their sensitivity and real-time automated analysis. In particular, PCI, SWAS, and the spectral exponent cannot differentiate intermediate levels of sedation, while Lempel-Ziv complexity and entropy measures are insensitive to small drug concentration changes, yet exhibit large within-state fluctuations (as they are sensitive to random processes and locally generated patterns). Furthermore, the performance of each technique vary across the different phases of anesthesia (e.g. SWAS accounts for the transition to *LOC*, but is not sufficient to explain the opposite transition to *Recovery*). With regards to automation, all of the mentioned studies have used data curation and manual intervention, either for artifact handling, or due to an algorithmic requirement (e.g. utilizing ICA or source modelling techniques). PCI also requires the use of TMS/EEG, which creates further complications in clinical settings. All of these aspects are major limitations for the creation of automated systems for real-time analysis. Finally, it is important to differentiate the reported results from group-level performances, with an emphasis on the ability of any metric to detect changes in individual participants (which is also a clinical end-goal). In this work, we have shown the capacity of deep learning to improve across all of the above issues.

Overall, beyond these theoretically- and empirically-based techniques, other methods of EEG analysis have also significantly contributed to anesthesia research. Most notably, methods of functional/effective connectivity (FC/EC) are often employed to assess how different regions of the brain interact under GA. To this end, many FC/EC techniques have been developed that rely on a variety of underlying assumptions, in terms of temporal correlations, causal interactions, and the directionality of information flow (data-driven or model-based methods). For example, (Lee *et al.* 2015) showed the disruption of the fronto-parietal feedback connectivity under GA, using transfer entropy measures on scalp EEG (information-based estimation of directed FC). This finding has been robustly found in studies including source-level analysis and model-based EC methods as well, across propofol, ketamine and other inhaled anesthetics (e.g. using dynamic causal modelling in (Boly *et al.* 2012; Muthukumaraswamy *et al.* 2015)). Other methods, exploring the connectivity of the brain using graph theoretical measures, have revealed properties such as increased local efficiency and decreased small-world properties for states under GA (Bonhomme *et al.* 2019). Of course, many of these methods require extensive EEG recordings and computational requirements (such as source reconstruction) to be applied, which makes them unsuitable for real-time analysis. Nevertheless, deep learning models can be informed by such findings, either by exploiting relevant areas and mechanisms of the brain under GA, or even by incorporating features that would otherwise be hardly extracted by the networks (e.g. measures of FC/EC, or even the utilization of architectures that handle such information, such as graph neural networks (Wu *et al.* 2019)).

7.3.3 Deep Learning Models

Over the past decade, machine learning has shown fundamental breakthroughs, both in scientific discovery, but also in the creation of state-of-the-art predictive models, which have been partly driven by advances in deep learning techniques. Specifically within biomedical engineering, deep learning models have been increasingly developed with an aim to remove the inherent subjectivity of the medical decision-making, automate certain tasks and services, and provide improved personalized diagnosis and treatment. However, predictive models do not always live up to the hype. While there is a lot of research on how to build and train different algorithms, we do not have a framework for the questions we want the algorithms to answer (especially in medicine, where we often lack straight-forward definitions and ground-truths). This problem is evident in our research, as we have several proxy measures of consciousness, but a limited understanding of the data and the behavior of the algorithms, which introduce a variety of biases. In addition, the obscurity of the decision making leads to further concerns about the appropriateness and trustworthiness of the models, which creates the need for extensive empirical evaluation. In Chapters 5 and 6, we explicitly tried to tackle each of these issues, in order to instill confidence in our findings.

In section 5.1.3, we mentioned a number of studies using deep learning-based EEG methods for the analysis and estimation of the DoA. Given that a comparative evaluation of the models was limited by the heterogeneity found in the anesthetic and methodological protocols, we proceeded to investigate several basic factors concerning the learning algorithms and the clinical ground-truths, in order to understand their effects on model performance. Our findings – aligned with the state-of-the-art literature described in section 5.6.6 – showed that models trained under a regression analysis over behavioral measures of unconsciousness, exhibited the best performance in terms of our clinical understanding and research end-goal. Here, we focus on the comparison of our work with studies incorporating cross-subject validation approaches, considering the problems discussed in section 7.2.5. Based on the above criteria, (Sun *et al.* 2019a) acquired a model with the best performance, which is comparable to our learning objective. However, since the majority of research – including (Sun *et al.* 2019a) – has focused on patient data under clinical settings, our evaluation here is constrained mainly within the capacity of the models for real-time estimation of the DoA (rather than an electrophysiological investigation of anesthetic states, which is easier under controlled experimental settings).

In order to understand the advantages and limitations of our DL-EEG model and the model found in (Sun *et al.* 2019b), we compare the anesthetic and methodological protocols used in each work. As we mentioned in section 5.6.6, (Sun *et al.* 2019b) developed a cNN model trained under a regression over Richmond Agitation-Sedation Scale (RASS) scores, achieving an average MAE_{RASS} of 0.2 (normalized, with $MAE_X = MAE / (X_{max} - X_{min})$). Similarly, our cross-study and cross-drug model described in section 6.5, trained under a regression over BR scores (Ramsay), acquired an average MAE_{BR} of 0.1 (normalized). In both

analyses, the models showed their feasibility for tracking levels of consciousness within a similar range of sedation (which included intermediate sub-anesthetic states), with the predicted values naturally interpreted, after discretization and mapping onto the respective scales (Ramsay or RASS scores). Beyond this framework, there were several differences in our works. Most importantly, (Sun *et al.* 2019b) used data from a significant number of ICU patients (174, with exclusion of patients after visual inspection of the EEG quality), which had been administered a variety of intravenous anesthetics and other types of medications (76% propofol, 20% dexmedetomidine and 4% ketamine, co-administered with opioids such as hydromorphone and fentanyl, and benzodiazepines, such as midazolam). Also, the model employed a larger window of analysis (4 sec), with the EEG data band-pass filtered from 0.5 - 20 Hz (higher frequencies were possibly filtered to avoid artifact contamination). Finally, the authors tested a number of cNN architectures, with minimal hyperparameter optimization, incorporating RNNs, spectrogram and band-power representations, with the cNN and cNN/RNN designs applied to the raw EEG obtaining the best performance.

Based on a rudimentary comparison of the main differences, we can detect several strengths on each model and approach. Most notably, the significant amount of clinical data in (Sun *et al.* 2019b), which spanned many hours for each subject, contributes to a robust validation of the model, but also to the clinical relevance of its performance. This was further supported by the use of a minimal number of EEG channels (using 4 frontal electrodes), which are important for clinical preparation. Of course, the inclusion of multiple agents increases the difficulty of the tasks under training, especially given the unknown interactions across the combination of drugs (albeit propofol is electrophysiologically closer to dexmedetomidine than ketamine, due to the strong presence of slow-delta oscillations and spindles (Patrick L. Purdon *et al.* 2015)). On the other hand, the methodology found in (Sun *et al.* 2019b) was significantly constrained by the sparsity of the RASS score assessment. Specifically, the authors assigned RASS scores to EEG segments spanning 1 hour, due to temporal ambiguities, despite the fact drug concentrations are typically adjusted manually during the procedure (contrary to our datasets, which were designed to maintain pharmacologically-steady states, typically upon attaining a targeted behavioral response). Moreover, the use of a large input window, alongside the use of RNNs (which acted as a smoothing function), created a delay in model response of 0.6 – 6 minutes, which can be significant for GA management (whereas our model exhibited stable performance, without response delays). Both of these problems potentially contributed to the large scale dynamics observed, which deviated from the ground-truth and could explain the high variance in patient performance (our model's lower MAE can likely be attributed to the quality of the recordings and the more robust labelling of the EEG epochs).

Overall, future research directions could be extended to both aspects of dataset acquisition and model improvement. For example, the idea to incorporate information over a number of consecutive epochs within our model's design (in analogy to the function of the RNN employed in (Sun *et al.* 2019b)), could be further explored with respect to model dynamics and performance. Moreover, training a model under differentiated targets of arousal

and awareness could be beneficial for feature learning, given the desired dissociation of the two dimensions (this would be relevant for tracking disconnected consciousness, or even connected consciousness in clinical settings with neuromuscular blocking agents. An experimentation with our model trained under both BR and DC scores did not reveal significant changes over our psychometrical analysis of Chapter 5). Another prospective analysis for model improvement could relate to representation learning and cNN interpretability, which are mutually connected (e.g. using convolutional sparse coding with rank-1 constraints, as in (La Tour *et al.* 2018)). Of course, the expansion of datasets that incorporate more anesthetic agents (including dissociated states) and ICU patients are crucial for the clinical translation of our findings. Such analysis could also expand to other state-based paradigms (e.g. sleep, or patients with DoC), towards the acquisition of signatures reflecting the full NCC.

