EEG-based methods for diagnosing awareness in Disorders of Consciousness

Idorenyin Amaunam

A thesis submitted for the degree of

Doctor of Philosophy

at the

School of Computer Science and Electronic Engineering

University of Essex

September 2025

Abstract

A brain–computer interface (BCI) is an advanced neurotechnological system that enables direct communication between the brain and the external environment by bypassing conventional neuromuscular pathways. This capability offers valuable insight into the assessment of awareness in acute clinical states. Clinically, diagnosing disorders of consciousness (DOC) remains a significant challenge, largely because current practice relies heavily on behavioral indicators of consciousness—markers that are often ambiguous and prone to misinterpretation. To address these limitations, electrophysiological and neuroimaging techniques have been explored, with electroencephalography (EEG) standing out for its non-invasiveness, portability, high temporal resolution, and robustness. As a result, EEG-based methods and BCI-inspired protocols have emerged as promising tools for improving the diagnosis and prognosis of DOC, particularly in detecting cognitive motor dissociation (CMD), a condition frequently overlooked by standard clinical scales. Despite this promise, the clinical translation of these approaches remains constrained, primarily due to a shortage of sufficiently powered validation studies.

In this thesis, I evaluate the effectiveness of several popular EEG-based methods and systematically compare the performance of deep and shallow classification models on a large, novel dataset acquired using a motor imagery (MI)-based command-following paradigm in accessing awareness with DOC patients. Specifically, I extracted measures including classification accuracy, brain rhythms, effective connectivity and the perturbational complexity index PCI, from MI, idling and functional electrical stimulation FES

epochs. These were contrasted against Coma Recovery Scale–Revised (CRS-R) scores, the current clinical gold standard. Furthermore, state-of-the-art deep learning models (EEGNet, DeepConvNet, and EEGConformer) were evaluated alongside a shallow classifier, employing leave-one-trial-out cross-validation scheme on the full and windowed trial segments. In addition, analysis was evaluated at the sessional level to account for variability in diagnostic states.

The findings confirm that EEG contain valuable information regarding the state of awareness of DOC patients. In particular, the classification accuracy and the μ-/β-band separability of MI power spectral density(PSD) features, as well as centro-parietal δ-band connectivity during MI and resting, correlate statistically significantly with CRS-R. Moreover, metric-specific thresholds separating awareness from non-awareness could be determined. I further provide useful insights on the ability of these metrics to detect CMD and rectify the false-negative vulnerability of CRS-R. At the same time, this work highlights the risk of statistical misuse of such metrics, which can lead to over-optimistic assessments of latent awareness. Furthermore, the thesis also reveals that deep learning architectures may be prone to overestimation of results when applied to DOC populations.

This research supports the potential of open-loop BCI DOC diagnosis and highlights the need for further development, validation and standardization to establish clinically deployable systems.

Keywords: disorders of consciousness, electroencephalography, brain-computer interface, command following, motor imagery, functional electrical stimulation, neuromarkers of awareness, statistical criteria

Acknowledgements

First and foremost, I sincerely thank GOD, my source, guide, and inspiration. He brought me to Essex, led me to my supervisor, sustained me and my family, and carried us through the many challenges of this research journey. To Him alone I owe my deepest gratitude.

I extend my profound appreciation to my supervisor, Dr. Serafeim Perdikis. To me, he is far more than a supervisor. From my very first day in Essex, he welcomed me, supported me both academically and personally, and guided me with patience and care. Joining the Perdikis Lab was not only purposeful but divinely orchestrated. I am grateful beyond words, Simis—thank you for being more than a teacher.

My heartfelt thanks also go to my lab teammate and senior colleague, Dr. Mushfika Sultana, for her kindness, guidance, and constant encouragement to strive harder, especially during my most difficult moments.

Special thanks are due to my supervisory chair, Dr. Junhua Li, for his honest and constructive feedback throughout my quarterly presentations. I am equally grateful to all members of the Brain-Computer Interface and Neural Engineering (BCI-NE) Laboratory for their support, insightful discussions, engaging seminars, and weekly lab meetings. It has been an honor to be part of such a stimulating and collaborative environment. I particularly acknowledge three people whose presence was always uplifting: Prof. Reinhold Scherer (Head of School), Dr. Caterina Cinel (Co-director of the BCI-NE group), and Mrs. Corinne Bacon (Executive Services Manager, CSEE). Thank you for your warmth and support.

I also wish to thank the University of Essex International College (UEIC) for giving me my first teaching opportunity in the UK. I am especially grateful to Mrs. Michelle Gibbons, Dr. Ryan Flitcroft, and Mrs. Stephanie King for their mentorship and unwavering support.

Beyond the lab, I was blessed with a wonderful community of mentors and friends who constantly looked out for me, including Dr. Jonathan Iworiso, Dr. Felix Ngobigha, Dr. Perekebode Amangele, Dr. Imeh Umoren, Dr. Godswill Agbaitoro, and Dr. Nestor Asiamah. Your guidance, encouragement, and reminders to stay focused were invaluable.

I extend my appreciation to my fellow PhD colleagues at CSEE. To my seniors, Dr. Milan Rybar, Dr. Leena Marreel, Dr. Yiyuan Han, Dr. Zilu Wang, and Ahmet Mercimek thank you for your generosity, advice, and shared knowledge. To my peers, Alberto Tates Puetate, Dr. Guangming Huang, and Dr. Zeel Pansara, I am grateful for your support, collaboration, and friendship. To the newer colleagues—Vishnu Kirar, Rizeng Wu, Jingyan Wang, and Shiguang Liu—thank you for keeping our lab atmosphere lively and collegial. I also extend my thanks to the wider PhD community at CSEE, especially Ise Anderson, Cristina Del Prete, Mohammadamin Jarrahi, and Tamim Almulla, for creating an inspiring and supportive academic environment. Serving as your representative was an honor.

This thesis would not have been possible without the study participants. Though I never met them personally, I am deeply grateful for their time, patience, and willingness to contribute their brain activity for research. I also thank my collaborators at the University Hospital of Lausanne, particularly Prof. Karin Diserens, Jane Jöhr, and Marina Lopes da Silva, for their dedication in collecting and sharing the dataset used in this work.

I am honored that Prof. Damien Coyle and Dr. Vito De Feo agreed to serve on my defense committee and evaluate this thesis. Thank you for your time and expertise.

I owe profound gratitude to TETFund Nigeria and Akwa Ibom State University (AKSU) for sponsoring my doctoral studies in the United Kingdom. My appreciation also goes to my colleagues at AKSU, especially Mrs. Abasiama Philip Chuks, Mr. Aloysius Akpanobong, and Dr. Godwin Ansa, for their encouragement and support.

I am grateful to my faith communities, including Deeper Life Bible Church, Colchester (Pastor Samuel Sevonukun) and the Redeemed Christian Church of God, Kingdom Life Zone, Ikot Abasi (Pastor Emmanuel Ikott), for their prayers and encouragement. Special thanks also to Mr. Kelvin Idehen and Mrs. Comfort Archibong for their constant support. I would also like to thank all my students at AKSU, the University of Essex, UEIC, and the University of Suffolk. Teaching you was both a privilege and a transformative experience for me. I equally appreciate everyone at The BlueCorner Experience for the inspiring fellowship and mindset-driven engagements.

Finally, my total gratitude goes to my family. To my parents, Mrs. Victoria Etukudoh and Mr. Akwaowo Nelson Amaunam, my mother-in-law, Mrs Esther Udofia, thank you for your unwavering belief in me. To my beloved wife, Emem, this PhD bears my name, but it truly belongs to you. You carried the weight of this journey with me—through the sleepless nights, the sacrifices, and the triumphs—and I am eternally grateful. To my children, Zach and Zoe, thank you for your patience, love, and for enduring my many absences. This thesis is dedicated to you three—Emem, Zach, and Zoe.

Contents

List of Figures

List of Tables

List of Abbreviations

ABI	Acquired Brain Injury.
AI	Artificial Intelligence.
ANOVA	Analysis of Variance.
BCI	Brain-Computer Interface.
CBA	Chance-level Based Accuracy.
cCMD	clinical Cognitive Motor Dissociation.
CMD	Cognitive Motor Dissociation.
CNN	Convolutional Neural Networks.
CRS-R	Coma Recovery Scale-Revised.
CSP	Common Spatial Filter.
CUDA	Compute Unified Device Architecture.
CuNN	CUDA Deep Neural Network.
DFT	Discrete Fourier Transform.
DOC	Disorders of Consciousness.

DTF	Directed Transfer Function.
EEG	Electroencephalography.
eMCS	emergence from Minimally Conscious State.
ERD/ERS	Event-Related Synchronization/Desynchronization.
ERP	Event-Related Potential.
FBCSB	Filter Bank Common Spatial Pattern.
FDR	False Discovery Rate.
FES	Functional Electrical Stimulation.
FFT	Fast Fourier Transform.
fMRI	functional Magnetic Resonance Imaging.
fNIRS	functional near-infrared spectroscopy.
FORCe	Fully Online and Automated Artifact Removal for BCI.
FS	Fisher Score.
GCS	Glasgow Coma Scale.
GPU	Graphics Processing Unit.
ICU	Intensive Care Unit.
iLIS	incomplete Locked-in Syndrome.
KVc	Kolmogorov complexity.

LDA	Linear Discriminant Analysis.
LIS	Locked-in Syndrome.
LOSO	Leave-one-session-out cross validation.
LOTO	Leave-one-trial-out cross validation.
LZc	Lempel-Ziv complexity.
MBT	Motor Behaviour Tool.
MCS	Minimally Conscious State.
MEG	magnetoencephalography.
MI	motor imagery.
ML	machine learning.
MVAR	Multivariate Autoregressive model.
PCI	Perturbational Complexity Index.
PET	Positron Emission Tomography.
PSD	Power Spectral Density.
QDA	Quadratic Discriminant Analysis.
ROI	Region of Interest.
SDTF	Short-time Directed Transfer Function.
SMR	Sensorimotor Rhythms.
SOTA	State-of-the-art.

SSVEP	Steady-State Visually Evoked Potentials.
TBI	Traumatic Brain Injury.
TMS	Transcranial Magnetic Stimulation.
UWS	Unresponsive Wakefulness Syndrome.
VS	Vegetative State.

Chapter 1

Introduction

1.1 Background

A Brain-Computer Interface (BCI) is a cutting-edge neuro-technological system designed to establish a direct communication channel between neural activity and external systems, circumventing traditional neuromuscular output pathways. This property makes BCIs particularly valuable for probing and assessing consciousness or awareness in acute clinical conditions. The concept of consciousness and/or awareness has a rich, multifaceted history spanning religion, philosophy, psychology, neuroscience, and, more recently, artificial intelligence. Its interpretation has evolved over centuries, shaped by cultural, spiritual, and scientific frameworks.

In Christian theology, awareness is often understood as spiritual proximity to God, moral uprightness, and the capacity for supernatural insight into human existence [11]. Similar perspectives are echoed in Islamic texts, where awareness (taqwa) encompasses moral consciousness and spiritual vigilance [12]. In Hindu philosophy, particularly in the Mandukya and Aitareya Upanishads [13], awareness is articulated through states of consciousness such as waking, dreaming, and deep sleep. Buddhism, by contrast, frames awareness primarily as mindfulness—a sustained, nonjudgmental attention to present ex-

perience [14].

In ancient philosophical traditions, the Greeks and Egyptians linked consciousness to the soul, and awareness to functions of the heart and brain. Aristotle, for example, explored these ideas through discussions on perception, mental imagery, intellect, and the soul [15, 16]. Later philosophers such as John Locke and David Hume expanded on these foundations, with Locke associating consciousness with the continuity of self and Hume highlighting its fleeting, perceptual nature. Psychology later introduced more structured distinctions, and neuroscience advanced these into empirically testable domains.

Sigmund Freud famously categorized awareness into three levels: the conscious, preconscious, and unconscious. Although largely based on introspective methods, his work marked a shift toward understanding consciousness as both a subjective phenomenon and a response to external stimuli—an early bridge between psycho-analytic and neuro-cognitive models. However, contemporary theories of consciousness [17] often divide the phenomenon into three major frameworks:

- Global Workspace Theory (GWT): posits that consciousness arises when information is broadcast across various neural systems, enabling coordinated and flexible behavior.
- Higher-Order Theories (HOT): suggest that consciousness depends on the brain's ability to represent its own mental states—a meta-cognitive mechanism.
- Integrated Information Theory (IIT): proposes that consciousness corresponds to the degree of integrated information generated by a system, reflecting its ability to unify diverse inputs into a coherent experience.

These theoretical models have been refined through advances in neuroscience and have influenced the development of artificial intelligence. In both fields, consciousness is increasingly understood as emerging from complex processes involving perception, cognition, and

internal representation—attributes once attributed to the soul or mind. This convergence of philosophical, religious, and scientific perspectives continues to inform the interdisciplinary inquiry into the neural correlates and computational models of consciousness, reinforcing the idea that the brain is the central organ of conscious experience.

1.2 Consciousness in Clinical Dimension

Consciousness and awareness, though integral to our daily life, are inherently subjective with interpretations varying across disciplines and individuals based on physiological, psychological, and clinical backgrounds [18]. Nevertheless, in the clinical setting and according to the latest widely acceptable definitions [1], awareness refers to the content of one's subjective experience, including external awareness–such as the ability to perceive, feel and be cognizant of events and stimuli [19], and internal awareness (e.g., stimuli-independent inner thoughts and speech). Consciousness (or "conscious awareness"), on the other hand, can be defined as the state of simultaneously exhibiting wakefulness/arousal (i.e., being awake, alert, vigilant and attentive) and awareness of one's self and surroundings.

Hence, as demonstrated by Laureys [20], the clinical spectrum of consciousness is directly related to the content of awareness. Individuals with high ability to perceive and understand exhibit higher levels of both consciousness and awareness, while patients who lack the innate ability to follow commands show significantly reduced levels of both. Disorders of Consciousness (DOC) is the clinical condition characterized by disruptions across both the awareness and consciousness components.

However, some clinical patients retain cognitive awareness while experiencing complete or partial motor paralysis. Such individuals are capable of thinking, understanding, perceiving, and hearing, but are unable to produce reliable voluntary movements—whether overt or subtle—that would signal their conscious state. This condition is referred to as Cognitive Motor Dissociation (CMD). Evidence suggests that CMD exists along a spec-

trum, ranging from a complete absence of motor output to the presence of weak or inconsistent voluntary movements. Patients who demonstrate partial or intermittently visible movements are classified as having clinical Cognitive Motor Dissociation (cCMD) [21].

Clinical diagnosis of DOC relies predominantly on behavioral assessment tools, which inherently require responses to externally delivered commands such as "open your eyes," "say your name," "move your foot" etc. Consequently, the diagnostic framework is deeply dependent on observable motor behavior.

DOC most commonly arises from Acquired Brain Injury (ABI) such as traumatic or cerebrovascular accidents, but may also result from neurodegenerative processes or pharmacological side-effects. Clinically, DOC is categorized into four principal states: coma [22], Vegetative State (VS)/Unresponsive Wakefulness Syndrome (UWS) [23], Minimally Conscious State (MCS)+ and MCS- [24], and emergence from Minimally Conscious State (eMCS) [24, 25]. Here, MCS+ denotes a higher level of preserved awareness compared to MCS-.

Although eMCS is often considered part of the DOC continuum, patients at this stage may demonstrate clear signs of regained awareness. It represents a transitional phase between MCS and full functional recovery, as illustrated in Figure 1.1. Patients remain classified within DOC until they meet the formal emergence criteria on the Coma Recovery Scale-Revised (CRS-R), specifically: the presence of functional communication or functional object use.

However, this classification is subject to debate because the CRS-R emergence criteria can fail to capture cognitive recovery in patients with significant praxic, aphasic, dysexecutive, or motor impairments. Such constraints may prevent the expression of the required behaviors despite meaningful improvements in consciousness. As a result, some patients may remain artificially positioned below the CRS-R threshold, not because of impaired awareness, but because their motor system is unable to manifest it.

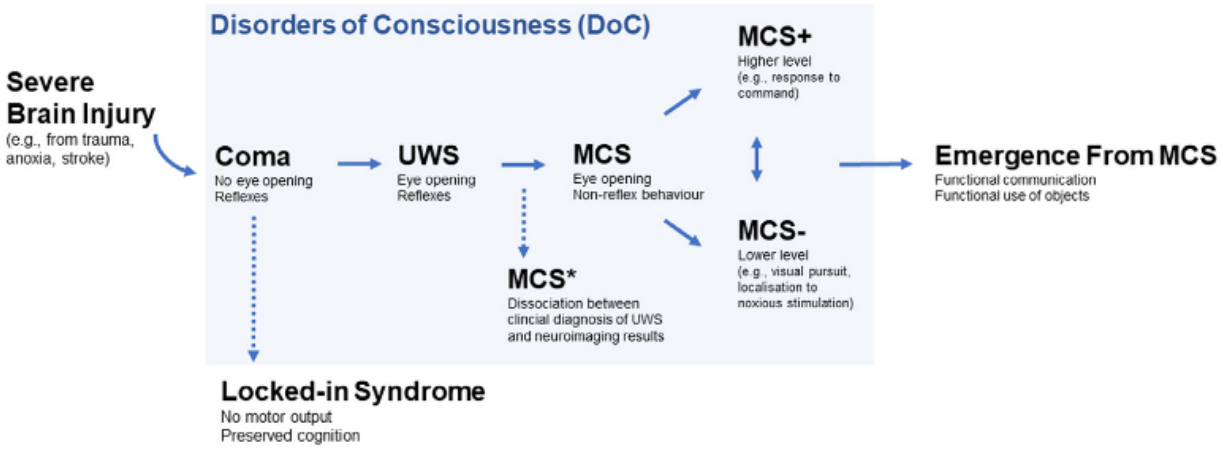

Figure 1.1: DOC spectrum rhythms adapted from [1]

1.3 Prognosis, Detection and Diagnosis of DOC

Prognosis of DOC—that is, the prediction of likely clinical outcomes—plays a critical role in guiding treatment decisions and informing end-of-life care or planning [26]. However, prognostication in this context remains a significant challenge due to the complex nature of consciousness and the limitations of current assessment methods [27, 26]. Accurate prediction often depends on preceding steps, such as detection and diagnosis.

Detection refers to the real-time identification of whether a patient is aware or conscious, particularly in cases where behavioral signs are absent or ambiguous. Diagnosis, on the other hand, involves classifying the patient's state of consciousness—such as coma, UWS or MCS- based on observable clinical or neurophysiological factors.

While these processes are conceptually distinct, in clinical practice they often rely on overlapping methods, including:

- Behavioral scales (e.g., CRS-R, Glasgow Coma Scale (GCS), Motor Behaviour Tool (MBT))
- Neuroimaging and Electrophysiology (e.g., Electroencephalography (EEG), functional Magnetic Resonance Imaging (fMRI), Positron Emission Tomography (PET))

- BCIs

Together, detection, diagnosis, and prognosis form an integrated framework that supports evidence-based decision-making and improving the management of DOC patients.

1.4 Neuro-imaging and Electrophysiology

Brain imaging and studies employing functional or electrophysiological brain imaging have emerged as the most promising approach to correcting the challenges of behavioral scales [28, 29]. They have provided mounting evidence that covert awareness may be concealed by a patient's inability to produce voluntary motor output spontaneously, or in response to commands and/or stimulation, a situation identified as CMD [30, 31]. Cases of CMD where the afflicted individual is known (or likely) to maintain high-level cognitive function are referred to as Locked-in Syndrome (LIS) (complete or incomplete, depending on whether there is residual motor activity, and/or spared communication channels).

fMRI initially dominated the field due to its high spatial resolution and early landmark studies [32, 33, 34]. More recently, the more practical, portable, and cost-effective EEG has gained increasing attention [35, 36, 37, 38]. Hybrid approaches combining multiple imaging modalities have also been proposed [39, 40]. Overall, while fMRI provided the foundation for early breakthroughs, EEG continues to gain widespread adoption due to its portability, affordability, reliability and superior temporal resolution; these properties are critical for producing impactful neuroimaging-based DOC diagnostic tools that are inexpensive, practical and logistically compliant with daily use in hospitals.

1.5 Brain-Computer Interfaces(BCIs)

A BCI is an advanced neurotechnological system that facilitates direct communication between the brain and external devices by bypassing conventional neuromuscular path-

ways. It operates by capturing brain signals, extracting and selecting informative features, and applying machine learning (ML) algorithms/models to interpret these features into actionable outputs. These outputs enable functions such as external device control, therapeutic interventions, and clinical decision-making, without requiring voluntary motor input. A closed-loop BCI system typically comprises the following sequential components:

- Signal acquisition: This initial stage involves recording neural signals from human subjects. Acquisition is mostly performed using non-invasive methods like EEG, magnetoencephalography (MEG), or functional near-infrared spectroscopy (fNIRS), which offer safer and more accessible alternatives, albeit with lower signal fidelity. Invasive techniques—such as intra-cortical or electrocorticographic electrodes, which provide high spatial and temporal resolution are also explored and relevant.

- Feature extraction: Given the complexity and dimensionality of raw brain signals, it is crucial to distill relevant information that reflects cognitive or motor intent. This step involves signal transformation techniques (e.g., time-frequency analysis, wavelet decomposition) to capture discriminative characteristics suitable for downstream analysis.

- Feature selection: Not all extracted features contribute equally to performance. In this stage, statistical or algorithmic methods are employed to identify a subset of features that offer optimal separability between classes or states, thereby improving classification accuracy and reducing computational burden.

- Classification and ML analysis: Machine learning models are trained to recognize patterns within the selected features. These models—ranging from linear or shallow classifiers to deep learning architectures—translate neural activity into predefined outputs corresponding to user intentions or states.

- Post-processing: The output of the classification stage is refined through smoothing,

filtering, or error correction techniques. This step ensures temporal consistency and enhances the reliability of the BCI system in real-time applications.

- Feedback: The final component involves returning the system's output to the user or an external device. Feedback may be provided through various modalities—such as visual cues, auditory signals, tactile stimulation, or direct control of assistive technologies—closing the control loop and facilitating user adaptation and learning.

1.6 Deep-learning and Accuracy of predictions

Historically, shallow models [41, 42] have dominated both classical computing and, more specifically, the ML classification paradigms. Within the BCI domain, these approaches provided the basis for decades of research. The landscape began to shift about a decade ago with the introduction of deep learning architectures. In the context of classical BCI, this transition was marked by the development of EEGNet [7], followed by DeepConvNet [8] as the study of perceptrons and layered architectures advanced into convolutional neural networks. More recently, the introduction of attention mechanisms in deep learning [43] has led to the rise of transformer-based architectures, with the EEGConformer [9] emerging as one of the most popular models in contemporary EEG and BCI research. Although deep learning models have consistently demonstrated superior classification accuracy compared to their shallow counterparts, they are highly data-intensive. Effective generalization and model stability typically require large quantities of high-quality data [44, 45, 46].

However, EEG recordings are characteristically noisy [47], highly variable across subjects [48, 49], and often limited in volume [45], which increases computational burden and necessitates extensive hyperparameter tuning [8, 50]. Because of the substantial cost associated with this offline optimization, such models are usually trained once and subsequently deployed in real-time (online) applications for diagnostic use. Yet, patient variability—particularly in DOC—poses a significant barrier to effective online deployment.

Pretrained models rarely generalize well to new patients, largely due to the absence of sufficiently large, high-intensity datasets within this clinical domain [34, 38, 51].

This limitation underpins the decision to conduct the analyses in this thesis offline, consistent with prevailing practices reported in the literature [36, 52, 53]. Although online decoding would be preferable—owing to its real-time responsiveness and potential therapeutic value—the classification accuracies achievable with the limited data typically available from DOC patients are insufficient for training a reliable online model. Prior studies indicate that a threshold below 70% classification accuracy leads to frustration and abandonment, while above 70-80% enables stable online operation [54, 55]. However, in the DOC population, and as observed in this work, such performance levels are rarely attained. This shortfall renders current online paradigms impractical for diagnostic use and raises concerns regarding the validity of closed-loop BCI implementations in DOC.

However, it is important to emphasize that online protocols represent a desirable future direction. Following the diagnosis of retained awareness, the logical next step would be to restore functional communication pathways—an overarching goal of BCIs. However, such interventions extend beyond the diagnostic scope addressed in this thesis.

This challenge underscores a fundamental limitation of traditional BCI systems, which must operate under strict real-time constraints, limited available trials, and pronounced inter- and intra-subject variability—all within an environment where the quantity and quality of data are intrinsically restricted.

1.7 Limitations to Clinical diagnosis of DOC

Over the years, several constraints have limited the accuracy and reliability of clinical diagnosis in patients with DOC. These include:

(a) An over-reliance on the behavioral assessment tools has shifted the diagnostic framework towards measuring observable behaviors alone, rather than actual awareness.

While these tools—such as the CRS-R—are clinically standardized, they provide limited sensitivity to CMD, making them prone to misdiagnosis.

(b) Although BCI-inspired and other neuro-imaging methods have shown promise in research contexts, they have not been widely adopted in clinical practice. This is primarily due to a lack of robust validation, limited reproducibility and discrepancies between reported results and real-world clinical utility.

(c) As a consequence of (b), many state-of-the-art techniques proposed in literature have not been independently replicated, leading to skepticism regarding the reliability of proposed neurophysiological biomarkers for awareness.

1.8 Motivations

The diagnosis of DOC traditionally relies on standardized behavioral scales such as the CRS-R and the GCS [3, 56, 57]. While widely adopted, these instruments have significant limitations and are particularly vulnerable to diagnostic errors, especially false negatives (type II errors) [58, 59, 60]. To address these shortcomings, alternative behavioral scales, such as the MBT [61], have been proposed. Nevertheless, the heavy reliance of all behavioral assessments on motor output contributes substantially to the 37–43% misdiagnosis rate reported in the literature [62, 56, 63]. Crucially, some patients may show no overt behavioral signs of awareness yet demonstrate clear evidence of command-following when assessed with neuroimaging [64, 32, 34, 65, 66]. This condition, known as CMD, underscores the fundamental limitations of behavior-based diagnostics.

To address these shortcomings, and improve DOC diagnostic accuracy, researchers have increasingly turned to neuroimaging-based approaches. Early breakthroughs were achieved with fMRI, whose high spatial resolution enabled landmark demonstrations of covert command-following [32, 33, 34]. However, despite its strengths, fMRI is costly,

logistically demanding, and not easily deployable at the bedside. In contrast, EEG has emerged as a more practical alternative due to its portability, lower cost, bedside applicability, and superior temporal resolution [35, 36, 37, 38]. Hybrid approaches that combine imaging modalities have also been explored [39, 40]. Despite these developments, a pressing clinical need remains for a neurophysiological diagnostic tool that: (i) does not rely on overt behavior, (ii) enables reliable detection of covert awareness at the bedside, and (iii) is scalable for routine clinical use.

EEG, with its non-invasiveness, high temporal fidelity, and bedside applicability, is well positioned to meet these requirements, and BCIs have emerged as a promising framework for detecting volitional brain activity in DOC patients. Yet, despite encouraging results, clinical adoption of BCIs for CMD diagnosis remains limited. A major barrier is the lack of standardized, validated metrics, which undermines reproducibility and prevents robust clinical translation. This leads to a critical research question: can existing state-of-the-art EEG-based methods be reliably replicated and validated to detect CMD in DOC patients using a novel dataset, while achieving results consistent with prior literature?

Beyond this challenge, the classical BCI pipeline— comprising signal acquisition, preprocessing, feature extraction, classification, post-processing, feedback, and application—is conceptually well established. However, considerable variability exists in the methodological choices within this pipeline. Some studies rely on shallow machine learning classifiers, whereas others employ deep learning architectures. This methodological heterogeneity has produced a diverse but fragmented body of evidence, raising persistent concerns about reproducibility, interpretability, and clinical relevance. These limitations highlight a second key research gap: (i) do deep learning methods confer unique diagnostic advantages over shallow classifiers for CMD detection in DOC?, (ii) do shallow models remain clinically relevant in the era of deep learning?, and (iii) how best to balance accuracy, interpretability, and reproducibility in clinical BCI applications. This motivates the second critical research question: do deep learning methods developed for motor imagery (MI)-based EEG

analysis offer added diagnostic value when applied to DOC?

1.9 Thesis Objectives

Motivated by these gaps, the overall goal of this thesis is to advance the diagnosis of awareness in DOC by validating different aspects of state-of-the-art BCI-inspired protocols on a novel patient dataset.

This work is thus structured around two main objectives;

1. **Objective 1 – Validation of state-of-the-art EEG-based diagnostic metrics**

    - Replicate and validate existing EEG-based metrics of diagnosing awareness (these include classification accuracy, brain rhythms and their separability between conditions, effective connectivity, spectral slope, Perturbational Complexity Index (PCI) using a novel dataset of 28 DOC patients recorded in the Intensive Care Unit (ICU) with a MI-based command-following EEG paradigm.
    - Assess the consistency of these metrics with clinical gold-standard assessments (CRS-R scores).
    - Assess the discriminant threshold between aware and non-ware.

2. **Objective 2 - Evaluation of deep learning approaches for EEG-based diagnosis**

    - Compare shallow classifiers with state-of-the-art deep learning architectures tailored for EEG analysis.
    - Assess differences in classification performance, interpretability, and clinical relevance.
    - Determine whether deep learning provides added diagnostic value beyond established shallow approaches in the context of DOC analysis.

1.10 Thesis Arrangement

The experiments presented in this thesis were conducted, and the thesis was written, with a view toward realizing the clinical adoption and usage of BCI-inspired neuro-imaging protocols in DOC populations. Thus, the structure, contents and contributions of each Chapter is as follows:

- **Chapter 2** provides a comprehensive literature review of relevant methods, algorithms, and theoretical foundations underpinning this thesis, with particular emphasis on EEG, BCI, DOC, and related methodologies. Each section is structured to present the necessary background while explicitly linking to the research objectives of this work.

- **Chapter 3** outlines the common dataset, experimental setup, protocols, data processing procedures, and methodological approaches employed throughout the thesis.

- **Chapter 4** presents the validation of diagnostic metrics, incorporating the specific methodological needs, associated results, and a critical discussion of the findings.

- **Chapter 5** introduces the deep learning architectures and models applied in this study, alongside the experimental results and an in-depth discussion specific to these approaches.

- **Chapter 6** provides a general discussion and synthesis of the research findings, followed by the conclusions and recommendations for future work. This chapter also outlines the key limitations of the present study.

1.11 Contributions to Knowledge

This thesis advances the diagnosis of awareness in DOC via BCI-inspired protocols by;

- Validating of EEG-based neuro-markers using a novel dataset from a real ICU environment.
- Establishment of quantitative thresholds for diagnosing awareness in DOC.
- Evaluation of deep learning against traditional EEG classifers, thus providing context-specific insights into the added value of deep learning.

1.12 Conclusion

The complexity of consciousness and awareness indicates that these constructs are more accurately represented as a continuum rather than as discrete, easily bounded states. Determining an individual's position along this continuum—which itself contains sub-spectra, as exemplified by the heterogeneity of the MCS [24, 1]—is inherently challenging and complicates behavioral diagnostic efforts. Although the CRS-R remains the clinical gold standard, its assessments are not immune to type I and type II errors.

These limitations have motivated the exploration of neuroimaging techniques augmented by advanced computational algorithms. However, rather than producing diagnostic convergence, these approaches have often increased variability, largely due to lack of validation across methods and datasets. For example, in ML–based classification, some researchers advocate for deep-learning architectures [8, 7, 9, 49, 67], whereas others report superior performance using shallow classifiers [55, 48, 67]. While the capabilities of deep learning are well documented, their advantages within EEG BCI-inspired diagnostic paradigms remain contested.

Therefore, there is a pressing need to rigorously validate the proposed metrics and systematically evaluate the perceived strengths—or lack thereof—associated with different components of the EEG-based diagnostic pipeline for DOC, which this thesis addresses.

Chapter 2

State-of-the-Art

This chapter presents an in-depth analysis and critical review of the relevant literature on awareness and consciousness. It specifically focuses on the role of BCIs in the assessment of consciousness, examining the methods, algorithms, and approaches employed in signal acquisition, mental task selection, and the refinement of these tasks for diagnostic purposes. Although most of these concepts were introduced in Chapter One, this chapter further provides a more comprehensive and detailed evaluation of the key research areas directly related to the objectives of this thesis.

2.1 EEG and the Brain

During human development, spontaneous neural activity in the brain can be observed between the 17th and 23rd weeks of fetal gestation [68]. It is therefore widely known that from the prenatal stage and throughout the human lifespan, the brain continuously generates and transmits electrical signals. These signals serve either as outputs driving physiological functions or as representations of the body's internal state [69, 70]. Consequently, the ability to record or measure these signals—either invasively or non-invasively—provides valuable insight into the functional condition of the human brain [47].

Neurons, the fundamental units of the brain, operate through the movement of key ions such as sodium (Na^+), potassium (K^+), calcium (Ca^{2+}), and chloride (Cl^-) across their membranes via specialized ion channels [71, 69]. At rest, the interior of a neuron maintains a negative electrical potential relative to the external environment, typically around –70 mV. This resting potential is maintained until a stimulus raises the membrane potential to a critical threshold, typically around –55 mV [69, 72]. Once this threshold is exceeded, voltage-gated sodium channels open, allowing Na^+ ions to flow into the cell, rapidly depolarizing the membrane to approximately +30 to +40 mV. This change triggers an action potential, which propagates along the axon, away from the cell body and toward the axon terminals, transmitting neural information as a wave of depolarization [73, 69, 72].

When synchronized across populations of neurons, these action potentials produce bioelectrical signals that can be detected on the scalp as weak voltage fluctuations using EEG [47, 74]. It is important to note that altered neuronal excitability can significantly affect brain function. Hyperexcitability, where neurons fire action potentials excessively in response to minimal stimuli, may reduce the threshold for activation and contribute to disorders such as spasticity or epilepsy [75]. Conversely, hypoactivity or unresponsiveness to stimuli can elevate the activation threshold, potentially leading to DOC, such as coma or UWS [65].

EEG is particularly suited for detecting these brain signals, which are typically in the microvolt to millivolt range [47, 76]. The effective sampling of these signals is constrained by the Nyquist theorem [77, 78], which states that the sampling frequency must be at least twice the highest frequency present in the signal to prevent aliasing [77, 78]. This makes EEG an effective non-invasive tool for monitoring real-time brain activity, particularly in clinical and research applications related to consciousness.

EEG a widely used neuroimaging technique for recording the electrical activity generated by neural tissues in the human brain. It operates by measuring voltage fluctuations on the scalp surface using conductive electrodes, typically arranged according to the con-

ventional 10–20 system. An example of such a device is shown in Figure 3.1F. Despite its utility, EEG is highly susceptible to artifacts, which makes robust signal-cleaning procedures essential when used in research or clinical practice [79, 80]. It also suffers from relatively poor spatial resolution, since the recorded signals often represent a summation of postsynaptic potentials referenced to an electrode [81]. Being non-invasive, EEG cannot capture deep brain activity such as subcortical potentials [82]. In practice, it can also be time-consuming to set up and may be uncomfortable for participants to wear. Nevertheless, these limitations are outweighed by its advantages. EEG remains the de facto non-invasive neuroimaging tool due to its portability, cost-effectiveness, reliability, and suitability for bedside use in clinical environments.

2.1.1 History of EEG and brain waves

Brain electrical activity was first discovered by English Physician and Scientist–Richard Caton in 1875 [83, 2]. Working with exposed cortical surfaces of cats and rabbits, Caton observed that a sensitive galvanometer detected weak electric currents varying in direction. These currents were recorded as they passed through electrodes placed on both the skull and the brain's gray matter. Although rudimentary, Caton's findings marked the first documentation of what would later be known as electroencephalographic EEG activity. Building on Caton's work, Adolf Beck advanced the field by identifying the rhythmic oscillatory nature of brain activity. He also observed that these rhythms diminished or disappeared when an animal was exposed to stimuli, such as light, or when it engaged in a task. This phenomenon, now referred to as desynchronization, arousal, activation, or alpha-blocking, highlighted the functional responsiveness of brain waves to sensory input and cognitive engagement. Later, Napoleon Cybulski contributed further by integrating a photographic recording system into the galvanometer setup. This innovation allowed brain electrical activity to be visualized in graphical form, enabling more precise analysis and

archiving of EEG signals.

However, the first successful recording of EEG signals from the intact human scalp was achieved in 1924 by Hans Berger, a German neurologist. Using modified radio equipment for signal amplification, Berger demonstrated that the brain's weak electrical currents could be recorded non-invasively, without penetrating the skull. He noted that the recorded electrical activity varied with physiological states such as sleep, anesthesia, hypoxia, and neurological disorders. Berger also introduced the term electroencephalogram EEG and systematically laid the scientific and methodological foundations for modern EEG technology and its applications in both clinical and research contexts [84, 85].

2.1.2 Brain waves classification

Brain waves typically appear as sinusoidal patterns and are most prominently observed during relaxed states, particularly when the subject has their eyes closed. These baseline oscillations often reflect a phenomenon known as desynchronization, which represents the attenuation of rhythmic activity in response to cognitive or sensory engagement. EEG signals are measured in volts, typically ranging from 0.5 to 100 μV when recorded on the scalp [47, 83]. Through power spectral analysis, particularly, Fast Fourier Transform (FFT) the raw EEG signal can be decomposed into its constituent frequency components. Although EEG signals exhibit a continuous spectral distribution, their dominant frequency content generally lies between 0 and 40 Hz. In this study, analysis was restricted to the 0–24 Hz frequency band, as higher frequencies do not reliably reflect meaningful EEG activity associated with motor-related processes and are more susceptible to noise and artifacts [86, 87]. Although conventional EEG wave classification extends into the gamma range (30–200 Hz), these higher-frequency components are particularly prone to contamination from muscle activity, electrical interference, and other non-neural sources [87, 80]. For this reason, gamma-band activity is not included or reported in this thesis.

Based on this distinction, brain waves are classified into four primary categories:

- Delta: 0 - 4 Hz.
- Theta: 4 - 8 Hz.
- Mu: 8 - 14 Hz.
- Beta: 14 - 24 Hz.

A typical illustration of this is presented below:

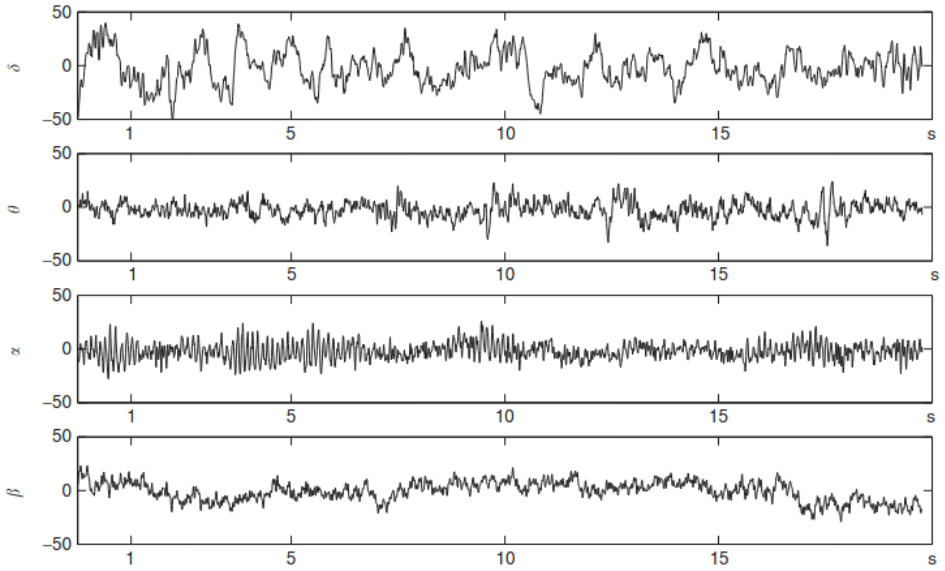

Figure 2.1: EEG brain rhythms adapted from [2] From top - δ(0-4)Hz, θ(4-8)Hz, μ(8-14)Hz, β(14-24)Hz

2.1.3 EEG Recording

As illustrated in Figure 2.2 the EEG, when compared to other neuroimaging techniques such as PET, fMRI, exhibits relatively poor spatial resolution. However, this limitation is

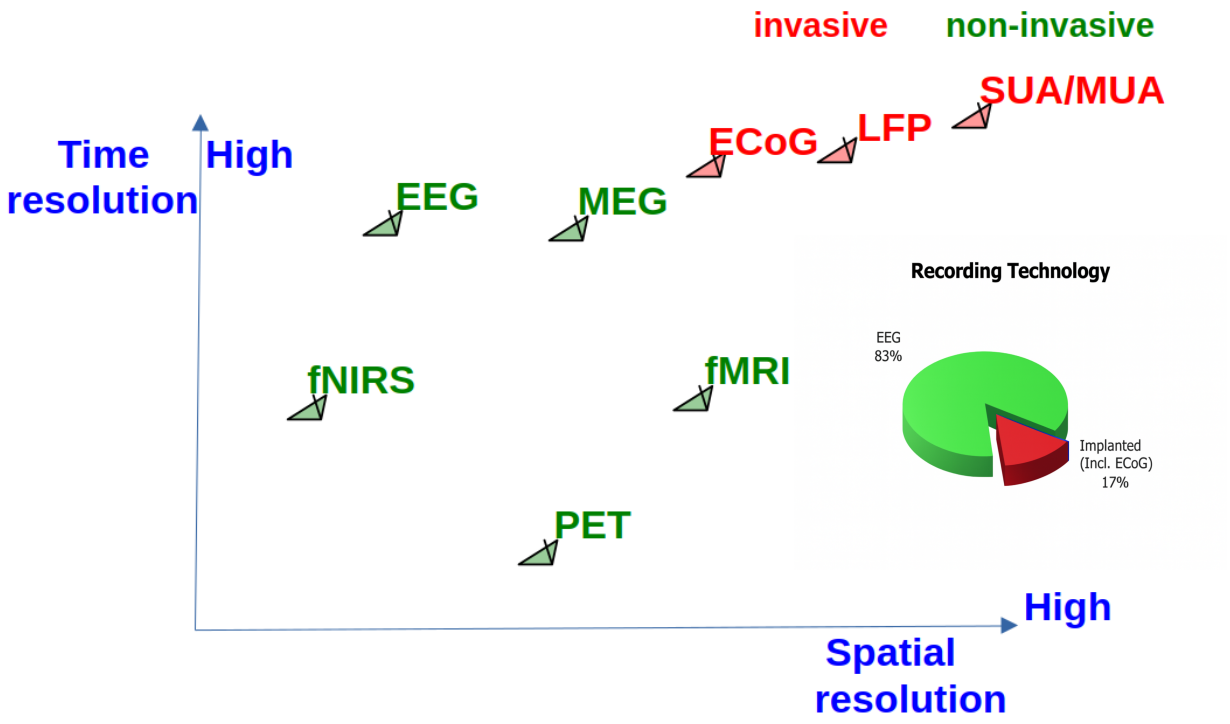

Figure 2.2: EEG linear map

offset by its superior temporal resolution, enabling the capture of neural dynamics that occur within milliseconds of stimulus onset. This makes EEG especially suitable for studying fast-paced brain processes in real time.

One of the strengths of EEG lies in its ability to simultaneously record electrical activity from multiple regions of the brain. To achieve this, electrodes are strategically arranged on the scalp—either covering the entire head or focusing on a specific Region of Interest (ROI). The human brain is anatomically divided into two hemispheres—left and right–and further segmented into functional regions such as the frontal cortex (responsible for decision-making, personality, and speech), the motor cortex (governing voluntary movements and fine motor control) and the parietal cortex (involved in sensory integration, spatial awareness, and hand-eye coordination) [72, 88].

EEG measures the brain's electrical activity by recording the potential difference between two electrodes, typically an active and a reference electrode. These signals are

extremely weak (in the microvolt range) and require amplification for accurate digitization and analysis. A standard EEG setup consists of several key components, including an EEG cap (with embedded electrodes), conductive gel or paste (to reduce impedance), amplifiers, measuring tape (for correct electrode placement), syringes (for gel application), and other accessories, as depicted in Figure 2.3. EEG caps can be either wet (requiring gel) or dry, with ongoing research exploring the relative performance of both. Electrode systems are typically multi-channel, ranging from as few as 2 up to 256 electrodes. For the purposes of this thesis, a 16-channel montage was employed.

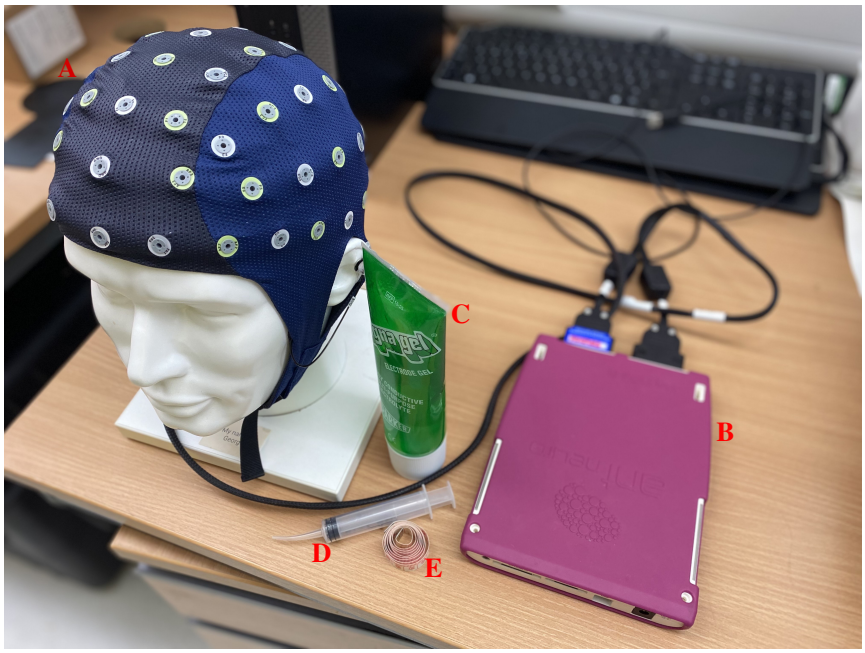

Figure 2.3: EEG recording equipment: (A) 64 channels antNeuro EEG cap, (B) antNeuro eego mylab amplifier, (C) Signa electrode gel, (D) syringe, and (E) measuring tape

To ensure high quality recordings, electrodes must be functional and capable of capturing clean signals. These electrodes—typically 1 to 3mm in diameter—must exhibit skin-electrode impedance below 5kΩ. Conductivity is enhanced through proper gelling and accurate electrode placement. Correct positioning is not only critical for signal fidelity but also for the accurate topographical interpretation of EEG data.

To standardize electrode placement, the International 10–20 System was introduced in 1958 by the International Federation of Societies for Electroencephalography and Clinical Neurophysiology [89]. In this system, electrodes are placed at intervals representing 10% or 20% of the total front-to-back or side-to-side distance of the skull. Electrode sites are labelled according to brain regions: F (frontal), C (central), T (temporal), P (parietal), and O (occipital). Odd numbers denote the left hemisphere, while even numbers refer to the right hemisphere [89] (see Figure 2.3A and Section 3.3). For the experiments in this thesis, electrodes were primarily concentrated over the motor cortex.

Maintaining low and balanced electrode impedance is essential, as high impedance can degrade signal quality and increase susceptibility to external noise. Impedance values should be below 5kΩ and balanced within 1kΩ across electrodes [83]. Impedance monitoring devices are typically used to ensure these standards are met.

Equally, EEG is a differential measurement of voltage, thus, the choice and location of the reference electrode are critical. The reference can be placed at a central scalp site, or on neutral sites such as the earlobe, mastoid, nose, or chin. However, no universal consensus exists regarding optimal reference placement, largely due to volume conduction—the spread of electrical activity from other physiological sources throughout the body [2].

EEG signals must be amplified to be compatible with acquisition hardware and to ensure an adequate signal-to-noise ratio for storage and analysis. Amplification is often accompanied by filtering, which can be performed either in hardware or as part of pre-processing. A high-pass filter is used to eliminate slow-drifting signals and DC offsets, while a low-pass filter removes high-frequency noise, ensuring that only relevant EEG components are retained for analysis.

2.1.4 Artifacts

One of the major challenges in EEG is the presence of artifacts, or unwanted noise, in the recorded signal. Since the primary ROI is the brain, any electrical activity arising from non-cerebral sources—whether from within the body or the external environment—that contaminates the EEG recording is considered an artifact. Artifacts are broadly classified into two categories: physiological and extraphysiological [47].

Physiological artifacts originate from the human body but are unrelated to brain activity. Common examples include ocular artifacts (e.g., eye movements and blinks), muscle artifacts (e.g., clenching of the jaw, chewing, or swallowing), and occasionally cardiac artifacts—the electrical activity of the heart—especially if the electrodes are positioned near large arteries or poorly shielded [47, 90, 91]. Although cardiac artifacts were unlikely to interfere in the experiments reported in this thesis, they remain a potential source of contamination in some EEG protocols.

In contrast, extraphysiological artifacts are induced by external sources such as recording equipment, electromagnetic interference, and environmental factors. For instance, powerline interference—typically observed at 50 Hz (or 60 Hz depending on the regional electrical standard)–can be introduced if electrode cables are inadequately shielded or routed near other electronic devices. Additional sources include movement artifacts, loose or poorly attached electrodes, and faulty amplifiers or analogue-to-digital conversion hardware. Physical disturbances like pressing on the electrodes during recording can also introduce significant signal distortions [47, 80, 92].

Due to the fact that artifacts can mimic or obscure genuine neural activity, their identification, mitigation, and removal are a critical pre-processing step in EEG analysis. Failing to manage artifacts properly can lead to erroneous interpretations of brain dynamics and, in clinical applications, may result in inappropriate or misleading conclusions. Therefore, a systematic and rigorous approach to artifact handling is essential for ensuring signal

quality and the validity of subsequent analyzes.

2.1.5 Artifact processing in EEG signals

The handling of artifacts in EEG signal processing typically falls into two broad strategies: artifact rejection and artifact correction (or cancellation) [47, 80, 92]. Artifact rejection involves identifying and discarding contaminated segments of the EEG signal, while artifact correction aims to retain as much of the original signal as possible by selectively removing the artifacts, effectively "cleaning" the data for further analysis.

Several techniques have been proposed in the literature for artifact removal, with varying degrees of complexity and effectiveness. The most widely used methods include:

- **Independent Component Analysis (ICA)**: ICA is a statistical technique that decomposes the mixed EEG signal into a set of independent components. By identifying and isolating components associated with artifacts (e.g., eye blinks, muscle activity), the remaining components can be recombined to reconstruct a cleaned EEG signal. ICA is particularly effective for short epochs and time-locked analyses [79, 93].

- **Regression-Based Artifact Removal**: This method estimates the artifact by applying a regression model using reference signals (e.g., EOG for eye movements). A common regressor is derived using complex regression coefficients, and the estimated artifact is then subtracted from the EEG signal [94, 95].

- **Use-of-filters**: Band-pass filters or common average referencing (CAR) are used to attenuate signals outside the desired frequency range. For example, powerline noise at 50/60 Hz can be removed using notch filters [96, 97]. However, filtering is less effective for broadband artifacts, as it can distort the signal of interest, especially when the artifacts overlap with meaningful brain frequencies.

- **Manual rejection of artifacts**: In this approach, EEG segments are visually inspected, and sections contaminated by artifacts are manually marked and excluded from further analysis. Although this method can be accurate, it is time-consuming and introduces subjectivity [98, 47].

Among automated pipelines for EEG preprocessing, regression-based and component-based techniques are most common. Although both assume linear, instantaneous mixing of source signals, they differ in their approaches: regression relies on observed physiological reference signals to model artifacts, whereas ICA assumes statistical independence and separates sources based on the underlying data structure.

In this thesis, ICA-based artifact removal was employed for the preprocessing and cleaning of EEG signals, following the framework described by Daly et al. [79]. This approach enabled effective isolation of neural activity from typical artifacts, thereby enhancing the quality of the signal for subsequent analysis.

2.1.6 EEG Feature Extraction

The design of EEG systems typically involves a preprocessing stage followed by feature extraction and classification. After preprocessing, relevant features must be extracted from the EEG signals before they can be classified to determine the task being performed or infer the user's mental state.

A feature is a measurable attribute or characteristic of the EEG signal that provides discriminative information for classification [99]. For example, the signal power at a specific electrode can serve as a feature. These features are usually assembled into a vector, referred to as a feature vector, that serves as input to ML classifiers.

The main sources of information used for feature extraction are:

- **Temporal information**: This captures the EEG signal values over specific time intervals or windows, illustrating how the signal evolves over time.

- **Spatial Information**: This involves selecting and analyzing signals from specific electrode sites or brain regions, providing location-based insights into brain activity.
- **Spectral information**: This relates to the power distribution of EEG signals across different frequency bands. In practice, this involves transforming the signal into the frequency domain and using the power in specific bands (e.g., delta, theta, alpha, beta) as features.

In modern EEG analysis pipelines, it is common to convert EEG signals into their spectral representations as feature extraction. Several methods are employed to achieve this, including:

1. **FFT**:

   This is a computationally efficient algorithm towards estimating the Power Spectral Density (PSD) of EEG signals. It is based on an optimized implementation of the Discrete Fourier Transform (DFT), which transforms time-domain signals into frequency-domain representations. The squared magnitude of the Fourier coefficients corresponds to the signal energy, and when averaged, yields the PSD. Compared to DFTs, PSD estimation provides more stable and precise results, albeit at the cost of reduced spectral resolution [100, 96]. Consequently, the FFT alone is insufficient for characterizing the temporal dynamics of EEG signals.

   To obtain PSD estimates, the spectrogram method is commonly employed. This involves segmenting the signal into overlapping windows and applying the FFT within each window to generate a time–frequency representation. A tapering window function, such as the Hann or Hamming window, is applied to reduce spectral leakage and mitigate edge effects [96]. The FFT is then computed for each windowed segment, squared, and averaged to estimate the PSD. The process proceeds as follows:

   Let $x[y]$ be the original EEG signal, where $y = 0, 1, \ldots, -(N-1)$ and $w[y]$ be the

window function, Then the n$^{\text{th}}$ windowed segment is given by:

$$x_n[y] = x[y + nR] \cdot w[y], \quad n = 0, 1, 2, \ldots \tag{2.1}$$

where:

- R is the hop size (i.e., the number of samples shifted for each window),
- n is the window index, and
- y is the sample index within each window.

For each segment $x_n[y]$ the FFT is then applied:

$$X_n[f] = \sum_{y=0}^{N-1} x[y + nR] \cdot w[y] \cdot e^{-j2\pi fy/N} \tag{2.2}$$

where:

- f is the frequency bin index (ranging from 0 to $N - 1$),
- $X_n[f]$ is the complex spectrum of the n$^{\text{th}}$ window at frequency f.

The spectrogram is obtained by computing the squared magnitude of the resulting complex spectrum:

$$S[n, f] = |X_n[f]|^2 \tag{2.3}$$

Here, $S[n, f]$ is a 2D matrix representing signal power over time (indexed by f). This time-frequency representation enables detailed analysis of transient brain activities and is particularly suited for processing EEG signals. In this thesis spectrogram was used to analyze time series in the frequency spectrum.

2.1.7 Brain Connectivity

Brain connectivity analysis plays a critical role in the diagnosis and understanding of DOC. Investigating the interactions and communication pathways within the brain can provide

valuable insights into the underlying neural mechanisms associated with various levels of consciousness in clinical populations. Brain connectivity is typically classified into three main categories:

1. **Structural connectivity**: This refers to the anatomical connections—primarily the white matter tracts—that physically link different brain regions [101]. Also known as the neural pathways, structural connectivity is commonly studied using diffusion magnetic resonance imaging (dMRI) and tractography [102]. It delineates the fixed channels through which neural signals are transmitted across the brain.

2. **Functional connectivity**: This examines the statistical dependencies or temporal correlations between neural signals from different brain regions, indicating areas that exhibit synchronized activity [103]. Unlike structural connectivity, it does not imply any direct or causal interaction but rather captures the degree of functional integration. Functional connectivity is often assessed through methods such as Pearson correlation, coherence analysis, and mutual information [104, 105].

3. **Effective connectivity**: This describes the directed, causal influence that one neural region exerts over another. It goes beyond co-activation to infer the directional flow of information within neural networks. Unlike structural and functional connectivity, effective connectivity aims to uncover the causal architecture of brain interactions [103]. It is typically measured using approaches such as Granger causality [106], dynamic causal modeling (DCM), and entropy-based methods [107]. Among these, Granger causality is widely applied in the literature and is employed in this thesis to compute effective connectivity, forming the foundation of the connectivity analyses presented herein.

Effective connectivity was originally developed as a method to analyze spike trains recorded from multi-electrode time series data [108]. It was conceptualized as a way to

understand how activity in one neural region influences another, either at the synaptic or cortical level [109]. While functional and effective connectivity may appear conceptually similar, they fundamentally differ in both methodology and neurophysiological interpretation. Functional connectivity captures statistical associations between signals, whereas effective connectivity aims to infer directed, causal relationships based on underlying generative models of brain dynamics. In essence, effective connectivity involves a fusion of two mathematical constructs: one describing how brain regions influence each other over time (temporal dynamics), and another characterizing the network structure of these directional influences. For example, if region A is directionally connected to region B, then the past activity of A provides predictive information about the future activity of B—indicating influence, though not necessarily causation in the classical sense.

Granger causality is one of the most widely used statistical frameworks for estimating effective connectivity. Suppose we have three variables: X_t, Y_t and W_t, to determine whether Y_t Granger-causes X_{t+1}, we compare two models:

1. A baseline model predicting X_{t+1} using past values of X_t and W_t.
2. An extended model predicting X_{t+1} using past values of X_t, Y_t and W_t.

If the extended model significantly improves prediction accuracy over the baseline model, then Y_t is said to Granger-cause X_{t+1}. This inference rests on two core conditions:

- **Temporal precedence**: Y_t must precede X_{t+1} in time
- **Informational uniqueness**: Y_t must contain predictive information not already present in X_t and W_t.

This relationship can be modeled using a linear autoregressive framework:

$$X_{t+1} = \sum_{i=1}^{p} \alpha_i X_{t-i} + \sum_{j=1}^{q} \gamma_j Y_{t-j} + \sum_{k=1}^{r} \beta_k W_{t-k} + \epsilon_t \tag{2.4}$$

where: p, q, r are the lag orders for X, Y, W respectively and ϵ_t represents the error term(assumed to be white noise).

While Granger causality offers a valuable statistical framework for inferring directional relationships, real-world brain systems are often nonlinear and characterized by complex multivariate interactions. In practice, estimating effective connectivity using Granger causality typically involves fitting a Multivariate Autoregressive model (MVAR) model of order p to the multi-channel time series, followed by comparing the prediction error variances of the full and reduced models to quantify directional influence. Significant connections are then identified through appropriate statistical testing. Thus, Granger causality–based connectivity analysis fundamentally relies on the accurate estimation of an underlying MVAR model as its first step. A general MVAR model is expressed as:

$$Z_t = \sum_{i=1}^{p} A_i Z_{t-i} + \epsilon_t \tag{2.5}$$

where: $Z_t = [X_t, Y_t, W_t, ...]^T$ is a vector of time series variables, A_i are the coefficient matrices capturing interactions at lag i and ϵ_t is a multivariate white noise process. The MVAR framework enables the decomposition of the system into eigenmodes, each characterized by its own oscillation frequency, damping time, and excitation profile. These eigenmodes—and their corresponding confidence intervals—can be estimated from the model parameters, providing insight into the dynamic structure of brain connectivity [110].

For this thesis, the model order p for the MVAR model, was selected to satisfy the inequality:

$$\frac{n(p+1)}{NWS} < 0.1 \tag{2.6}$$

where:

- n: Number of samples
- p: Model order

- S: Sampling rate
- N: Number of trials
- W: Window size

This criterion ensures a sufficient ratio between the number of model parameters and the available data to avoid overfitting and maintain model reliability.

While the MVAR framework effectively models complex systems—including indirect interactions between brain regions—it operates in the time domain and assumes stationary signals. However, many brain systems exhibit frequency-specific causal interactions that are not fully captured by time-domain Granger causality. To address this, the Directed Transfer Function (DTF) extends Granger causality into the spectral domain. DTF mathematically defined as:

$$\mathrm{DTF} j \to i(f) = \frac{|Hij(f)|^2}{\sum_k |H_{ik}(f)|^2}, \quad H(f) = \left(I - \sum_{i=1}^{p} A_i e^{-j2\pi fi} \right)^{-1} \tag{2.7}$$

Here:

- $H(f)$ is the transfer function derived from MVAR coefficients
- $\mathrm{DTF}_{j \to i}(f)$ represents the normalized causal influence from channel j to channel i at frequency f.

DTF values range from 0 to 1, indicating the strength of directional influence across frequency bands. Peaks in specific frequency ranges reveal frequency-selective causal pathways between brain regions.

Having explored the foundational concepts of EEG, brain rhythms, artifacts, feature extraction and connectivity, it is essential to contextualize these elements within a modern, integrative framework. The BCI paradigm provides such a platform—offering a rapidly evolving interdisciplinary domain that combines signal processing, neuroscience, engineering, and ML. In BCI systems, artifacts pose a significant challenge, as they can degrade

performance by introducing non-neural interference into the signal. Since BCIs are predominantly based on brain activity recorded through EEG (or other neuroimaging modalities), robust preprocessing is crucial. This step ensures that the extracted features accurately represent the user's cognitive or motor intentions, a necessity particularly in real-time applications, which BCIs encourage. Effective feature extraction serves as the bridge between raw brain signals and the accurate classification of user intent. By transforming complex EEG data into meaningful representations, feature extraction enables ML models to interpret neural patterns with higher precision and reliability.

Traditionally, BCIs have focused on localized features, such as power in specific electrode channels. However, emerging approaches based on brain connectivity provide a more holistic view of brain function. Measures of functional and effective connectivity capture the interaction between spatially distributed brain regions, offering deeper insight into the dynamic networks of the brain. These methods are particularly valuable for decoding complex cognitive states, as they reflect coordinated activity across neural systems.

Incorporating connectivity-based features into BCI frameworks not only enhances system responsiveness but also improves interpretability—a key factor for clinical, assistive, and neurorehabilitation applications. By leveraging the brain's intrinsic communication pathways, such approaches can elevate BCI performance and broaden its application scope.

2.2 BCIs in Consciousness/Awareness

BCIs which encompasses both invasive and non-invasive approaches- most notably EEG- are gaining increasing relevance in contemporary society due to its diverse applications, clinical potential and associated benefits. Active BCIs have been used to control external devices, ranging from entertainment scenarios such as gaming [111, 112, 113] and virtual racing [114, 115, 116], to high-demand tasks including military applications [117, 118]. In the healthcare domain, they provide essential tools for enabling communication in patients

with severe disabilities [64, 119, 120, 121] and supporting motor restoration [122, 123, 124]. In parallel, passive BCIs are increasingly applied to monitor user states, cognitive workload, and patient conditions in clinical settings [125, 126].

The emerging framework of BCIs is fundamentally based on signal processing techniques. Each step previously discussed -EEG acquisition, artifact handling, feature extraction, and connectivity analysis - constitutes an integral component of this comprehensive pipeline, collectively ensuring the reliability, efficiency, and effectiveness of BCI systems. While artifacts must be meticulously minimized to preserve signal integrity, features need to be extracted in a way that accurately reflects the user's cognitive or motor intent. Additionally, connectivity analysis offers a deeper understanding of the brain's internal communication pathways, enabling the detection of distributed neural patterns that enrich the interpretation of user states and experiences.

As outlined in the preceding chapter, a BCI fundamentally serves as a communication and control bridge between the brain and external systems that does not involve muscles. By employing ML algorithms to decode brain signals into actionable outputs, facilitate interaction with external devices or interfaces—without requiring overt muscular activity or physical movement. While not primarily an imaging paradigm, BCI systems function as real-time neurophysiological interfaces structured around three fundamental components: signal acquisition, user control (or volitional intent), and the feedback mechanism [127, 128]. These core elements collectively distinguish BCIs from other neurotechnological systems. This intersection of components positions BCI as a particular promising approach for the detection, diagnosis, and even prediction of awareness in patients with DOC.

Beyond the choice of neuroimaging modality, the efficacy and clinical value of consciousness-assessment systems are strongly shaped by the design of the experimental protocol. This encompasses the method of signal acquisition, the mode of patient engagement or stimulation, and the interpretability and clinical relevance of the resulting data [129, 60]. The specific neuroimaging technology used in this thesis to acquire neural

data is EEG.

A BCI on the other hand is defined as a system that records and decodes neural signals reflecting user intent, translates these signals into commands for an external device or application, and provides real-time sensory feedback to the user—thereby establishing a closed control loop [127]. The feedback component is essential for facilitating user learning and adaptive control [128].

Although many studies in DOC research employ BCI-based paradigms, the majority do not incorporate feedback mechanisms [32, 34, 36, 130, 131, 132, 38]. Notable exceptions include the closed-loop implementations reported by Coyle et al. [133] and Guger et al. [134]. However, these remain isolated cases, and the diagnostic benefit or added clinical value of the closed-loop component was not clearly established. The omission of closed-loop systems is primarily due to the limited and inconsistent ability of DOC patients to engage in sustained learning or volitional control. As a result, closed-loop BCI protocols are typically considered more appropriate for post-diagnostic rehabilitation or communication following demonstrable recovery of awareness, rather than for the initial detection of consciousness.

2.2.1 Signal Acquisition

Signal acquisition can be conducted via either invasive or non-invasive methods [127, 128, 135]. Invasive techniques involve the direct implantation of electrodes on or within the brain to capture high-fidelity neural signals [136, 137, 138]. In contrast, non-invasive methods—such as EEG or fNIRS—place sensors externally on the scalp or skin surface [139, 140]. While invasive methods often offer higher spatial and temporal resolution, they are associated with surgical risks and ethical considerations [141]. Non-invasive approaches, although less precise, are safer, more accessible, and widely used in clinical and bedside applications. Both modalities offer distinct advantages and limitations, and their suitability depends on the specific use case, patient condition, and ethical constraints. However, in

DOC literature, the non-invasive approach seems to be mainly preferred.

2.2.2 Mental task

By mental task or user control, I refer to the degree of the participant's involvement in generating brain signals. This can be understood as the specific thought patterns or cognitive activities that produce the neural activity used for analysis. Also known as the BCI paradigm, user control can be either active—where the participant intentionally engages in specific cognitive tasks—or passive, where brain signals are extracted and analyzed without requiring deliberate participation. An additional intermediate category has also been proposed, particularly in the context of awareness diagnosis [53, 59].

Active paradigms are designed to deliberately prompt participants to engage in goal-directed mental tasks. For example, patients may be instructed to perform specific cognitive operations (e.g., MI or mental arithmetic), after which the system evaluates whether the expected neural correlates are present. These paradigms aim to detect volitional brain activity and are frequently applied in command-following assessments, where patients are asked to comprehend instructions and produce discernible neural responses, or to discriminate between target and non-target stimuli. Such approaches are particularly valuable when assessing individuals with severely limited motor or cognitive capabilities.

Commonly employed paradigms include MI [142, 120], Steady-State Visually Evoked Potentials (SSVEP) [143], P300 [144], auditory stimulation [59], slow cortical potentials [64], and hybrid modalities [145].

2.2.3 Feedback and Refinement

The feedback mechanism is a key distinguishing feature in the classification of BCI systems. It refers to the way a BCI communicates information back to the user, thereby establishing a closed-loop protocol within the system. Without feedback, learning and effective control

of a BCI are often unsuccessful. However, as mentioned earlier in Sections 1.9 and 2.2, while learning is not the objective of this work, DOC patients typically do not exhibit the stable, sustained volitional control or learning capacity [53, 146] required for effective feedback-based interaction, even over intervals longer than 1 s [52, 147, 133]. Therefore, feedback-driven protocols are generally viewed as appropriate only after awareness has been reliably established—serving rehabilitative or communicative purposes—rather than as tools for the initial detection of consciousness.

Nevertheless, in a true BCI, feedback is delivered to the user in real time, typically in the form of control over external devices or as an adaptive mechanism to promote mutual learning, motivation, and sustained engagement. A system that incorporates this mechanism is termed a closed-loop BCI, whereas one that lacks feedback is referred to as an open-loop BCI.

Feedback can be implemented through three primary modalities:

1. **Visual feedback**: Achieved via screen-based cues such as bars, cursor movements, or dynamic visual effects. For instance, a bar may progressively grow taller or shorter depending on how well the participant achieves the mental task target.

2. **Auditory feedback**: Delivered through sound-based cues, such as tonal variations, voice prompts, or verbal instructions that inform the user whether the desired target has been achieved.

3. **Tactile feedback**: Implemented using wearable devices or equipment that provide vibratory or haptic signals to convey system behavior, activity, or state.

In this thesis, Functional Electrical Stimulation (FES) was applied (overt hand extension movement) as a motivational paradigm and to keep the patients engaged. This is based on the assumption that externally induced muscle contractions might engage afferent sensory–motor pathways comparable to those recruited during voluntary movement

or motor imagery. If this were the case, the resulting feedback could theoretically evoke Sensorimotor Rhythms (SMR) in the corresponding motor cortical areas, thereby serving as an additional marker of motor-system integrity. However, as shown in Chapter 4, the empirical results did not support this assumption, indicating that FES-evoked responses do not reliably mirror the neural patterns associated with voluntary motor imagery.

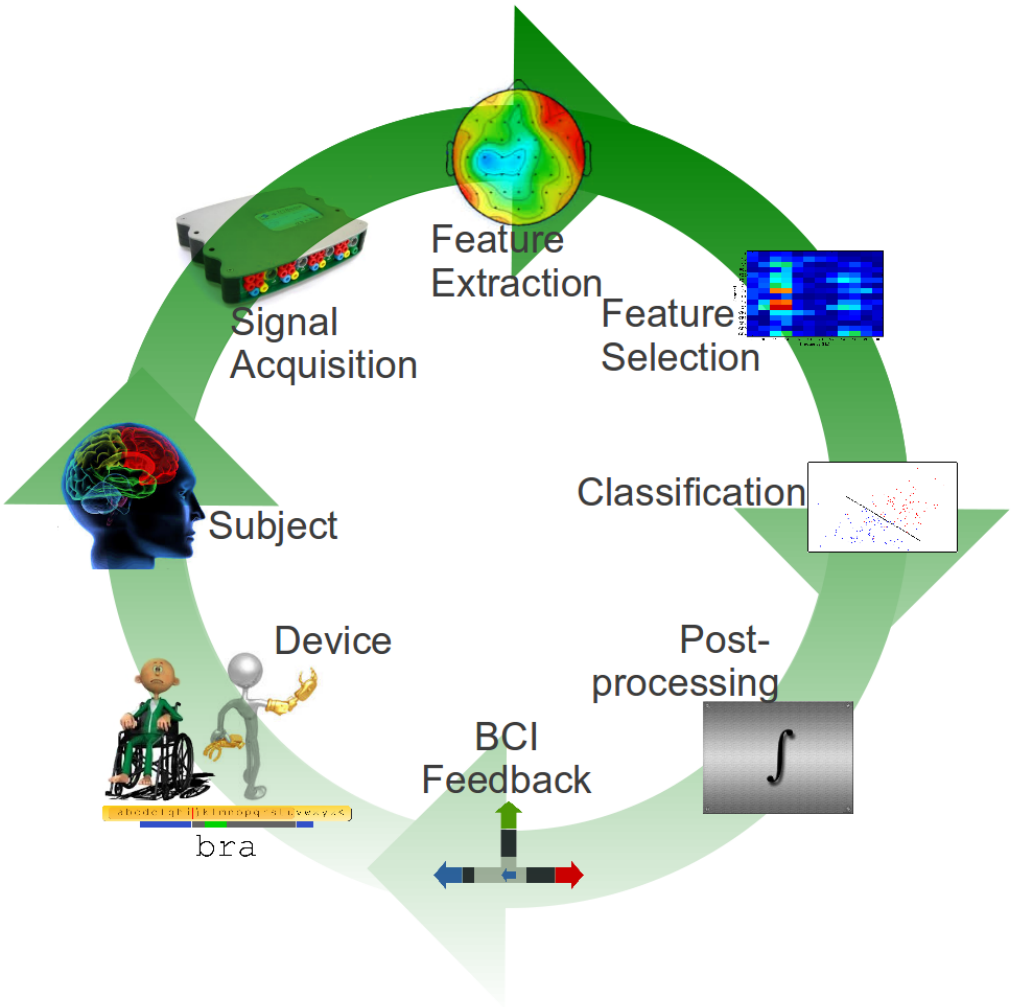

Figure 2.4: Closed-loop BCI Architecture

Aside the feedback protocol, the refinement and effective utilization of feedback play a crucial role in the development and application of BCI systems. Fundamentally, the

interpretability and transparency of the BCI architecture are paramount to its successful deployment and broader adoption. Figure 2.4 illustrates the standard components of a typical BCI architecture: - signal acquisition, pre-processing, feature extraction, feature selection, classification, post-processing, feedback, and end-usage/application. Of these components, only feature selection and classification remain to be discussed, as the others have been covered in preceding sections.

2.2.4 Feature Selection

As illustrated in Figure 2.4, classification is a fundamental component of the BCI system architecture. However, not all extracted features are suitable for direct input into classifiers [55, 41]. This necessitates dimensionality reduction, a critical step in most ML pipelines, aimed at enhancing class separability and improving computational efficiency. The challenge of handling high-dimensional data is well-recognized in the machine learning community [148]. Feature selection addresses this challenge by reducing dimensionality through the elimination of irrelevant or redundant features—those that contribute little to class discrimination. This not only reduces computational complexity and enhances classification accuracy, but also helps prevent overfitting [149]. Importantly, feature selection achieves this without compromising the integrity of the original data's principal attributes.

In essence, feature selection is the process of identifying an optimal subset of features from the original dataset that maximizes inter-class separability. While both feature extraction and feature selection are methods of dimensionality reduction, feature selection maintains the original data structure, whereas feature extraction transforms the data into a new feature space. Feature selection is particularly valued for its ability to enhance pattern recognition, improve classifier performance, and reduce model complexity. According to literature, feature selection methods can be broadly categorized into approaches based on:

- Statistical analysis [150, 151],
- Manifold learning [152, 153],
- Information theory [154, 155], and
- Rough set theory [156, 157].