7.4 Conclusions

Our thesis began by introducing the scientific endeavour towards a neuroscientific understanding of consciousness, which has many implications for science, medicine and society in general. In this work, we set the basis for the development of a model that allows us to investigate and estimate anesthetic-induced states and levels of unconsciousness. Given the absence of strong theories and markers of consciousness, we focused on a data-driven analysis of the electrophysiological signatures under general anesthesia, by incorporating state-based paradigms and minimal assumptions from current clinical DoA measures.

While EEG provides a simple, non-invasive, and widely accessible way to record the brain activity of individuals and patients in hospitals, the complexity of the signals has presented many limitations in the past, with respect to the computational approaches and expertise required for its interpretation. Nevertheless, deep learning techniques have recently shown great promise for EEG decoding tasks, as they offer automatic end-to-end feature learning. Here, we demonstrated the effectiveness of deep learning in discovering and utilizing relevant electrophysiological features that characterize the different states and levels of unconsciousness. Specifically, we have presented a realization of a unified cNN architecture and a standard pre-processing pipeline, that allows us to analyze the spatio-temporal structure of the raw data, without the need for manual curation (artifact handling) or prior knowledge of the signals (or features of interest). Our results have been further supported by the recent DL-EEG literature, with similar findings on cNN models acquiring state-of-the-art performance.

We have also demonstrated the capacity of the models to tackle problems of contemporary GA research. In particular, our 3D cNN was able to enrich the information of behavioral measures under a regression analysis (and correct for ground-truth inaccuracies), which provided us with a predictive model that sensitively discriminated a wide range of anesthetic depths. This was confirmed by our ability to track the electrophysiological changes reflecting the drug concentrations (which followed the anesthetic protocols), the transitional

dynamics of EEG (consistently with the observed DoA), as well as the clinical outcomes of individuals. We further validated these findings using an independent behavioral assessment paradigm, and across the anesthetic agents of propofol, ketamine and xenon. Lastly, we found preliminary evidence for common cross-drug signatures of unconsciousness.

As we hypothesized in the Introduction, the development of such model can also contribute to the production of improved, real-time, automated DoA monitoring systems, which can advance patients' safety. The problems and complications found in contemporary anesthesia practice, stem from the imbalanced administration of medications, due to our inability in reliably estimating the individualized anesthetic effects. In this work, we showed the ability of our model to progressively detect the anesthetic effects in the brain, with higher granularity than both behavioral and pharmacological indices, using only 1-second segments of the EEG. Additionally, we reasoned for the advantages of our model over alternative methods (theoretically- or learning-based), in terms of knowledge discovery and their capacity for real-time EEG analysis.

Of course, while we have empirical evidence for the performance of our model, various challenges remain open. Further research can be conducted in our work, mainly with regards to the role of specific hyperparameters of our 3D cNN architecture design, and their interactions to model performance and interpretability. Such analysis could further promote the discovery of knowledge with neuroscientific or clinical interest, towards the transparency and trustworthiness of deep learning.

Appendix

A1. The propofol concentrations, as measured from the blood samples of the participants in the *Cambridge Anesthesia Dataset* (section 5.2.1).

Table A1. Drug concentrations (ng/ml) for *Mild Sedation*, *Moderate Sedation* and *Recovery*

Subjects	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
Mild Sedation	204	246	525	878	604	482	311	542	144	529
Moderate Sedation	506	689	1032	1521	1437	947	806	1149	433	1167
Recovery	299	224	236	483	351	266	312	397	243	363
Subjects	S11	S12	S13	S14	S15	S16	S17	S18	S19	S20
Mild Sedation	200	385	482	723	493	394	272	712	394	394
Moderate Sedation	555	1018	1029	791	978	800	749	800	795	800
Recovery	197	171	287	303	226	265	207	395	148	435

A2. The *lifetime*, *study* and *relative-change* scores of the participants in the *Michigan Anesthesia Dataset* (section 5.2.1). Each score is calculated as the average of the three subscales - *disembodiment*, *transcendence of time and space*, and *complex imagery* – and represents a different ground-truth for measuring levels of disconnected consciousness (DC).

Table A2. The psychometric scores of the three ground-truth measures

Subjects	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15
Lifetime Score	0.32	0.10	0.14	0.20	0.22	0.30	0.18	0.28	0.07	0.13	0.06	0.00	0.14	0.05	0.10
Study Score	0.79	0.85	0.45	0.66	0.59	0.85	0.85	0.63	0.93	0.48	0.83	0.77	0.58	0.37	0.83
Relative-Change	0.47	0.75	0.31	0.46	0.57	0.54	0.67	0.34	0.86	0.35	0.77	0.77	0.44	0.32	0.73

References

Absalom, A.R. *et al.* (2009). Pharmacokinetic models for propofol- Defining and illuminating the devil in the detail. *British Journal of Anaesthesia* [Online] **103**:26–37. Available at: http://dx.doi.org/10.1093/bja/aep143.

Acharya, J.N. *et al.* (2016). American Clinical Neurophysiology Society Guideline 3: A Proposal for Standard Montages to Be Used in Clinical EEG. *The Neurodiagnostic Journal* [Online] **56**:253–260. Available at: https://www.tandfonline.com/doi/full/10.1080/21646821.2016.1245559.

Alkanhal, I., Kumar, B.V.K.V. and Savvides, M. (2019). Automatic Seizure Detection via an Optimized Image-Based Deep Feature Learning. *Proceedings - 17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018*:536–540.

Alkire, M.T. (2008). Loss of Effective Connectivity During General Anesthesia. *International Anesthesiology Clinics* [Online] **46**:55–73. Available at: https://journals.lww.com/00004311-200804630-00006.

AlMeer, M.H. and Abbod, M.F. (2019). Deep Learning in Classifying Depth of Anesthesia (DoA). In: *Springer Nature Switzerland*. Springer International Publishing, pp. 160–169. Available at: http://dx.doi.org/10.1007/978-3-030-01054-6_80.

Amini, A. *et al.* (2018). Spatial Uncertainty Sampling for End-to-End Control. [Online]. Available at: https://arxiv.org/abs/1805.04829v2 [Accessed: 16 May 2022].

Artificial Intelligence - Accelerate the Power with Neural Networks | Blog [Online]. Available at: https://dimensionless.in/artificial-intelligence-accelerate-the-power-with-neural-networks/ [Accessed: 15 May 2022].

Aru, Jaan *et al.* (2012). Distilling the neural correlates of consciousness. *Neuroscience & Biobehavioral Reviews* [Online] **36**:737–746. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0149763411002107.

Aru, J. *et al.* (2012). Local Category-Specific Gamma Band Responses in the Visual Cortex Do Not Reflect Conscious Perception. *Journal of Neuroscience*.

Aurlien, H. *et al.* (2004). EEG background activity described by a large computerized database. *Clinical Neurophysiology* [Online] **115**:665–673. Available at: https://linkinghub.elsevier.com/retrieve/pii/S138824570300378X.

Avidan, M.S. *et al.* (2008). Anesthesia Awareness and the Bispectral Index. *New England Journal of Medicine* [Online] **358**:1097–1108. Available at: http://www.nejm.org/doi/abs/10.1056/NEJMoa0707361.

Avidan, M.S. *et al.* (2011). Prevention of Intraoperative Awareness in a High-Risk Surgical Population. *New England Journal of Medicine* [Online] **365**:591–600. Available at: http://www.nejm.org/doi/abs/10.1056/NEJMoa1100403.

Bai, S., Kolter, J.Z. and Koltun, V. (2018). An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. [Online]. Available at: http://arxiv.org/abs/1803.01271.

Bai, Y. *et al.* (2015). Permutation Lempel-Ziv complexity measure of electroencephalogram in GABAergic anaesthetics. *Physiological Measurement* **36**:2483–2501.

Barachant, A., Andreev, A. and Congedo, M. (2013). The Riemannian Potato: an automatic and adaptive artifact detection method for online experiments using Riemannian geometry. *Proceedings of TOBI Workshop IV*.

Barr, G. *et al.* (1999). Nitrous oxide does not alter bispectral index: study with nitrous oxide as sole agent and as an adjunct to i.v. anaesthesia. *British Journal of Anaesthesia* [Online] **82**:827–830. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0007091217384878.

Bashivan, P. *et al.* (2015). Learning Representations from EEG with Deep Recurrent-Convolutional Neural Networks. *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings* [Online]. Available at: http://arxiv.org/abs/1511.06448.

Bastos, A.M. and Schoffelen, J.-M. (2016). A Tutorial Review of Functional Connectivity Analysis Methods and Their Interpretational Pitfalls. *Frontiers in Systems Neuroscience* [Online] **9**. Available at: http://journal.frontiersin.org/Article/10.3389/fnsys.2015.00175/abstract.

Bengio, Y., Courville, A. and Vincent, P. (2013). Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence* [Online] **35**:1798–1828. Available at: http://ieeexplore.ieee.org/document/6472238/.

Bigdely-Shamlo, N. *et al.* (2015). The PREP pipeline: standardized preprocessing for large-scale EEG analysis. *Frontiers in Neuroinformatics* [Online] **9**. Available at: http://journal.frontiersin.org/Article/10.3389/fninf.2015.00016/abstract.

Bishop, C.M. (2006). *Patterns Recognition and Machine Learning*.

Boly, M. *et al.* (2012). Connectivity Changes Underlying Spectral EEG Changes during Propofol-Induced Loss of Consciousness. *Journal of Neuroscience* [Online] **32**:7082–7090. Available at: http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.3769-11.2012.

Boly, M. *et al.* (2013). Consciousness in humans and non-human animals: Recent advances and future directions. *Frontiers in Psychology* **4**:1–20.

Bonhomme, V. *et al.* (2019). General Anesthesia: A Probe to Explore Consciousness. *Frontiers in Systems Neuroscience* [Online] **13**. Available at: https://www.frontiersin.org/article/10.3389/fnsys.2019.00036/full.

Bonhomme, V. *et al.* (2016). Resting-state Network-specific Breakdown of Functional Connectivity during Ketamine Alteration of Consciousness in Volunteers. *Anesthesiology* **125**:873–888.

Brown, E.N., Pavone, K.J. and Naranjo, M. (2018). Multimodal General Anesthesia. *Anesthesia & Analgesia* [Online] **127**:1246–1258. Available at: http://journals.lww.com/00000539-201811000-00023.

Bruhn, J. *et al.* (2006). Depth of anaesthesia monitoring: what's available, what's validated

and what's next? *British Journal of Anaesthesia* [Online] **97**:85–94. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0007091217351875.

Bruhn, J., Röpcke, H. and Hoeft, A. (2000). Approximate Entropy as an Electroencephalographic Measure of Anesthetic Drug Effect during Desflurane Anesthesia. *Anesthesiology* [Online] **92**:715–726. Available at: https://pubs.asahq.org/anesthesiology/article/92/3/715/39710/Approximate-Entropy-as-an-Electroencephalographic.

Buzsáki, G., Anastassiou, C.A. and Koch, C. (2012). The origin of extracellular fields and currents — EEG, ECoG, LFP and spikes. *Nature Reviews Neuroscience* [Online] **13**:407–420. Available at: http://www.nature.com/doifinder/10.1038/nrn3241.

Byrd, J. and Lipton, Z.C. (2018). What is the Effect of Importance Weighting in Deep Learning? *36th International Conference on Machine Learning, ICML 2019* [Online]. Available at: http://arxiv.org/abs/1812.03372.

Casali, A.G. *et al.* (2013). A Theoretically Based Index of Consciousness Independent of Sensory Processing and Behavior. *Science Translational Medicine* [Online] **5**. Available at: https://www.science.org/doi/10.1126/scitranslmed.3006294.

Chambon, S. *et al.* (2018). A Deep Learning Architecture for Temporal Sleep Stage Classification Using Multivariate and Multimodal Time Series. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **26**:758–769.

Chan, M.T.V. *et al.* (2013). BIS-guided Anesthesia Decreases Postoperative Delirium and Cognitive Decline. *Journal of Neurosurgical Anesthesiology* [Online] **25**:33–42. Available at: https://journals.lww.com/00008506-201301000-00005.

Chatelle, C. *et al.* (2012). Brain-computer interfacing in disorders of consciousness. *Brain injury* [Online] **26**:1510–22. Available at: http://orbi.ulg.ac.be/handle/2268/162403.

Chennu, S. *et al.* (2016). Brain Connectivity Dissociates Responsiveness from Drug Exposure during Propofol-Induced Transitions of Consciousness. *PLoS Computational Biology* **12**:1–17.

Chennu, S. *et al.* (2014). Spectral Signatures of Reorganised Brain Networks in Disorders of Consciousness Ermentrout, B. ed. *PLoS Computational Biology* [Online] **10**:e1003887. Available at: https://dx.plos.org/10.1371/journal.pcbi.1003887.

de Cheveigné, A. and Simon, J.Z. (2008). Sensor noise suppression. *Journal of Neuroscience Methods* [Online] **168**:195–202. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0165027007004621.

Cho, K.-O. and Jang, H.-J. (2020). Comparison of different input modalities and network structures for deep learning-based seizure detection. *Scientific Reports* [Online] **10**:122. Available at: http://www.nature.com/articles/s41598-019-56958-y.

Choi, D. *et al.* (2019). On Empirical Comparisons of Optimizers for Deep Learning. [Online]. Available at: https://arxiv.org/abs/1910.05446v3 [Accessed: 12 May 2022].

Choi, H.-I., Noh, G.-J. and Shin, H.-C. (2020). Measuring the Depth of Anesthesia Using Ordinal Power Spectral Density of Electroencephalogram. *IEEE Access* [Online] **8**:50431–50438. Available at: https://ieeexplore.ieee.org/document/9034018/.

Cole, S.R. and Voytek, B. (2017). Brain Oscillations and the Importance of Waveform Shape. *Trends in Cognitive Sciences* [Online] **21**:137–149. Available at: http://dx.doi.org/10.1016/j.tics.2016.12.008.

Colombo, M.A. *et al.* (2019). The spectral exponent of the resting EEG indexes the presence of consciousness during unresponsiveness induced by propofol, xenon, and ketamine. *NeuroImage* [Online] **189**:631–644. Available at: https://doi.org/10.1016/j.neuroimage.2019.01.024.

Comsa, I.M., Bekinschtein, T.A. and Chennu, S. (2019). Transient Topographical Dynamics of the Electroencephalogram Predict Brain Connectivity and Behavioural Responsiveness During Drowsiness. *Brain Topography* [Online] **32**:315–331. Available at: http://dx.doi.org/10.1007/s10548-018-0689-9.

Csanád Csáji, B. (2001). Approximation with MSc thesis Contents : [Online]:1–45. Available at:

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.101.2647&rep=rep1&type=pdf.

Del Cul, A., Baillet, S. and Dehaene, S. (2007). Brain Dynamics Underlying the Nonlinear Threshold for Access to Consciousness. *PLOS Biology* [Online] **5**:e260. Available at: https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.0050260 [Accessed: 3 May 2022].

Deep Learning Made Easy with Deep Cognition | by Favio Vázquez | Becoming Human: Artificial Intelligence Magazine [Online]. Available at: https://becominghuman.ai/deep-learning-made-easy-with-deep-cognition-403fbe445351 [Accessed: 16 May 2022].

DeGrave, A.J., Janizek, J.D. and Lee, S.-I. (2021). AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence* [Online] **3**:610–619. Available at: http://www.nature.com/articles/s42256-021-00338-7.

Deiss, O. *et al.* (2018). HAMLET: Interpretable Human And Machine co-LEarning Technique. [Online]. Available at: http://arxiv.org/abs/1803.09702.

Dennett, D.C. (2003). Who's on first? Heterophenomenology explained. *Journal of Consciousness Studies* [Online] **10**. Available at: https://philpapers.org/rec/DENWOF [Accessed: 20 September 2022].

Destrebecqz, A. and Peigneux, P. (2005). Methods for studying unconscious learning. In: *Progress in Brain Research*. pp. 69–80. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0079612305500062.

Drazkowski, J.F. (2011). Book Review: Niedermeyer's Electroencephalography: Basic Principles, Clinical Applications, and Related Fields. *Neurology* [Online] **77**:1209–1209. Available at: https://www.neurology.org/lookup/doi/10.1212/WNL.0b013e31822f0490.

Drover, D.R. *et al.* (2002). Patient State Index. *Anesthesiology* [Online] **97**:82–89. Available at: https://pubs.asahq.org/anesthesiology/article/97/1/82/40133/Patient-State-IndexTitration-of-Delivery-and.

Dubost, C. *et al.* (2019). Selection of the Best Electroencephalogram Channel to Predict the

Depth of Anesthesia. *Frontiers in Computational Neuroscience* **13**:1–13.

EEG [Online]. Available at: https://www.medicine.mcgill.ca/physio/vlab/biomed_signals/eeg_n.htm [Accessed: 17 December 2018].

Eichhorn, J. *et al.* (2019). A Randomized Controlled Trial of Anesthesia Guided by Bispectral Index Versus Standard Care: Effects on Cognition, AANA Journal, April 2019. *AANA Journal* **87**:115–123.

Ellerkmann, R.K. *et al.* (2010). The Entropy Module® and Bispectral Index® as Guidance for Propofol-Remifentanil Anaesthesia in Combination with Regional Anaesthesia Compared with a Standard Clinical Practice Group. *Anaesthesia and Intensive Care* [Online] **38**:159–166. Available at: http://journals.sagepub.com/doi/10.1177/0310057X1003800125.

Engel, A.K. and Singer, W. (2001). Temporal binding and the neural correlates of sensory awareness. *Trends in Cognitive Sciences* [Online] **5**:16–25. Available at: https://linkinghub.elsevier.com/retrieve/pii/S1364661300015680.

Fellner, M.C. *et al.* (2016). Spurious correlations in simultaneous EEG-fMRI driven by in-scanner movement. *NeuroImage* [Online] **133**:354–366. Available at: https://pubmed.ncbi.nlm.nih.gov/27012498/ [Accessed: 3 February 2022].