In this thesis, Fisher Score (FS) (statistical) analysis was used to assess the discriminability of features across mental states. The FS was computed using the standard expression:

$$FS = \frac{|\mu_1 - \mu_2|}{\sqrt{s_1^2 + s_2^2}}, \tag{2.8}$$

where μ_1, μ_2 are the means and s_1, s_2 are the standard deviations of the two mental classes contrasted (in this thesis either MI vs Rest or FES vs Rest).

2.2.5 Classification

Artificial Intelligence (AI) and ML, are rapidly becoming ubiquitous tools, transcending disciplinary boundaries and reshaping the paradigms of data-driven research and application. As their adoption broadens, ML models are increasingly viewed not just as applied mathematical tools, but as an interplay of both science and art—requiring domain insight, design precision, and iterative refinement.

Broadly, ML frameworks are classified into two main categories: supervised learning and unsupervised learning. Supervised learning involves learning from labeled datasets, where each input is paired with a corresponding output label, enabling the model to infer relationships for future prediction tasks [158]. In contrast, unsupervised learning seeks to discover hidden structures or patterns in unlabeled data, operating without predefined outputs [159]. These paradigms differ fundamentally in their objectives and methodologies.

Unsupervised learning techniques commonly include clustering (e.g., k-means, hierarchical clustering) and dimensionality reduction (e.g., PCA, t-SNE), which reveal latent

data groupings or compress data representations. Supervised learning, on the other hand, typically focuses on classification and regression tasks. This thesis was conducted within the supervised learning framework, specifically employing classification to predict discrete labels based on EEG-derived features.

Classification is the process of predicting whether a data instance belongs to a particular category or class [160]. For instance, in BCI applications, classification may involve predicting whether a subject is performing left-hand or right-hand motor imagery based on EEG features, or if the patient is aware or not based on collated EEG features. Several standard methods are employed to assess the performance and generalizability of classification models:

1. **Hold-Out Validation**: The dataset is randomly split into two parts—typically, two-thirds are used for training the model, and the remaining one-third for testing its performance.

2. **k-Fold Cross-Validation**: The dataset is divided into k mutually exclusive, equally sized subsets (folds). The model is trained on k–1 folds and tested on the remaining fold. This process is repeated k times, each time with a different fold used for testing. The overall classification accuracy is computed as the average accuracy across all k trials.

3. **Leave-One-Out Cross-Validation (LOOCV)**: A special case of k-fold cross-validation where k equals the number of samples. Each individual data point is used once as the test set, while the rest serve as the training set. Though computationally expensive, LOOCV is particularly useful for small datasets, as it maximizes the use of available data for training. In this thesis, LOOCV was initially considered for classification analysis. However, due to the use of sliding windows in the computation of PSDs, adjacent samples in the dataset are temporally correlated and therefore not statistically independent. Over a given trial, multiple windowed segments may

correspond to the same class label, and treating these as independent samples would risk biasing the evaluation. To address this, classification was performed using leave-one-trial-out (LOTO) cross-validation, even when classification was conducted at the single-sample (PSD window) level rather than only at the whole-trial level. In effect, this adjustment ensures that all correlated samples within a trial are excluded together during testing, thereby preventing data leakage across training and test sets. Under these conditions, LOTO cross-validation becomes functionally equivalent to LOOCV for the purposes of this analysis.

2.3 Neuro-imaging and BCI-inspired diagnostic paradigms for DOC

As earlier stated, the diagnosis of DOC is primarily conducted using standardized clinical assessment tools such as the CRS-R and the GCS [3, 56, 57], which are shown in Figure 2.5. These tools consist of ordinal scales that assign scores based on the patient's ability to perform specific tasks in response to structured and standardized stimuli. The scoring is determined by the degree of behavioral responsiveness, with higher scores indicating greater levels of consciousness.

The CRS-R, for example, comprises sub-scales assessing auditory, visual, motor, oro-motor, communication, and arousal functions. Within each sub-scale, items are arranged hierarchically, and scores represent progressively more complex or intentional behaviours. However, because these are ordinal measures, the intervals between scores are not necessarily equal—meaning that a score of 2 does not quantitatively double the level of consciousness reflected by a score of 1.

Despite their widespread acceptability and use, behavioral scales have several methodological limitations:

A

JFK COMA RECOVERY SCALE ©2004
Record Form

This form should only be used in association with the "CRS-R ADMINISTRATION AND SCORING GUIDELINES" which provide instructions for standardized administration of the scale.

Patient:	**Diagnosis:**
Date of onset:	**Date of Admission:**

Date									
Assessment	**1**		**2**		**3**		**4**		**5**	
AUDITORY FUNCTION SCALE	**#**	**TCC**	**#**	**TCC**	**#**	**TCC**	**#**	**TCC**	**#**	**TCC**
4 – Consistent Movement to Command▪										
3 – Reproducible Movement to Command▪										
2 – Localization to Sound										
1 – Auditory Startle										
0 – None										
VISUAL FUNCTION SCALE	**#**	**TCC**	**#**	**TCC**	**#**	**TCC**	**#**	**TCC**	**#**	**TCC**
5 – Object Recognition▪										
4 – Object localization: Reaching*										
3 – Visual Pursuit*										
2 – Fixation*										
1 – Visual Startle										
0 – None										
MOTOR FUNCTION SCALE	**#**	**TCC**	**#**	**TCC**	**#**	**TCC**	**#**	**TCC**	**#**	**TCC**
6 – Functional Object Use†										
5 – Automatic Motor Response*										
4 – Object Manipulation*										
3 – Localisation to Noxious Stimulation*										
2 – Flexion Withdrawal										
1 – Abnormal Posturing										
0 – None										
OROMOTOR/VERBAL FUNCTION SCALE	**#**	**TCC**	**#**	**TCC**	**#**	**TCC**	**#**	**TCC**	**#**	**TCC**
3 – Intelligible Verbalization▪										
2 – Vocalization/Oral Movement										
1 – Oral Reflexive Movement										
0 – None										
COMMUNICATION SCALE	**#**	**TCC**	**#**	**TCC**	**#**	**TCC**	**#**	**TCC**	**#**	**TCC**
2 – Functional: Accurate†										
1 – Non-functional: Intentional▪										
0 – None										
AROUSAL SCALE	**#**	**TCC**	**#**	**TCC**	**#**	**TCC**	**#**	**TCC**	**#**	**TCC**
3 – Attention										
2 – Eye Opening w/o Stimulation										
1 – Eye Opening with Stimulation										
0 – Unarousable										
TOTAL SCORE										

* Denotes Minimally Conscious State Minus (MCS-)
▪ Denotes Minimally Conscious State Plus (MCS+)
† Denotes emergence from Minimally Conscious State (eMCS)
TCC Test Completion Code

B

Figure 2.5: Behavioral scale (A) CRS-r [3] and (B) GCS [4].

- Non-linearity: score progression does not reflect linear or proportional changes in neural function.

- Motor and sensory dependence: accurate scoring requires preserved motor output or sensory input, which may be absent in patients with severe impairments, who nevertheless maintains awareness [32].

- Subjectivity and variability: inter-rater variability and inconsistent interpretation of ambiguous responses can affect diagnostic accuracy [161, 162, 163].

- Behavioral inference: these tools infer consciousness from overt behavior, potentially overlooking covert awareness in patients unable to respond overtly [34, 164].

DOC, most commonly arising from ABI such as traumatic or cerebrovascular accidents, but also from neurological diseases or pharmacological side-effects, is primarily classified into four categories: coma [22], VS or UWS [23], MCS+/MCS- [24], and eMCS [24, 25]. Coma is clinically identified by the absence of both the components of awareness and consciousness. UWS is characterized by preserved wakefulness but absent awareness. MCS- reflects minimal but reproducible signs of awareness, while MCS+ indicates more complex but still inconsistent evidence of awareness. Last, eMCS is identified by the recovery of functional communication and/or object use.

However, the clinical golden standard diagnostic tools used for these diagnosis have significant limitations, including being prone to both type I and, especially, type II errors [58, 59, 60]. In an effort to address the limitations of these main assessment tools, alternative behavioral scales have been proposed such as the Motor Behavior Tool MBT [61]. However, the reliance on assessment through behavior contributes largely to the 37-43% of clinical misdiagnosis as reported in literature [62, 56, 63]. The emphasis of clinical evaluation on the behavioral outputs in DOC patients has opened the door to the possibility of many false-negative diagnoses, namely, patients who may lack behavioral signs

of awareness, but can otherwise (usually, through neuroimaging-based assessment) exhibit clear evidences of command-following [64, 32, 34, 65, 66]. This condition has been termed CMD [32]. In recent clinical spheres, the CMD has also been proposed as a DOC classification [25].

To improve DOC diagnostic accuracy, researchers have explored various techniques, with brain imaging emerging as the most promising approach. Although fMRI initially dominated the field due to its high spatial resolution and early landmark studies [32, 33, 34], the more practical, portable, and cost-effective EEG has gained increasing attention [35, 36, 37, 38]. Hybrid approaches combining multiple imaging modalities have also been proposed [39, 40]. Overall, while fMRI provided the foundation for early breakthroughs, EEG continues to gain widespread adoption due to its portability, affordability, reliability and superior temporal resolution; these properties are critical for producing impactful neuroimaging-based DOC diagnostic tools that are inexpensive, practical and logistically compliant with daily use in hospitals.

Beyond the imaging modality itself, another critical distinction lies in the protocol design [129, 60]. The category of paradigms that can be termed as "active" attempt to assess awareness by prompting patients to perform various cognitive tasks and subsequently determining whether the anticipated brain activity is observed. The most demanding among these, ask patients to follow commands, leading to BCI-inspired protocols [60]. In contrast, passive paradigms assess spontaneous or, more often, stimulus-evoked brain responses without active participation or high-order cognitive requirements, such as evoked potentials, EEG reactivity and sleep patterns [165, 129]. Intermediate paradigms have also been proposed, involving some minimal cognitive processing [53, 59], but not as cognitively challenging as with command following and BCI-inspired ones, where subjects must sense, comprehend and respond to verbal commands and/or discriminate target from non-target stimuli. While passive paradigms minimize the patient's cognitive workload and hence may be better suited for detecting MCS, active paradigms are thought to provide a more

definitive indicator of consciousness and awareness [60]. The use of ML-based metrics is gaining prominence, often in conjunction with BCI-inspired protocols such as SMR and P300 [129, 60].

A variety of open-loop BCI protocols have been proposed, mainly mirroring the popular BCI paradigms originally designed for communication and control [127]; namely, command-following based on MI and SMRs [166, 167, 168] and event-related potentials such as P300 [169, 60], but not limited to these [59, 170, 38]. It is important to clarify that these approaches are not fully-fledged BCI systems as formally defined in the literature. A true BCI operates as a closed-loop system, continuously acquiring neural signals, decoding them in real time, generating outputs and delivering feedback to the brain. In contrast, the BCI-based diagnostic paradigms for DOC lack this closed-loop feedback mechanism, borrowing only the open-loop BCI procedures. In BCI for control, these are meant to collect the EEG data labeled with the underlying mental task that are needed to produce (by and large, with supervised ML requiring such labeled data) the BCI's mental task decoder model; the latter is necessary for subsequent closed-loop BCI operation [114, 171, 172]. In the diagnosis of DOC, the ML decoder and other relevant statistical models are built so as to recognize the neural correlates of the requested cognitive tasks; evidence of these indicates that the patient has been able to follow commands, and must therefore present with some level of spared awareness. For this reason, I find that the term BCI for these systems is a misnomer, and prefer to refer to them here as "BCI-inspired" paradigms. However, evidence of consciousness derived with an open-loop BCI paradigm does imply that re-establishment of communication with the patient may be possible through the corresponding closed-loop protocol, with relatively minor modifications and by means of the same decoder trained for diagnosis.

BCI-inspired diagnostic paradigms have gained traction largely due to their potential to reveal CMD in patients who are unable to exhibit reliable behavioral responses. The pioneering works of Owen et al. [32] and Monti et al. [34] first demonstrated that high-

resolution fMRI could rectify the misdiagnosis of patients classified as UWS by standard clinical scores. Building on these findings, Cruse at al. [36] employed a bedside EEG-based protocol in which patients diagnosed with VS were instructed to imagine moving their right hand and toes. Remarkably, 9 out of 16 patients exhibited task-specific brain activity consistent with command following, leading to their classification as CMD. These findings paved the way for the development of several other SMR-based BCI-inspired diagnostic paradigms [132, 130, 133, 131, 173]. Notably, Coyle et al. [133] recruited four MCS patients and proposed an auditory feedback paradigm that implemented a closed-loop BCI, although no evidence of learning was provided. However, extending such protocols to other diagnostic categories—particularly UWS, which represents the critical boundary between awareness and unawareness—remains a substantial challenge. Similarly, Eliseyev et al. [173] examined a closed-loop BCI with auditory feedback in ICU patients as well as healthy controls. Their findings indicated that only conscious participants were able to engage with and control the BCI task, underscoring the difficulty of translating closed-loop paradigms to unresponsive patient populations. For instance, Holler et al. [130] reported difficulties in diagnosing DOC based solely on classification accuracy, primarily due to the absence of reliable ground truth in this patient population. In contrast, Guger et al. [131] argued that using MI protocols, classification accuracy could serve as a practical, quantifiable metric for assessing covert awareness; however, the results of this study were derived by an extremely limited sample. Chatelle et al. [132], emphasizing the commercial translation of BCI-based diagnostic tools, reported a lack of correlation between BCI-derived responses and clinical behavioral assessments, highlighting the complexity of reliably interpreting these neural signals in clinical settings.

Bodien et al. [174] investigated command-following in DOC patients using both EEG and fMRI. Participants engaged in MI tasks, and the study revealed that 25% of the patients who lacked behavioral evidence of command following exhibited covert signs of consciousness, thus, CMD. Similarly, Edlow et al. [38] employed an auditory cue-based

command-following protocol for the ICU applied to patients with severe Traumatic Brain Injury (TBI). Their findings suggested that EEG can reliably detect CMD in acute settings, potentially enhancing diagnostic accuracy and preventing premature withdrawal of care. In Claassen et al. [52], a prognostic method for predicting outcomes in DOC was proposed. Goldfine et al. [175] focused on the diagnostic utility of EEG in differentiating VS and MCS. In a recent, comprehensive review, Galiotta et al. [60] highlighted the lack of standardized analytical methods and protocols in BCI-inspired research targeting DOC populations, recommending that future developments prioritize diagnostic applications. This recommendation aligns with the perspective of Annen et al. [176], who additionally noted that the absence of oculomotor control in most DOC patients limits the effectiveness of visually guided BCI systems. They further argued that task-based BCI paradigms, despite their challenges, provide more robust indicators of consciousness. Finally, they advocated for the use of empirically adjusted chance levels instead of theoretical ones, as the latter fails to account for reallocation effects commonly observed in DOC patients.

2.4 Deep-learning Models in EEG Classification

Traditionally, EEG-based BCIs have relied on handcrafted features, where researchers explicitly design mathematical rules to extract relevant information from neural signals. Classification, feature extraction, and spatial filtering techniques—which improve the signal-to-noise ratio in EEG—have been central to this process. Shallow approaches such as Linear Discriminant Analysis (LDA), Common Spatial Filter (CSP), Filter Bank Common Spatial Pattern (FBCSB), have built upon these principles and remain widely used due to their effectiveness, computational efficiency, and relative simplicity. More recent shallow methods, particularly those based on Riemannian geometry, bypass assumptions of covariance equality by directly modelling covariance structures of EEG signals before applying linear projections like LDA. These methods have shown strong robustness and improved

generalization in classification tasks.

In parallel, deep learning models—including EEGNet [7], DeepConvNet [8], and EEG-Conformer [9]—have emerged as compelling alternatives. Unlike shallow models, they automatically learn hierarchical feature representations from raw or minimally processed data, and in many cases outperform traditional approaches in terms of classification accuracy. However, their reliance on large-scale datasets remains a critical limitation in BCI, where data is inherently scarce, heterogeneous, and subject-specific. Consequently, while deep models often report strong accuracy, their reproducibility and consistency—particularly in relation to neurophysiological interpretability—remain subjects of ongoing debate. This tension between shallow and deep models continues to shape the trajectory of BCI research, particularly in clinically sensitive domains such as DOC diagnosis.

Most shallow learning models in BCI, such as CSP, and Riemannian geometry methods are grounded in covariance-based approaches. They compute covariance matrices for the assigned classes, identify directions of maximum separability—often by solving an eigenvalue problem—and apply spatial filtering to enhance discriminability. The extracted log-variance features are then projected onto a linear subspace, where classification is achieved by maximizing inter-class differences along the most separable directions. These methods are often combined with linear classifiers like LDA that exploit class mean differences under a shared covariance assumption.

By contrast, EEGNet introduces a lightweight deep learning architecture specifically tailored for EEG analysis. It incorporates depthwise and separable convolutions for spatial and temporal filtering, employs dropout and batch normalization for regularization, and concludes with a dense classification layer. Although compact, EEGNet mimics aspects of spatial filtering used in shallow paradigms while maintaining a minimal parameter count, making it computationally efficient. DeepConvNet, on the other hand, adopts a hierarchical feature-extraction strategy directly from raw EEG signals. It integrates temporal convolutions for frequency-specific features, spatial convolutions to capture inter-channel

dependencies, non-linear pooling operations, and a dense classification layer. This architecture enables the modeling of complex non-linear patterns, but it is significantly heavier than EEGNet and considerably more computationally demanding.

The introduction of attention mechanisms in deep learning [43] has further advanced model architectures by enabling the capture of long-range dependencies through vector embeddings in high-dimensional space. Building on this principle, EEGConformer integrates convolutional operations with self-attention. Specifically, it applies initial convolutional layers to extract local temporal features, introduces self-attention layers to capture global dependencies, and then fuses both feature types before passing them to a dense classification layer. This hybrid design leverages the strengths of both convolution and attention, allowing for more effective modeling of spatial-temporal dynamics in EEG data. While the attention mechanism substantially improves representational capacity, like other deep learning models, EEGConformer remains highly data-dependent—an ongoing limitation in BCI research, where sufficiently large datasets are rarely available.

2.5 Conclusion

The current gold standard for clinical diagnosis of DOC—the CRS-R—like other behavioral assessment tools, fails to reliably identify patients with CMD. This limitation leads directly to both Type I and Type II diagnostic errors, contributing to the 34–40% misdiagnosis rate widely reported in the literature [177, 56, 63]. These shortcomings highlight the need for diagnostic protocols that do not rely on overt behaviour but instead target neural markers of cognitive processing—thereby enabling more accurate detection of awareness.

This realization motivated the introduction of neuroimaging methods into the DOC diagnostic landscape, including fMRI and EEG. However, the substantial financial, logistical, and technical constraints associated with bedside fMRI, coupled with the superior temporal advantages of EEG for neural/cognitive assessment [178, 179], have made EEG

the most widely adopted modality. Yet, signal acquisition alone is insufficient. The field required computational tools capable of transforming raw neural data into interpretable outputs, giving rise to the use of BCIs within the DOC domain—systems originally designed for communication, rehabilitation, and motor learning [127, 64, 144].

Beyond signal acquisition, a functional BCI requires an associated imaging paradigm—such as MI, SSVEP, or P300—and a feedback mechanism that relays system outputs back to the user in real time. These paradigms, broadly categorized as active, passive, or intermediate, are central to system design because they determine the cognitive demands placed on the user and the interpretability of the resulting neural responses. In the context of DOC, active paradigms have consistently been shown to provide stronger evidence of preserved awareness.

However, because most DOC patients cannot sustain the volitional control required for learning-based or feedback-driven communication, closed-loop systems cannot be at present implemented at the bedside. As a result, the majority of neuroimaging protocols used for diagnosis in DOC research do not constitute true BCIs; rather, they are more accurately described as BCI-inspired systems, employing BCI-like decoding methods in open-loop without the defining feedback component.

Chapter 3

General Materials and Methods

In this chapter, I present the general materials and methods employed throughout this thesis. It is important to note that specialized methods tailored to the validation of analyzed metrics or the comparison of analytical pipelines are described in detail within their respective chapters. Accordingly, this chapter is limited to outlining the participant data, the experimental tools used for data collection, the experimental protocol of the BCI-inspired paradigm, and the preprocessing steps applied during experimental analysis.

3.1 Introduction

This chapter outlines the methodological framework that underpins this thesis. The first and second objectives are to validate, as well as compare state-of-the-art metrics and analytical pipelines on a novel dataset, employing a BCI-inspired MI–EEG paradigm. To this end, data collection was conducted under formal ethical approval, followed by systematic preprocessing, before subsequent integration into method-specific pipelines.

Data acquisition remains a central challenge in BCI research. High-quality recordings require participants to engage with EEG systems for extended durations, which can be burdensome and contribute to limited recruitment and retention. Additionally, the inherent

variability of neural activity—both within and across individuals—further complicates the establishment of robust datasets. These constraints frequently lead researchers to rely on publicly available repositories which, while valuable, are often suboptimal for specific protocols or experimental designs.

In this work, these limitations are addressed through the collection of a novel dataset obtained at the ICU of Lausanne University Hospital, using a protocol developed specifically for this study. This dataset forms the basis for all subsequent analyses.

It is important to note that BCI experiments are inherently collaborative, typically involving engineers, clinicians, neuroscientist, neuro-psychologists, etc. Accordingly, the dataset used in this thesis had already been acquired by clinicians at Lausanne University Hospital using the study design described here and subsequently shared with the University of Essex BCI laboratory through my supervisor, who designed, implemented the protocol, experimental setup, took part in the data collection and co-owns the data. The data are therefore co-owned by both institutions. All raw data were anonymised at the point of collection and uploaded to a secure, access-controlled server at the University of Essex.

The remainder of this chapter introduces the general methodological framework, beginning with a description of the data collection procedures and associated ethical approvals.

3.2 Participants and data

Twenty-eight (28) patients (8 female, age 53.5±16.6, range 20-75) with varying levels of awareness within the spectrum of DOC, admitted to the Acute Neurorehabilitation Unit of the University Hospital of Lausanne (CHUV), Switzerland were analyzed in this study. The experimental protocol complied with the Helsinki Declaration and was approved by the ethical committee of the Canton of Vaud, Switzerland (No. 142/09). Written consent to participate in the study was obtained from relatives of the participants. At the end of the first stage of pre-processing, data from one participant were discarded because over

20% of the subject's trials (intervals where the subject is commanded to engage into some motor/cognitive task) could not be cleaned by means of a state-of-the-art EEG artifact removal method [79]. I report in this thesis 179 "runs" (continuous blocks of multiple command-following trials) in 62 recording sessions acquired from these 28 DOC patients. Multiple sessions within subjects correspond to recordings on different days, dictated by the patients' acute clinical status, which led in several cases to higher CRS-R scores in subsequent session (see Table 3.1). This consideration further motivates the use of session-wise analyses throughout this thesis, as a single clinical assessment of a patient's level of awareness can be reasonably assumed to correspond only to the EEG data acquired within the same recording session. All patients were assessed by clinical neurologists and underwent repeated behavioral evaluations using the CRS-R [3] during their hospital stay. Based on these assessments, the neurology team assigned each recording session to one of seven categories derived from the corresponding CRS-R scores: Coma, UWS, MCS-, MCS+, eMCS, LIS, and Healthy. Although LIS is not considered a DOC state, it is frequently used in the literature as a conscious comparison group, particularly in studies aimed at characterizing or detecting CMD and related DOC conditions.

Table 3.1: Demographics and clinical assessment of patients

S/N	Subject, Session	Sex/ Age	Etiology	Runs	CRS-R Score	CRS-R Diagnosis	CRS-R Subscale Scores					
							Auditory (max 4)	Visual (max 5)	Motor (max 6)	Oro-motor (max 3)	Communication (max 2)	Arousal (max 3)
1	S1,1	M/62	TBI	4	13	MCS-	2	3	3	3	0	2
2	S2,1	F/75	CVI	2	5	UWS	0	1	2	1	0	1
3	S3,1	M/59	CVI	3	3	UWS	0	1	0	1	0	1

Continued on next page

Table 3.1 (continued)

4	S3,2			3	11	MCS-	2	3	2	2	0	2
5	S3,3			3	15	MCS+	3	3	3	3	1	2
6	S4,1	F/32	CVI	3	9	MCS+	3	2	2	1	0	1
7	S4,2			3	11	MCS+	3	3	3	1	0	2
8	S4,3			2	11	MCS+	3	3	2	1	0	2
9	S4,4			2	7	MCS-	0	2	2	1	0	2
10	S4,5			3	7	MCS-	0	2	2	1	0	2
11	S4,6			3	8	MCS-	0	3	2	1	0	2
12	S4,7			1	8	MCS-	0	3	2	1	0	2
13	S4,8			3	11	MCS+	3	3	2	1	0	2
14	S5,1	M/54	Anoxic	3	17	MCS+	3	3	5	3	2	1
15	S6,1	M/75	CVI	3	20	eMCS	4	4	6	2	2	1
16	S7,1	M/33	TBI	2	3	UWS	0	0	0	1	0	2
17	S8,1	F/48	CVI	3	18	eMCS	3	4	6	3	0	2
18	S8,2			3	18	eMCS	3	4	6	3	0	2
19	S9,1	M/66	Healthy	3	23	Healthy	4	5	6	3	2	3
20	S9,2			3	23	Healthy	4	5	6	3	2	3
21	S10,1	M/71	Anoxic	3	6	UWS	0	1	2	2	0	1
22	S10,2			3	10	MCS-	2	2	2	3	0	1
23	S10,3			3	11	MCS+	2	3	2	3	0	1
24	S10,4			1	17	MCS+	3	3	5	3	1	2
25	S10,5			2	17	MCS+	3	3	5	3	1	2
26	S11,1	F/24	TBI	4	7	UWS	2	1	2	1	0	1
27	S11,2			4	11	MCS-	2	3	3	1	0	2
28	S12,1	M/56	Anoxic	2	15	MCS+	3	2	5	2	1	2
29	S13,1	F/53	CVI	3	10	MCS+	3	2	2	1	0	2
30	S13,2			3	13	MCS+	3	2	5	1	0	2
31	S13,3			3	13	MCS+	3	2	5	1	0	2
32	S14,1	M/58	TBI	3	14	MCS+	3	3	3	2	1	2
33	S14,2			3	17	MCS+	3	5	3	3	2	1
34	S14,3			3	17	MCS+	3	5	3	3	2	1
35	S14,4			3	22	eMCS	4	5	6	3	2	2
36	S15,1	M/51	Anoxic	3	20	eMCS	4	4	6	3	1	2

Continued on next page

Table 3.1 (continued)

37	S15,2			3	20	eMCS	4	4	6	3	1	2
38	S16,1	M/60	CVI	3	23	eMCS	4	5	6	3	2	3
39	S17,1	M/24	TBI	3	21	MCS+	4	5	5	3	2	2
39	S17,2		TBI	3	21	MCS+	4	5	5	3	2	2
41	S18,1	F/20	Anoxic	3	21	eMCS	4	5	6	3	2	1
42	S18,2			3	22	eMCS	4	5	6	3	2	2
43	S19,1	M/54	CVI	4	6	UWS	0	1	2	1	0	2
44	S20,1	M/66	CVI	3	14	MCS+	3	3	5	1	0	2
45	S21,1	M/73	TBI	2	10	MCS-	2	3	2	1	0	2
46	S21,2			2	11	MCS-	2	3	3	1	0	2
47	S22,1	M/70	CVI	3	18	LIS	4	5	2	2	2	3
48	S22,2			3	18	LIS	4	5	2	2	2	3
49	S22,3			3	18	LIS	4	5	2	2	2	3
50	S23,1	M/23	TBI	3	5	UWS	0	0	3	1	0	1
51	S24,1	F/50	CVI	3	4	UWS	0	0	2	1	0	1
52	S25,1	M/43	TBI	3	0	Coma	0	0	0	0	0	0
53	S26,1	M/59	Anoxic	3	6	UWS	1	0	2	1	0	2
54	S27,1	M/72	CVI	4	15	MCS+	3	3	4	2	1	2
55	S28,1	M/66	TBI	3	10	MCS+	1	1	5	2	0	1
56	S28,2			3	12	MCS-	2	2	5	2	0	1
57	S28,3			3	13	MCS+	2	2	5	3	0	1
58	S28,4			3	14	MCS+	2	3	5	3	0	1
59	S28,5			3	12	MCS-	2	2	5	2	0	1
60	S28,6			3	13	MCS-	2	2	5	2	0	2
61	S28,7			3	13	MCS-	2	2	5	2	0	2
62	S28,8			3	13	MCS-	2	2	5	2	0	2

Subjects with blue-colored rows are those recorded with 512 Hz (g.USBamp) sampling frequency. The rest were recorded with 500 Hz (g.Nautilus). TBI = traumatic brain injury, CVI = cerebovascular injury.

3.3 Experimental apparatus

Figure 3.1 illustrates the study's experimental setup. Patients were lying down on their beds in the hospital ward of the acute neurorehabilitation unit during each recording session. An active, 16-channel EEG electrode montage concentrated on the motor cortex according to the 10-20 EEG sensor positioning system (Figure 3.1 D&F) was used for data acquisition. EEG data were captured either with a g.Nautilus wireless amplifier at 500 Hz sampling rate or with a wired g.USBamp amplifier at 512 Hz (g.tec Medical Engineering, Austria, Figure 3.1 B). In both cases, a [0.1, 100] Hz hardware bandpass filter was enabled. Electrode position FCz was used as ground and the right earlobe as reference in all cases. Two bipolar FES electrodes (Figure 3.1 E) were placed on the patient's origin and insertion points of the extensor digitorum muscle of the most affected side, as determined by the supervising neurologist, so that a full hand extension movement was achieved. A laptop PC (Figure 3.1 C) was used to present the protocol and for the temporary storage of EEG data.

3.4 Experimental protocol

As shown in Table 3.1, data was collected in 1 to 8 sessions (average 2.2±1.90) for each patient. Each session consisted of 1 to 4 runs (average 2.9±0.57, the minimum target set was 3 runs but it was not always possible due to the patient's condition) and each run lasted approximately 6 minutes. Short breaks were interleaved between consecutive runs. Each of these runs had 15 "motor trials" and 15 "rest trials", presented in random order. An inter-trial interval was imposed between trials whose duration randomly varied between 4 s and 6 s. Before starting each session, patients were awakened and asked to attempt unilateral hand movements of the most affected side or remain at rest. Each trial lasted 4 seconds and began with an auditory cue via loudspeakers delivered in French to

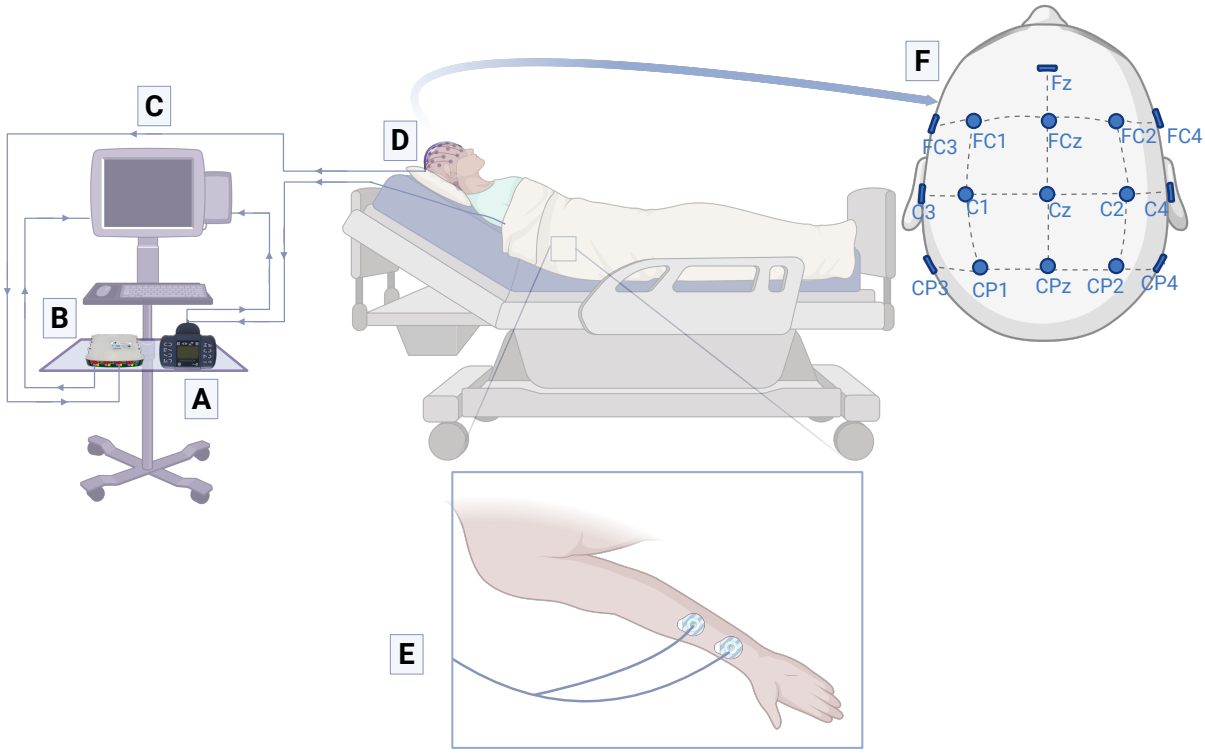

Figure 3.1: Experimental setup:(A) FES device,(B) g.USBamp amplifier,(C) Protocol presentation/recording computer,(D) EEG cap and g.Nautilus amplifier,(E) FES electrodes,(F) EEG layout [5].

the patient: "Ne bougez pas" ("Do not move") for rest trials, or "Bougez" ("Move") for motor trials, respectively. During movement trials, patients were instructed to attempt a hand and palm extension movement, even if they were unable to overtly produced any movement, and repeat the attempts until the trial's end.

The movement trials were immediately followed by FES. The stimulation was transmitted via a MotionStim 8 FES device (Medel, Germany) at a fixed frequency of 35 Hz. It lasted 2 s and was delivered as a 1 s linear ramp-up of the stimulator's duty cycle, from 10% to 100%, then 1 s of continuous stimulation. The FES amplitude varied between 8 and 15mA. It was defined individually for each patient and session via 'trial-and-error' at

the beginning of the session, starting at 5 mA and increasing the amplitude slowly by 1 mA steps until the had extension movement could be achieved, thus using the minimum possible amplitude and the least obtrusive stimulation that is able to produce overt movements. Figure 3.2 shows the corresponding visualization of the experimental protocol.

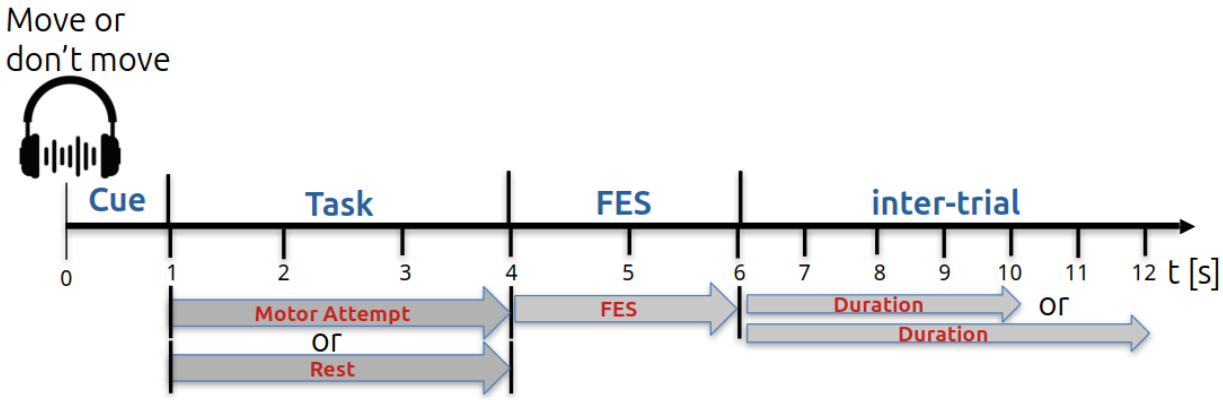

Figure 3.2: Experimental trial timeline visualization.

3.5 Preprocessing, Feature extraction and selection

The data was processed on a per-session basis to account for variability in awareness levels across sessions of the same participant, given the acute condition of the recruited patients. This variability is evident in the clinical CRS-R scores and diagnosis shown in Table 3.1. For instance, subject $S3$ exhibited increasing levels of awareness in all three of his recorded sessions, with substantial improvement between the first two sessions, critical enough to alter the neurological, CRS-R-based diagnosis from UWS, to MCS- and, eventually, MCS+. The movement, rest, and corresponding FES epochs were then extracted and analyzed for each metric. The extracted epochs were labeled and subjected to artifact removal using the Fully Online and Automated Artifact Removal for BCI (FORCe) method [79]. The first second of the movement and rest epochs was discarded in order to eliminate any effect of the auditory cues on the EEG, so that only the last 4 s of these epochs was fed to

subsequent processing. Sessions with less than 80% successfully cleaned EEG trials were excluded due to excessive artifacts, leading to the removal of 1 subject from the analyzed dataset (3 sessions, 9 runs, not reported in Table 3.1). For the remaining subjects, less than 1% of the epoch-of-interest data were found to contain non-rectifiable artifacts. All participatory recorded data are reported in Table 3.1.

Each run was treated with DC removal for all channels and a cross-neighbour Laplacian spatial filter was applied to enhance signal localization. When neighboring channels were unavailable, the remaining available neighbors were used for the Laplacian derivation. To evaluate SMR, the PSD of the EEG signal was computed for each channel individually in 1 s overlapping, sliding windows with shift of 62.5 ms or 100 ms (for 512 Hz and 500 Hz sampling frequency, respectively), using the Welch periodogram method [180]. The internal Welch PSD parameterization involved 0.5 s windows with 50% overlap for a sampling frequency of 512 Hz (g.USBamp) and 40% overlap for 500 Hz (g.Nautilus). This resulted in all cases in a 2 Hz frequency band resolution of the PSD spectrum which was confined to the [4,48] Hz range of interest, covering the θ (4-8 Hz), μ (8-12 Hz), and β (16-30 Hz) bands, which are neurophysiologically relevant with respect to motor and cognitive tasks.

The extracted PSD is used to evaluate whether SMR and other brain rhythms contain information on awareness, but also form the candidate spatio-spectral feature space giving rise to classification accuracy as a possible neuromarker. However, the extracted PSDs are only applicable to the first objective, as the second objective pipelines are fed pre-processed EEG signals directly. Towards this end, PSD features undergo a feature selection step before being fed to the training of a ML decoder, replicating the open-loop processing of a MI BCI. Specifically, for each session, the 6 best features, defined as those exceeding the 5^{th} percentile after FS ranking were selected for classification. I employ a conventional definition of FS as $FS = \frac{|\mu_1 - \mu_2|}{\sqrt{s_1^2 + s_2^2}}$, where μ_1, μ_2 are the means and s_1, s_2 are the standard deviations of the two mental classes contrasted (either MI vs Rest or FES vs Rest). I excluded features corresponding to (particularly vulnerable to artifacts) channel Fz, and

frequencies beyond the physiologically significant for motor tasks μ (8-14 Hz) and low β (18-24 Hz) bands.

3.6 Conclusion

Despite the broader challenge of data scarcity in BCI research, the dataset used in this thesis offers significant novelty due to its scale and its inclusion of patients spanning the entire DOC spectrum, alongside healthy controls, all recorded in a real-time ICU environment. The imaging paradigm and measurement protocol were specifically designed to elicit only the epochs relevant for extracting SMR for subsequent analysis. Moreover, in accordance with current best practices, the dataset underwent rigorous preprocessing, including noise removal, PSD extraction, and feature generation tailored to the BCI-inspired methodology employed in this work. At the end, analysis was conducted for each metric as shown in the Table 3.2;

Table 3.2: Summary of metrics evaluated across MI, FES, and Rest conditions.

✓ Metric	MI	FES	Rest
Classification Accuracy (LDA, EEGNet, DeepConvNet, EEGConformer)	✓	✓	
Effective Connectivity (DTF)	✓	✓	✓
μ-band Separability	✓	✓	
β-band Separability	✓	✓	
Perturbational Complexity Index (Lempel–Ziv & Kolmogorov)	✓	✓	✓
Spectral Slope	✓	✓	✓
Feature Significance (After FDR/Bonferroni correction in 1, 3, 5 & 7 Hz)	✓		
CRS-R thresholds (10, 12, 13)			
Chance-level Based Accuracy (CBA)	✓	✓	

Chapter 4

Validating state-of-the-art EEG-based Metrics for Diagnosing Awareness

BCI-inspired neuroimaging paradigms have increasingly generated empirical evidence supporting the use of neuro-imaging techniques for the detection of CMD. Yet, despite these advances, clinical translation remains elusive, largely due to methodological and validation gaps. The experiments presented in this chapter are designed to answer the first research objective. It is also designed as a necessary step toward bridging the divide in clinical adoption, providing systematic validation of candidate EEG-based metrics in contexts directly relevant to clinical practice. This chapter presents the experimental analyses, evaluation procedures, results, and an in-depth discussion of the findings that collectively advance the case for clinically deployable BCI-based diagnostics.

4.1 Introduction

Victims of traumatic brain injury or cerebrovascular accidents are often prone to neurological complications. Among these, one of the most debilitating is a disorder of consciousness (DOC), which impairs a patient's awareness of themselves and their environment.

Tragically, DOC can also arise from non-traumatic causes, such as pharmaceutical abuse.

Within neuroimaging and BCI-inspired frameworks, the diagnosis of DOC is informed by a set of established neurophysiological markers. These include ML-based classification accuracy [168, 131, 53] contrasting a motor task with an alternative condition (rest in this thesis), brain connectivity measures [181, 182, 183] (specifically effective connectivity in this work), separability metrics [184, 185, 186], the PCI [187, 188, 189], spectral slope [190, 191, 192], feature significance analyses [193, 194, 195], and behavioral thresholds derived from the CRS-R. Although this list is not exhaustive, it encompasses the most widely adopted approaches currently used in the field.

Classification accuracy refers to the performance obtained when labeled neural data from two experimental conditions—typically a motor-related task and a control condition (e.g., rest)—are input into either shallow [53, 196] or deep-learning classifiers [197, 198]. The resulting classification accuracy is evaluated against chance-level bounds, and performance exceeding these bounds is interpreted as evidence of preserved awareness. In parallel, functional and effective connectivity analyses examine interactions between distributed brain regions. Depending on the regions involved, these interaction patterns can be neuroscientifically interpreted to infer motor intent, task engagement, or the presence of conscious processing. Connectivity-based measures have therefore proven to be valuable indicators in clinical contexts. Separability analysis (though sometimes proposed under different names like decodability, classification separability, representational differentiation) aims to quantify the discriminability of task-related neural activity—often expressed in terms of SMR—relative to a control condition such as rest. Awareness is inferred when the discriminant capacity between conditions reaches statistical significance. Notably, separability is conceptually related to feature selection in ML, although in this context the discriminant power is assessed explicitly for statistical significance between conditions rather than being used solely to optimise classifier performance.

Beyond these approaches, the Perturbational Complexity Index (PCI) has been pro-

posed as an alternative neurophysiological marker of awareness, although it has been applied predominantly in studies involving anesthetized subjects. In this framework, neural time-series data are transformed—often via time–frequency methods such as the Hilbert transform—binarised, and then analysed to compute spatiotemporal complexity. The resulting index ranges from 0 to 1, with higher values indicating greater levels of consciousness. Spectral slope approaches model neural dynamics using logarithmic frequency representations and polynomial fitting to estimate the degree of continuous cortical interaction. A steeper (i.e., more negative) spectral slope indicates a relative dominance of low-frequency activity and has been associated with reduced arousal or levels of consciousness, whereas a flatter spectral slope reflects relatively greater high-frequency activity and is typically linked to higher arousal states [199]. A further statistical method involves segmenting neural data into task and control conditions and applying parametric hypothesis testing under the null assumption that both conditions arise from the same distribution. After correcting for multiple comparisons using False Discovery Rate (FDR), awareness is inferred when the alternative hypothesis is accepted. To strengthen robustness, some studies require significance to persist across consecutive frequency bins (e.g., up to 7 Hz), a procedure commonly referred to as feature significance analysis.

In clinical practice, the evaluation of consciousness continues to rely heavily on CRS-R scores, despite their known methodological and diagnostic limitations. To address this, classification outcomes in this thesis are assessed using permutation-based chance-level bounds, herein termed CBA, providing a statistically rigorous basis for interpreting decoding performance beyond arbitrary accuracy thresholds. While neuroimaging and BCI-inspired approaches have shown increasing promise, their uptake in clinical settings remains limited. This continued dependence on the CRS-R appears to stem less from its adequacy and more from the lack of alternative methods that are both sufficiently validated and reproducible. As a result, many proposed techniques fail to converge toward clinical adoption, hindered by heterogeneous protocols and the absence of robust validation frameworks. By

adopting a statistically principled evaluation strategy, this thesis aims to contribute toward bridging this gap between methodological innovation and clinical applicability.

In this context, the present chapter seeks to systematically validate several state-of-the-art EEG-based metrics for diagnosing awareness in DOC, including classification of mental-state versus rest conditions, brain rhythm analysis, effective connectivity, perturbational complexity index (PCI), spectral slope, and complexity measures.

4.2 Data Analysis

I report in this Chapter a two-class classification accuracy of both MI vs Rest and FES vs Rest task pairs, pooling all PSD samples of all corresponding epochs in a session. Classification accuracy is estimated using a leave-one-trial-out cross-validation approach with LDA decoders. A random classification accuracy (chance) level with 99% confidence is determined by performing 100 repetitions of the cross-validated classification procedure, each time randomly permuting the class labels [6]. Increasing the number of permutations improves the resolution of the estimated null distribution and is, in principle, advantageous for characterizing chance-level effects. However, permutation testing primarily aims to establish a stable estimate of the null distribution rather than to achieve arbitrarily high precision in p-value estimation. Given the high computational burden imposed by the size of the dataset, I resorted to using 100 permutations which are sufficient to derive a chance-level threshold with $\alpha = 0.99$ by means of the acquired set of 100 random accuracy samples. The chance level is then obtained as the 99^{th} percentile of the resulting distribution of random classification accuracy values.

Previous research supports the use of PCI as a diagnostic metric for assessing awareness in DOC patients [187, 188, 200, 201]. Although PCI is conventionally derived from Transcranial Magnetic Stimulation (TMS) responses and most often studied with respect to loss of awareness and consciousness following anaesthesia administration, given the obvious

analogy with awareness in DOC, I applied an identical approach to the data. Lempel-Ziv complexity (LZc) [202] quantifies the rate at which new patterns appear within a time series, and is widely employed to assess the complexity of an EEG signal [187, 203, 188]. I computed the LZc for each channel. The process involved calculating the absolute value of the Hilbert transform of the EEG signal and obtaining its mean for each channel. The signal was then binarized by assigning a value of 1 where the signal exceeded the mean, and 0 otherwise, following the approach proposed by [188]. For each channel, a binary sequence was generated and a corresponding dictionary of unique substrings was constructed. The LZc was defined as the number of distinct substrings in this dictionary. To account for inherent signal structure, LZc is normalized by dividing the LZc of the original signal by the average LZc computed from 50 randomly shuffled versions of the same signal.

Additionally, given that LZc captures the rate of emergence of new patterns, I also opted to estimate a variant of the LZc known as Kolmogorov complexity (KVc) [204] which reflects the re-constructibility of a substring or sequence. With the same binarized data used in the computation of LZc, I assessed whether new segments of the binary sequence could be reconstructed from previously encountered parts. Each time a new segment could not be generated from the existing sequence, the complexity counter was incremented. The final complexity value was normalized using the theoretical upper bound for information growth, given by n/log2(n) (where n is the sequence length).

I also computed effective connectivity across the task conditions to investigate the directional information flow between brain regions during the respective mental tasks. Specifically, I employed DTF [205] to estimate the frequency-domain connectivity, using MVAR [206] within windows corresponding to the length of each trial. The model order was set at 44, determined using Schwarz's Bayesian Information Criterion [207].

The optimal model order was then estimated using the function `arorder.m` [110, 208], defined as: $[popt, sbc] = arorder(v, pmin, pmax, selector)$ where v is the windowed EEG data, `popt` is the optimal order and `sbc` is a vector of Schwarz Bayesian Criterion(SBC)

values whose minimum corresponds to the selected order. This process is typically repeated across multiple data windows, and the final model order is chosen based on the most frequently occurring optimal value or the one minimizing model variance across windows. After adopting this process, a model order of "44" was chosen and used for the effective connectivity analysis in this research.

DTF connectivity was computed over the frequency range of [1,40] Hz, at a resolution of 1 Hz. This approach yielded directed connectivity matrices for all possible pairs of EEG channels (excluding Fz), with each matrix capturing the pairwise interactions across frequency bands.

Finally, following [190], I computed the "spectral slope" for each channel. This metric is quantified as the slope (i.e., first order coefficient) of the linear polynomial that represents the linear fit to a channel's log-transformed PSD spectrum. The spectral slope is first computed separately for each epoch, and the final metric value is acquired as the average of all epochs of the same type in a session. The linear fit is estimated with regular, maximum-likelihood linear regression.

4.3 Data Evaluation

To assess the clinical relevance of the aforementioned popular EEG-based metrics for awareness in DOC using the novel dataset, I applied a battery of statistical approaches across the board, along with certain metric-specific methodologies, especially with regard to the use of classification accuracy in this context. It must be highlighted that, in this line of research, one is always faced with the inevitable issue of the absence of any "hard" and reliable ground truth. In other words, there is no widely accepted method or measure of any sort that can infer a patient's level of awareness precisely and without reasonable doubt, so that it can be used to exactly evaluate the suitability of the examined EEG-based metrics. The CRS-R clinical scale is, among the measured variables available, the one that is most

well-established and considered (here, and in general) to best approximate the unavailable ground truth; however, its known deficiencies, especially concerning false negative diagnosis (i.e., inability to detect CMD), are precisely why objective neuroimaging-based metrics are sought. Given this issue, the general principle followed for evaluation is to adopt CRS-R and the thereby derived neurological diagnosis as a "virtual ground truth", but also take into account their limitations. More specifically, I require that an EEG metric exhibits a notion of statistically significant agreement with CRS-R in order to be accepted as relevant, but I do not demand absolute or even strong agreement, and interpret individual session deviations from the projected CRS-R values on their own merits.

More elaborately, to examine the relationship between the selected metrics and total CRS-R scores per session, I computed the corresponding correlation coefficient and its significance with $a = 0.05$. The statistical significance of these correlations was assessed with Student's t-tests to accept or reject the hypothesis that the linear fit to the data has a slope β=0, and thus no correlation between the dependent and independent variables.

For connectivity, I computed the correlation with CRS-R individually for each source, target channel, and frequency band triplet and applied FDR correction for multiple comparisons. Only DTF connectivity from channel C3 to channel CP1 within sub-bands of the δ band (specifically, 1, 2 and 4 Hz) survived this test for the MI and Rest conditions (not for FES). Based on this finding, I subsequently submit to further analysis only this particular effective connectivity variable, namely, the full-δ-band C3 $\rightarrow$ CP1 DTF connectivity, where full-δ is derived as the average of the corresponding original DTF estimates in $[1, 4]$ Hz.

Additionally, for metrics exhibiting significant correlation with CRS-R, I studied several partitions of the CRS-R scores into two distinct groups, "aware" vs "unaware", each based on a CRS-R threshold $th \in [7, 19]$ with $CRSR <= th$ mapped to the "unaware" group and $CRSR > th$ to the "aware" one. Alternatively, I also tested such partitions based on the neurologists' diagnosis (which, in turn, was mainly based on the CRS-R score, but also took into account other potentially available scales such as MBT and GCS, as well as

the overall clinical picture of the patient); specifically, for this type of grouping, patients diagnosed as Coma or UWS are always placed into the unaware group, while LIS, eMCS and the healthy control are always placed in the aware group. Finally, the three possible partitions resulting from placing MCS- and MCS+ in either of the two groups are defined. For each of these partitions, an unpaired, two-sided t-test comparing the two groups (aware vs unaware) was then performed on all metrics.

I repeated this analysis using the sub-scale component items of CRS-R scores (*Auditory, Visual, Motor, Oro-Motor, Communication, Arousal*), which enables us to identify approximate levels of EEG metric values at which awareness can be distinguished from non-awareness, and link these with CRS-R and its sub-scales. Furthermore, to assess whether there were significant differences across the finer, CRS-R-based diagnostic groups for each metric, a one-way Analysis of Variance (ANOVA) was also conducted. This test was applied independently to all evaluated metrics to determine whether the means differed significantly between the seven formal awareness groups in Table 3.1.

Last but not least, in spite of the fact that the lack of actual ground truth does not allow for formal and precise assessment of the examined metrics' diagnostic value, I find it useful towards judging their overall potential to define methodologies able to produce a final binary inference on a subject's current possession of concealed awareness. To do so, I derive a metric-specific threshold, above which the subject is inferred to be, to some degree, aware. In other words, this threshold determines a cut-off value of the metric in question indicating a possible case of CMD. Only metrics that have been adequately shown to be relevant in previous analysis (i.e., showing at least one of the following signs: i. significant correlation with CRS-R, ii. significant ANOVA outcome, iii. existence of significant aware vs unaware group splits) are subjected to this process. For this subset of metrics, the threshold is extracted by splitting the total metric's range across the whole dataset into 100 equidistant levels and checking whether, for any of these levels, the distribution differences of the two groups formed by considering all sessions above/below the level in question

are statistically significant with an unpaired, two-sided t-test. Effectively, the underlying assumption of this approach is that, if a threshold able to significantly discriminate CMD from DOC exists, it must correspond to the value that creates the most separable groups possible on its left (DOC, unaware) and right (CMD, aware). If multiple metric values satisfy this condition, the one with the smallest p-value is selected as the CMD-critical threshold. Equipped with this tool, I present each metric's prediction on the CMD status of each session, and thus also the percentage of sessions diagnosed with CMD by each metric.

This threshold-based approach to make CMD predictions is compared to another three such methods: First, simple CRS-R thresholding, using three thresholds on the total CRS-R scores (namely, 10, 12 and 13) justified by the literature and by the preceding group-splitting analysis. Second, specifically for the classification accuracy metric, by labeling as CMD all sessions where the derived accuracy exceeds the session-wise chance-level with 99% confidence. As previously mentioned, I term this approach, which I introduced in my previous work [6] and is very similar to the statistical inference applied by [38], as CBA (Chance-level-Based Accuracy). Lastly, I also make CMD predictions using the individual PSD feature values using a similar approach to that described in [34, 193]. This involves statistically testing each candidate PSD feature to determine whether its values tend to differ significantly between a target (in this case MI) and a control (here, Rest) mental task. A session is classified as aware/CMD by this feature significance (FS) method when any particular feature pops up as significantly differing between the two conditions after correction for multiple comparisons. Two correction methods are applied, conservative Bonferroni correction and the less strict FDR correction (FS-BF and FS-FDR, respectively). To make this approach less susceptible to false positive decisions, I further test the cases that a number N of consecutive frequency bands of the PSD on a particular channel must be singificant in order to infer CMD. I report results for $N = 1, 3, 5, 7$, and $a = 0.05$ throughout. Although there is no way to strictly quantify the accuracy of any of

these methods to detect CMD, these comparisons provide interesting insights, especially with regard to the tendency of each criterion to be conservative or optimistic in diagnosing covert awareness.

4.4 Results

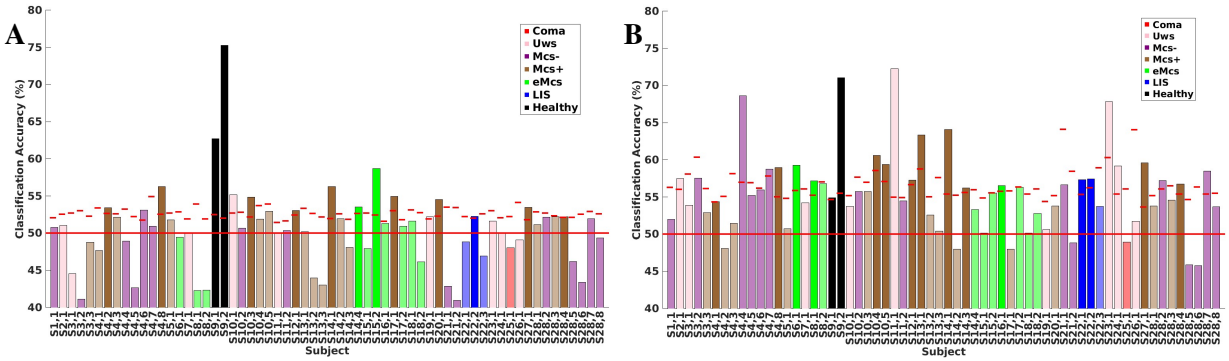

Figure 4.1: Classification Accuracy (A) MI vs Rest and (B) FES vs Rest with leave-one-trial-out cross-validation for all subject sessions. The horizontal red line illustrates the expected value of the random chance level for 2-class problems, while the session-wise lines indicate the permutation-based chance level with 99% confidence [6].

Figure 4.1A presents the 2-class classification accuracy for the MI vs Rest and Figure 4.1B for the FES vs Rest taskset, with red lines on top of each bar indicating the session-wise random classification level. It is observed that, with regard to MI epochs, the accuracy of 17 (out of total 62, 27%) sessions across 12 different (out of total 28, 43%) subjects exceeds the chance level, whereas for FES epochs the corresponding figure concerns 25 sessions (40%) across 18 subjects (64%). As argued in my previous work [6], I posit that the condition of exceeding the dataset-specific chance classification level (called here CBA) may be a more reliable measure of awareness in DOC than the absolute value and other common usages of classification accuracy in this context. Notably, coma patient $S25$ who unequivocally had no awareness (Table 3.1), did not exceed the chance level by large

margins in either MI or FES epoch classification. Conversely, the healthy control subject demonstrated above-chance and consistently high, well-above the chance level, classification accuracy in both MI- and FES-based decoding. Hence, the CBA metric exhibits sound response for the available "negative" (coma) and "positive" (healthy) controls. This does not hold true for the conventional $1/N$ (50% in this case) chance-level threshold often employed in the literature [134], which is shown to be over-optimistic.

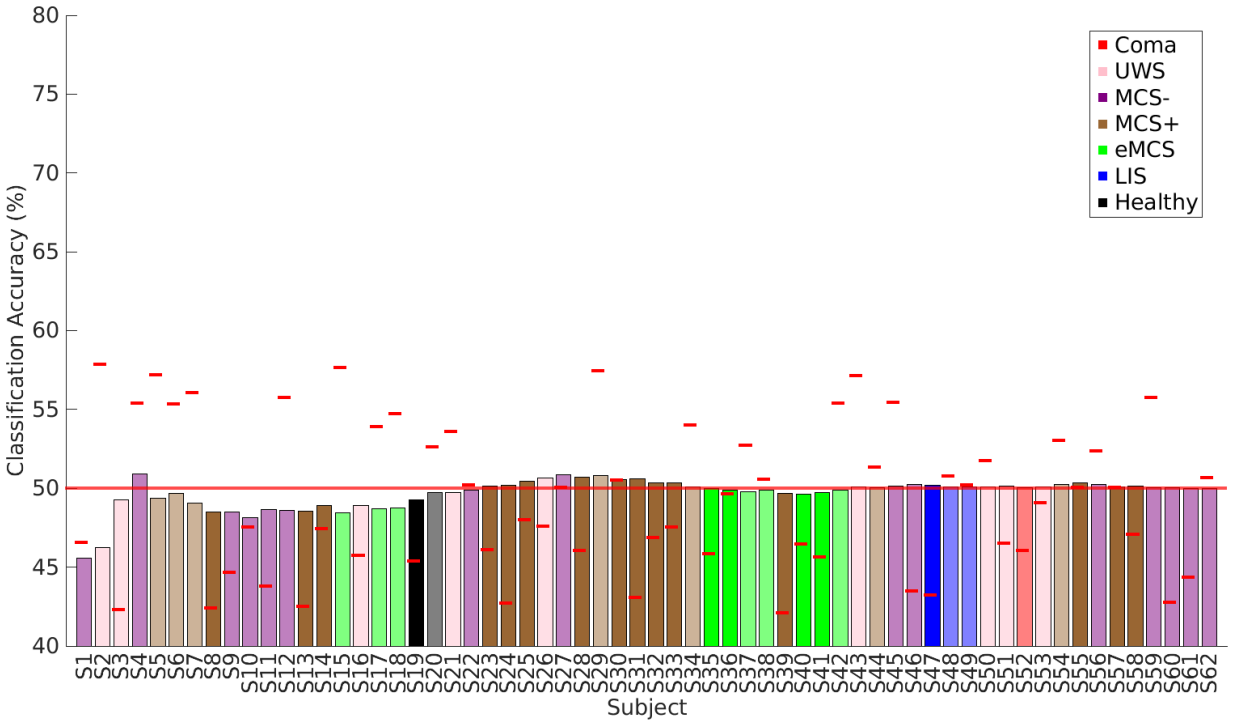

Figure 4.2: Classification Accuracy MI vs Rest with leave-one-session-out cross-validation for all subject sessions. The horizontal red line illustrates the expected value of the random chance level for 2-class problems, while the session-wise lines indicate the permutation-based chance level with 99% confidence [6].

Similar to Figure 4.1, but employing a Leave-one-session-out cross validation (LOSO) cross-validation across all subject sessions, Figure 4.2 depict the two-class classification accuracies for the MI versus Rest taskset. Red horizontal markers above each bar represent the session-specific chance-level thresholds derived from random permutations. Accuracy

exceeded chance in 34 of 62 sessions (51.6%). Interestingly, the coma patient $S52$, clinically confirmed to be unaware, was classified as aware, whereas the second session of healthy control $S20$ was classified as unaware. Overall, the LOSO accuracies fluctuate around the 50% baseline expected for a two-class problem, suggesting that the classification results largely reflect chance-level performance rather than reliable task-specific decoding.

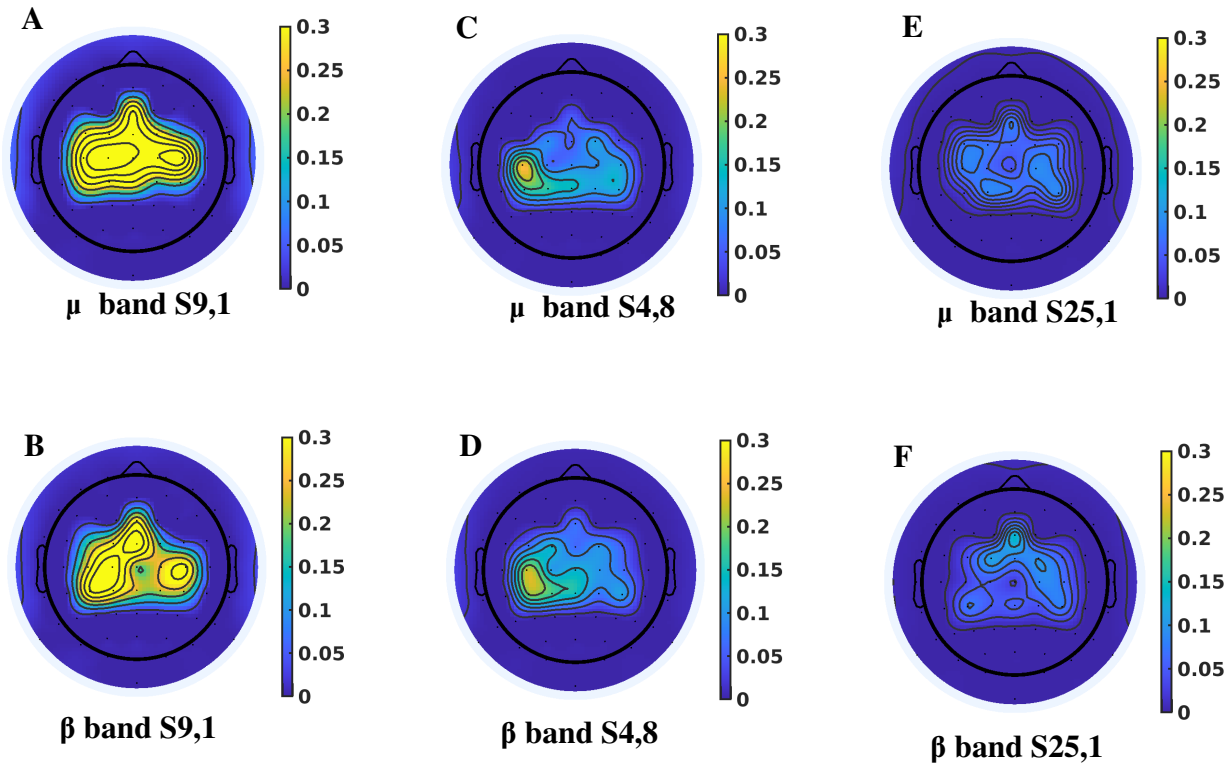

Figure 4.3: Topographic SMR distribution of three different sessions. (A-B) Healthy control session $S9, 1$. (C-D) Session $S4, 8$ of potential CMD patient. (E-F) Coma patient control session $S25, 1$. Top panels show the topographic distribution of μ rhythms and bottom panels of β rhythms. The strength of SMR is represented by the average Fisher Score separability of PSD features (Fisher-95, as described in Chapter 3, 3.5) between MI and Rest within the μ (8-14 Hz) (top) and β (bottom) bands, as colour-coded in the colourbars.

Figure 4.3 illustrates the topographic SMR distribution of 3 sessions of interest, two corresponding to the control participants (healthy S9, coma S25) and one to the diagnosed MCS patient S4 that is found by the CBA method (and other EEG metrics, see below) to possess covert awareness. These scalp maps reinforce the expectation that participants with high accuracy and, in general, signs of awareness (above-chance accuracy), were indeed able to generate neurophysiologically relevant SMR brain patterns and, thus, were likely following the protocol's commands. On the contrary, the coma patient who is not expected to follow commands or engage in motor attempts, simultaneously exhibits well below-chance classification accuracy and complete absence of SMR modulation. This fully corroborates the soundness of the classification outcome. Furthermore, the patterns of both the healthy control S9 and of the assumed CMD patient S4 show the anticipated for motor tasks of the upper limbs laterality within the MI-relevant bands (μ, β) [209, 114]. It is also noteworthy that, as also expected, there seems to exist direct proportionality between the strength and spread of Event-Related Synchronization/Desynchronization (ERD/ERS) and accuracy across all sessions. This additionally highlights that even DOC patients that likely maintain latent awareness may be limited in their ability to elicit and maintain strong SMR modulation, equivalent to that observed in most healthy subjects performing MI. However, spared SMRs, despite feeble, do follow the same trends as in healthy brains, both in terms of somatotopic organisation and of frequency specificity.

Figure 4.4 presents the ERD/ERS heatmaps for the corresponding subjects whose topographic maps are shown in Figure 4.3. These heatmaps were computed from a single experimental run for each session. A key advantage of ERD/ERS analysis is that it expresses task-related changes in EEG power within specific frequency bands relative to a baseline period, typically defined as the few seconds preceding event onset and recorded from the same electrode derivations [210]. This normalization facilitates the visualization and interpretation of task-induced modulations in neural oscillatory activity.

In the absence of any definite ground truth, I make use of the correlation between the

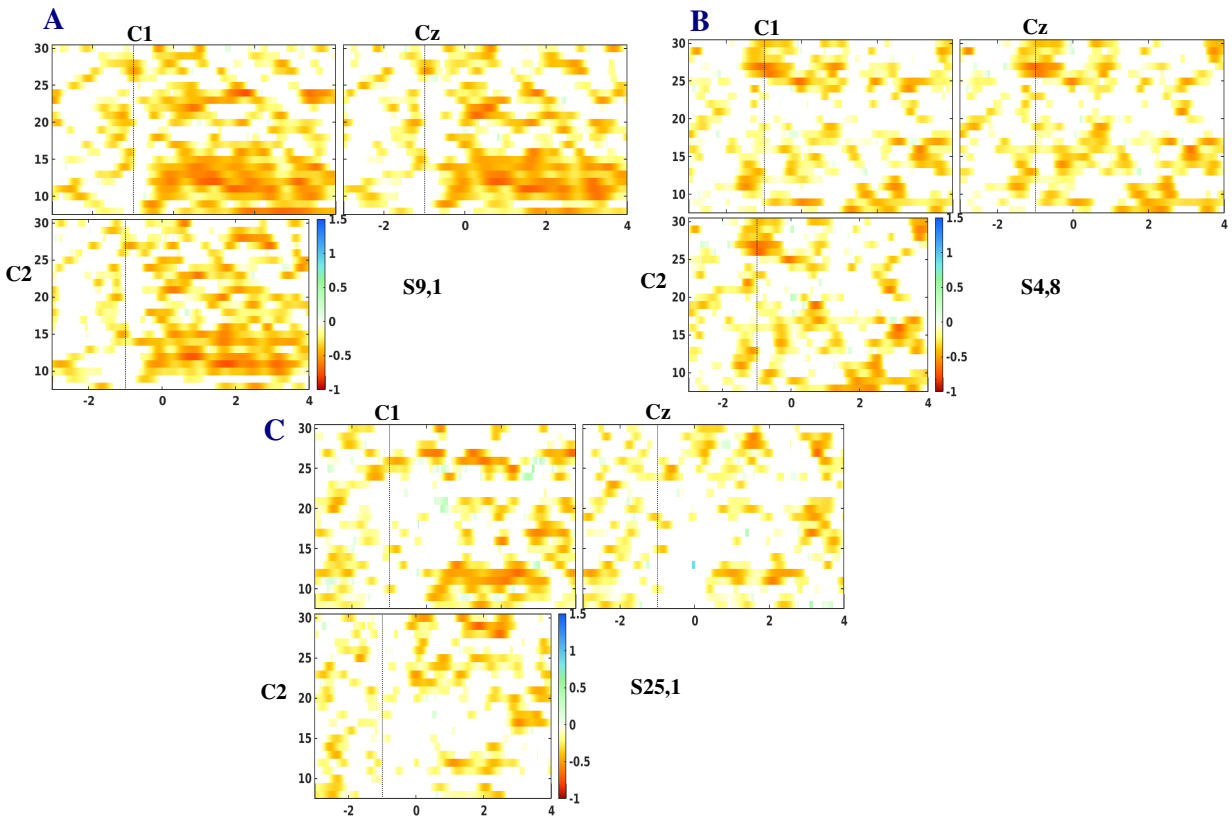

Figure 4.4: ERD/ERS maps of the three different sessions. (A) Healthy control $S9, 1$. (B) Subject $S4, 8$, (C) Coma patient $S25, 1$.

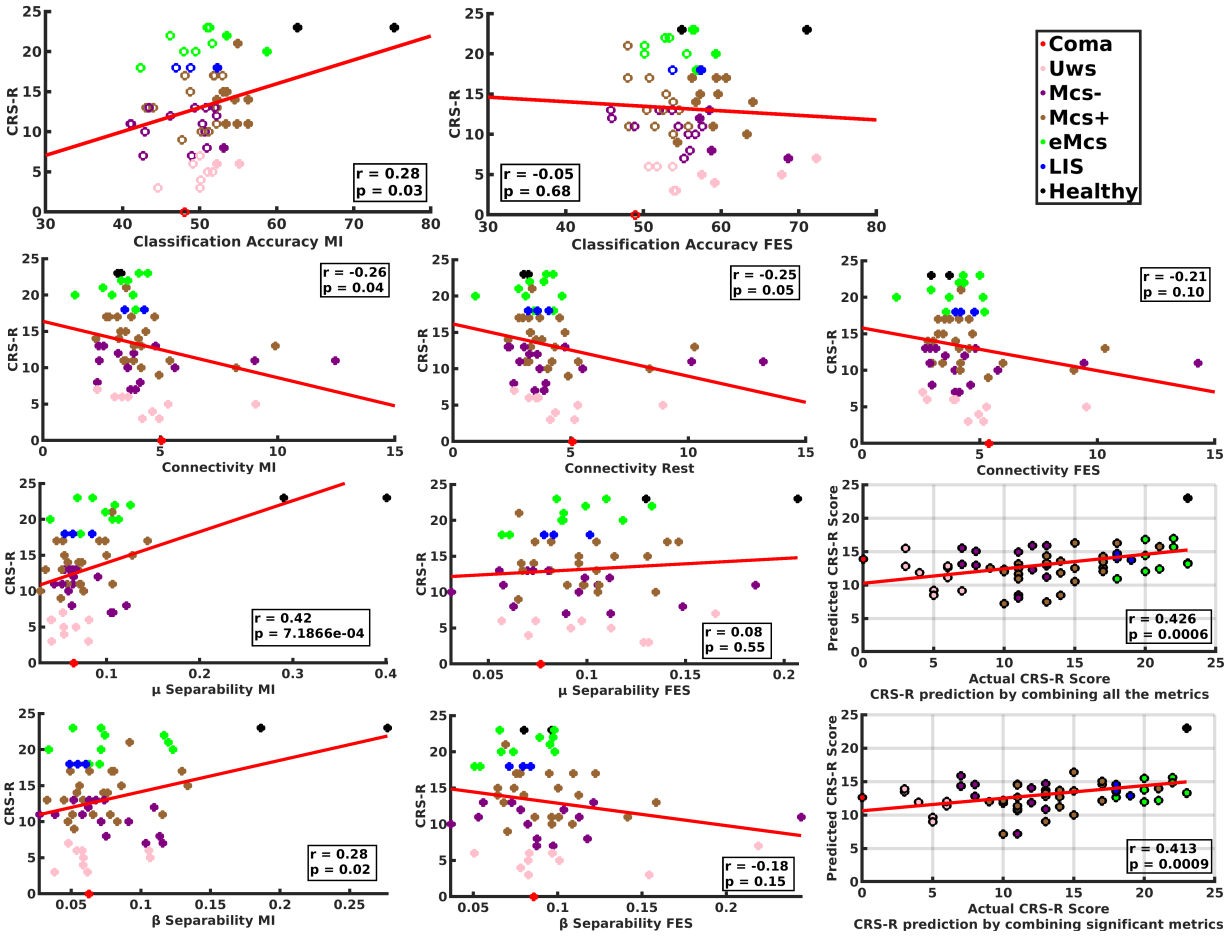

Figure 4.5: Correlation between CRS-R and various metrics as shown by the panel titles using MI, FES, and Rest epochs, as well as regression-based prediction of CRS-R using the metrics. The CRS-R-based diagnosis of each subject-session is colour-coded as indicated by the legend. For the classification accuracy panels, subject-session points whose accuracy exceeds the chance threshold (see Figure 4.1) are represented with filled markers, while those below the threshold are unfilled.

evaluated EEG-based metrics and the CRS-R scores to assess their prowess in reflecting awareness. Strong, statistically significant correlation suggests that a given metric tends to agree with CRS-R, while lack thereof indicates the contrary. For metrics fulfilling the basis of significant correlation with CRS-R, I subsequently discuss further the extent to

which specific occasions of disagreement can be regarded as evidence that the metric in question can rectify CRS-R's false negative vulnerability, or simply highlight 'the said metric's own shortcomings. In addition, CRS-R scores were estimated using a regression-based framework in which the derived metrics were combined as an ensemble. Model performance was evaluated under two conditions: (i) using the full set of metrics and (ii) restricting the model to statistically significant metrics only. The resulting correlations between predicted and actual CRS-R scores are presented in the bottom two rows of Figure 4.5, with corresponding p-values of 0.0006 and 0.0009 for the full and reduced metric sets, respectively.