Ferreira, A.L. *et al.* (2021). Implementation of Neural Networks to Frontal Electroencephalography for the Identification of the Transition Responsiveness/Unresponsiveness During Induction of General Anesthesia. *Irbm* [Online] **1**:1–8. Available at: https://doi.org/10.1016/j.irbm.2021.02.004.

Frassle, S. *et al.* (2014). Binocular Rivalry: Frontal Activity Relates to Introspection and Action But Not to Perception. *Journal of Neuroscience* [Online] **34**:1738–1747. Available at: https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.4403-13.2014.

Gaskell, A.L. *et al.* (2017). Frontal alpha-delta EEG does not preclude volitional response

during anaesthesia: Prospective cohort study of the isolated forearm technique. *British Journal of Anaesthesia* [Online] **119**:664–673. Available at: http://dx.doi.org/10.1093/bja/aex170.

Ghassemi, M., Oakden-Rayner, L. and Beam, A.L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet. Digital health* [Online] **3**:e745–e750. Available at: http://dx.doi.org/10.1016/S2589-7500(21)00208-9.

Giacino, J.T. and Smart, C.M. (2007). Recent advances in behavioral assessment of individuals with disorders of consciousness. *Current opinion in neurology* [Online] **20**:614–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17992078.

Glen, J.B. and White, M. (2014). A comparison of the predictive performance of three pharmacokinetic models for propofol using measured values obtained during target-controlled infusion. *Anaesthesia* **69**:550–557.

Goebel, R. *et al.* (2018). Explainable AI: The new 42? *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* [Online] **11015 LNCS**:295–303. Available at: https://link.springer.com/chapter/10.1007/978-3-319-99740-7_21 [Accessed: 12 May 2022].

Gotman, J. (2013). High frequency oscillations: The new EEG frontier. *Epilepsy & Behavior* **28**:309–310.

de Graaf, T.A., Hsieh, P.-J. and Sack, A.T. (2012). The 'correlates' in neural correlates of consciousness. *Neuroscience & Biobehavioral Reviews* [Online] **36**:191–197. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0149763411001072.

Gramfort, A. *et al.* (2014). MNE software for processing MEG and EEG data. *NeuroImage* [Online] **86**:446–460. Available at: http://dx.doi.org/10.1016/j.neuroimage.2013.10.027.

Gray, C.M. *et al.* (1989). Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature* [Online] **338**:334–337.

Available at: https://pubmed.ncbi.nlm.nih.gov/2922061/ [Accessed: 3 May 2022].

Gu, Y., Liang, Z. and Hagihira, S. (2019). Use of Multiple EEG Features and Artificial Neural Network to Monitor the Depth of Anesthesia. *Sensors* [Online] **19**:2499. Available at: https://www.mdpi.com/1424-8220/19/11/2499.

Gunaydin, B. and Babacan, A. (1998). Cerebral hypoperfusion after cardiac surgery and anesthetic strategies: a comparative study with high dose fentanyl and barbiturate anesthesia. *Annals of thoracic and cardiovascular surgery : official journal of the Association of Thoracic and Cardiovascular Surgeons of Asia*.

Haggard, P. (2008). Human volition: towards a neuroscience of will. *Nature Reviews Neuroscience* [Online] **9**:934–946. Available at: http://www.nature.com/articles/nrn2497.

Hajinoroozi, M. *et al.* (2016). EEG-based prediction of driver's cognitive performance by deep convolutional neural network. *Signal Processing: Image Communication* [Online] **47**:549–555. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0923596516300832.

Hans, P. *et al.* (2005). Comparative effects of ketamine on Bispectral Index and spectral entropy of the electroencephalogram under sevoflurane anaesthesia. *British Journal of Anaesthesia* [Online] **94**:336–340. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0007091217357045.

Hartmann, K.G., Schirrmeister, R.T. and Ball, T. (2018). EEG-GAN: Generative adversarial networks for electroencephalograhic (EEG) brain signals. [Online]. Available at: http://arxiv.org/abs/1806.01875.

Heilmeyer, F.A. *et al.* (2019). A Large-Scale Evaluation Framework for EEG Deep Learning Architectures. *Proceedings - 2018 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2018*:1039–1045.

Hernandez, L.G. (2017). A Comparison of Deep Neural Network Algorithms for Recognition of EEG Motor Imagery Signals. *Handbook of Discrete and Computational Geometry,*

Third Edition:723–762.

Herzog, M.H., Kammer, T. and Scharnowski, F. (2016). Time Slices: What Is the Duration of a Percept? *PLoS Biology* **14**:1–12.

Higashi, H., Tanaka, T. and Tanaka, Y. (2014). Smoothing of spatial filter by graph Fourier transform for EEG signals. *2014 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA 2014*:1–8.

Hoenig, J. and Toakley, J.G. (1959). The Diagnosis of Stupor. *European Neurology* [Online] **137**:128–144. Available at: https://www.karger.com/Article/FullText/134234.

Höller, Y. *et al.* (2011). Preserved oscillatory response but lack of mismatch negativity in patients with disorders of consciousness. *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology* [Online] **122**:1744–1754. Available at: https://pubmed.ncbi.nlm.nih.gov/21377413/ [Accessed: 3 May 2022].

Honderich, T. (2008). *The Oxford Companion to Philosophy*.

Howbert, J.J. *et al.* (2014). Forecasting seizures in dogs with naturally occurring epilepsy. *PLoS ONE*.

Huang, J.R. *et al.* (2013). Application of multivariate empirical mode decomposition and sample entropy in EEG signals via artificial neural networks for interpreting depth of anesthesia. *Entropy* **15**:3325–3339.

Hudetz, A.G. (2012). General Anesthesia and Human Brain Connectivity. *Brain Connectivity* [Online] **2**:291–302. Available at: http://www.liebertpub.com/doi/10.1089/brain.2012.0107.

Hudetz, A.G. *et al.* (2016). Propofol anesthesia reduces Lempel-Ziv complexity of spontaneous brain activity in rats. *Neuroscience Letters* [Online] **628**:132–135. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0304394016304268.

Huizenga, H.M. *et al.* (2001). Simultaneous MEG and EEG source analysis. *Physics in Medicine and Biology* [Online] **46**:1737–1751. Available at:

https://iopscience.iop.org/article/10.1088/0031-9155/46/7/301.

Jas, M., Engemann, D.A., *et al.* (2017). Autoreject: Automated artifact rejection for MEG and EEG data. *NeuroImage* **159**:417–429.

Jas, M., La Tour, T.D., *et al.* (2017). Learning the morphology of brain signals using alpha-stable convolutional sparse coding. *Advances in Neural Information Processing Systems* **2017-Decem**:1100–1109.

Jasper, H.H. (1948). Charting the sea of brain waves. In: *Science*.

Jayaram, V. and Barachant, A. (2018). MOABB: trustworthy algorithm benchmarking for BCIs. *Journal of neural engineering* [Online] **15**:066011. Available at: http://www.ncbi.nlm.nih.gov/pubmed/30177583.

Jiang, G.J.A. *et al.* (2015). Sample Entropy Analysis of EEG Signals via Artificial Neural Networks to Model Patients' Consciousness Level Based on Anesthesiologists Experience. *BioMed Research International* [Online] **2015**:1–8. Available at: http://www.hindawi.com/journals/bmri/2015/343478/.

Jiang, X., Bian, G.-B. and Tian, Z. (2019). Removal of Artifacts from EEG Signals: A Review. *Sensors* [Online] **19**:987. Available at: http://www.mdpi.com/1424-8220/19/5/987.

Jiang, Y. *et al.* (2006). A gender- and sexual orientation-dependent spatial attentional effect of invisible images. *Proceedings of the National Academy of Sciences* [Online] **103**:17048–17052. Available at: https://pnas.org/doi/full/10.1073/pnas.0605678103.

Jordan, M.I. and Mitchell, T.M. (2015). Machine learning: Trends, perspectives, and prospects. *Science* [Online] **349**:255–260. Available at: https://www.science.org/doi/10.1126/science.aaa8415.

Juel, B.E. *et al.* (2018). Distinguishing Anesthetized from Awake State in Patients: A New Approach Using One Second Segments of Raw EEG. *Frontiers in Human Neuroscience* [Online] **12**. Available at: http://journal.frontiersin.org/article/10.3389/fnhum.2018.00040/full.

Kammoun, A. *et al.* (2022). Generative Adversarial Networks for face generation: A survey. *ACM Computing Surveys* [Online]:1291. Available at: https://dl.acm.org/doi/10.1145/1122445.1122456.

Kangas, L.J. *et al.* (1997). Neurometric assessment of adequacy of intraoperative anesthetic. *IEEE International Conference on Neural Networks - Conference Proceedings* **4**:2475–2479.