Figure 4.5 summarizes the key correlations results. The top-left panel shows a significant positive correlation between MI vs Rest classification accuracy and CRS-R ($r = 0.28, p = 0.03$). Notably, several subjects with above-threshold classification accuracy exhibited low CRS-R scores, potentially indicating false negatives, as hypothesized in the literature [32, 62, 56]. This finding supports the hypothesis that covert awareness may exist in patients with impaired motor function. In contrast, the FES vs Rest classification accuracy showed no significant correlation with CRS-R ($r = -0.05, p = 0.68$). The second row of Figure 4.5 illustrates the correlation between δ-band C3 $\rightarrow$ CP1 effective connectivity and CRS-R scores across MI, FES, and Rest conditions. While effective connectivity during FES showed no significant relation to CRS-R, in both MI and Rest significant negative correlations are found ($r = -0.26, p = 0.043$ and $r = -0.25, p = 0.046$, respectively).

The observed negative correlations have, in some contexts, been interpreted as reflecting pathological hyperconnectivity, such as in neuroticism, where neural networks become excessively coupled yet lack meaningful information exchange [211]. In contrast, Demertzi et al. [212] argue that negative correlations in brain activity arise from neural inhibitory mechanisms that regulate global and local network interactions. From this perspective, anticorrelations are considered a physiological marker of preserved consciousness, whereas

their absence may indicate a loss of subjective experience.

The final two rows of Figure 4.5 show the correlation between brain rhythm separability (quantified as the FS between the distributions of a particular brain rhythm–channel and frequency band pair–for the MI vs Rest or FES vs Rest mental classes) and CRS-R scores. Significant positive correlations were observed for MI vs Rest in both the μ-band ($r = 0.42, p = 7.19 \times 10^{-4}$) and β-band ($r = 0.28, p = 0.02$), suggesting that higher sensorimotor rhythm modulation during MI may reflect greater levels of residual awareness. In contrast, separability in the FES vs Rest condition showed no significant correlation with CRS-R ($r = 0.08$, $p = 0.55$ and $r = -0.18$, $p = 0.15$, respectively) in either frequency band. For the sake of completeness, Figure 4.6 illustrates the corresponding plots for metrics that did not satisfy any of the statistical criteria assessing the degree of affinity with CRS-R set here (complexity, slope of spectrum).

Complementing Figure 4.5 to provide a full overview, Table 4.1 reports the correlation (to four decimal places) between all metrics with CRS-R scores across MI, FES and Rest conditions. Overall, complexity and slope metrics did not show significant correlations in any condition. However, four key metrics (classification accuracy, C3→CP1 δ-band connectivity, μ-, and β-band separability) exhibited significant correlations with CRS-R during MI. One-way ANOVA analysis for these five metrics revealed significant group differences with regard to the 7 CRS-R-based diagnostic groups defined (Table 4.1). In contrast, MI-based Lempel-Ziv complexity yielded no significant correlation with CRS-R, leading to rejection of the hypothesis that group means differ significantly. Lempel-Ziv complexity during Rest is the only metric and condition combination yielding a marginally non-significant correlation with CRS-R ($p = 0.0524$). As shown in both Figure 4.5 and Table 4.1, none of the examined metrics during FES correlated significantly with CRS-R. The ANOVA results also suggest that accuracy and separability carry information on the differentiation of the levels of awareness, but, again, only during MI. Effective connectivity does not exhibit significant differences across diagnostic groups despite its

Table 4.1: Overview of metrics' correlations with CRS-R and one-way ANOVAs with CRS-R group as factor. Blue-colored cells indicate statistical significance.

Metric	**Correlation**		**ANOVA**	
	CRS-R		**CRS-R grouping**	
	p-value	**r-value**	**p-value**	**F statistic**
MI				
Classification Accuracy	0.0307	0.2700	1.378e06	8.57
Effective Connectivity(C3→CP1, δ-band)	0.0430	-0.2579	0.8430	0.4
$\boldsymbol{\mu}$ Separability	7.1866e-04	0.4182	1.40341e-16	33.7
$\boldsymbol{\beta}$ Separability	0.0191	0.2969	4.57512e-08	11.08
Slope	0.1800	0.1725	0.8372	0.46
Lempel-Ziv Complexity	0.6436	0.0599	0.0524	2.24
Kolmogorov Complexity	0.8675	-0.0216	0.1264	1.75
FES				
Classification Accuracy	0.6623	-0.0566	0.2816	1.28
Effective Connectivity(C3→CP1, δ-band)	0.1007	-0.2104	0.9154	0.34
$\boldsymbol{\mu}$ Separability	0.5184	0.0836	0.0696	2.09
$\boldsymbol{\beta}$ Separability	0.1417	-0.1888	0.7882	0.52
Slope	0.1824	0.1716	0.6922	0.46
Lempel-Ziv Complexity	0.3166	0.1293	0.0741	2.05
Kolmogorov Complexity	0.3652	0.1170	0.2156	1.44
Rest				
Effective Connectivity(C3→CP1, δ-band)	0.0464	-0.2539	0.8125	0.49
Slope	0.2622	0.1446	0.7751	0.54
Lempel-Ziv Complexity	0.4126	0.1059	0.1239	1.76
Kolmogorov Complexity	0.6385	0.0609	0.2528	1.35

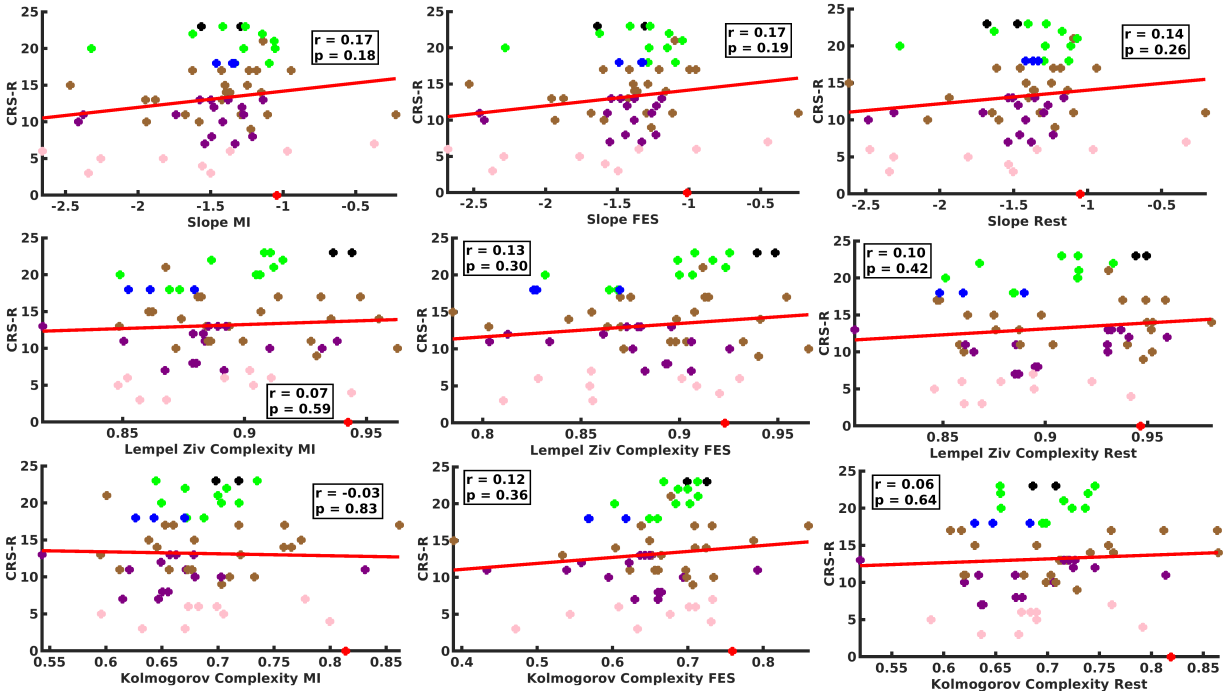

Figure 4.6: Correlation between CRSR and average effective connectivity for various metrics as shown by the panel titles using MI, FES, and Rest epochs. The CRSR-based diagnosis of each subject-session is colour-coded and the same as indicated by the legend in Figure 4.5.

significant correlation with CRS-R.

Table 4.2 reports correlations between CRS-R sub-scales (Auditory, Visual, Motor, Oro-motor, Communication, Arousal) and the different metrics benchmarked here. The MI classification accuracy correlation was significant only for the oro-motor and communication sub-scales. MI and Rest effective connectivity correlated significantly with the oro-motor and communication sub-scales, too. MI μ separability showed robust significance across all six sub-scales, and MI β separability was significant for motor, oro-motor, and marginally non-significant for communication. Finally, during FES, neither μ- nor β-band separability correlates significantly with CRS-R, although β separability comes close with respect to the oro-motor sub-scale.

Table 4.2: Correlation of CRS-R unit sub-scales with different metrics. Blue-coloured cells indicate significant correlation.

Unit Scale	**Classification Accuracy**				**Connectivity - C3→CP1, δ-band**						**Separability**							
	MI		**FES**		**MI**		**Rest**		**FES**		**μ MI**		**β MI**		**μ FES**		**β FES**	
	p-value	r-value	p-value	r-value	p-value	r-value	p-value	r-value	p-value	r-value	p-value	r-value	p-value	r-value	p-value	r-value	p-value	r-value
Auditory	0.0959	0.2134	0.5751	-0.0726	0.2383	-0.1520	0.2553	-0.1467	0.3597	-0.1183	0.0300	0.2758	0.2735	0.1413	0.5499	0.0774	0.4530	-0.0971
Visual	0.0988	0.2116	0.5793	-0.0718	0.1060	-0.2073	0.1112	-0.2043	0.2363	-0.1526	0.0055	0.3488	0.0627	0.2378	0.8460	0.0252	0.1976	-0.1659
Motor	0.1420	0.1886	0.6136	0.0654	0.0600	-0.2402	0.0654	-0.2355	0.1097	-0.2051	0.0052	0.3510	0.0270	0.2809	0.7594	0.0397	0.1856	-0.1703
Oro-motor	0.0360	0.2668	0.6052	0.0669	0.0124	-0.3157	0.0163	-0.3039	0.0337	-0.2701	0.0101	0.3245	0.0180	0.2996	0.8819	0.0193	0.0511	-0.2489
Communication	0.0160	0.3048	0.8250	0.0287	0.0379	-0.2644	0.0309	-0.2743	0.0928	-0.2154	0.0044	0.3574	0.0520	0.2479	0.1727	0.1754	0.3840	-0.1125
Arousal	0.1865	0.1700	0.3762	0.1143	0.8431	0.0257	0.8331	0.0273	0.8398	0.0262	0.0027	0.3740	0.0686	0.2328	0.3550	0.1195	0.3994	-0.1089

Figure 4.7 presents an analysis aimed at identifying metrics that could discriminate between awareness and unawareness. Towards this, using again CRS-R as a surrogate for ground truth, session data are split into two groups based on some CRS-R threshold, and the hypothesis that a metric's values in the two resulting groups come from the same distribution is tested. Rejecting the hypothesis implies that the said metric tends to take on different values depending on the level of awareness, and thus could form a useful diagnostic predictor. Previous studies have proposed CRS-R thresholds of 10 [213] and 12 [3, 214] for splitting groups, motivating the investigation of a broader range of thresholds in [7, 19], including an assessment of individual CRS-R sub-scales. The classification accuracy of MI showed significant differences at thresholds 13, 18, and 19. Effective connectivity during MI, Rest and FES conditions was significant across thresholds 11–16 for both MI and Rest, and 11-13 for FES. μ- and β-band separability in the MI condition were significant from thresholds 13 to 19, while β separability in the FES condition reached significance only at thresholds 11 and 12. No significant effects were observed at any threshold for FES-based classification accuracy, and μ separability. Notably, μ-band separability during MI became significant at the threshold of 13 with $p < 0.01$, but not at lower thresholds, suggesting this may represent a critical boundary for distinguishing covert awareness.

Further similar analysis with CRS-R sub-scales (Figure 4.7), revealed that MI classification accuracy was significant only at higher thresholds, particularly near the upper bound. It must be noted that too small or high CRS-R thresholds reduce the reliability of these findings, as the imbalance between the two groups grows accordingly. Both μ- and β-band separability MI consistently showed significant group differences across various sub-scales. Additionally, effective connectivity during MI and Rest conditions also demonstrated significant differences specifically with the motor, oro-motor, and communication sub-scales. Interestingly, while β-band separability in the FES condition was significant when considering total CRS-R scores, it did not reach significance with most individual sub-scales. In addition to this, FES-based classification accuracy, μ- and β-band separability did not

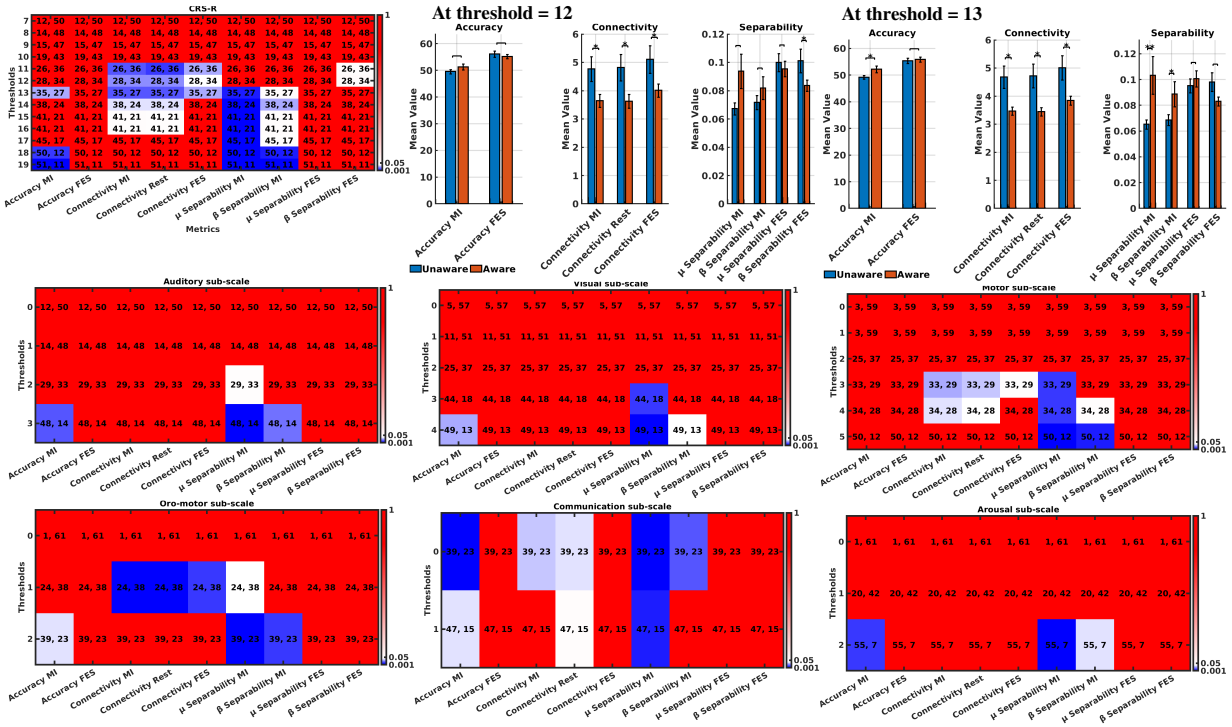

Figure 4.7: Metric and condition ability to discriminate between awareness and unawareness. The heatmaps illustrate the p-values of unpaired, two-sided t-tests as coded in the colourbars, where red shades correspond to no statistically significant difference ($p \gg 0.05$), blue to statistically significant difference ($p \ll 0.05$) and white indicates marginal significance ($p \approx 0.05$). Each cell visualises the p-value of an unpaired, two-sided t-test of the hypothesis that the means of two groups of values of the metric/condition indicated by the cell's column label come from the same distribution. The metric's value derived by a single subject-session is assigned to group "unaware" if the corresponding session is associated with a CRS-R value $CRSR \leq th$ and to group "aware" otherwise ($CRSR > th$). The CRS-R threshold th used for each t-test is shown in the corresponding cell's row label. The numbers (X, Y) in each cell show the cardinality of each compared group pair for the particular test. The top-left panel splits groups according to the total CRS-R. The middle and right panel of the first row show the mean and standard deviation of the group values for each metric for CRS-R thresholds $th = 12$ and $th = 13$, respectively. Statistical significance is shown with asterisks (* for $\alpha = 0.05$, ** for $\alpha = 0.01$). The second and third row show heatmaps where group splitting relies on a CRS-R sub-scale as appears in the title.

exhibit significant two-group difference with any CRS-R sub-scale.

Table 4.3: Statistical significance (unpaired, two-sided t-test) of differences between "unaware" and "aware" groups in terms of Classification Accuracy, C3→CP1, δ-band Connectivity and μ/β-band Separability for different group splitting criteria. Blue-coloured cells indicate significant group differences.

Unit Scale	**Classification Accuracy**		**Connectivity - C3→CP1, δ-band**			**Separability**			
	MI	**FES**	**MI**	**Rest**	**FES**	**μ MI**	**β MI**	**μ FES**	**β FES**
	p-value	**p-value**	**p-value**	**p-value**	**p-value**	**p-value**	**p-value**	**p-value**	**p-value**
CRS-R $\leq$ 10 **vs** CRS-R $>$ 10	0.3131	0.0692	0.2878	0.3465	0.4053	0.1472	0.6507	0.9743	0.5469
CRS-R $\leq$ 12 **vs** CRS-R $>$ 12	0.2012	0.5302	0.0206	0.0205	0.0372	0.0607	0.3134	0.5848	0.0487
CRS-R $\leq$ 13 **vs** CRS-R $>$ 13	0.0268	0.6967	0.0125	0.0129	0.0263	0.0065	0.0449	0.5269	0.0945
Coma, UWS, MCS- **vs** MCS+, LIS, eMCS, Healthy	0.0261	0.8349	0.2578	0.1966	0.2769	0.0771	0.2930	0.6503	0.2334
Coma, UWS **vs** MCS-, MCS+, LIS, eMCS, Healthy	0.8320	0.3861	0.4615	0.4800	0.6260	0.1349	0.2859	0.7598	0.2500
Coma, UWS, MCS-, MCS+ **vs** LIS, eMCS, Healthy	0.219Although conventional4	0.6533	0.1194	0.1096	0.2413	0.0033	0.0470	0.5732	0.1417

Table 4.3 summarizes the results concerning the ability of different EEG metrics to statistically significantly separate between aware and unaware groups, as these can be ostensibly defined by means of different CRS-R thresholds or diagnosis-based groupings. As shown, the commonly proposed CRS-R threshold of 10 did not lead to statistically significant group differences for any metric. At threshold 12, significance emerged for MI, FES and Rest effective connectivity, as well as β-band separability under FES. Threshold 13 exhibited stronger and broader significance, including MI classification accuracy, μ

and β separability, though excluding FES β separability. In the group analysis using the neurological diagnosis grouping to distinguish between awareness and unawareness, MI classification accuracy was significant when MCS- and MCS+ were assigned to either groups. Conversely, both μ and β separability showed significant awareness discrimination ability when MCS- and MCS+ were grouped together.

Figure 4.8A compares the percentage of sessions in the dataset where the participant would have been diagnosed as exhibiting covert awareness (CMD) with different metrics and statistical approaches as explained in the Methods. Table 4.4 provides the respective information per session. It is evident that inferring CMD on the grounds of individual spatiospectral components found to be significantly different between MI and Rest through common statistical testing seems extremely optimistic, as 90-100% of sessions are classified as CMD, even after correction (Bonferroni or FDR), and after demanding a large range of clustered bands to be simultaneously significant. When requiring 7 Hz-wide clusters of significance, the percentage of CMD drops with this method drops to approximately 65%, which, however, may be inconsistent with the method's neurophysiological basis, as the range of μ- and β sub-bands usually modulated by the MI of able-bodied users is typically narrower [114]. The type of statistical correction used seems to have little effect on the diagnosis with this methodology.

Attempting to infer CMD by means of thresholding the total CRS-R score is more conservative, as it yields 70% CMD for the threshold of 10 examined in the literature [213] and approximately 40-60% for the thresholds of 12 and 13 that I found here to best split many EEG metrics in two groups. Moving to the EEG-derived metrics thresholded based on optimal two-group separation, the classification accuracy obtained from classifying MI against Rest is consistent with the CRS-R thresholds and detects 60% CMD in the dataset, while the equivalent for FES vs Rest is much lower, at 10%. This is the case for all other EEG measures, too, with β separability for MI slightly higher at 20%, suggesting that the thresholding method used here is rather conservative. The CBA statistical approach on

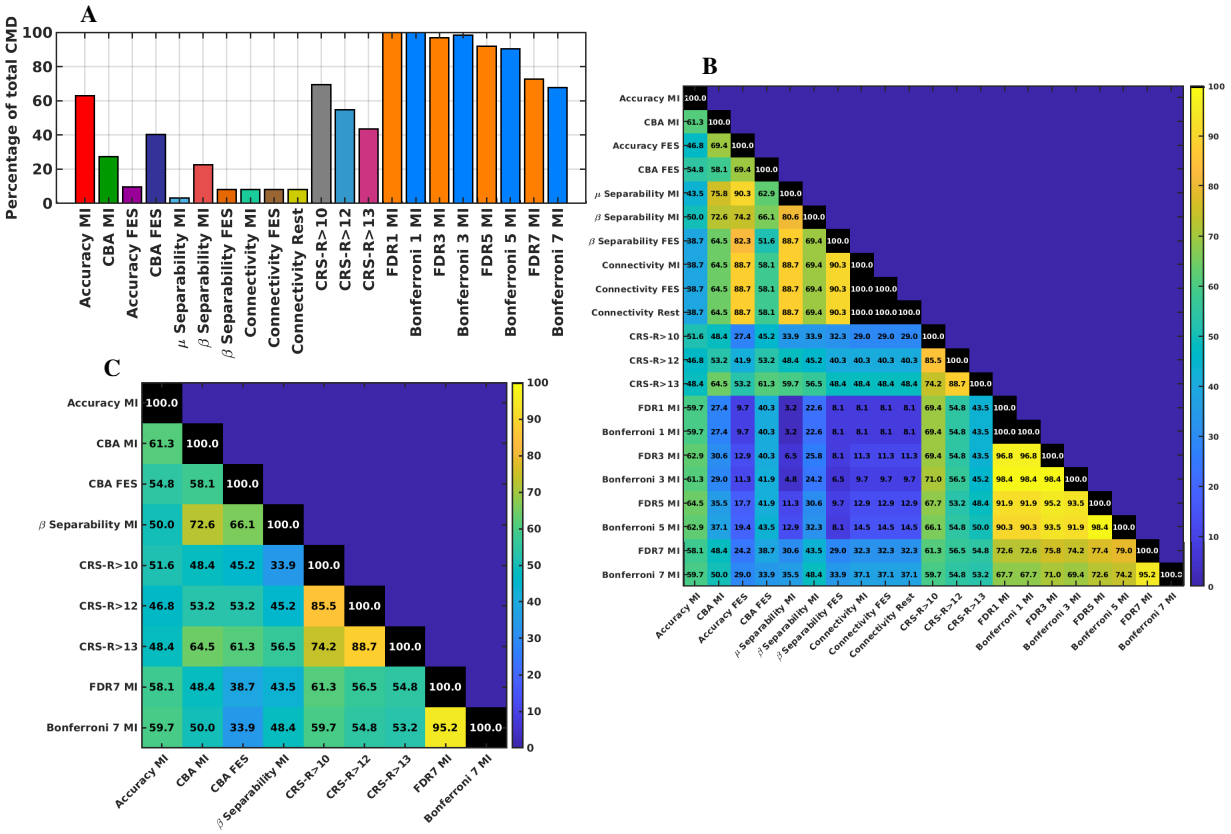

Figure 4.8: Diagnosis of CMD (covert awareness) by different metrics. (A) Percentage of sessions diagnosed as CMD by each metric. (B) Agreement rate heatmap for all metric pairs encoded as percentage of sessions with common CMD inference, as colour-coded by the colourbar. (C) Reduced version of the agreement rate heatmap including metrics exhibiting at least one agreement rate in the range [20, 80].

accuracy produces CMD detection estimates in the range 30-40%, thus seemingly being more conservative than CRS-R, but less than the approach of thresholding each EEG metric with the sample-wide group-separating thresholds.

Table 4.4: CMD diagnosis by metrics

Sub,	CRS-R	MI	CBA	FES	CBA	Separability			Connectivity			CRS-R			FDR				Bonferroni			
Ses	Diagnosis		MI		FES	μ MI	β MI	β FES	MI	FES	Rest	>10	>12	>13	1	3	5	7	1	3	5	7
S1,1	MCS-	X										X	X		X	X	X	X	X	X	X	X
S2,1	UWS	X			X		X								X	X	X	X	X	X	X	X
S3,1	UWS							X							X				X			
S3,2	MCS-								X	X	X	X			X	X	X	X	X	X	X	X
S3,3	MCS+											X	X	X	X	X	X	X	X	X	X	X
S4,1	MCS+				X										X	X	X	X	X	X	X	
S4,2	MCS+	X	X									X			X	X	X	X	X	X	X	X
S4,3	MCS+	X						X				X			X	X	X	X	X	X	X	X
S4,4	MCS-			X	X		X								X	X	X		X	X	X	
S4,5	MCS-						X								X	X	X	X	X	X	X	X
S4,6	MCS-	X	X												X	X	X	X	X	X	X	X
S4,7	MCS-	X			X		X								X	X	X		X	X	X	
S4,8	MCS+	X	X		X		X					X			X	X	X	X	X	X	X	X
S5,1	MCS+	X										X	X	X	X	X	X	X	X	X	X	X
S6,1	eMCS				X							X	X	X	X	X	X	X	X	X	X	
S7,1	UWS	X													X	X	X	X	X	X	X	X
S8,1	eMCS				X							X	X	X	X	X	X	X	X	X	X	X
S8,2	eMCS											X	X	X	X	X	X	X	X	X	X	X
S9,1	Healthy	X	X		X	X	X					X	X	X	X	X	X	X	X	X	X	X
S9,2	Healthy	X	X	X	X	X	X					X	X	X	X	X	X	X	X	X	X	X
S10,1	UWS	X	X				X								X	X	X	X	X	X	X	X
S10,2	MCS-	X													X	X	X	X	X	X	X	X
S10,3	MCS+	X	X									X			X	X	X	X	X	X	X	X
S10,4	MCS+	X			X		X					X	X	X	X	X	X	X	X	X	X	X
S10,5	MCS+	X			X							X	X	X	X	X	X	X	X	X	X	X
S11,1	UWS	X		X	X			X							X	X	X		X	X	X	
S11,2	MCS-	X										X			X	X	X		X	X		

Continued on next page

Table 4.4 (continued)

S12,1	MCS+	X	X		X		X					X	X	X	X	X	X	X	X	X	X	X
S13,1	MCS+	X		X	X				X	X	X				X	X	X	X	X	X	X	X
S13,2	MCS+											X	X		X	X	X		X	X	X	
S13,3	MCS+							X	X	X	X	X	X		X	X			X	X		
S14,1	MCS+	X	X	X	X							X	X	X	X	X	X	X	X	X	X	X
S14,2	MCS+	X										X	X	X	X	X	X		X	X	X	
S14,3	MCS+				X							X	X	X	X	X	X		X	X	X	
S14,4	eMCS	X	X									X	X	X	X	X	X	X	X	X	X	X
S15,1	eMCS						X					X	X	X	X	X	X	X	X	X	X	X
S15,2	eMCS	X	X									X	X	X	X	X	X	X	X	X	X	X
S16,1	eMCS	X			X							X	X	X	X	X	X	X	X	X	X	X
S17,1	MCS+	X	X									X	X	X	X	X	X	X	X	X	X	X
S17,2	MCS+	X										X	X	X	X	X	X	X	X	X	X	X
S18,1	eMCS	X					X					X	X	X	X	X	X	X	X	X	X	X
S18,2	eMCS						X					X	X	X	X	X	X	X	X	X	X	X
S19,1	UWS	X	X												X	X	X	X	X	X	X	X
S20,1	MCS+	X	X									X	X	X	X	X	X	X	X	X	X	X
S21,1	MCS-														X	X	X	X	X	X	X	X
S21,2	MCS-							X	X	X	X	X			X				X	X		
S22,1	LIS				X							X	X	X	X	X	X		X	X	X	
S22,2	LIS	X	X		X							X	X	X	X	X	X	X	X	X	X	
S22,3	LIS								X	X	X	X	X	X	X	X	X		X	X	X	
S23,1	UWS	X		X	X										X	X	X	X	X	X	X	X
S24,1	UWS	X			X										X	X			X	X		
S25,1	Coma														X	X	X	X	X	X	X	X
S26,1	UWS														X	X	X	X	X	X	X	X
S27,1	MCS+	X	X		X							X	X	X	X	X	X	X	X	X	X	X
S28,1	MCS+	X													X	X	X	X	X	X	X	X
S28,2	MCS-	X			X		X					X			X	X	X	X	X	X	X	X
S28,3	MCS+	X										X	X		X	X	X		X	X	X	
S28,4	MCS+	X	X		X							X	X	X	X	X			X	X		
S28,5	MCS-											X			X	X	X		X	X	X	
S28,6	MCS-											X	X		X	X	X	X	X	X	X	X

Continued on next page

Table 4.4 (continued)

S28,7	MCS-	X			X							X	X		X	X	X	X	X	X	X	X
S28,8	MCS-											X	X		X	X	X		X	X	X	

cell marked indicates metric diagnosed the session as aware and empty indicates otherwise, that is unaware. Also, Sub, Ses indicates Subject, Session, MI implies Classification Accuracy MI and FES indicates Classification Accuracy FES.

Figure 4.8B depicts a heatmap expressing the percentage of agreement between all pairs of the diagnostic metrics considered here. The high agreement rate between the CRS-R-thresholding methods, and those based on statistically testing individual PSD features and correcting with FDR or Bonferroni is anticipated, as these share the same underlying inference mechanism and only differ on the strictness imposed for detecting CMD (i.e., cut-off CRS-R value and number of consecutive bands required, respectively). Given that these feature-testing measures are the most optimistic ones, while, on the contrary, all EEG-based metric except for Accuracy MI and CBA are rather stringent and pessimistic (at least with the way thresholds have been defined), it is not surprising that these two clusters of approaches show no particular agreement. For the same reason, as the EEG-based metrics are shown to be very parsimonious in inferring CMD, the high agreement in-between these clusters most likely reflects the large amount of negative inferences in both cases, rather than a true agreement of these different aspects of EEG activity in detecting awareness. It is thus more meaningful to examine a smaller version of the heatmap involving only the metrics with comparable overall CMD inference rates in the range 20-80%, as shown in Figure 4.8C. Disregarding again the expected high agreement between pairs of congener metrics (CRS-R thresholding, 7 Hz-wide clusters of significant PSD features on the same channel after Bonferroni or FDR correction), the 32 remaining pairs exhibit a fairly narrowly distributed agreement around 50% (precisely, 52.9% $\pm$ 8.7%) with minimum at 33.9% (two pairs: CRS-R-10 with β separability MI, and CBA FES with Bonferroni PSD significance) and maximum at 72.6% between CBA and β separability MI. The overall

moderate agreement rate advocates in favour of the hypothesis that these metrics may express different facets of awareness and, in that regard, be complementary to each other.

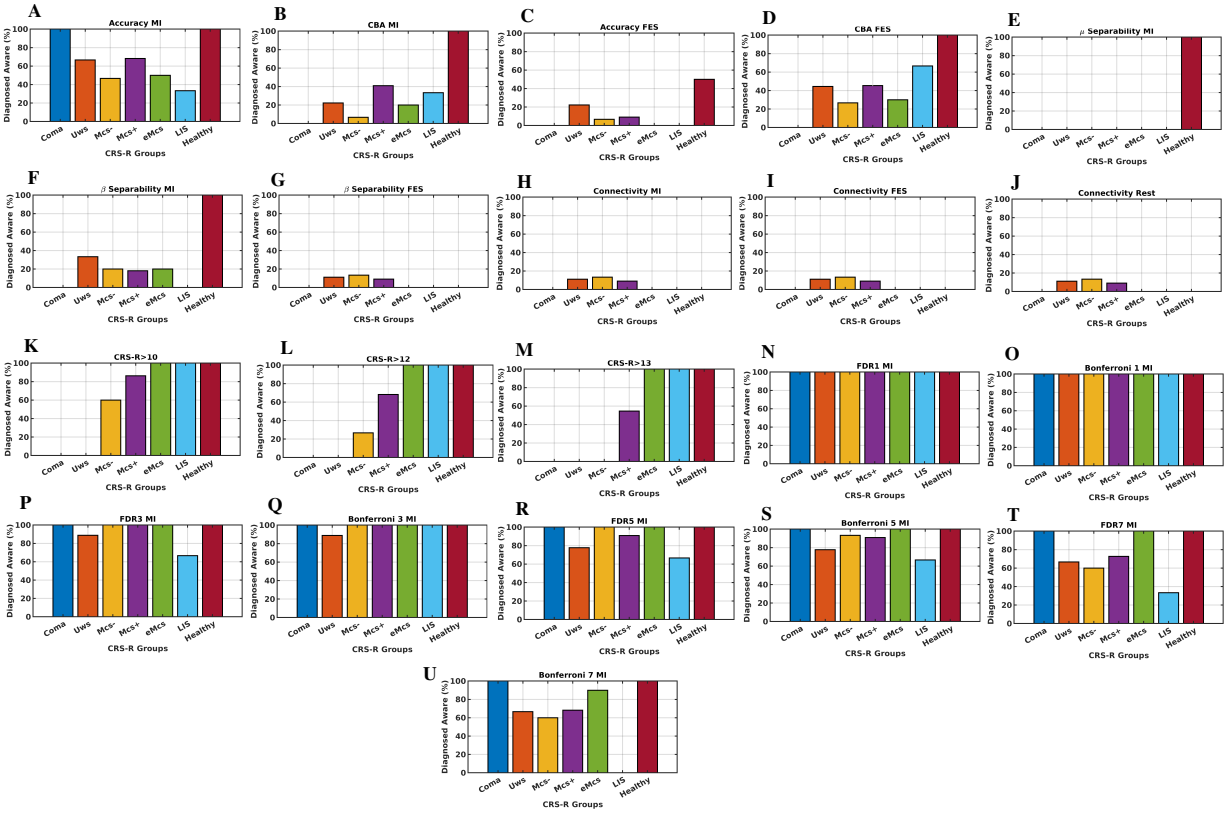

Figure 4.9: Percentage of predicted awareness across CRS-R clinical groups per metric - (A) Accuracy MI, (B) CBA MI, (C) Accuracy FES, (D) CBA FES, (E) μ separability MI, (F) β separability MI, (G) β separability FES, (H) Effective Connectivity MI, (I) Effective Connectivity FES, (J) Effective Connectivity Rest (K) CRS-R >10 (L) CRS-R >12 (M) CRS-R >13 (N) FDR1 MI (O) Bonferroni 1 MI (P) FDR3 MI (Q) Bonferroni 3 MI (R) FDR5 MI (S) Bonferroni 5 MI (T) FDR7 MI (U) Bonferroni 7 MI.