Kaplan, L. and Bailey, H. (2000). Bispectral index (BIS) monitoring of ICU patients on continuous infusion of sedatives and paralytics reduces sedative drug utilization and cost. *Critical Care* [Online] **4**:P190. Available at: http://ccforum.biomedcentral.com/articles/10.1186/cc910.

Kearse, L.A. *et al.* (1994). Bispectral Analysis of the Electroencephalogram Correlates with Patient Movement to Skin Incision during Propofol/Nitrous Oxide Anesthesia. *Anesthesiology* [Online] **81**:1365–1370. Available at: https://pubs.asahq.org/anesthesiology/article/81/6/1365/34746/Bispectral-Analysis-of-the-Electroencephalogram.

Khan, F.H. *et al.* (2018). A Patient-Specific Machine Learning based EEG Processor for Accurate Estimation of Depth of Anesthesia. In: *2018 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE, pp. 1–4. Available at: https://ieeexplore.ieee.org/document/8584828/ [Accessed: 7 October 2021].

King, J.-R. *et al.* (2017). Encoding and Decoding Neuronal Dynamics: Methodological Framework to Uncover the Algorithms of Cognition. *(To Appear)*.

Klockars, J.G.M. *et al.* (2012). Spectral Entropy as a Measure of Hypnosis and Hypnotic Drug Effect of Total Intravenous Anesthesia in Children during Slow Induction and Maintenance. *Anesthesiology* [Online] **116**:340–351. Available at: https://pubs.asahq.org/anesthesiology/article/116/2/340/13016/Spectral-Entropy-as-a-Measure-of-Hypnosis-and.

Koch, C. (2004). Chapter Five - What Are the Neuronal Correlates of Consciousness? *The Quest for Consciousness - A Neurobiological Approach* [Online]:1–7. Available at:

https://books.google.com/books/about/The_Quest_for_Consciousness.html?hl=el&id=7L9qAAAAMAAJ [Accessed: 13 May 2022].

Koch, C. *et al.* (2016). Neural correlates of consciousness: progress and problems. *Nature Reviews Neuroscience* [Online] **17**:307–321. Available at: http://www.nature.com/doifinder/10.1038/nrn.2016.22.

Korotcov, A. *et al.* (2017). Comparison of Deep Learning With Multiple Machine Learning Methods and Metrics Using Diverse Drug Discovery Data Sets. *Molecular Pharmaceutics* [Online] **14**:4462–4475. Available at: https://pubs.acs.org/doi/abs/10.1021/acs.molpharmaceut.7b00578 [Accessed: 4 October 2021].

Korshunova, I. *et al.* (2017). Towards improved design and evaluation of epileptic seizure predictors. *IEEE Transactions on Biomedical Engineering* [Online] **9294**:1–1. Available at: http://ieeexplore.ieee.org/document/7915772/.

Kreuer, S. *et al.* (2003). Narcotrend Monitoring Allows Faster Emergence and a Reduction of Drug Consumption in Propofol–Remifentanil Anesthesia. *Anesthesiology* [Online] **99**:34–41. Available at: https://pubs.asahq.org/anesthesiology/article/99/1/34/39454/Narcotrend-Monitoring-Allows-Faster-Emergence-and.

Kreuzer, M. (2017). EEG Based Monitoring of General Anesthesia: Taking the Next Steps. *Frontiers in Computational Neuroscience* [Online] **11**:1–7. Available at: http://journal.frontiersin.org/article/10.3389/fncom.2017.00056/full.

Krishnaswamy, P. *et al.* (2017). Sparsity enables estimation of both subcortical and cortical activity from MEG and EEG. *Proceedings of the National Academy of Sciences* [Online] **114**. Available at: https://pnas.org/doi/full/10.1073/pnas.1705414114.

Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM* [Online] **60**:84–90. Available at: https://linkinghub.elsevier.com/retrieve/pii/S2212017314001224.

Krkic, M. (1996). EEG-based assessment of anaesthetic depth using neural networks. In: *IEE Colloquium on Artificial Intelligence Methods for Biomedical Data Processing*. IEE, pp. 10–10. Available at: https://digital-library.theiet.org/content/conferences/10.1049/ic_19960645.

Kunimoto, C., Miller, J. and Pashler, H. (2001). Confidence and Accuracy of Near-Threshold Discrimination Responses. *Consciousness and Cognition* [Online] **10**:294–340. Available at: https://linkinghub.elsevier.com/retrieve/pii/S105381000090494X.

Kwak, N.-S., Müller, K.-R. and Lee, S.-W. (2017). A convolutional neural network for steady state visual evoked potential classification under ambulatory environment Schwenker, F. ed. *PLOS ONE* [Online] **12**:e0172578. Available at: https://dx.plos.org/10.1371/journal.pone.0172578.

Lai, C.Q. *et al.* (2019). Arrangements of Resting State Electroencephalography as the Input to Convolutional Neural Network for Biometric Identification. *Computational Intelligence and Neuroscience* [Online] **2019**:1–10. Available at: https://www.hindawi.com/journals/cin/2019/7895924/.

Lalitha, V. and Eswaran, C. (2007). Automated Detection of Anesthetic Depth Levels Using Chaotic Features with Artificial Neural Networks. *Journal of Medical Systems* [Online] **31**:445–452. Available at: http://link.springer.com/10.1007/s10916-007-9083-y.

Lan, J.-Y. (2012). Intelligent modeling and control in anesthesia. *Journal of Medical and Biological Engineering* [Online] **32**:293. Available at: http://jmbe.bme.ncku.edu.tw/index.php/bme/article/view/1370/908.

Laufs, H. *et al.* (2003). EEG-correlated fMRI of human alpha activity. *NeuroImage* [Online] **19**:1463–1476. Available at: https://linkinghub.elsevier.com/retrieve/pii/S1053811903002866.

Lawhern, V.J. *et al.* (2018). EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering* [Online] **15**:056013. Available at: https://iopscience.iop.org/article/10.1088/1741-2552/aace8c.

Lee, U. *et al.* (2015). Disruption of Frontal-Parietal Communication by Ketamine, Propofol, and Sevoflurane. *Anesthesiology* **118**:1264–1275.

Lei, X. and Liao, K. (2017). Understanding the influences of EEG reference: A large-scale brain network perspective. *Frontiers in Neuroscience* **11**:1–11.

Lemon, R.N. and Edgley, S.A. (2010). Life without a cerebellum. *Brain* [Online] **133**:652–654. Available at: https://academic.oup.com/brain/article-lookup/doi/10.1093/brain/awq030.

Liang, Z. *et al.* (2015). EEG entropy measures in anesthesia. *Frontiers in Computational Neuroscience* **9**:1–17.

Linassi, F. *et al.* (2018). Isolated forearm technique: a meta-analysis of connected consciousness during different general anaesthesia regimens. *British Journal of Anaesthesia* [Online] **121**:198–209. Available at: https://doi.org/10.1016/j.bja.2018.02.019.

Liu, Q. *et al.* (2015). EEG Signals Analysis Using Multiscale Entropy for Depth of Anesthesia Monitoring during Surgery through Artificial Neural Networks. *Computational and Mathematical Methods in Medicine* **2015**.

Liu, Q. *et al.* (2019). Spectrum Analysis of EEG Signals Using CNN to Model Patient's Consciousness Level Based on Anesthesiologists' Experience. *IEEE Access* [Online] **7**:53731–53742. Available at: https://ieeexplore.ieee.org/document/8695791/.

Lobes of the Brain | Introduction to Psychology [Online]. Available at: https://courses.lumenlearning.com/waymaker-psychology/chapter/reading-parts-of-the-brain/ [Accessed: 17 December 2018].

Lopez, S. *et al.* (2017). An analysis of two common reference points for EEGS. In: *2016 IEEE Signal Processing in Medicine and Biology Symposium, SPMB 2016 - Proceedings*.

Lotte, F. *et al.* (2018). A Review of Classification Algorithms for EEG-based Brain-Computer Interfaces: A 10-year Update. *Journal of Neural Engineering* [Online]:0–20.

Available at: http://iopscience.iop.org/article/10.1088/1741-2552/aab2f2.

Luby, B.J. (1998). The nature of consciousness: Philosophical debates. *Journal of the History of the Behavioral Sciences* [Online] **34**:433–434. Available at: https://onlinelibrary.wiley.com/doi/10.1002/(SICI)1520-6696(199823)34:4%3C433::AID-JHBS38%3E3.0.CO;2-V.

Luu, P. *et al.* (2001). Localizing Acute Stroke-related EEG Changes: *Journal of Clinical Neurophysiology* [Online] **18**:302–317. Available at: http://journals.lww.com/00004691-200107000-00002.

Machine Learning Has Transformed Many Aspects of Our Everyday Life, Can It Do the Same for Public Services? – Capgemini Worldwide [Online]. Available at: https://www.capgemini.com/2016/05/machine-learning-has-transformed-many-aspects-of-our-everyday-life/ [Accessed: 17 December 2018].