Figure 4.9 contrasts the CMD inference of all metrics considered with the clinical categorization of patients into seven main groups (coma, UWS, MCS-, MCS+, eMCS, LIS, healthy). The reader is reminded that this clinical diagnosis was performed by the clinical neurologists for each session on the basis, mainly, of the CRS-R scores, but took into

account also the MBT, the individual CRS-R subscales, and other information on the patient's medical record and physical condition. Using statistical significance of individual features with correction (Figures 4.9N-U) is shown again to be overly optimistic, likely leading to many false positives as the coma patient S25 and all, or nearly all, sessions thought to reflect UWS (S2, S3, S7, S10, S11, S19, S23, S24, S26) are found to be aware even with the strictest requirements. This method fails in a few cases to diagnose awareness in the patients believed to be LIS, thus being prone not only to many false positives, but simultaneously also to false negatives. Thresholding the total CRS-R score (Figures 4.9K-M) seems to strike an excellent balance, where the coma and UWS patients are diagnosed as unaware, the healthy control and the LIS patients as aware, and awareness is detected in the spectrum of MCS with increasing percentages, as reasonably expected. Of course, the performance of CRS-R thresholding here is biased, as the diagnosis was done mainly based on CRS-R to begin with. Connectivity, μ separability MI and β-separability FES are evidently again extremely conservative, raising further suspicions that the statistical procedure used for defining the thresholds may be too strict; however, β-separability MI (Figure 4.9F), a metric showcasing robust relation with CRS-R even when removing outliers (see Discussion), exhibits sound behaviour with the same thresholding approach. Specifically, it succeeds in diagnosing coma and most of the UWS cases, as well as to identify the healthy participant. However, it misses the LIS case and finds awareness in a few UWS patients (which, however, could also be an advantage, see Discussion). Overall, it is more conservative in the spectrum of MCS compared to CRS-R. It must be noted though that the latter effect does not necessarily constitute erroneous prediction: in the absence of ground truth and with the known CRS-R limitations, there is no way to conclude beyond doubt whether it is CRS-R that is overly optimistic, or separability that is too pessimistic (or, whether both are true to some extent). Classification accuracy for MI (Figure 4.9A) and FES (Figure 4.9C) are definitely far off being good indicators of awareness, for opposite reasons: accuracy MI is extremely optimistic detecting awareness for the coma patient and

most of the UWS cases, while accuracy FES lies at the opposite extremity misdiagnosing the LIS and one of the healthy control sessions, despite finding covert awareness in a few UWS patients. CBA MI and FES, on the other hand, emerge as the methods-of-choice, closely following the trends and logic of the CRS-R scores and the resulting diagnostic predictions. More elaborately, both these methods diagnose correctly the controls (coma, healthy), get the (believed to be) LIS patient right in about half of their sessions, find awareness in about 5-45% of the MCS cases (with CBA in the MI condition being the more conservative of the two) and in 20% (MI) or 40% (fes) of the labeled UWS sessions. Overall, CBA MI and FES, closely followed by β separability MI, are the metrics that can be distinguished for complying well with the CRS-R-based diagnostics.

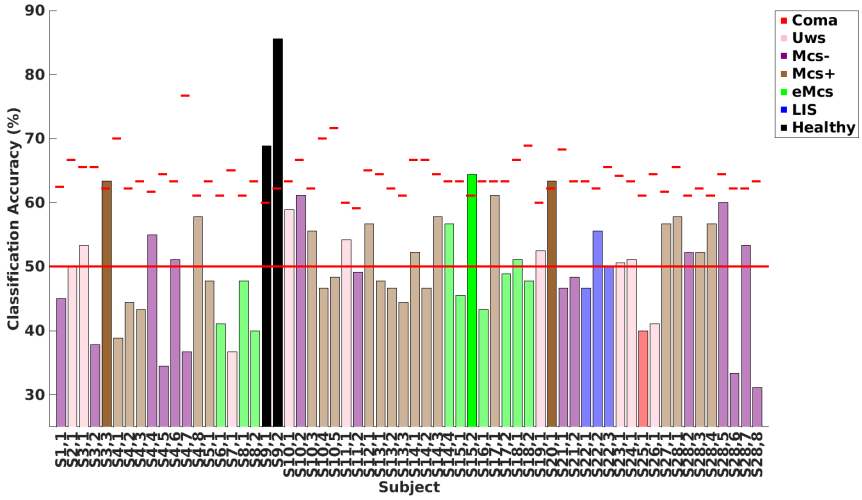

Figure 4.10: Classification Accuracy MI vs Rest with leave-one-trial-out cross-validation for all subject sessions extracting a single PSD sample from each 4 s trial. The horizontal red line illustrates the expected value of the random chance level for 2-class problems, while the session-wise lines indicate the permutation-based chance level with 99% confidence [6].

Additionally, analyses were repeated after excluding healthy control sessions, restricting the dataset to DOC patients only. This was done to assess robustness as discussed in section 4.5 to characterize variability in consciousness and awareness within the DOC

population, rather than to discriminate patients from healthy individuals, and to assess the impact of this restriction on both classification and regression analyses. The overarching aim of this supplementary analysis was to evaluate whether the correlation between a given EEG metric and CRS-R remains robust even when removing the extreme consciousness points. This approach tests the utility of the metric in capturing more subtle awareness differences (e.g., within the narrower UWS-MCS+ spectrum). As shown in Figure 4.11, Figure 4.12, and Tables 4.5 and 4.6, most previously observed significant effects diminish when healthy controls are removed, suggesting that the quest for better neuromarkers of awareness must continue.

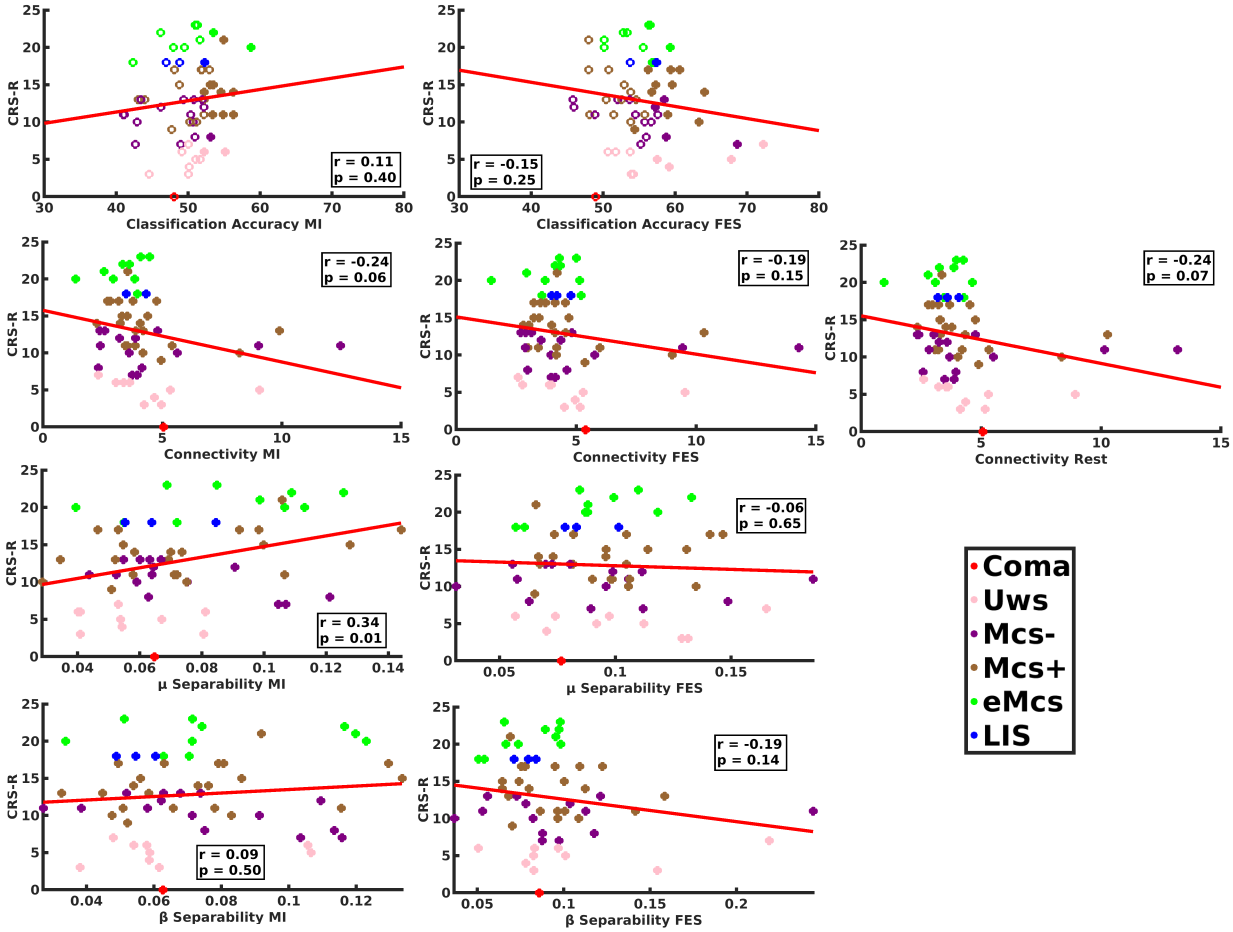

Figure 4.11: Correlation after excluding the healthy control between CRS-R and various metrics as shown by the panel titles using MI, FES, and Rest epochs. The CRS-R-based diagnosis of each subject-session is colour-coded as indicated by the legend. For the classification accuracy panels, subject-session points whose accuracy exceeds the chance threshold (see Figure 4.1) are represented with filled markers, while those below the threshold are unfilled.

Table 4.5: Overview of correlations with CRS-R and one-way ANOVAs with CRS-R group as factor after removing the healthy control. Blue-colored cells indicate significance.

Metric	**Correlation**		**ANOVA**	
	CRS-R		**CRS-R grouping**	
	p-value	**r-value**	**p-value**	**F statistic**
MI				
Classification Accuracy	0.4009	0.1104	0.1124	1.88
Effective Connectivity(C3→CP1, δ-band)	0.0598	-0.2445	0.8161	0.44
$\boldsymbol{\mu}$ Separability	0.0095	0.3322	0.2331	1.42
$\boldsymbol{\beta}$ Separability	0.4005	0.1105	0.7131	0.58
Slope	0.1743	0.1777	0.5927	0.75
Lempel-Ziv Complexity	0.8246	-0.0292	0.1639	1.65
Kolmogorov Complexity	0.6925	-0.0521	0.0928	2
FES				
Classification Accuracy	0.2542	-0.1495	0.5512	0.8
Effective Connectivity(C3→CP1, δ-band)	0.1504	-0.1879	0.9261	0.27
$\boldsymbol{\mu}$ Separability	0.6959	-0.0515	0.9022	0.31
$\boldsymbol{\beta}$ Separability	0.1392	-0.1932	0.6896	0.61
Slope	0.1605	0.1835	0.5789	0.77
Lempel-Ziv Complexity	0.7041	0.0500	0.2078	1.49
Kolmogorov Complexity	0.5543	0.0779	0.2062	1.5
Rest				
Effective Connectivity(C3→CP1, δ-band)	0.0676	-0.2376	0.7954	0.47
Slope	0.2625	0.1469	0.6750	0.63
Lempel-Ziv Complexity	0.7366	0.0443	0.2015	1.51
Kolmogorov Complexity	0.6408	0.0615	0.1793	1.59

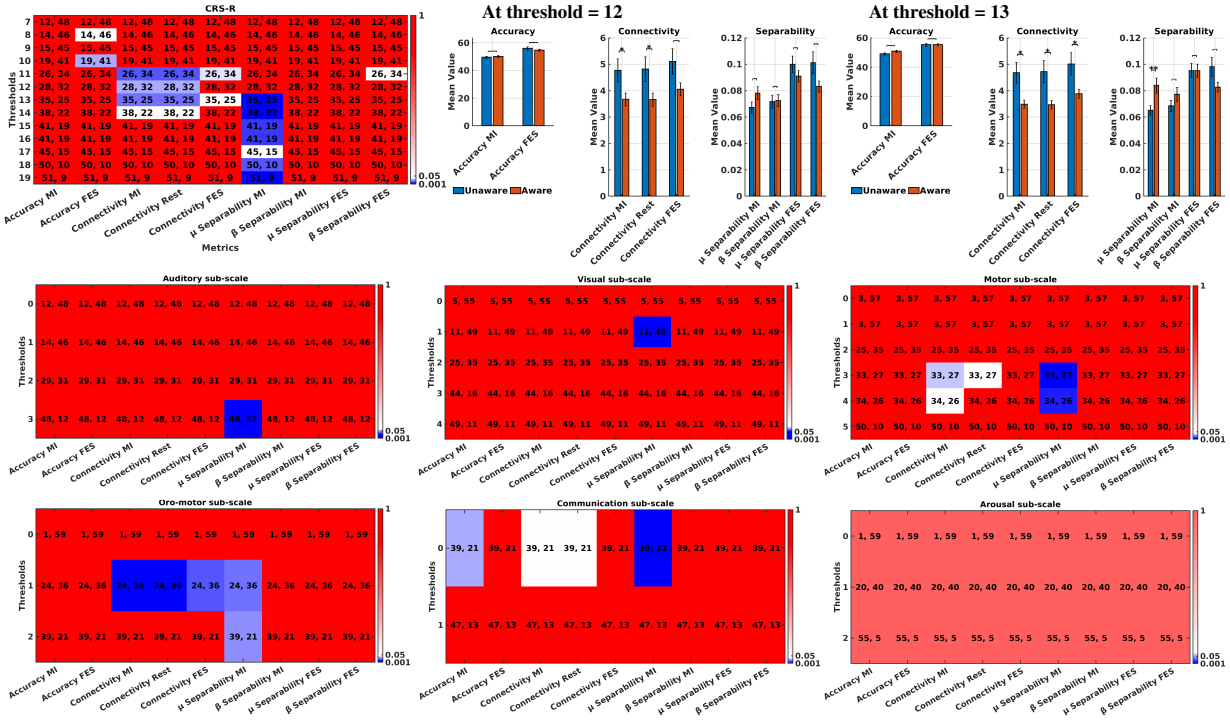

Figure 4.12: Metric and condition ability to discriminate between awareness and unawareness after excluding the healthy control. The heatmaps illustrate the p-values of unpaired, two-sided t-tests as coded in the colourbars, where red shades correspond to no statistically significant difference ($p \gg 0.05$), blue to statistically significant difference ($p \ll 0.05$) and white indicates marginal significance ($p \approx 0.05$). Each cell visualises the p-value of an unpaired, two-sided t-test of the hypothesis that the means of two groups of values of the metric/condition indicated by the cell's column label come from the same distribution. The metric's value derived by a single subject-session is assigned to group "unaware" if the corresponding session is associated with a CRS-R value $CRSR \leq th$ and to group "aware" otherwise ($CRSR > th$). The CRS-R threshold th used for each t-test is shown in the corresponding cell's row label. The numbers (X, Y) in each cell show the cardinality of each compared group pair for the particular test. The top-left panel splits groups according to the total CRS-R. The middle and right panel of the first row show the mean and standard deviation of the group values for each metric for CRS-R thresholds $th = 12$ and $th = 13$, respectively. Statistical significance is shown with asterisks (* for $\alpha = 0.05$, ** for $\alpha = 0.01$). The second and third row show heatmaps where group splitting relies on a CRS-R sub-scale as appears in the title.

Table 4.6: Statistical significance (unpaired, two-sided t-test) of differences between "unaware" and "aware" groups in terms of Classification Accuracy, C3→CP1, δ-band Connectivity and μ/β-band Separability for different group splitting criteria, after removing the healthy control from the dataset. Blue-coloured cells indicate significant group differences.

Unit Scale	**Classification Accuracy**		**Connectivity- C3→CP1, δ-band**			**Separability**			
	MI	**FES**	**MI**	**Rest**	**FES**	**μ MI**	**β MI**	**μ FES**	**β FES**
	p-value	**p-value**	**p-value**	**p-value**	**p-value**	**p-value**	**p-value**	**p-value**	**p-value**
CRS-R $<$ 10 **vs** CRS-R $\geq$ 10	0.4893	0.1186	0.7112	0.7608	0.8372	0.2792	0.7146	0.8635	0.3140
CRS-R $<$ 12 **vs** CRS-R $\geq$ 12	0.2465	0.3094	0.0119	0.0110	0.0245	0.0642	0.2592	0.6668	0.0454
CRS-R $<$ 13 **vs** CRS-R $\geq$ 13	0.2012	0.5302	0.0206	0.0205	0.0372	0.0607	0.3134	0.5848	0.0487

4.5 Discussion

The main purpose of this study has been to put to the test some of the most popular, state-of-the-art EEG-based metrics and statistical inference methods that have been proposed for the diagnosis of covert awareness in DOC patients, using a sizable, novel dataset collected with a standard command-following MI paradigm [168, 6]. Although, in the absence of factual, precise and reliable ground truth any relevant judgment remains rather inconclusive, by leveraging statistical criteria relying on the CRS-R clinical scale–widely acceptable as the closest surrogate to the unavailable ground truth–such as a metric's statistically significant correlation with CRS-R and the existence of CRS-R thresholds that lead to two statistically distinct and separable groups of metric values, my results indirectly provide evidence in support of the case that some of these approaches do indeed have diagnostic potential. More specifically, I found that classification accuracy during MI, effective connectivity in the δ band between scalp locations C3→CP1 during MI, Rest and potentially also FES, as well as μ- and β-band PSD separability during MI showed clear signs of association with CRS-R. On the contrary, a range of other metrics that have been proposed as possible neuromarkers of awareness and consciousness, such as Lempel-Ziv or Kolmogorov complexity, as well as the spectral slope did not pass the battery of tests imposed here in any of the examined task conditions (MI, FES, Rest). As complexity measures have been mainly implicated with awareness in the context of general anaesthesia and brain stimulation, the respective lack of association with CRS-R suggests that the neural underpinnings of awareness may in fact differ fundamentally between DOC and anaesthetized individuals, as well as between spontaneous and evoked brain activity. Overall, the merits of command following protocols and of several EEG-based metrics that can be derived on this basis in assessing awareness are largely confirmed by the data.

However, it needs to be underlined that the quality of the relevant evidence is often weak. For example, the correlation with CRS-R for several metrics becomes statistically

insignificant after removing the healthy control (Figure 4.11 and Table 4.5), who represents the high-consciousness extremity in the dataset. This indicates that currently used EEG metrics can be successful in capturing pronounced differences in awareness, but probably lack, in general, the sensitivity to reliably distinguish between more subtly different states within the spectrum of awareness. The separability of the μ-band rhythms between MI and Rest is the most robust metric in that regard, aligning with literature which posits that the separability of brain rhythms reflects cortical activation and may serve as a reliable indicator of awareness [215, 168]. Effective connectivity and β-band separability also remain fairly consistent when disregarding the able-bodied subject, as shown by comparing Table 4.3 to Table 4.6, Table 4.1 to Table 4.5, Figure 4.5 to Figure 4.11 and Figure 4.7 to Figure 4.12.

It is interesting to contrast the results extracted in the "active" MI epochs with those obtained from the "passive" FES intervals, as both these conditions are expected to elicit SMRs in cortical regions that are largely monitored by the same EEG channels, but differ significantly in the type of cognitive processing involved (motor planning and execution versus sensory stimulation) and, obviously, also with respect to the element of volition (active vs passive) which can be linked to awareness. The ensemble of the relevant results shows that, in contrast to the MI task, none of the salient EEG metrics could be associated with FES, with the marginal exception of β-band separability. Nevertheless, the classification during FES against Rest yielded subjects with above-chance accuracy (CBA FES), and, in fact, more than the equivalent metric during MI (CBA MI); overall, CBA FES was found to be similar (in spite of being less parsimonious) to CBA MI in detecting CMD for patients within the MCS spectrum. Both these methods were able to predict correctly all controls and showed meaningful trends, such as that more patients are found to be aware when the clinical diagnosis was MCS+ than when it was MCS- (Figures 4.9B and D). A possible explanation for the lack of direct association with CRS-R scoring is that the FES condition elicits separable, sensory-induced SMRs that, unlike in the active MI condition, may not

directly reflect conscious awareness, or only to a lesser degree (i.e., states of relative awareness with reduced cognitive capacity akin to the higher-degree motor dissociation defined by Edlow et al. [38]. Since the sensory SMR responses do not reflect intentional activity and volitional control, could potentially lead to overestimation of awareness [216]. This distinction aligns with the findings of Monti et al. [34], who emphasized that intentional cortical modulation provides a more reliable indicator of consciousness. My results further reinforce this view and support the utility of MI-based, command following paradigms in awareness detection as opposed to protocols relying entirely on evoked responses of the brain to stimulation. On the other hand, the sound and similar to CBA MI performance of CBA FES in inferring CMD across different CRS-R-based diagnostic groups, shows that passive protocols have their own merits, and are probably promising for diagnosing CMD in cases where the patient's spared cognitive abilities are compromised [59, 217].

In command following MI paradigms and, more generally, in BCI-inspired protocols for DOC diagnosis, classification accuracy has so far undoubtedly been the prevalent metric [60]. Often, thresholds defined on accuracy to detect awareness are arbitrary, or loosely justified from a statistical standpoint [134], which may lead to highly controversial conclusions [218]. Here, although I provide weak evidence that the raw classification accuracy may indeed carry information on awareness based on its correlation with CRS-R and other CRS-R-based statistical criteria (notably, the relation accuracy-CRS-R virtually disappeared after removing the healthy control from the dataset), I argue against absolute and universal accuracy thresholds. Instead, I recommend dataset/session-specific thresholds based on the statistical estimation of chance levels. In particular, I have used a method termed here as CBA which computes a chance level separately for each session at the 99% level of significance using repeated classification with random permutation of data labels, which has been found to be more reliable than accepting the hypothesis of a binomial distribution of classification decisions [219].

By means of CBA, I found that evaluating classification accuracy against chance-level

bounds offers a more parsimonious and better balanced (in terms of false positive and false negative errors) assessment compared to all other methods. Based on my results, CBA is probably able to prevent possible overestimation of CMD (false positives) that I show may be likely with the assessment of significance at the individual feature level, as sometimes proposed in the relevant literature [193]. In addition to this, CBA seems to be less conservative than most of the other EEG metrics examined (see Figure 4.8A and Figure 4.9), thus largely avoiding also false negatives.

Importantly, CBA is shown to be very robust when considering the (admittedly, few) control cases present in the dataset, the only instances where the available clinical diagnosis can actually be taken as the ground truth. In detail, the subject diagnosed with coma (S25), which is firmly determined to lack awareness beyond any doubt by the clinicians, produced classification accuracy below the CBA threshold during both MI and FES. Conversely, the healthy control (S9), as well as the participant thought to be LIS (S22), were always found to be aware. With the exception of β-band PSD separability that only fails to detect awareness in the ostensible LIS case, all other EEG-based metrics (including the accuracy estimates thresholded for optimal two-group separability rather than with the chance level, and the approaches using statistical significance of individual PSDs) failed this critical test. Furthermore, these metrics were shown to be either too conservative or too optimistic in finding covert awareness in the spectrum spanning UWS to eMCS. Overall, motivated by my results and similar recommendations in the literature [219, 38], I postulate that CBA-type of evaluation should be the standard statistical inference approach when BCI-inspired paradigms (command following or stimulation-based alike) are employed, which rely on classification accuracy as the main diagnostic metric.

However, even in this case, there is a need of standardization of the criteria for selecting methodological details and chance-level thresholds. For example, Edlow et al. [38] adopt an almost identical to CBA procedure, but differ in the definition of the accuracy value that represents the chance level at certain confidence. In this work, for a certain confidence

level α, I choose as session-wise chance-level accuracy threshold the corresponding percentile of the distribution of classification accuracy outcomes obtained with repetitive random permutation. This is illustrated in Figure 4.1 and Figure 4.10 as the session-specific limit denoted with red on top of each bar, where the bar height represents the actual accuracy estimate of a session. Specifically, I have used here the 99^{th} percentile of the random permutation distribution, corresponding to confidence $\alpha = 0.01$. I find that this approach and confidence threshold are both straightforward (in terms of their statistical interpretation) and, importantly, conservative enough to prevent over-optimistic CMD diagnosis. The CBA-equivalent of Edlow et al. [38] for the same confidence $\alpha = 0.01$ used here produces very similar, but not identical results: the control participants are identified correctly as aware (healthy, S9) or unaware (coma, S25), and 16 (instead of 17 with my own definition) out of 62 sessions are flagged as CMD. Out of these 16, 14 are the exact same ones diagnosed as CMD with my own method; two are not included by my own CBA MI, and one session is classified as aware by the CBA of Edlow et al. [38], but not by mine. For $\alpha = 0.05$, which is the less strict value reported in Edlow et al. [38], the thereby definition of CBA is again able to correctly infer both controls in the dataset (coma, healthy) and fully agrees with my own CBA MI regarding the 17 cases found to be CMD with my approach. However, it additionally finds latent awareness in another 7 sessions, for a less conservative 39% CMD detection rate compared to 27% here. While, again, there is no way to prove or disprove one or the other method or threshold, it becomes clear that there is great need in this field to reach a consensus and standardize both the general statistical framework and the confidence levels that should be used.

Figure 4.10 depicts the classification accuracy MI versus Rest when extracting a single PSD estimate from the 4 s MI epoch of each trial (i.e., without sliding, overlapping windows). It is easy to see that in this case only 4 sessions are found to exceed the chance level (though, all four are meaningful cases, corresponding to the healthy control, two MCS and one eMCS session). There are two reasons behind this effect. First, trial-wise classification

greatly reduces the classification problem's sample size, as a result of which the chance level estimates with 99% confidence become particularly high. Second, DOC patients, even when aware and vigilant, are shown to be largely unable to sustain MI for long periods of time, while the resulting SMR modulation is less consistent than that of healthy individuals; PSD estimation on the full trial length may thus average out short within-trial intervals of high-quality ERD/ERS, with other trial segments exhibiting no SMR modulation whatsoever, thus reducing the overall separability of the PSD features between MI and Rest and, therefore, also the classification accuracy. Based on these findings, it seems not advisable to extract EEG neuromarkers of awareness in intervals much greater than 1 s at least as far as command following protocols are concerned that require greater vigilance, attention and cognitive ability. On a crucial technical note, classification accuracy should still be evaluated with leave-one-trial-out, rather than leave-one-sample-out cross-validation (i.e., EEG feature vectors coming from close nearby–potentially overlapping–short windows must always remain in the same fold, since they are not independent), as otherwise there is great risk of inflated accuracy [220] and, therefore, also false positive CMD decisions.

A strong relation between an accuracy-based index and separability, as found in Figures 4.8 and 4.9 for CBA MI and both μ- and β-rhythm separability during MI, is by no means peculiar; in fact, it is anything but, as the classified features must be separable enough to be classifiable [114]. What is very interesting, however, is that separability aligns much more with CBA MI (μ 75.8%, β 72.6%) than it does with the absolute classification accuracy during MI (43.5% and 50.0%, respectively). In addition to this, CMD diagnosis was also very similar between these two approaches (Figure 4.9). This finding further substantiates the conclusions reached here, in my previous work and in the literature [168, 219, 6, 38] suggesting that separability indices and session-specific chance-level thresholding on accuracy are better indicators of covert awareness than other types of thresholding of the raw accuracy value. One reasonable explanation for the possible unsuitability of accuracy used as index of awareness without considering chance levels is that

it tends to saturate around the expected value (for 2-class problems) of 50%. DOC patients, including those thought to potentially have CMD and be able to follow commands, will rarely manage adequate SMR modulation so as to achieve accuracy levels well above the statistical chance level, as also found in the data (Figure 4.1A). Hence, thresholding must happen in the crowded region around and below the chance level (here, around 50%-60%) where the estimated accuracy is bound to be unreliable, especially given the small amount of data usually available. This problem is unknown in the main branch of BCI literature and practice concerned with assistive technology [114, 171], where patients are either able to achieve accuracy well-above the chance threshold or are not considered at all as users of the technology. In the DOC context, thresholding with the chance level threshold of the specific dataset at hand and using conservative confidence intervals (i.e., as with the CBA MI approach investigated here with formal 99% significance level) provides a statistically sound and more trustworthy way to define a meaningful, data-driven, session-specific threshold within this region, potentially addressing to some extent the high uncertainty of the diagnostic problem.

My analysis also identified CRS-R thresholds that seem to create two statistically significantly separable groups in many of the investigated EEG-based measures (Table 4.3). When this was the case, I took it as evidence of existing association between the respective metrics and CRS-R. A useful byproduct of this analysis is that it also elucidates the issue of finding optimal thresholds for detecting CMD by means of the traditional (purely clinical) CRS-R approach. Although a range of values emerged across different metrics, I noticed that the CRS-R threshold of 13 consistently appeared as a salient value. This observation contrasts with commonly reported thresholds in the literature, such as 10 [221] and 12 [3, 214]. Furthermore, it led to CMD predictions closer to those obtained with CBA MI than the other two thresholds. Although it is quite clear that there can be no definitive and universal threshold separating awareness from unawareness on the grounds of CRS-R alone, this observation could better inform clinical diagnosis where CRS-R scoring is part

of the evidence taken into account. Similarly, my findings imply the existence of a transitional awareness range between 12 and 14 in the CRS-R scale, thus highlighting a possible focus group for future research.

Another interesting result is that, when the group splitting was based on the neurologists' diagnosis rather than on numerical CRS-R thresholds (Table 4.3), the classification accuracy metric during MI became statistically significant only when the splitting criterion was set to be between MCS- and MCS+. In contrast, μ and β separability during MI were significant only when both MCS- and MCS+ were placed together in the "unaware"/non-CMD group. When comparing coma and UWS against all other diagnoses, no metric could be significantly split into two groups. Two conclusions can be drawn from these observations. First, it seems that the examined EEG-based metrics, accuracy included, may be in fact limited in their ability to detect what definitely is the most crucial transition for a patient, namely, the transition from UWS to the MCS. Consequently, this is another argument in favour of CBA, as CBA does not rely on any fixed threshold on accuracy, but on how the derived accuracy compares to the chance level in the same session. Second, it seems that different EEG metrics may in fact better specialise in different regions of the spectrum of awareness, thus showing complementarity to each other.

Sub-scale-specific analysis of the CRS-R revealed that the motor sub-scale correlated significantly with two metrics, the oro-motor sub-scale with six, and the communication sub-scale with five, whereas auditory, visual, and arousal sub-scales each correlated with three metric. Of note, when significant correlation was observed, it was for the same EEG metrics that correlated with the total CRS-R. The prominence of motor and language/communication-related sub-scales in my findings corroborates that EEG activity in DOC predominantly reflects the participant's engagement with the process of following commands, rather than the processing of the afferent signals caused by the stimuli present in my protocol (auditory cue stimulation, vibrotactile during FES). The relatively lower degree of association of the motor CRS-R sub-scale with EEG compared to the commu-

nication and oro-motor sub-scales can be explained by the fact that, during delivery of the CRS-R, the motor responses of the patient may be compromised, which is clearly the main motivation of resorting to EEG protocols in the first place. This further underscores the notion that at least certain crucial aspects of latent awareness may be better assessed by active, command-following paradigms, rather than passive ones.

The main motivation for adopting neuroimaging techniques in the assessment of awareness has been the educated belief that common clinical instruments such as GCS and CRS-R have a tendency towards Type II errors, particularly because they heavily rely on motor responses that may be impaired or completely absent in patients with DOC (the condition referred to as CMD). The hope is that motor intentions will be evident in the patient's brain activity even in the absence of overt motor output, as has been shown in relevant BCI studies, and thus prove that a patient is following commands; thus, they must be aware and conscious, at least to some extent. I argue that the fixation on coping with false negatives in this field may have encouraged a certain amount of neglect for Type I errors (false positives) and for the fact that any CMD inference method must carefully trade both types of errors off. This has also potentially led to a bias towards methods that may not be parsimonious enough in detecting awareness. For example, I have highlighted above how diagnosing CMD on the basis of single EEG feature values being significantly different between two conditions, as for example in the work of Curley et al. [193], may be overoptimistic (hence, likely commit several Type I/false positive errors) even after correcting for multiple comparisons or applying additional constraints (such as the requirement of several adjacent PSD bands being simultaneously significant on a particular EEG channel). Although some discount of false positives may be justified given that false negatives could have much more severe repercussions for patient care in this context (e.g., unfortunate decisions to remove life support on the grounds of undetected latent awareness), future EEG paradigms should be able to deal with both situations. One recommendation for assessing a method's performance on both types of errors is to include, not only healthy (false neg-

ative) controls, as is often the case [174], but also coma patients as definitive false positive controls in the cohort studied. Another fruitful avenue is to include qualitative, instead of only quantitative/statistical criteria when making inference. For example, a certain PSD feature found to be significantly different between hand MI and Rest must be also tested for neurophysiological relevance (i.e., whether it lies within the anticipated for MI μ or β frequency bands, and on scalp locations corresponding to the contra/ipsi-lateral to the MI employed regions of the sensorimotor cortex, see Figure. 4.3). A significant PSD feature with irrelevant to MI characteristics (e.g., 40 Hz on a peripheral channel) could very likely be the result of artifact contamination or represent some other kind of noise. This sort of tests can be automated [172].

Given the inevitable focus on avoiding false negatives, it is important to ask the question how the EEG metrics identified here as relevant perform in this domain. In other words, how successful they can be in identifying covert awareness in patients thought, based on the clinical team's opinion, to belong to some point in the spectrum of MCS. As shown by my analysis on Figure 4.9, most metrics seem to operate at either of two extremities: either they find CMD in most of the clinically assessed as MCS sessions, but at the great expense of pervasive false positives for nearly all coma and UWS ones; or, on the contrary, they are unable to detect latent awareness in MCS, while getting the coma and UWS right. It is thus worth to focus only on the four methods proven to strike a good balance between Type I and II errors and be able to perfectly classify both the aware and unaware controls: CRS-R thresholding, CBA MI and FES, and β separabiity during MI. To begin with, CRS-R thresholding with values in the region 10-13 is shown to be a tough competitor to beat for the EEG metrics, as it successfully reveals awareness in all eMCS sessions for all three thresholds, as also in 50-90% of cases in the MCS+ category, and many in the MCS- class, which is only missed for the most conservative CRS-R threshold of 13. CBA MI and FES, as well as β separability, apart from potentially misdiagnosing 20-40% of the UWS sessions, exhibit similar CMD detection rates with one another for all MCS levels, and these rates

are in most cases the same or lower than those obtained with directly thresholding the total CRS-R score. Hence, if the clinical diagnosis can be trusted, at a first glance, no superiority of the best EEG approaches over simple CRS-R thresholds can be established. Of course, it must be underlined again that the clinical diagnosis is mainly made on the basis of CRS-R, which therefore has an unfair advantage in these comparisons.

However, exactly because CRS-R (and, as a consequence, a CRS-R-based diagnosis) cannot be fully trusted, it is reasonable to assume that some of the sessions diagnosed as UWS might have in fact been CMD. There are, in particular, several interesting patient cases in the dataset with this profile (Table 4.4 and also Figures 4.1 and 4.5): S4, S10 and S19 for CBA MI, S2, S4, S11, S23, S24, S28 for CBA FES and S2, S4, S10 and S28 for β-band separability during MI. While for half of these eight patients (S2, S19, S23, S24) there is a single session marked by a disagreement between CRS-R and EEG with no way to tell which one is right, the remaining four patients (S4 MCS+, and S10, S11, S28 UWS) can be distinguished by the fact that they have performed several sessions, where only the EEG metrics are able to detect signs of awareness early on; only after one or more sessions (where, except for S4, there is an improving trend in recovering awareness) does the CRS-R scale manage to confirm CMD, already spotted by the EEG much earlier. It is therefore clear that the greatest added value of EEG-based diagnosis of DOC is indeed their ability to correct potential false negatives in the highly uncertain and contested region between UWS and MCS. Another important conclusion from this analysis is that repeated testing can be critical for diagnosing CMD, as also argued by Halder et al. [217]. In addition to this, it appears that different EEG metrics may act best in tandem in order to uncover CMD that has been missed by the clinical assessment.

4.6 Conclusion

The absence of a definitive ground truth in DOC research necessitates reliance on the CRS-R as the best available clinical reference. However, this reliance inevitably renders any derived findings open to debate, particularly given that the primary motivation of this field is to address the very limitations of the CRS-R, notably its susceptibility to type I and type II diagnostic errors. In the absence of an alternative, widely accepted standard—as also argued by [221]—it is currently more appropriate to focus on developing diagnostic frameworks that distinguish awareness from unawareness, rather than attempting to predict precise scores, spectral positions, or fine-grained consciousness states. Such predictive efforts are constrained by the scarcity of reliable ground truth and limited dataset availability. In this context, and contrary to the commonly adopted classification threshold of 10, this work proposes a more conservative threshold of 13, which demonstrated greater consistency across multiple independent metrics.

In further support of the diagnostic potential of EEG-based, BCI-inspired approaches, this study aligns with prior work [176, 60] in emphasizing the advantage of active paradigms for eliciting neural signatures associated with volitional engagement and cognitive processing. Nevertheless, the extraction of task-related epochs should be restricted to durations of approximately 1 s, reflecting probably the limited clinical capacity of DOC patients to sustain motor imagery beyond this timeframe.

Importantly, the findings of this work also suggest that the field's strong focus on mitigating type I errors in CRS-R-based diagnosis may have inadvertently reduced sensitivity to type II errors, which are equally detrimental in clinical decision-making. Existing neurophysiological metrics appear to specialize unevenly, with some favoring the detection of covert awareness (type II correction) and others emphasizing the exclusion of false positives (type I correction). As a result, these approaches remain limited in their ability to capture transitional states along the awareness continuum. To address this limitation, this

thesis introduces the CBA framework, conceptually aligned with [38], which avoids reliance on fixed thresholds and instead evaluates classified neural responses relative to empirically derived chance levels.

Taken together, the results presented here provide strong evidence that EEG signals encode meaningful markers of consciousness in DOC patients. Specifically, classification accuracy, separability in the μ and β frequency bands, and δ-band effective connectivity between central and parietal regions demonstrate notable concordance with CRS-R assessments. Among these, the proposed CBA framework offers a more robust and principled means of differentiating awareness from unawareness, particularly in the absence of an absolute diagnostic ground truth.

Chapter 5

Comparing Shallow vs Deep Architectures for EEG-based DOC decoding

Deep learning models have demonstrated remarkable success in fields such as computer vision, medical imaging, natural language processing, and other domains where the task involves identifying complex patterns, hierarchies, and subtle features within high-dimensional datasets. Their ability to automatically extract and represent features at multiple levels of abstraction has made them the dominant paradigm in these areas. However, can these advantages be directly translated to EEG-based BCI applications for the diagnosis of DOC? While DOC diagnosis similarly requires decoding intricate neural patterns, the data available in this domain are often limited in size, heterogeneous, and highly variable across sessions and patients. These constraints raise concerns about the feasibility and reliability of deep learning approaches in such clinical contexts. This chapter addresses this question by comparing/juxtaposing the shallow model employed in Chapter 4 against three state-of-the-art deep learning architectures—EEGNet, DeepConvNet, and EEGConformer—thereby enabling a systematic comparison of their relative strengths and

limitations in the context of DOC diagnosis.

5.1 Introduction

The real appeal of BCI lies in their ability to facilitate direct interaction with neural activity independent of muscular control. However, their rapid adoption—particularly in clinical contexts—has at times outpaced the establishment of methodological standards and validated protocols. Regardless of the application, whether for control, monitoring, gaming, or clinical diagnosis, the performance of the classifier is critical: reliable accuracy is indispensable for practical utility.

In recent years, the advent of deep learning has transformed classification across multiple fields, including computer vision, medical imaging, and natural language processing. Within the BCI domain, deep learning architectures began to emerge roughly a decade ago, gradually complementing and, in some instances, challenging the dominance of traditional shallow models. Despite promising advances, their impact in BCI applications has so far been moderate, constrained largely by data limitations and variability inherent to neuro-imaging.

In this chapter, I extend this line of inquiry by applying the state-of-the-art deep learning models with architecture designed specifically for EEG signals—EEGNet [7], DeepConvNet [8], and EEGConformer [9]—to the EEG DOC patient dataset used in this thesis. Their performance is systematically compared against the validated shallow model presented in Chapter 4, providing a direct comparison of their relative strengths and limitations in this challenging clinical setting.

Prior to conducting this comparative analysis, it was anticipated that deep learning models would struggle to achieve reliable performance with the present dataset, particularly under within-subject cross-validation schemes, given the limited number of trials per class (typically fewer than 100). Nevertheless, session-wise deep learning analyses were con-

ducted to enable a direct comparison with the corresponding shallow ML results presented in Chapter 4. Furthermore, EEGNet has been explicitly designed to perform robustly in low-data EEG settings [7], and both DeepConvNet and EEGConformer have been widely adopted in the literature, with several studies reporting strong performance even in constrained data scenarios [222, 223, 224, 225, 226, 227, 228, 229, 230, 231, 232, 233]. For these reason, their inclusion is essential to provide a comprehensive and contextually grounded evaluation.

Accordingly, model performance was evaluated using multiple cross-validation strategies, namely: (i) Leave-one-trial-out cross validation (LOTO), implemented in a session-wise manner with entire trials as inputs; (ii) LOSO, also trial-based, where all trials from one session were held out; and (iii) a leave-one-sample-out scheme, conducted session-wise using 1-s windows as inputs.

5.2 Methods

For the analysis presented in this chapter, the participants (subjects), experimental apparatus, and protocol are the same as described in Sections 3.2, 3.3, and 3.4, as well as in the introductory paragraph of Section 3.5. Consistent with the findings reported in Chapter 4, where classification of the FES condition against rest did not yield significant results, the present analysis is constrained to the MI vs Rest condition.

Following the artifact removal method described by [79] and the preproccessing method described in Section 3.5, each run was first cleaned, subjected to DC offset removal across all channels, and processed with a cross-neighbor Laplacian spatial filter to improve signal localization. In cases where neighboring channels were unavailable, the Laplacian derivation was computed using the remaining available neighbors.

Unlike shallow models, which rely on handcrafted features such as PSDs across frequency bands, deep learning models directly process the raw EEG signals. Accordingly,

the continuous EEG data were segmented into MI and rest epochs, which were then fed directly to the classifiers. Prior to classification, each epoch was normalized across all trials per channel using Z-score standardization, in line with the requirements of deep learning models.

For each input format, data splits were generated using precomputed leave-one-trial-out cross-validation schemes. In the context of this thesis, a trial refers to either (i) a 4-second window of a given task condition or (ii) an entire run (i.e., a continuous block of multiple command-following tasks). By contrast, a sample denotes a 1-second segment of a task condition, or a single run further subdivided into overlapping 1-second windows. This distinction is necessary because two complementary approaches—trial-level and sample-level analysis—were adopted in this chapter.

In the first approach, the entire 4-second trial was directly fed into the classification pipeline as a single input instance. In the second approach, each trial was decomposed into overlapping 1-second sliding windows, with a shift of 62.5 ms or 100 ms for 512 Hz and 500 Hz sampling rates, respectively. This yielded 32 samples per trial at 512 Hz and 50 samples per trial at 500 Hz. A trial in this context was therefore defined as the collection of all its constituent samples.

For each fold of cross-validation, one trial (i.e., all 32 or 50 samples, depending on the sampling rate) was held out as the test set, while the remaining trials were used for training. Predictions were obtained for all held-out samples and subsequently aggregated via a softmax operation to yield a trial-level prediction. Finally, accuracies were averaged across folds to compute the session-level accuracy. This approach is consistent with the methodology described in the second paragraph of 3.5 and produced the results reported in 4.4.

The EEG data were organized into an array of shape (N, C, T) where N represents the number of trials, C the number of channels, and T the number of time samples. To ensure compatibility with two-dimensional convolutional neural networks, the array was reshaped

to $(N, 1, C, T)$. This conforms to the expected input format of Conv2D layers, defined as $(batch, in-channels, height, width)$, analogous to image data. Here, the batch dimension corresponds to the number of trials (N), the height to the number of channels (C), and the width to the number of time samples (T). The additional singleton dimension (1) specifies the input channel, indicating that EEG represents a single modality, similar to a grayscale image. Each input sample was associated with its corresponding label, and the sampling frequency was also provided as model input.

The training optimization and hyperparameters was standardized across all three models. Model optimization was performed using the Adam optimizer with a learning rate of 0.0002 and exponential decay parameters β_1= 0.05 and β_2= 0.999. Training was conducted with a batch size of 64, while testing was performed using a batch size of 1. A dropout rate of 0.5 was applied to mitigate overfitting, and early stopping with checkpointing was employed to preserve the best-performing model weights. For reproducibility, random seeds were fixed across Python, NumPy, and PyTorch to control for sources of randomness such as weight initialization, data shuffling, and dropout masks, ensuring that experiments could be exactly replicated on the same hardware. In addition, CUDA Deep Neural Network (CuNN) was run in deterministic mode, and Compute Unified Device Architecture (CUDA) optimizations were enabled to enforce reproducible behavior. At the end of training, as expected, the number of trainable parameters varied substantially across architectures. EEGNetv4 was the most lightweight model, with 4,090 parameters at 500 Hz and 4,202 parameters at 512 Hz. By comparison, DeepConvNet was larger, comprising 562,927 trainable parameters, while the EEGConformer contained 225,642 parameters.

All experiments were executed on a single NVIDIA GeForce GTX 1050 Graphics Processing Unit (GPU) (driver version 577.57) using the CUDA backend for GPU-accelerated computations. Final model performance is reported as the mean classification accuracy across all cross-validation folds.

5.2.1 EEGNetv4

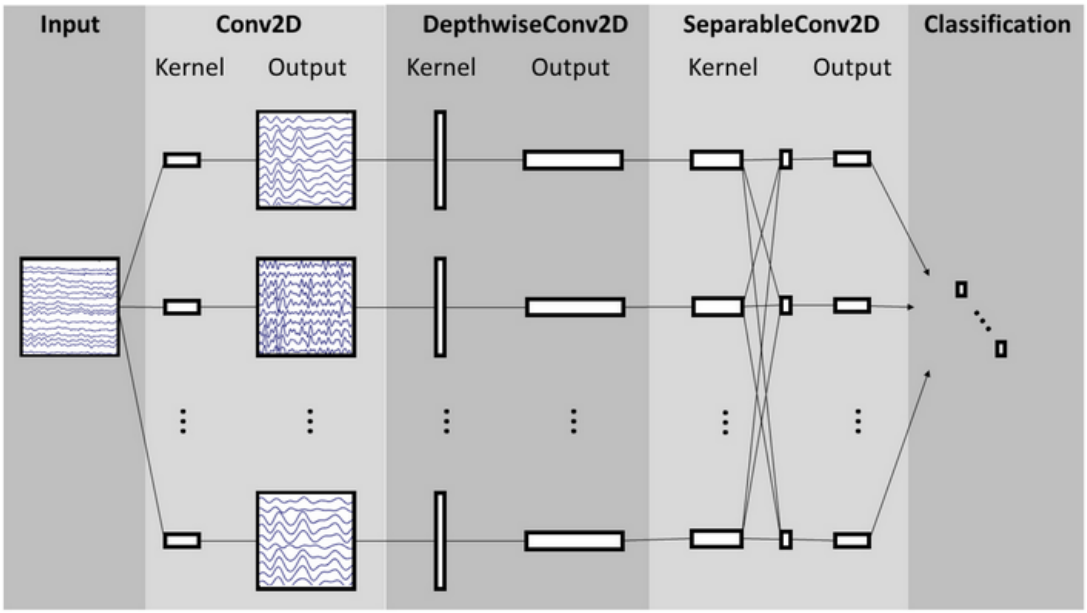

Figure 5.1: The EEGNetv4 architecture adapted from [7]

The EEGNetv4 pipeline used in this thesis is adapted from [7], and as illustrated in Figure 5.1, represents a compact Convolutional Neural Networks (CNN) architecture specifically designed for EEG data. Its structure incorporates both temporal and spatial convolutions (Conv2D and DepthwiseConv2D), conceptually comparable to conventional shallow EEG analysis. However, the features are learned automatically, and the network mitigates overfitting by constraining the number of trainable parameters.

The temporal convolution is defined as;

$$\text{Conv2d}(1 \to F_1,\ \text{kernel} = (1, L_t),\ \text{padding='same-ish'}) \tag{5.1}$$

This operation examines each channel independently over a temporal window of 4 seconds or 1 second (for sample/window-wise analysis), which is sufficiently large to capture oscillatory context while avoiding overly global representations. The convolution automatically learns temporal features across all frequency bands, eliminating the need for manual

band selection. Batch normalization is then applied to stabilize activations and accelerate training.

The depthwise spatial convolution across channels is formulated as:

$$\text{Conv2d}(F_1 \to F_1 \times D,\ \text{kernel} = (C, 1),\ \text{groups} = F_1) \tag{5.2}$$

This layer learns spatial patterns by combining electrodes for each temporal feature independently. Depthwise convolution ensures that each temporal feature has its own spatial filter, reducing parameter count and preventing unnecessary feature mixing. Batch normalization is again applied, followed by the Exponential Linear Unit (ELU) activation, chosen for its smooth nonlinearity and robust gradient propagation, which is particularly suitable for the small-amplitude activations characteristic of EEG. Temporal downsampling is performed via average pooling, which reduces computational load and provides invariance to small temporal shifts. Dropout is also applied to mitigate overfitting by randomly deactivating neurons during training.

The subsequent separable temporal convolution block further refines the learned features. A depthwise temporal convolution is applied:

$$\text{Conv2D}(F_1 \times D \to F_1 \times D, \text{kernel} = (1, L_{\text{sep}}), \text{groups} = F_1 \times D) \tag{5.3}$$

with kernel length approximately 0.125 seconds (i.e., $L_{\text{sep}} = \text{round}(0.125 \times \text{sfreq})$). This operation refines each temporal feature map independently, capturing short-range temporal dynamics efficiently without inter-channel mixing.

A pointwise (1×1) convolution follows:

$$\text{Conv2D}(F_1 \times D \to F_2), \quad F_2 = F_1 \times D \text{ (by default)} \tag{5.4}$$

which integrates information across feature maps and increases representational capacity. This is succeeded by the same sequence of operations as the first block: batch normalization, ELU activation, average pooling for temporal downsampling, and dropout regularization.

For classification, the resulting feature maps are flattened and passed through a linear layer, ensuring dimensional consistency after pooling and appropriate device placement to prevent CPU/GPU mismatches. Additional preprocessing safeguards include:

- Leave-One-trial-Out (LOTO) cross-validation, with metrics aggregated across sessions.
- Z-score normalization using training set statistics, transferred to the test set; degenerate standard deviations are clamped to maintain numerical stability.
- Temporal padding to ensure that the length of the temporal axis is at least `pool1 × pool2`, preventing underflow during pooling.
- Logging and skipping of invalid or empty folds to avoid propagation of NaN values.

Overall, a widely cited advantage of EEGNet is its strong performance in low-data regimes despite being a deep learning model. As originally demonstrated by Lawhern et al. [7], EEGNet achieves parameter efficiency through a compact architectural design with a markedly reduced number of trainable parameters, thereby mitigating overfitting. This characterization is further supported by Roy et al. [234], who position EEGNet as a benchmark architecture for training deep models on single-subject EEG data. Similarly, Schirrmeister et al. [8] report that EEGNet performs competitively, and often comparably, to more complex architectures when data availability is limited. Additional studies highlight its versatility [235], reproducibility, and suitability for transfer learning [236].

Nevertheless, EEGNet assumes relative signal stationarity within fixed temporal windows, which may limit its adaptability in dynamic or real-time applications [237, 238].. This limitation is not prohibitive in the context of the present work, as the primary objective is the development of accurate diagnostic BCI-inspired protocols for DOC, rather than real-time deployment. Furthermore, EEGNet has been shown to be sensitive to hyperparameter selection [239] and may exhibit reduced performance when applied to highly

complex or heterogeneous feature representations. To achieve optimal performance, prior work recommends rigorous preprocessing, including artifact rejection, band-pass filtering, and regularization strategies such as dropout (typically set to 0.5), weight decay, and cross-validation [7].

Conceptually, EEGNet closely mirrors traditional EEG signal processing pipelines: its learned temporal filters resemble band-pass filtering, depthwise spatial convolutions capture spatial patterns analogous to common spatial patterns (CSP), and separable temporal convolutions refine short-term temporal dynamics. This strong inductive bias, coupled with extreme parameter efficiency—often fewer than 5,000 trainable parameters per time window, compared to hundreds of thousands in many deep architectures—renders EEGNet a robust and practical baseline for deep learning in EEG research, particularly in data-constrained clinical settings.

5.2.2 Deep ConvNet

The DeepConvNet pipeline described in this thesis is adapted from [8] with minimal modifications. As illustrated in Figure 5.2, the architecture consists of four sequential convolutional blocks. The first block performs temporal convolutions to learn frequency-selective filters, followed by a spatial convolution that captures distributed spatial patterns across electrodes. Pooling is subsequently applied to reduce temporal resolution and normalize the resulting feature maps. Both the deep and shallow ConvNet variants automatically adapt their kernel and pooling sizes to the sampling frequency of the data—500 Hz and 512 Hz, respectively. Time-consistent "same-ish" padding was applied, and trials were right-padded as necessary to survive repeated pooling operations. For the DeepConvNet, temporal kernel and pooling sizes were scaled as follows:

- Temporal kernel size: $k\text{-}time = \text{round}(10 \times sf/250)$, yielding approximately 20–21 at 500–512 Hz. Odd values (≥ 3) were enforced to allow symmetric padding.

- Pooling size: $p\text{-}time = \max(2, \text{round}(3 \times sf/250))$, yielding approximately 6 at 500–512 Hz.

The four convolutional blocks are described as follows:

- **Block 1 (temporal + spatial)**: A temporal convolution with kernel $(1, k)$ learns frequency-selective filters, conceptually analogous to traditional band-pass filtering but optimized in a data-driven manner. This is followed by a spatial convolution with kernel $(C, 1)$, which captures spatial relationships across electrodes, functionally similar to CSP. Each convolutional stage is followed by batch normalization, an ELU activation, max pooling (kernel = $(1, p)$) for temporal downsampling and invariance, and dropout regularization to mitigate overfitting.

- **Blocks 2–4**: Each subsequent block consists of a temporal convolution (kernel = $(1, k)$), followed by batch normalization, ELU activation, max pooling, and dropout. The number of channels increases progressively across blocks ($50 \rightarrow 100 \rightarrow 200$), expanding the network's representational capacity with depth.

- **Classifier**: The final feature maps are flattened and passed through a linear classification layer mapping to two classes.

In this architecture, temporal convolutions function as learned band-pass filters, while spatial convolutions extract physiologically meaningful montages through contrastive weighting of electrodes. The combination of ELU activations with batch normalization stabilizes training on low signal-to-noise EEG data. Max pooling reduces the sequence length and provides temporal invariance; however, because four successive pooling layers are applied, the temporal axis must remain at least p^4 in length. Accordingly, trials were padded during preprocessing to prevent underflow.

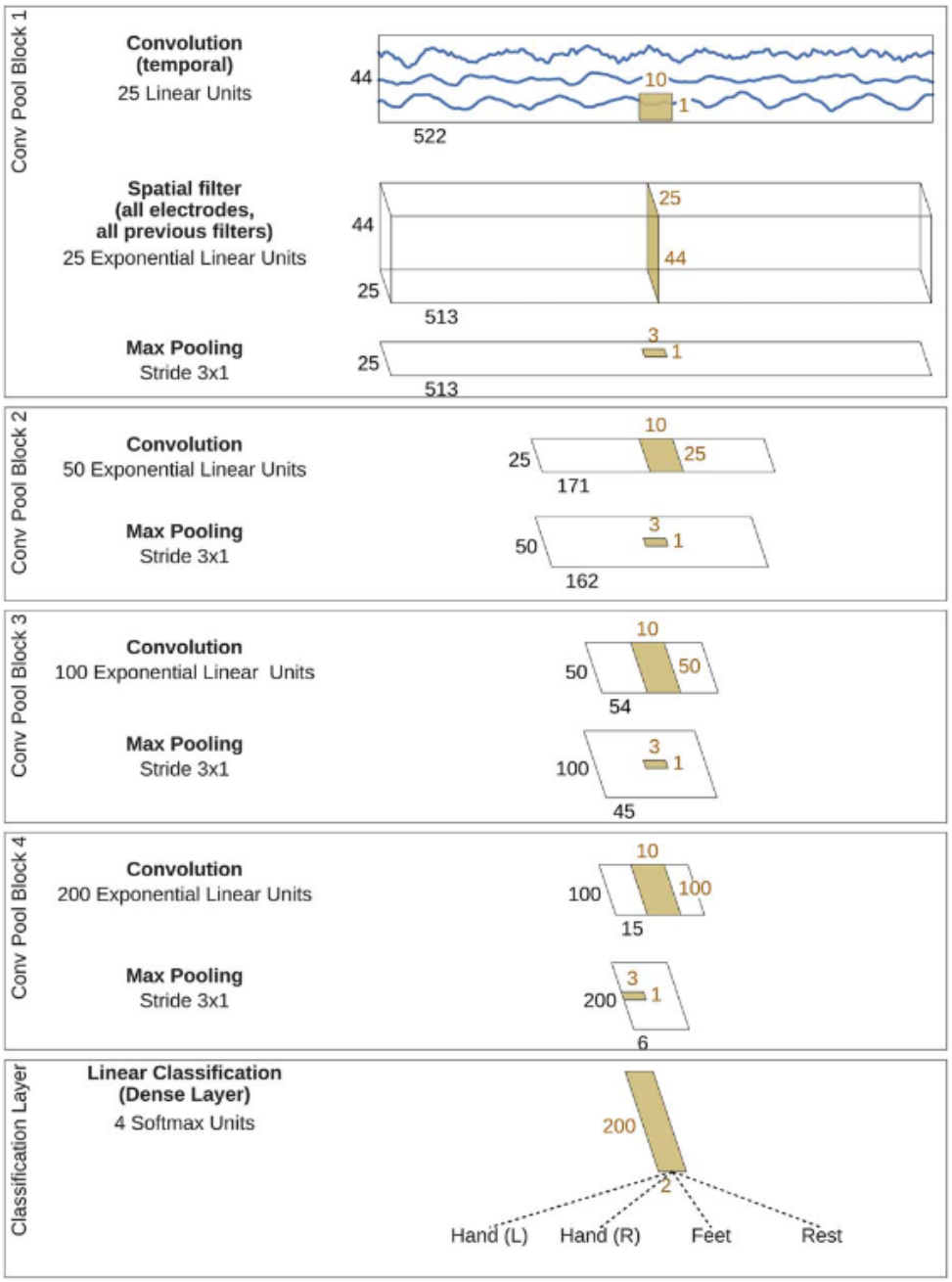

Figure 5.2: The DeepConvNet architecture adapted from [8]

DeepConvNet is generally more susceptible to overfitting than EEGNet when applied to small datasets [7, 234], primarily due to its substantially larger parameter count [8]. Its effective performance is therefore strongly contingent on the use of robust regular-

ization strategies, including batch normalization, weight decay, dropout, careful hyper-parameter optimization, transfer learning, and appropriate within-subject or cross-subject cross-validation schemes [239]. Despite these challenges, DeepConvNet represents one of the earliest architectures to demonstrate end-to-end deep learning directly on raw EEG signals [235] and has served as a benchmark model in studies involving larger datasets.

Prior work suggests that when the number of trials per subject is limited (e.g., fewer than approximately 100 trials), EEGNet may provide a more appropriate and stable alternative [7, 234]. In contrast, DeepConvNet, with an approximate parameter count of 570,000 and comparatively weaker inductive bias, is better suited to scenarios where sufficient data are available to support effective generalization. However, several studies [222, 231] continue to employ DeepConvNet even in small-sample settings, typically in conjunction with extensive regularization and validation strategies.

5.2.3 EEGConformer

Based on the description in [9], the EEGConformer architecture was adapted for this work with minimal modifications. As illustrated in Figure 5.3, the networkk consists of a convolutional front-end, followed by self-attention layers, and concludes with a classification module. Short 1-D or 2-D convolutions were applied over time-and in some cases, across channels-to extract local patterns, such as transient rhythms or event-related deflections. These convolutional layers act as learned band-pass/time–frequency filters, generating compact feature maps that serve as input to subsequent attention modules.

Each Conformer block integrates multiple complementary components which are summarized as follows:

- **Feed-Forward Network (FFN) with residual connection:** Expands and reprojects features, enhancing the network's representational capacity.
- **Multi-Head Self-Attention (MHSA):** Captures global temporal dependencies across

the entire trial, allowing the network to model delayed or phase-related interactions. The multi-head design enables simultaneous attention to multiple temporal scales and patterns.

- **Depthwise Separable Convolution Module:** Captures local time–frequency structure and short-range temporal continuity that attention mechanisms alone may overlook.
- **Layer Normalization and Residual Connections:** Stabilize training, preserve gradient flow, and facilitate the construction of deeper networks.
- **Attention-Convolution Synergy:** By combining MHSA and convolutional modules, the network simultaneously models long-range temporal dependencies and local EEG motifs, reflecting the dual temporal structure characteristic of EEG signals.

The output sequence from the final Conformer block is pooled and flattened before being passed through a linear layer mapping to two classes. This operation transforms sequence-level representations into trial-level logits suitable for classification. Convolutions in both the front-end and depthwise modules capture localized EEG features, such as ERD/ERS rhythms, while attention layers integrate information across the entire trial. Residual connections and layer normalization further enhance stability, gradient propagation, and generalization, making EEGConformer particularly well-suited for low signal-to-noise EEG data.

EEGConformer constitutes a substantial methodological progression in EEG decoding, reflecting a shift from exclusively convolutional architectures toward hybrid models that integrate neuro-inspired feature extraction with transformer-based attention mechanisms [9]. This design alleviates the restricted receptive fields inherent to conventional convolutional networks and enables the modeling of long-range temporal dependencies in neural signals [9]. At the same time, the reduced inductive bias of transformer components increases the model's flexibility, but also renders it considerably more data-intensive.

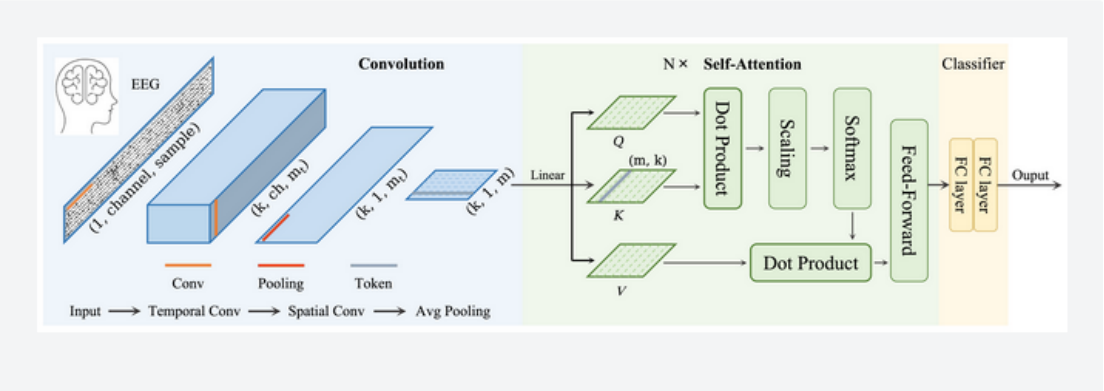

Figure 5.3: The EEGConformer architecture adapted from [9]

The architecture has attracted considerable attention due to its strong theoretical foundations, competitive performance across a range of EEG decoding benchmarks, and the introduction of attention-driven representations into BCI research. Nevertheless, EEGConformer is computationally demanding, typically comprising between 0.5 and 1 million trainable parameters [9], and is therefore more susceptible to overfitting, particularly in limited-data settings [236]. Its performance is also highly dependent on careful hyperparameter tuning. Despite these challenges, studies have shown that, when properly trained, EEGConformer can achieve state-of-the-art accuracy in major EEG decoding tasks [9]. Moreover, its compatibility with emerging self-supervised and large-scale pretraining paradigms has further increased its adoption [136], even in studies with relatively modest sample sizes [223, 224, 227].

Consequently, although EEGConformer offers substantial representational power, its optimal use lies in specialized applications—ideally leveraging transfer learning—where capturing global temporal structure and complex interdependencies is essential.

In summary, the table below provides a comparative overview of the three models.

Table 5.1: Comparison of EEGNet, DeepConvNet, and EEGConformer architectures for EEG-based decoding.

Feature	EEGNet [7]	DeepConvNet [8]	EEGConformer [9]
Core Design Philosophy	**Domain-optimised compact CNN.** Uses depthwise and separable convolutions to hardcode efficient EEG-specific feature (temporal → spatial) processing [7].	**Generic deep CNN.** A standard deep learning stack applied to EEG for hierarchical feature learning [8].	**Hybrid CNN–Transformer.** Combines CNN local feature extraction with transformer global self-attention for long-range context modelling [9].
Parameter Efficiency	**Extremely high** (approximately 2K–5K parameters)[7]. The key strength. Minimal trainable weights.	**Very low** (approximately 570K parameters). Large, dense layers with many trainable weights [8].	**Moderate to low** (approximately 500K–1M parameters) [9]. More efficient than DeepConvNet but larger than EEGNet. Complexity depends on transformer depth.
Inductive Bias	**Strong and explicit.** Architecture mirrors standard EEG analysis pipelines (temporal filtering → spatial filtering) [7].	**Weak and generic.** Learns hierarchical representations but must discover EEG-specific structure from data [8].	**Hybrid.** Strong local bias from CNN front-end; weak global bias from transformer attention mechanisms [9].
Performance with small datasets (<100 trials per class)	**Excellent.** Default recommendation. Low parameter count naturally resists overfitting and works well with basic regularization [7].	**Poor (unless heavily engineered).** Highly prone to overfitting [7]. Requires aggressive regularization, data augmentation, or pre-training [239].	**Poor from scratch, excellent with pre-training.** Transformer components are data-hungry and require large-scale pre-training followed by fine-tuning [9, 236].

Continued on next page

Feature	EEGNet [7]	DeepConvNet [8]	EEGConformer [9]
Key Strengths	1. Robust benchmark baseline 2. High efficiency and training stability 3. Generalizable across EEG paradigms 4. Well-suited for transfer learning [7, 240, 236]	1. High representational capacity 2. Strong performance when data is abundant 3. Serves as a foundational proof-of-concept for end-to-end deep EEG 4. Early-layer filters are interpretable [8, 235]	1. Superior performance on complex tasks 2. Ability to model long-range temporal dependencies 3. Attention mechanisms provide enhanced interpretability 4. Ideal for pre-training and self-supervised learning paradigms [9, 43, 236]
Primary Criticisms	1. Limited capacity for very complex/long-range features 2. Can be simplistic for tasks needing deep hierarchy [67]	1. Severely overfits with small data-hungry 2. Computationally expensive 3. Parameter-inefficient for EEG-specific 4. Deep layers are less interpretable [7, 234]	1. Computationally heavy to train 2. Requires large-scale pre-training for effectiveness 3. Complex hyperparameter tuning 4. Attention maps are not automatically clear []
Interpretability	**Good (first layers).** Temporal kernels = bandpass filters; spatial kernels = scalp topographies. Deeper layer fusion is less clear [7].	**Good (first layers).** Similar to EEGNet for early layers. Higher-layer features become abstract and hard to map [8].	**Novel but complex.** Attention weights show which time points interact, potentially revealing connectivity, but require careful validation [9]
Ideal use cases	1. Standard baseline for any new task 2. Studies with limited data 3. Rapid prototyping and transfer learning [234, 240]	1. Offline analysis of large-scale EEG datasets 2. Research on advanced regularization techniques 3. As a pre-trained feature extractor (transfer learning) [8, 239]	1. Complex tasks with long-range dependencies (sleep staging, seizure prediction, cognitive state) 2. Downstream tasks with access to large pre-training 3. When attention-based interpretability is a goal [9]

Continued on next page

Feature	EEGNet [7]	DeepConvNet [8]	EEGConformer [9]
Recommended	You need a reliable, efficient, and reproducible baseline that works "out of the box" with limited data	You have a very large dataset or are explicitly studying how to regularize deep CNNs for EEG	You have resources for pre-training/fine-tuning and your task explicitly benefits from global context modeling

The cross-validation accuracies obtained from three state-of-the-art deep learning models were compared against the PSD-based LDA classifier introduced in Chapter 4. Cross-validation was conducted at two levels of data partitioning: (i) **trial-wise**, in which entire trials were excluded from training, and (ii) **sample-wise**, where individual temporal windows within trials were held out.

5.3 Results

Figure 5.4 shows a 2-class classification accuracy for MI vs Rest across all subjects using full trials, comparing three state-of-the-art deep learning models (EEGNet, DeepConvNet, and EEGConformer) with a shallow model. All three deep learning models achieved accuracies above the 50% random chance level. However, when assessed against the stricter binomial threshold, only 29 sessions (47%) with EEGNet exceeded this level, compared to all 62 sessions (100%) with both DeepConvNet and EEGConformer. In contrast, the shallow model surpassed the permutation-based chance level in only 5 sessions (8.1%). Notably, coma subject $S25, 1$ exceeded the random chance bound with EEGNet but not the binomial threshold, while both DeepConvNet and EEGConformer classified this subject above both thresholds. Similarly, the healthy subject $S9, 1$ and $S9, 2$ failed to exceed the binomial threshold with EEGNet but did so with the other two deep models, suggesting less consistent performance with EEGNet. Interestingly, several DOC patients outperformed the healthy control when analyzed with the deep learning models—a characteristic absent in the shallow model.

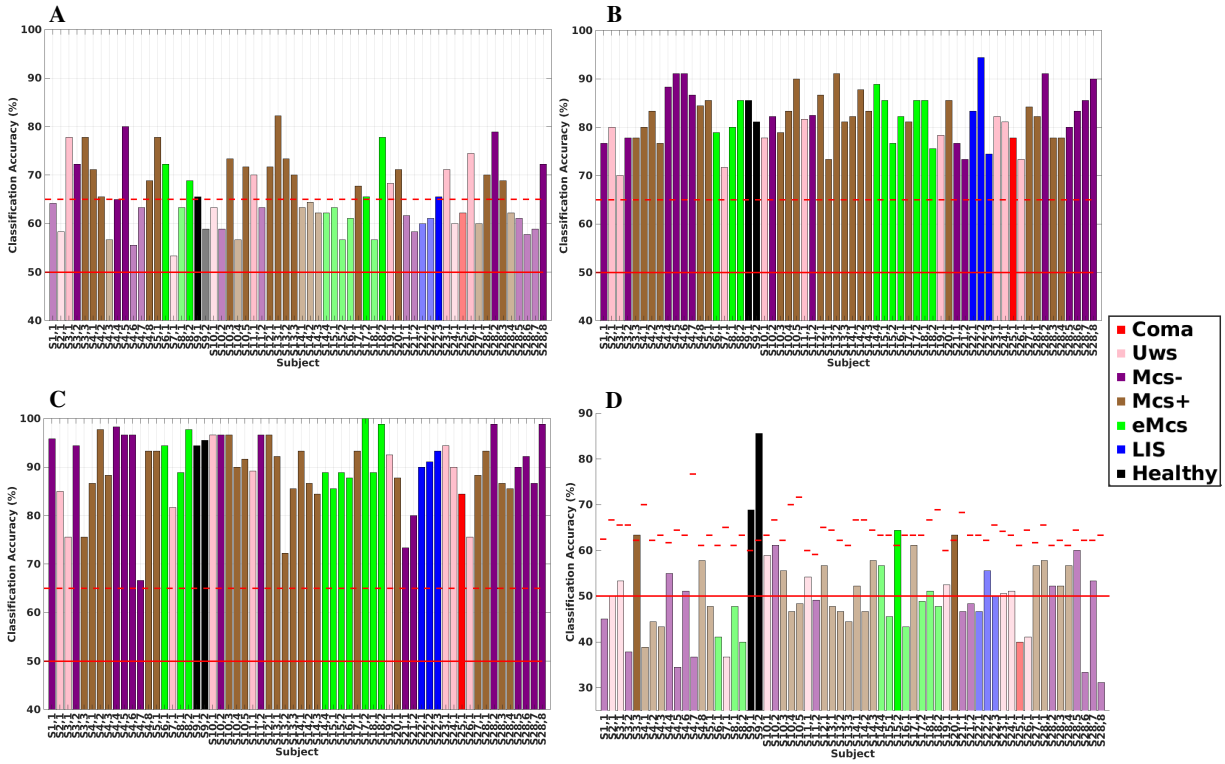

Figure 5.4: Classification Accuracy MI vs Rest with leave-one-trial-out cross-validation for all subject sessions with full trials (A) EEGNet (B) DeepConvNet (C) EEGConformer (D) Shallow Model. The horizontal red line illustrates the expected value of the random chance level for 2-class problems, the horizontal dashed line indicates a 45 trials/2-class BCI binomial chance level with 99% confidence [10] while the session-wise lines in (D) indicate the permutation-based chance level with 99% confidence [6]

Figure 5.5 presents the results obtained using windowed segments of the EEG signals, which were processed by the deep learning models and, in the case of the shallow model, as PSD features. With this approach, 55 sessions (89%) exceeded the binomial chance level using EEGNet, 51 sessions (82%) with DeepConvNet, and all 62 sessions (100%) with EEGConformer. In sharp contrast, only 17 sessions (27%) surpassed the permutation based threshold with the shallow model. Unlike the full-trial analysis, where the coma subject failed to exceed the binomial threshold under EEGNet, the windowed-segment approach

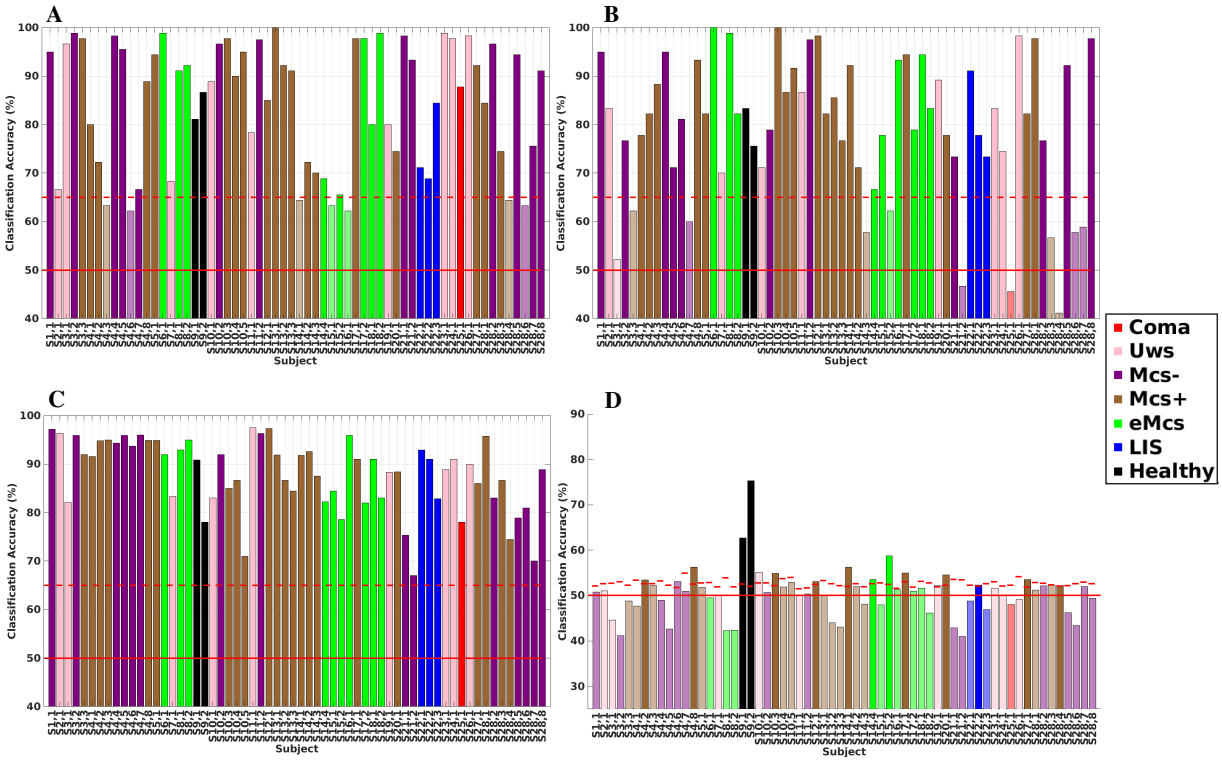

Figure 5.5: Classification Accuracy MI vs Rest with leave-one-trial-out cross-validation for all subject sessions with windowed smaples (A) EEGNet (B) DeepConvNet (C) EEGConformer (D) Shallow Model. The horizontal red line illustrates the expected value of the random chance level for 2-class problems, the horizontal dashed line indicates a 45 trials/2-class BCI binomial chance level with 99% confidence [10] while the session-wise lines in (D) indicate the permutation-based chance level with 99% confidence [6]

revealed the opposite pattern: the subject failed only with DeepConvNet. Furthermore, several DOC patients were classified with higher accuracy than healthy control by the deep models. Notably, DeepConvNet achieved both levels of control but also classified a number of DOC patients more accurately than healthy subject. This inconsistency was absent in the shallow model, which consistently classified the coma subject below both random and permutation-based thresholds, while placing the healthy subject above the DOC group.