MacLean, K.A. *et al.* (2012). Factor Analysis of the Mystical Experience Questionnaire: A Study of Experiences Occasioned by the Hallucinogen Psilocybin. *Journal for the Scientific Study of Religion* [Online] **51**:721–737. Available at: https://onlinelibrary.wiley.com/doi/10.1111/j.1468-5906.2012.01685.x.

Mahmood, U. *et al.* (2021). Detecting Spurious Correlations With Sanity Tests for Artificial Intelligence Guided Radiology Systems. *Frontiers in Digital Health* **3**:85.

Maksimow, A. *et al.* (2006). Increase in high frequency EEG activity explains the poor performance of EEG spectral entropy monitor during S-ketamine anesthesia. *Clinical Neurophysiology* **117**:1660–1668.

Mashour, G.A. and Avidan, M.S. (2015). Intraoperative awareness: Controversies and non-controversies. *British Journal of Anaesthesia* [Online] **115**:I20–I26. Available at: http://dx.doi.org/10.1093/bja/aev034.

Mashour, G.A. and Hudetz, A.G. (2017). Bottom-up and top-down mechanisms of general anesthetics modulate different dimensions of consciousness. *Frontiers in Neural Circuits* **11**:1–6.

Melloni, L. *et al.* (2007). Synchronization of Neural Activity across Cortical Areas Correlates with Conscious Perception. *Journal of Neuroscience* [Online] **27**:2858–2865. Available at: https://www.jneurosci.org/content/27/11/2858 [Accessed: 3 May 2022].

Melnik, A. *et al.* (2017). Systems, Subjects, Sessions: To What Extent Do These Factors Influence EEG Data? *Frontiers in Human Neuroscience* [Online] **11**:1–20. Available at: http://journal.frontiersin.org/article/10.3389/fnhum.2017.00150/full.

Merriam-Webster (2012). *MERRIAM-WEBSTER'S ONLINE DICTIONARY*.

Mhuircheartaigh, R.N. *et al.* (2013a). Slow-wave activity saturation and thalamocortical isolation during propofol anesthesia in humans. *Science Translational Medicine* **5**.

Mhuircheartaigh, R.N. *et al.* (2013b). Slow-wave activity saturation and thalamocortical isolation during propofol anesthesia in humans. *Science Translational Medicine* **5**.

Micallef, J. *et al.* (2002). Effects of Subanesthetic Doses of Ketamine on Sensorimotor Information Processing in Healthy Subjects. *Clinical Neuropharmacology* [Online] **25**:101–106. Available at: http://journals.lww.com/00002826-200203000-00008.

Miller, S.M. (2014). Closing in on the constitution of consciousness. *Frontiers in Psychology*.

Miotto, R. *et al.* (2018). Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics* [Online] **19**:1236–1246. Available at: https://academic.oup.com/bib/article/19/6/1236/3800524.

Montages [Online]. Available at: http://eegatlas-online.com/index.php/en/montages [Accessed: 17 December 2018].

Murphy, M. *et al.* (2011). Propofol Anesthesia and Sleep: A High-Density EEG Study. *Sleep* [Online] **34**:283–291. Available at: https://academic.oup.com/sleep/article-lookup/doi/10.1093/sleep/34.3.283.

Musso, F. *et al.* (2011). Ketamine effects on brain function - Simultaneous fMRI/EEG during a visual oddball task. *NeuroImage* [Online] **58**:508–525. Available at:

https://pubmed.ncbi.nlm.nih.gov/21723949/ [Accessed: 9 October 2021].

Muthukumaraswamy, S.D. *et al.* (2015). Evidence that subanesthetic doses of ketamine cause sustained disruptions of NMDA and AMPA-mediated frontoparietal connectivity in humans. *Journal of Neuroscience* **35**:11694–11706.

Muthukumaraswamy, S.D. (2013). High-frequency brain activity and muscle artifacts in MEG/EEG: a review and recommendations. *Frontiers in Human Neuroscience* [Online] **7**:1–11. Available at: http://journal.frontiersin.org/article/10.3389/fnhum.2013.00138/abstract.

Myles, P. *et al.* (2004). Bispectral index monitoring to prevent awareness during anaesthesia: the B-Aware randomised controlled trial. *The Lancet* [Online] **363**:1757–1763. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0140673604163009.

Neyshabur, B., Tomioka, R. and Srebro, N. (2015). In search of the real inductive bias: On the role of implicit regularization in deep learning. *3rd International Conference on Learning Representations, ICLR 2015 - Workshop Track Proceedings*:1–9.

Ní Mhuircheartaigh, R. *et al.* (2013). Slow-Wave Activity Saturation and Thalamocortical Isolation During Propofol Anesthesia in Humans. *Science Translational Medicine* [Online] **5**. Available at: https://www.science.org/doi/10.1126/scitranslmed.3006007.

Nobili, L. *et al.* (2012). Local aspects of sleep: observations from intracerebral recordings in humans. *Progress in brain research* [Online] **199**:219–232. Available at: https://pubmed.ncbi.nlm.nih.gov/22877668/ [Accessed: 3 May 2022].

Nolan, H., Whelan, R. and Reilly, R.B. (2010). FASTER: Fully Automated Statistical Thresholding for EEG artifact Rejection. *Journal of Neuroscience Methods* [Online] **192**:152–162. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0165027010003894.

Nunez, P.L. (1988). Methods to improve spatial resolution of EEG. In: *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, pp. 972–973 vol.2. Available at:

http://ieeexplore.ieee.org/document/95290/.

O'Shea, A. *et al.* (2017). Neonatal Seizure Detection using Convolutional Neural Networks. [Online]. Available at: http://arxiv.org/abs/1709.05849.

Oizumi, M., Albantakis, L. and Tononi, G. (2014). From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0 Sporns, O. ed. *PLoS Computational Biology* [Online] **10**:e1003588. Available at: https://dx.plos.org/10.1371/journal.pcbi.1003588.

Pandit, J.J. *et al.* (2014). 5th National Audit Project (NAP5) on accidental awareness during general anaesthesia: summary of main findings and risk factors † ‡. *British Journal of Anaesthesia* [Online] **113**:549–559. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0007091217307468.

Park, Y. *et al.* (2011). Seizure prediction with spectral power of EEG using cost-sensitive support vector machines. *Epilepsia* [Online] **52**:1761–1770. Available at: https://onlinelibrary.wiley.com/doi/10.1111/j.1528-1167.2011.03138.x.

Parvizi, J. (2001). Consciousness and the brainstem. *Cognition* [Online] **79**:135–160. Available at: https://linkinghub.elsevier.com/retrieve/pii/S001002770000127X.

Pascanu, R., Mikolov, T. and Bengio, Y. (2012). Understanding the exploding gradient problem. *Proceedings of The 30th International Conference on Machine Learning* [Online]:1310–1318. Available at: http://jmlr.org/proceedings/papers/v28/pascanu13.pdf.

Patlatzoglou, K. *et al.* (2018). Deep Neural Networks for Automatic Classification of Anesthetic-Induced Unconsciousness. In: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. pp. 216–225. Available at: http://link.springer.com/10.1007/978-3-030-05587-5_21.

Patlatzoglou, K. *et al.* (2020). Generalized Prediction of Unconsciousness during Propofol Anesthesia using 3D Convolutional Neural Networks. In: *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society*

(EMBC). IEEE, pp. 134–137. Available at: https://ieeexplore.ieee.org/document/9175324/.

Patrick L. Purdon, P.. *et al.* (2015). *Clinical Electroencephalography for Anesthesiologists-Part I- Background and Basic Signatures*. Vol. 123.

Pavlov, Y.G. *et al.* (2021a). #EEGManyLabs: Investigating the replicability of influential EEG experiments. *Cortex* **144**:213–229.

Pavlov, Y.G. *et al.* (2021b). #EEGManyLabs: Investigating the replicability of influential EEG experiments. *Cortex* [Online] **144**:213–229. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0010945221001106 [Accessed: 31 January 2022].

Pedregosa, F. *et al.* (2012). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* [Online] **12**:2825–2830. Available at: http://dl.acm.org/citation.cfm?id=2078195%5Cnhttp://arxiv.org/abs/1201.0490.

Pennachin, C. and Goertzel, B. (2007). Contemporary Approaches to Artificial General Intelligence. In: *Artificial General Intelligence*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 1–30. Available at: http://link.springer.com/10.1007/978-3-540-68677-4_1.

Piantoni, G., Halgren, E. and Cash, S.S. (2016). The Contribution of Thalamocortical Core and Matrix Pathways to Sleep Spindles. *Neural Plasticity* [Online] **2016**:1–10. Available at: http://www.hindawi.com/journals/np/2016/3024342/.

Pockett, S. and Holmes, M.D. (2009). Intracranial EEG power spectra and phase synchrony during consciousness and unconsciousness. *Consciousness and cognition* [Online] **18**:1049–1055. Available at: https://pubmed.ncbi.nlm.nih.gov/19775914/ [Accessed: 3 May 2022].

Purdon, P.L. *et al.* (2013). Electroencephalogram signatures of loss and recovery of consciousness from propofol. *Proceedings of the National Academy of Sciences* [Online] **110**. Available at: https://pnas.org/doi/full/10.1073/pnas.1221180110.

Purdon, P.L. *et al.* (2015). The Ageing Brain: Age-dependent changes in the electroencephalogram during propofol and sevoflurane general anaesthesia. *BJA: British Journal of Anaesthesia* [Online] **115**:i46–i57. Available at: https://academic.oup.com/bja/article/115/suppl_1/i46/234305 [Accessed: 4 January 2022].

Van Putten, M.J.A.M., Olbrich, S. and Arns, M. (2018). Predicting sex from brain rhythms with deep learning. *Scientific Reports* [Online] **8**:1–7. Available at: http://dx.doi.org/10.1038/s41598-018-21495-7.

Radek, L. *et al.* (2018). Dreaming and awareness during dexmedetomidine- and propofol-induced unresponsiveness. *British Journal of Anaesthesia* [Online] **121**:260–269. Available at: https://linkinghub.elsevier.com/retrieve/pii/S000709121830268X.

Renka, R.L. and Cline, A.K. (1984). A triangle-based C^1 interpolation method. *Rocky Mountain Journal of Mathematics* [Online] **14**. Available at: https://projecteuclid.org/journals/rocky-mountain-journal-of-mathematics/volume-14/issue-1/A-triangle-based-C1-interpolation-method/10.1216/RMJ-1984-14-1-223.full.

Robbins, K.A. *et al.* (2020). How Sensitive Are EEG Results to Preprocessing Methods: A Benchmarking Study. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* [Online] **28**:1081–1090. Available at: https://ieeexplore.ieee.org/document/9047940/.

Robert, C. *et al.* (2002). Monitoring anesthesia using neural networks: A survey. *Journal of Clinical Monitoring and Computing* **17**:259–267.

Roy, R.J. and Sharma, A. (1994). Depth of Anesthesia using Neural Networks. *IFAC Proceedings Volumes* [Online] **27**:197–202. Available at: http://dx.doi.org/10.1016/S1474-6670(17)46204-5.

Roy, Y. *et al.* (2019). Deep learning-based electroencephalography analysis: a systematic review. *Journal of Neural Engineering* [Online] **16**:051001. Available at: https://iopscience.iop.org/article/10.1088/1741-2552/ab260c.

Saadeh, W., Khan, F.H. and Altaf, M.A. Bin (2019). Design and Implementation of a Machine Learning Based EEG Processor for Accurate Estimation of Depth of Anesthesia. *IEEE Transactions on Biomedical Circuits and Systems* **13**:658–669.

Salah, D. and Alansary, A.M. (2019). Impact of sub-anesthetic dose of ketamine on post spinal hypotension in cesarean delivery. *Open Anesthesia Journal* **13**:86–92.

Sandberg, K. *et al.* (2010). Measuring consciousness: Is one measure better than the other? *Consciousness and Cognition* [Online] **19**:1069–1078. Available at: https://linkinghub.elsevier.com/retrieve/pii/S1053810009001998.

Sanders, R.D. *et al.* (2012). Unresponsiveness ≠ unconsciousness. *Anesthesiology* **116**:946–959.

Sanei, S. and Chambers, J.A. (2007). *EEG Signal Processing*. [Online]. West Sussex, England: John Wiley & Sons Ltd,. Available at: http://doi.wiley.com/10.1002/9780470511923.

Sanqing Hu *et al.* (2010). On the Recording Reference Contribution to EEG Correlation, Phase Synchorony, and Coherence. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* [Online] **40**:1294–1304. Available at: http://ieeexplore.ieee.org/document/5398939/.

Sarasso, S. *et al.* (2015). Consciousness and complexity during unresponsiveness induced by propofol, xenon, and ketamine. *Current Biology* [Online] **25**:3099–3105. Available at: http://dx.doi.org/10.1016/j.cub.2015.10.014.

Schartner, M. *et al.* (2015). Complexity of multi-dimensional spontaneous EEG decreases during propofol induced general anaesthesia. *PLoS ONE* **10**:1–21.

Schirrmeister, R.T. *et al.* (2017). Deep learning with convolutional neural networks for EEG decoding and visualization. *Human brain mapping* [Online] **38**:5391–5420. Available at: http://www.ncbi.nlm.nih.gov/pubmed/28782865%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5655781.

Schmidt, T. and Vorberg, D. (2006). Criteria for unconscious cognition: Three types of dissociation. *Perception & Psychophysics* [Online] **68**:489–504. Available at: http://link.springer.com/10.3758/BF03193692.

Schnakers, C. *et al.* (2009). Diagnostic accuracy of the vegetative and minimally conscious state: Clinical consensus versus standardized neurobehavioral assessment. *BMC Neurology* [Online] **9**:35. Available at: https://bmcneurol.biomedcentral.com/articles/10.1186/1471-2377-9-35.

Schreckenberger, M. *et al.* (2006). Corrigendum to "The thalamus as the generator and modulator of EEG alpha rhythm: A combined PET/EEG study with lorazepam challenge in humans" [NeuroImage 22 (2004) 637–644]. *NeuroImage*.

Schuller, P.J. *et al.* (2015). Response of bispectral index to neuromuscular block in awake volunteers. *British Journal of Anaesthesia* **115**:i95–i103.

Schultz, T.L. (2012). Technical tips: mri compatible eeg electrodes: Advantages, disadvantages, and financial feasibility in a clinical setting. *Neurodiagnostic Journal*.

Seeber, M. *et al.* (2019). Subcortical electrophysiological activity is detectable with high-density EEG source imaging. *Nature Communications* [Online] **10**:1–7. Available at: http://dx.doi.org/10.1038/s41467-019-08725-w.

Sergent, C., Baillet, S. and Dehaene, S. (2005). Timing of the brain events underlying access to consciousness during the attentional blink. *Nature Neuroscience 2005 8:10* [Online] **8**:1391–1400. Available at: https://www.nature.com/articles/nn1549 [Accessed: 3 May 2022].

Shah, V. *et al.* (2017). Optimizing channel selection for seizure detection. In: *2017 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*. IEEE, pp. 1–5. Available at: http://ieeexplore.ieee.org/document/8257019/.

Shamim Hossain, M. *et al.* (2019). Applying deep learning for epilepsy seizure detection and brain mapping visualization. *ACM Transactions on Multimedia Computing, Communications and Applications* **15**:1–17.

Shen, Y. (2020). Machine Learning Based Epileptic Seizure Detection for Responsive Neurostimulator System Optimization. *Journal of Physics: Conference Series* [Online] **1453**:012089. Available at: https://iopscience.iop.org/article/10.1088/1742-6596/1453/1/012089 [Accessed: 15 May 2022].

Shi, W. *et al.* (2020). Non-Canonical Microstate Becomes Salient in High Density EEG During Propofol-Induced Altered States of Consciousness. *International Journal of Neural Systems* [Online] **30**:2050005. Available at: https://www.worldscientific.com/doi/abs/10.1142/S0129065720500057.

Siclari, F. *et al.* (2017). The neural correlates of dreaming. *Nature neuroscience* [Online] **20**:872–878. Available at: http://www.ncbi.nlm.nih.gov/pubmed/28394322.

Silverstein, B.H. *et al.* (2015). P3b, consciousness, and complex unconscious processing. *Cortex; a journal devoted to the study of the nervous system and behavior* [Online] **73**:216–227. Available at: https://pubmed.ncbi.nlm.nih.gov/26474391/ [Accessed: 3 May 2022].

Sitt, J.D. *et al.* (2014). Large scale screening of neural signatures of consciousness in patients in a vegetative or minimally conscious state. *Brain* [Online] **137**:2258–2270. Available at: https://pubmed.ncbi.nlm.nih.gov/24919971/ [Accessed: 3 May 2022].

Sleigh, J. *et al.* (2019). Electroencephalographic slow wave dynamics and loss of behavioural responsiveness induced by ketamine in human volunteers. *British Journal of Anaesthesia* [Online] **123**:592–600. Available at: https://doi.org/10.1016/j.bja.2019.07.021.

Srivastava, N. *et al.* (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research* **15**:1929–1958.

Steriade, M., Timofeev, I. and Grenier, F. (2001). Natural waking and sleep states: a view from inside neocortical neurons. *Journal of neurophysiology* [Online] **85**:1969–1985. Available at: https://pubmed.ncbi.nlm.nih.gov/11353014/ [Accessed: 3 May 2022].

Stober, S. *et al.* (2015). Deep Feature Learning for EEG Recordings. [Online]. Available at:

http://arxiv.org/abs/1511.04306.

Stober, S., Cameron, D.J. and Grahn, J. a (2014). Using Convolutional Neural Networks to Recognize Rhythm Stimuli from Electroencephalography Recordings. *Neural Information Processing Systems (NIPS) 2014*:1–9.

Stokes, M.G., Wolff, M.J. and Spaak, E. (2015). Decoding Rich Spatial Information with High Temporal Resolution. *Trends in Cognitive Sciences* [Online] **19**:636–638. Available at: http://dx.doi.org/10.1016/j.tics.2015.08.016.

Studerus, E., Gamma, A. and Vollenweider, F.X. (2010). Psychometric Evaluation of the Altered States of Consciousness Rating Scale (OAV) Bell, V. ed. *PLoS ONE* [Online] **5**:e12412. Available at: https://dx.plos.org/10.1371/journal.pone.0012412.

Subramanian, S. *et al.* (2020). Detecting Loss and Regain of Consciousness during Propofol Anesthesia using Multimodal Indices of Autonomic State.

Sun, H. *et al.* (2019a). Automated tracking of level of consciousness and delirium in critical illness using deep learning. *npj Digital Medicine* [Online] **2**:1–8. Available at: http://dx.doi.org/10.1038/s41746-019-0167-0.

Sun, H. *et al.* (2019b). Automated tracking of level of consciousness and delirium in critical illness using deep learning. *npj Digital Medicine* **2**:1–8.

Tacke, M. *et al.* (2020). Machine learning for a combined electroencephalographic anesthesia index to detect awareness under anesthesia Erdoes, G. ed. *PLOS ONE* [Online] **15**:e0238249. Available at: http://dx.doi.org/10.1371/journal.pone.0238249.

Tan, C. *et al.* (2017). Multimodal Classification with Deep Convolutional-Recurrent Neural Networks for Electroencephalography. In: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*.

Tarnal, V., Vlisides, P.E. and Mashour, G.A. (2016). The Neurobiology of Anesthetic Emergence. *Journal of Neurosurgical Anesthesiology* [Online] **28**:250–255. Available at: file:///C:/Users/Carla Carolina/Desktop/Artigos para acrescentar na qualificação/The

impact of birth weight on cardiovascular disease risk in the.pdf.

Tatum, W.O. (2014). *Handbook of EEG Interpretation*.

Tekell, J.L. *et al.* (2005). High Frequency EEG Activity during Sleep: Characteristics in Schizophrenia and Depression. *Clinical EEG and Neuroscience* **36**:25–35.

La Tour, T.D. *et al.* (2018). Multivariate convolutional sparse coding for electromagnetic brain signals. *Advances in Neural Information Processing Systems* **2018-Decem**:3292–3302.

Towle, V.L. *et al.* (1993). The spatial location of EEG electrodes: locating the best-fitting sphere relative to cortical anatomy. *Electroencephalography and Clinical Neurophysiology* [Online] **86**:1–6. Available at: https://linkinghub.elsevier.com/retrieve/pii/001346949390061Y.

Truong, D. *et al.* Deep Convolutional Neural Network Applied to Electroencephalography : Raw Data vs Spectral Features. :2–5.

Turner, J. *et al.* (2017). Deep Belief Networks used on High Resolution Multichannel Electroencephalography Data for Seizure Detection. [Online]. Available at: http://arxiv.org/abs/1708.08430.

Vanschoren, J. *et al.* (2014). OpenML: networked science in machine learning. [Online]. Available at: http://arxiv.org/abs/1407.7722.

Varvel, J.R., Donoho, D.L. and Shafer, S.L. (1992). Measuring the predictive performance of computer-controlled infusion pumps. *Journal of Pharmacokinetics and Biopharmaceutics* [Online] **20**:63–94. Available at: http://link.springer.com/10.1007/BF01143186.

Vlisides, P.E. *et al.* (2017). Neurophysiologic Correlates of Ketamine Sedation and Anesthesia. *Anesthesiology* **127**:58–69.

Vlisides, P.E. *et al.* (2018). Subanaesthetic ketamine and altered states of consciousness in humans. *British Journal of Anaesthesia* [Online] **121**:249–259. Available at:

https://doi.org/10.1016/j.bja.2018.03.011.

Völker, M. *et al.* (2017). Deep Transfer Learning for Error Decoding from Non-Invasive EEG. [Online]. Available at: http://arxiv.org/abs/1710.09139.

Vuilleumier, P. *et al.* (2000). Distinct behavioral and EEG topographic correlates of loss of consciousness in absences. *Epilepsia* [Online] **41**:687–693. Available at: https://pubmed.ncbi.nlm.nih.gov/10840400/ [Accessed: 3 May 2022].

Wang, F. *et al.* (2018). Data Augmentation for EEG-Based Emotion Recognition with Deep Convolutional Neural Networks. In: pp. 82–93. Available at: http://link.springer.com/10.1007/978-3-319-73600-6_8.

Warnaby, C.E. *et al.* (2017). Investigation of Slow-wave Activity Saturation during Surgical Anesthesia Reveals a Signature of Neural Inertia in Humans. *Anesthesiology* **127**:645–657.

Weiser, T.G. *et al.* (2008). An estimation of the global volume of surgery: a modelling strategy based on available data. *The Lancet* [Online] **372**:139–144. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0140673608608788.

What Is Overfitting? | IBM [Online]. Available at: https://www.ibm.com/cloud/learn/overfitting [Accessed: 8 May 2022].

Wilaiprasitporn, T. *et al.* (2020). Affective EEG-Based Person Identification Using the Deep Learning Approach. *IEEE Transactions on Cognitive and Developmental Systems* [Online] **12**:486–496. Available at: https://ieeexplore.ieee.org/document/8745473/.

Wildes, T.S. *et al.* (2019). Effect of Electroencephalography-Guided Anesthetic Administration on Postoperative Delirium Among Older Adults Undergoing Major Surgery. *JAMA* [Online] **321**:473. Available at: http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.2018.22005.

Wu, Z. *et al.* (2019). A Comprehensive Survey on Graph Neural Networks. [Online] **XX**:1–22. Available at: http://arxiv.org/abs/1901.00596.

Wyart, V. and Tallon-Baudry, C. (2008). Neural Dissociation between Visual Awareness and Spatial Attention. *Journal of Neuroscience* [Online] **28**:2667–2679. Available at: https://www.jneurosci.org/content/28/10/2667 [Accessed: 3 May 2022].

Yakura, H. *et al.* (2018). Malware analysis of imaged binary samples by convolutional neural network with attention mechanism. *CODASPY 2018 - Proceedings of the 8th ACM Conference on Data and Application Security and Privacy* **2018-January**:127–134.

Yang, B. *et al.* (2018). Automatic ocular artifacts removal in EEG using deep learning. *Biomedical Signal Processing and Control* [Online] **43**:148–158. Available at: https://doi.org/10.1016/j.bspc.2018.02.021.

Yao, D. *et al.* (2005). A comparative study of different references for EEG spectral mapping: The issue of the neutral reference and the use of the infinity reference. *Physiological Measurement*.

Yao, D. *et al.* (2019). Which Reference Should We Use for EEG and ERP practice? *Brain Topography* [Online] **32**:530–549. Available at: https://doi.org/10.1007/s10548-019-00707-x.

Yao, Y., Plested, J. and Gedeon, T. (2018). *Deep Feature Learning and Visualization for EEG Recording Using Autoencoders*. [Online]. Springer International Publishing. Available at: http://dx.doi.org/10.1007/978-3-030-04239-4_50.

Yaron, I. *et al.* (2022). The ConTraSt database for analysing and comparing empirical studies of consciousness theories. *Nature Human Behaviour*.

Zacharias, N. *et al.* (2020). Ketamine effects on default mode network activity and vigilance: A randomized, placebo-controlled crossover simultaneous fMRI/EEG study. *Human Brain Mapping* **41**:107–119.

Zanner, R., Pilge, S., Kochs, E.F., *et al.* (2009). Time delay of electroencephalogram index calculation: analysis of cerebral state, bispectral, and Narcotrend indices using perioperatively recorded electroencephalographic signals. *British Journal of Anaesthesia* [Online] **103**:394–399. Available at: https://pubmed.ncbi.nlm.nih.gov/19648154/

[Accessed: 5 October 2021].

Zanner, R., Pilge, S., Kochs, E. F., *et al.* (2009). Time delay of electroencephalogram index calculation: Analysis of cerebral state, bispectral, and Narcotrend indices using perioperatively recorded electroencephalographic signals. *British Journal of Anaesthesia* **103**:394–399.

Zhang, H. *et al.* (2020). EEGdenoiseNet: A benchmark dataset for deep learning solutions of EEG denoising. [Online]:1–25. Available at: http://arxiv.org/abs/2009.11662.

Zhang, P. *et al.* (2019). Learning Spatial-Spectral-Temporal EEG Features With Recurrent 3D Convolutional Neural Networks for Cross-Task Mental Workload Assessment. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*.

Zhang, X.S. *et al.* (2001). Discrimination of anesthetic states using mid-latency auditory evoked potential and artificial neural networks. *Annals of Biomedical Engineering*.

Zikov, T. *et al.* (2006). Quantifying Cortical Activity During General Anesthesia Using Wavelet Analysis. *IEEE Transactions on Biomedical Engineering* [Online] **53**:617–632. Available at: http://ieeexplore.ieee.org/document/1608511/.