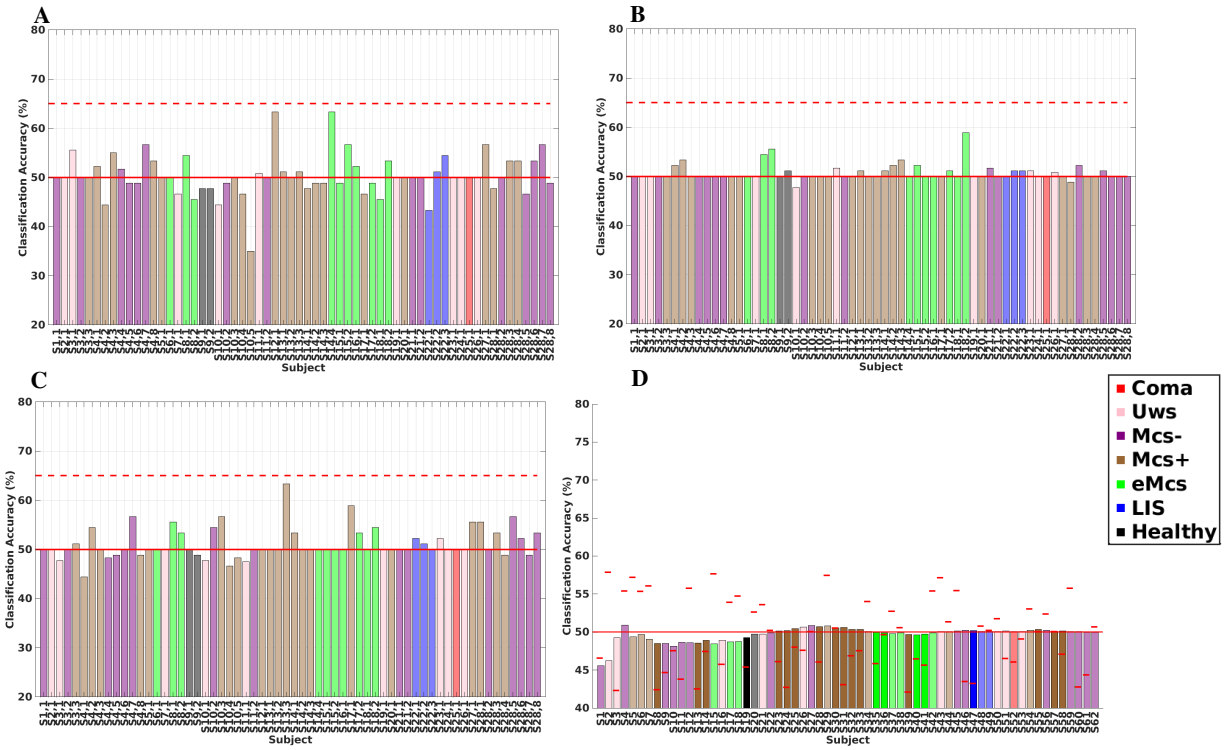

Figure 5.6: Classification Accuracy MI vs Rest with leave-one-session-out cross-validation for all subject sessions with full sessional runs (A) EEGNet (B) DeepConvNet (C) EEG-Conformer (D) Shallow Model. The horizontal red line illustrates the expected value of the random chance level for 2-class problems, the horizontal dashed line indicates a 45 trials/2-class BCI binomial chance level with 99% confidence [10] while the session-wise lines in (D) indicate the permutation-based chance level with 99% confidence [6]

Figure 5.6 presents the classification results obtained using a leave-one-session-out cross-

validation scheme. This strategy was adopted under the hypothesis that training on all remaining sessions would provide sufficient data for the models to generalise to the held-out session. However, across all three deep learning architectures—and, as expected, the shallow classifier—no session achieved performance exceeding the corresponding random or binomial chance-level thresholds. Consequently, classification performance across all sessions is consistent with chance-level prediction.

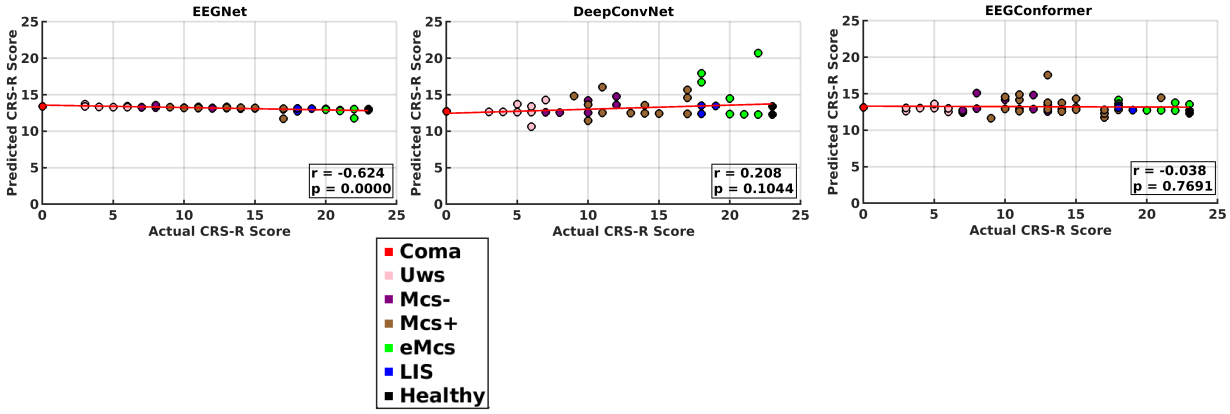

Figure 5.7: Regression predictive performance of CRS-R with classification accuracies of the models as the metric.

To further investigate model behaviour, Figure 5.7 presents the regression performance of these models under the leave-one-session-out cross-validation scheme, using classification accuracy as the predictive variable. Consistent with the classification results, the regression outputs exhibit substantial variability and appear largely random. Notably, the coma control (red marker) is consistently assigned predicted values above the threshold of 10, thereby being incorrectly classified as aware, while the healthy control is repeatedly predicted at levels comparable to the coma patient. These patterns further underscore the lack of reliable generalisation under this validation framework.

5.4 Discussion

The sole aim of this chapter has been to compare state-of-the-art deep learning pipelines with a classical shallow model, using the sizeable novel dataset employed in this thesis and derived by a standard command-following MI paradigm [6, 168]. In other words, I have attempted to study whether the classification accuracy metric extracted with deep neural network models yields an added value when it comes to inferring CMD in DOC. The classification accuracy results obtained with both the full-trial and the sample/window-based analysis with deep learning models are very consistent and exhibit the same characteristics. First, all deep learning models resulted in much higher average accuracy compared to LDA; this is true also for the vast majority of the individual session results. Second, this effect is clearly correlated with the size and expressiveness of the model, with the EEGConformer (which is equipped with both transformer-like, far-sighted, attention-based features and the ability to extract spatial/spectral/temporal dependencies through convolution) outperforming the DeepConvNet (which is only capable of the latter), while they both significantly outperform the shallower EEGNet architecture. As a result, the accuracy of most of the individual sessions exceeds the statistical significance threshold for most of the sessions and, again, increasingly so for the deeper and more expressive models. It must be noted that chance levels were extracted here with a binomial distribution assumption for classification outcomes despite this assumption's known limitations, due to the great computational burden of using random shuffling of labels with deep models.

While at a first, superficial glance the above observations may lead to the conclusion that deep learning can indeed have great impact in diagnosing command-following behaviour in DOC, the other emerging commonalities of these results tell an entirely different story. Specifically, the accuracy extracted with the deep models with respect to the control subject/sessions seems to be erratic and completely unreasonable. On the one hand, in all cases but two (EEGNet trial-wise and DeepConvNet sample-wise) the deep models found

the coma control participant, $S25, 1$, to show signs of awareness, as the corresponding accuracy is well above chance. Even more strangely, the coma patient accuracy results are in fact shown to exceed the performance of the healthy control and of patients clinically diagnosed to be somewhere in the MCS spectrum and, thus, aware to some extent. On the other hand, the results become even more counterintuitive when it comes to the healthy control sessions, $S9, 1$ and $S9, 2$. Not only does the EEGNet miss the (beyond any doubt) full-awareness state of the healthy control in the trial-wise classification scenario, but, overall, the sensible outcomes of the shallow model, where the healthy control sessions vastly exceed the accuracy of all other subjects, is not replicated at all by the deep architectures: in all these cases, the performance of the healthy subject lies somewhere close to the average accuracy of the studied population.

First, it is worth to highlight that, also the results of this chapter, underline the need for large control groups on both sides of the awareness spectrum (coma, healthy) that I have inferred in the previous chapter, since, absent these controls, the findings here could have been easily misinterpreted as being favourable to the use of deep learning in DOC. Importantly, these paradoxical control-session findings cannot be blamed on the limitations induced by the binomial assumption to extract chance levels, as the the erratic behaviour regards the absolute accuracy values and not only their relation to the chance level definition used.

Clearly, on the grounds of a complete failure of deep neural network models to correctly classify the control patients, and of their overall overoptimistic tendency to diagnose almost all sessions as CMD (which I have postulated is sign of a weak diagnosing capacity, prone to false positives, also for other EEG metrics in chapter 5), it is my strong opinion that these results cannot be taken at face value. It is thus crucial to interpret these outcomes and identify potential causes explaining the unanticipated behaviour. Towards this goal, I have come up with two hypotheses, which may have acted individually or in tandem to produce such confusing findings.

First, it is observed that the aforementioned correlation between the depth and expressiveness of the models and the average accuracy obtained indicates overfitting caused by data leakage. However, this cannot be a typical case of data leakage, as all results were acquired with LOTO cross-validation (and the corresponding code has been subjected to meticulous testing for bugs), keeping dependent samples in the same fold and thus preventing data leakage overestimating accuracy, exactly like it has been done for the LDA model. Nevertheless, it is possible that data leakage could still have occurred because adjacent and, in general, nearby trials of EEG are only a few seconds apart and thus still somewhat auto-correlated; hence, still, not entirely independent. While LOTO cross-validation is shown to be enough for eliminating any effect of such small dependency for shallow models with low expressiveness (and, thus, no big tendency to overfit), the profound learning capacity and expressiveness of deep models probably renders them able to discover very subtle similarities of even slightly correlated samples that appear in both the training and testing folds, so that data leakage is not avoided. This hypothesis is consistent with the observed effect where larger models showcased higher accuracy.

Second, I conjecture that the difference of feeding PSD samples to the shallow model and time-domain signals to the deep models may also, partly or fully, explain the ostensible overfitting effect. Furthermore, all 16 recorded channels were fed to the deep models, while the too frontal channel Fz was rejected for the shallow model analysis, which could have also played a role. The PSD input to the shallow model effectively guarantees that LDA only attempts to classify SMR rhythms between Rest and MI, in accordance with the relevant neurophysiological background in the context of a command-following protocol involving MI tasks. On the contrary, the deep models, being able to process the raw data and implementing their own, embedded feature extraction, are not constrained to exploiting any particular feature domain and brain phenomenon. This opens the door to the exploitation of features that may be differentially modulated within the periods labeled as either Rest or motor attempt, but in fact have no relation to command following and, consequently,

to awareness. For example, it could be that remnants of the auditory cue in the trial data, which are different between the two tasks ("Bougez", "Ne bougez pas"), despite the fact that the first second of each epoch was discarded for this very reason, are classified with high accuracy giving the false impression that MI is successfully distinguished from idling, even for the coma patient. Another possibility is that artifacts which were not completely removed by the FORCe method and are correlated with the class labels generate "fake" separability, unrelated to the command-following task, especially in view of the fact that the highly vulnerable to artifacts Fz channel was maintained in the processing pipeline in this case. Of note, class-correlated artifacts are a known issue in SMR BCI. For example, able-bodied participants often tend to assume specific body postures when performing specific MI tasks (e.g., lean their head/body right for right-hand MI, and left for left-hand, which can induce muscular and movement artifacts). Similar effects might have occurred in this protocol, especially since many subjects tended react to the FES stimulation with movements, often lasting several seconds. Again, since state-of-the-art artifact removal has been performed, I believe that these potential issues had no effect on the LDA classifier, but artifact residuals may have been spotted by the deep models to inflate their accuracy estimates.

I further note that my original intuition suggested that the deep learning models would fail to exhibit higher (or, at least, considerably higher) accuracy, because, in order to enable direct comparisons with the shallow model results, all classification tasks were forced to only exploit session-specific data (i.e., a mere 45 trials per class); I thus believed that the overfitting tendency of large models in conjunction with the provision of very small data would prevent these models to achieve high classification accuracy, which is obviously not the case. Irrespective of the aforementioned discussion where I lay out why I am convinced that the–eventually–high accuracy observed is anyway bogus, I must stress that a, probably, fairer evaluation of the added value deep learning can contribute to this particular application would entail leave-one-subject-out cross-validation, so that the big

dataset used here can be leveraged and provide, hopefully, adequately big data for the deep models to learn maybe subtle, but actually consciousness-related patterns that a shallow model will miss, while also avoiding data leakage due to trial dependencies. This investigation has been left as future work due to time limitations of training many versions of deep architectures. Nevertheless, it must be already discussed that a big question pertaining to this transfer learning approach is the extent to which SMR patterns are transferrable among individuals and, especially, DOC patients. Current transfer learning and domain adaptation approaches applied to MI BCI seem to yield improvements, but those seem to be, so far, rather limited [241, 242]. In any case, future work along this line of research is expected to elucidate these points in the DOC context.

Figure 4.10 (in this chapter, Figure 5.4D) depicts the classification accuracy MI versus Rest when extracting a single PSD estimate from the 4 s MI epoch of each trial (i.e., without sliding, overlapping windows). It is easy to see that in this case only 4 sessions are found to exceed the chance level (though, all four are meaningful cases, corresponding to the healthy control, two MCS and one eMCS session). There are two possible reasons behind this effect. First, trial-wise classification greatly reduces the classification problem's sample size, as a result of which the chance level estimates with 99% confidence become particularly high. Second, DOC patients, even when aware and vigilant, are shown to be largely unable to sustain MI for long periods of time, while the resulting SMR modulation is less consistent than that of healthy individuals; PSD estimation on the full trial length may thus average out short within-trial intervals of high-quality ERD/ERS, with other trial segments exhibiting no SMR modulation whatsoever, thus reducing the overall separability of the PSD features between MI and Rest and, therefore, also the classification accuracy. Based on these findings, it seems not advisable to extract EEG neuromarkers of awareness in intervals much greater than 1 s at least as far as command following protocols are concerned that require greater vigilance, attention and cognitive ability.

Finally, as evidenced by Figures 5.6 and 5.7, the deep learning architectures generally

failed to discriminate MI from rest under this validation scheme, even when trained with comparatively more data within the LOSO transfer learning scheme. This outcome suggests that the underlying neural representations are insufficiently separable in this context, a limitation that is likewise observed with the shallow classifier. Moreover, any weak or session-specific separability that may exist does not appear to transfer across sessions, further limiting the generalisability of these models.

5.5 Conclusion

Deep learning models have demonstrated remarkable success in domains characterized by large, well-annotated datasets. However, in fields such as BCI and BCI-inspired EEG research, data availability is often severely limited and marked by substantial inter- and intra-subject variability. Under such conditions, deep architectures are generally assumed to perform suboptimally. EEGNet, a domain-specific architecture, challenges this assumption to some extent by incorporating strong inductive biases tailored to EEG signals. In parallel, numerous studies have reported promising results using generic convolutional neural networks and, more recently, transformer-based architectures. These mixed findings have broadened the range of acceptable modeling frameworks and underscored the need for systematic validation of deep learning models on clinically realistic EEG datasets.

As demonstrated by the results presented in this chapter, model expressiveness increases with architectural complexity. At face value, this expressiveness might be misinterpreted as evidence of meaningful awareness-related signal decoding. However, examination of control conditions reveals that these models fail to generalise reliably, indicating substantial overfitting, even when data augmentation strategies embedded within the architectures are employed.

Moreover, increasing the amount of training data through LOSO cross-validation does not improve robustness; instead, predictive performance becomes increasingly unstable

and approaches randomness. This outcome highlights the lack of transferable discriminative features within the highly variable within-subject structure of DOC EEG data. Consequently, in this setting, shallow models demonstrate superior reliability and interpretability compared to more complex deep learning architectures.

Chapter 6

Conclusion & Future Work

The diagnosis of DOC remains a significant clinical challenge, with current behavioral scales prone to diagnostic errors. However, the findings presented here confirm that EEG carries valuable information regarding the state of awareness in DOC patients. Specifically, classification accuracy and the μ/β-band separability of MI PSD features, together with centro-parietal δ-band connectivity during MI and resting states, show statistically significant correlations with CRS-R scores. Furthermore, metric-specific thresholds could be established to reliably distinguish awareness from non-awareness. These results provide important insights into the ability of EEG-based metrics to detect CMD and to reduce the false-negative limitations of CRS-R. At the same time, this work cautions against potential statistical misuse of these metrics, which could otherwise lead to over-optimistic interpretations of latent awareness.

6.1 Introduction

As discussed in Chapter 1, the primary aim of this research was to evaluate the effectiveness of widely used EEG-based approaches for diagnosing awareness, alongside state-of-the-art deep learning models for neural feature classification. This evaluation was conducted using

a large, novel dataset acquired from 28 DOC patients in a real-world ICU setting. In line with prior literature, an active paradigm was adopted—specifically, a MI-based command-following EEG protocol. The focus of this thesis is strictly diagnostic; prognostic outcomes related to recovery or survival were intentionally excluded.

A range of contemporary EEG-derived metrics was extracted, including classification accuracy, brain rhythm power and separability across task conditions, effective connectivity, spectral slope, and the PCI. These metrics were computed from MI, rest, and FES epochs. While not exhaustive, this set encompasses, to the best of my knowledge, the most promising neurophysiological markers currently proposed for awareness detection in DOC. All metrics were subsequently contrasted with CRS-R scores, the current clinical gold standard for assessing awareness following ABI. In addition, classification performance achieved using deep learning pipelines was systematically compared against that of conventional shallow models within the DOC population.

The results demonstrate that EEG signals contain clinically meaningful markers of awareness. As shown by the regression analyses in Chapter 4, an ensemble of relevant EEG-based metrics significantly predicts session-wise CRS-R scores. Importantly, this work also indicates that the neural mechanisms underlying DOC differ fundamentally from those characterizing other neurological conditions, supporting the view that DOC should be investigated as a distinct clinical entity. Even within this focused framework, substantial inter- and intra-subject variability persists, which limits the effectiveness of transfer learning approaches. This variability in transfer learning context, is particularly detrimental to shallow classifiers, which struggle to capture sufficiently stable patterns to generalize to unseen patients, as evidenced by the results in Chapter 4.

A persistent challenge in DOC research is the limited availability of large-scale datasets. Although the dataset used in this thesis—comprising 28 subjects and 62 recording sessions—is novel and comparatively large for this domain, it remains insufficient to support robust generalization, especially for data-hungry deep learning architectures. This

limitation leaves many findings open to debate. The challenge is further compounded by the absence of a reliable ground truth, which forces continued reliance on the CRS-R, despite its known limitations. Together, data scarcity and pronounced neural variability directly constrain the applicability of deep learning methods in this field.

Another key observation from this work is that the strong emphasis on correcting type II errors in CRS-R-based diagnosis may have inadvertently led to insufficient attention to type I errors. While minimizing false negatives is critical given the clinical implications of withdrawal-of-care decisions, a clinically acceptable diagnostic framework must address both error types. This consideration motivated the introduction of the CBA framework in this thesis. In the absence of a trusted ground truth, CBA provides a principled diagnostic approach by evaluating neural classification performance relative to empirical chance levels rather than fixed thresholds. The results in Chapter 4 demonstrate that CBA offers superior robustness for distinguishing awareness from unawareness.

Based on the findings of this research, it is recommended that ML pipelines for DOC diagnosis prioritize well-regularized shallow models, given the high variability and limited data volumes characteristic of this domain. Furthermore, chance-level–based classification accuracy emerges as one of the most reliable indicators of awareness. Looking forward, meaningful progress will require coordinated, multi-centre data collection efforts to establish shared DOC datasets. While ethical considerations and data governance remain significant challenges, such collaboration is essential. Ultimately, convergence on robust diagnostic markers must precede any large-scale efforts toward predictive modelling or prognosis in DOC.

6.2 Contributions to Knowledge

This thesis advances the diagnosis of awareness in DOC using BCI-inspired protocols by:

- **Validating EEG biomarkers in a real ICU environment**: Multiple state-of-the-art

EEG-based diagnostic metrics are validated using a novel dataset spanning the full DOC spectrum, collected in a real-world ICU setting.

- **Integrating multi-metric biomarkers for consciousness assessment**: Rather than relying on a single biomarker, this work systematically evaluates and contrasts a range of complementary EEG-based metrics to assess awareness.

- **Establishing quantitative thresholds for awareness**: The study proposes data-driven, metric-specific thresholds to discriminate between aware and unaware states, contributing toward more objective and clinically interpretable criteria for consciousness assessment.

- **Evaluating deep learning against shallow EEG classifiers**: A methodological comparison of deep learning architectures and conventional shallow classifiers is conducted, providing context-specific insight into their relative performance in DOC diagnosis.

- **Providing context-aware insights into classification in DOC**: Instead of presuming the superiority of deep learning approaches, this work offers empirical evidence on the limitations and applicability of deep models in clinical EEG-based DOC assessment.

6.3 General Conclusion

In this thesis, As was discussed in Chapter 1, the motivating aim of this research was to evaluate the effectiveness of several widely used EEG-based methods for the diagnosis of awareness, as well as state-of-the-art deep learning models with classifying neural features, by applying them to a large novel dataset collected from 28 DOC patients in a real ICU environment. And as proposed in literature, I adopted the active paradigm-the MI-based command-following EEG protocol. Also, I focus exclusively on diagnosis, without discussing here the prognosis of emergence from DOC and/or survival. Specifically, I extracted

state-of-the-art EEG-based metrics, such as classification accuracy, brain rhythms and their separability between mental task conditions, effective connectivity, spectrum slope and PCI from MI, rest and FES epochs. Although this is not an exhaustive list, it contains, to the best of my knowledge, the most promising metrics introduced so far in this line of research. I then contrasted these measures to CRS-R scores, the current gold-standard clinical measure of awareness following ABI. I also compared the classification accuracies achieved via the deep learning pipelines to the shallow pipeline with respect to DOC population.

Overall, my findings reinforce the conclusion that EEG signals carry meaningful markers of consciousness in DOC patients. I deduce that, classification accuracy, μ- and β-band separability, and δ-band effective connectivity between central and parietal brain regions show clear evidence of agreement with the CRS-R assessment and ability to discriminate groups with different levels of awareness when applied to MI and, in some cases, rest intervals, but not during FES epochs. On the contrary, measures like Kolmogorov and Lempel-Ziv complexity [188, 243] or the slope of the EEG spectrum [190, 203] that have been proposed by other studies as putative neuromarkers of awareness and/or consciousness were not found to be relevant here. Importantly, extending my previous preliminary work [6], I postulate that, in order to avoid overly optimistic results regarding the inference of latent awareness and CMD, the usage of classification accuracy as a marker of awareness in DOC should be determined by means of subject/session dataset-specific statistical bounds that are common in BCI and ML literature, such as the statistical thresholds of "chance-level" classification, rather than on generic, population-wide thresholds on the accuracy value. I also posit that, given the peculiarities of EEG datasets—which may increase the risk of data leakage—together with the methodological differences between architectures (i.e., providing pre-computed PSD features to shallow models versus feeding raw time-domain signals, sometimes containing artifact-laden channels, to deep learning models), shallow architectures may in fact offer greater reliability and dependability for diagnosing DOC populations. Their relative robustness may stem from the reliance on

neurophysiologically relevant features, rather than on representations learned directly from potentially noisy raw signals.

This research supports the potential of open-loop BCI DOC diagnosis and highlights the need for further development, validation and standardization to establish clinically deployable systems.

6.4 Limitations and Future Work

While this work offers valuable insights into the diagnosis of awareness in DOC populations, several limitations should be considered. First, the dataset is unbalanced in terms of the CRS-R-based diagnostic categories at first test and very limited in the inclusion of controls (with only one healthy control and one clinically diagnosed coma patient). These limitations reflect logistical constraints of bedside studies in the hospital and my dependence on the uncontrolled influx of traumatic and cerebrovascular brain injury incidents with DOC at the acute neurorehabilitation unit of CHUV, Lausanne. For the same reason, although a target of 3 runs per session was pursued, the goal was not achieved in all cases (there exist two sessions with only 1 run and 8 sessions with two runs, see Table 3.1). These imbalances may impact the generalization of results. Future studies with larger and more balanced control groups and individual participant data are necessary to validate these findings. Along the same lines, although with 28 recruited patients over 62 EEG sessions this investigation is well-powered relatively to the majority of related studies, much larger recruitment in the future will help remove uncertainty from the conclusions reached here [174].

Admittedly, while I strongly believe that DOC diagnostic analyses are best conducted on a session-based and subject-specific basis—since this accounts for inter- and intra-subject reallocation effects—alternative cross-validation schemes could still be explored. In particular, cross-subject validation may prove more beneficial for deep learning models,

given their capacity to generalize across larger and more diverse training sets. Another possible strategy would be to use healthy control sessions as the training data while reserving DOC sessions for testing. Deep models could potentially leverage it to learn feature-specific discriminant patterns more effectively.

The single most important caveat of this work, and of this overall line of research in general [130], is undoubtedly the lack of ground truth. Researchers in this field are called to evaluate and enhance the effectiveness of imaging-based paradigms for the diagnosis of DOC with no strict and formal procedure to secure the integrity of and compare the novel methods' predictions on a participant's state of awareness. This is a particularly significant problem also with regard to the introduction of elaborate machine learning and deep learning approaches (the application of the latter being also limited by the conventionally small size of neuroimaging datasets in DOC research), as most of these models can only be trained in a supervised way and thus require strict ground truth labels. In the current landscape, CMD is not often judged by the appearance of known (as previously studied in conscious, able-bodied populations) neural markers of a certain task, state, or response [32, 60]. However, this methodology is very vague, often, qualitative and open to interpretations; even when quantified it becomes an issue of defining thresholds, the determination of which is similarly debatable, as also shown here.

Here, I attempted to gauge the value of candidate EEG correlates of awareness through their actual mathematical correlation and various other measures of association between these metrics and CRS-R. While at first sight this may seem counterintuitive given that it is CRS-R itself that EEG paradigms for DOC are trying to rectify (mainly, its suspected vulnerability with false negative errors), there is a sound reasoning behind this approach which is based on the well accepted effectiveness of CRS-R in the medical world: the idea has been to require from an EEG metric to show significant correlation with CRS-R, the current golden standard, as proof of relevance, and subsequently also examine to what extent these metrics seem, first, to confirm individual cases of MCS (i.e., agree with CRS-

R on the existence of covert awareness in marginal occasions) or refute CRS-R-based UWS diagnoses, thus potentially being able to correct false negative errors. Clearly, without actual, precise knowledge on the patient's awareness, this type of analysis will always be highly conjectural.

One possible escape from this deadlock is to longitudinally monitor patient outcomes [38, 174]. Specifically, emergence from DOC or even improvement of responsiveness captured by CRS-R and other scores constitutes fairly strong evidence in favour of the hypothesis that an early CMD diagnosis was correct; similarly, a plateau in or deterioration of a patient's clinically assessed awareness (absent concomitant diseases) would indicate the opposite. Another promising approach entails exploiting the looming complementarity of EEG metrics highlighted in this thesis. The assumption would be that agreement between several metrics and/or fusion of individual metric predictions could greatly increase the confidence in EEG-based diagnosis [213]. Furthermore, different metrics can become the features of a machine learning model predicting CMD or even a specific degree of awareness. However, this approach requires large datasets that are logistically hard to collect [174].

Bibliography

[1] Naji Alnagger, Paolo Cardone, Charlotte Martial, Steven Laureys, Jitka Annen, and Olivia Gosseries. The current and future contribution of neuroimaging to the understanding of disorders of consciousness. *La Presse Médicale*, 52(2):104163, 2023.

[2] Katarzyna Blinowska and Piotr Durka. Electroencephalography (eeg). *Wiley encyclopedia of biomedical engineering*, 10:9780471740360, 2006.

[3] Joseph T Giacino, Kathleen Kalmar, and John Whyte. The jfk coma recovery scale-revised: measurement characteristics and diagnostic utility. *Archives of physical medicine and rehabilitation*, 85(12):2020–2029, 2004.

[4] Graham Teasdale and Bryan Jennett. Assessment of coma and impaired consciousness: a practical scale. *The lancet*, 304(7872):81–84, 1974.

[5] BioRender.com. Doc command-following recording protocol, 2025. Adapted from BioRender.com templates Accessed: 2025-02-27.

[6] Idorenyin Amaunam, Christoph Schneider, Marina Lopes da Silva, Jane Jöhr, Karin Diserens, and Serafeim Perdikis. A discussion of statistical criteria for assessing awareness with smr bci after brain injury. In *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 3709–3714. IEEE, 2023.

[7] Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. Eegnet: a compact convolutional neural network

for eeg-based brain–computer interfaces. *Journal of neural engineering*, 15(5):056013, 2018.

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

[9] Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. Eeg conformer: Convolutional transformer for eeg decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2022.

[10] Gernot Müller-Putz, Reinhold Scherer, Clemens Brunner, Robert Leeb, Gert Pfurtscheller, et al. Better than random? a closer look on bci results. *International journal of bioelectromagnetism*, 10(1):52–55, 2008.

[11] Old Testament. The holy bible. *Old Testament*, 2018.

[12] Muhammad Marmaduke Picthall. The koran. *The Koran*, 1953.

[13] Archibald Edward Gough. *The philosophy of the Upanishads and ancient Indian metaphysics*. Routledge, 2013.

[14] UČENJA BUDDHE. Majjhima-nikaya.

[15] Aristotle Aristotle. *On the soul*. Alex Catalogue, 2000.

[16] Pavel Gregoric. *Aristotle on the common sense*. Oxford University Press, 2007.

[17] Anil K Seth and Tim Bayne. Theories of consciousness. *Nature reviews neuroscience*, 23(7):439–452, 2022.

[18] Adam Zeman. What do we mean by "conscious" and "aware"? *Neuropsychological rehabilitation*, 16(4):356–376, 2006.

[19] Olivia Gosseries, Audrey Vanhaudenhuyse, Marie-Aurélie Bruno, Athena Demertzi, Caroline Schnakers, Mélanie M Boly, Audrey Maudoux, Gustave Moonen, and Steven Laureys. Disorders of consciousness: coma, vegetative and minimally conscious states. *States of consciousness: Experimental insights into meditation, waking, sleep and dreams*, pages 29–55, 2011.

[20] Steven Laureys. The neural correlate of (un) awareness: lessons from the vegetative state. *Trends in cognitive sciences*, 9(12):556–559, 2005.

[21] Polona Pozeg, Jane Jöhr, John O Prior, Karin Diserens, and Vincent Dunet. Explaining recovery from coma with multimodal neuroimaging. *Journal of neurology*, 271(9):6274–6288, 2024.

[22] Jerome B Posner. *Plum and Posner's diagnosis of stupor and coma*, volume 71. OUP USA, 2007.

[23] Steven Laureys, Gastone G Celesia, Francois Cohadon, Jan Lavrijsen, José León-Carrión, Walter G Sannita, Leon Sazbon, Erich Schmutzhard, Klaus R von Wild, Adam Zeman, et al. Unresponsive wakefulness syndrome: a new name for the vegetative state or apallic syndrome. *BMC medicine*, 8:1–4, 2010.

[24] Joseph T Giacino, Stephen Ashwal, Nancy Childs, Ronald Cranford, Bryan Jennett, Douglas I Katz, James P Kelly, Jay H Rosenberg, JOHN Whyte, Ross D Zafonte, et al. The minimally conscious state: definition and diagnostic criteria. *Neurology*, 58(3):349–353, 2002.

[25] Karin Diserens, Ivo A Meyer, Jane Jöhr, Alessandro Pincherle, Vincent Dunet, Polona Pozeg, Philippe Ryvlin, Dafin F Mureşanu, Robert D Stevens, and Nich-

olas D Schiff. A focus on subtle signs and motor behavior to unveil awareness in unresponsive brain-impaired patients: the importance of being clinical. *Neurology*, 100(24):1144–1150, 2023.

[26] Yamei Yu, Fanxia Meng, Li Zhang, Xiaoyan Liu, Yuehao Wu, Sicong Chen, Xufei Tan, Xiaoxia Li, Sheng Kuang, Yu Sun, et al. A multi-domain prognostic model of disorder of consciousness using resting-state fmri and laboratory parameters. *Brain Imaging and Behavior*, 15:1966–1976, 2021.

[27] Piergiuseppe Liuzzi, Antonello Grippo, Silvia Campagnini, Maenia Scarpino, Francesca Draghi, Annamaria Romoli, Bahia Hakiki, Raisa Sterpu, Antonio Maiorelli, Claudio Macchi, et al. Merging clinical and eeg biomarkers in an elastic-net regression for disorder of consciousness prognosis prediction. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 30:1504–1513, 2022.

[28] S. L. Hauger, A.-K. Schanke, S. Andersson, C. Chatelle, C. Schnakers, and M. Løvstad. The clinical diagnostic utility of electrophysiological techniques in assessment of patients with disorders of consciousness following acquired brain injury: A systematic review. *Journal of Head Trauma Rehabilitation*, 32(3):185–196, May 2017.

[29] V. Galiotta, I. Quattrociocchi, M. D'Ippolito, F. Schettini, P. ò, S. Sdoia, R. Formisano, F. Cincotti, D. Mattia, and A. Riccio. EEG-based Brain-Computer Interfaces for people with Disorders of Consciousness: Features and applications. A systematic review. *Front Hum Neurosci*, 16:1040816, 2022.

[30] Nicholas D Schiff. Cognitive motor dissociation following severe brain injuries. *JAMA neurology*, 72(12):1413–1415, 2015.

[31] William H Curley, Peter B Forgacs, Henning U Voss, Mary M Conte, and Nicholas D

Schiff. Characterization of EEG signals revealing covert cognition in the injured brain. *Brain*, 141(5):1404–1421, March 2018.

[32] Adrian M Owen, Martin R Coleman, Melanie Boly, Matthew H Davis, Steven Laureys, and John D Pickard. Detecting awareness in the vegetative state. *science*, 313(5792):1402–1402, 2006.

[33] Mélanie Boly, Martin R Coleman, MH Davis, Adam Hampshire, Daniel Bor, Gustave Moonen, Pierre A Maquet, John D Pickard, Steven Laureys, and Adrian M Owen. When thoughts become action: an fmri paradigm to study volitional brain activity in non-communicative brain injured patients. *Neuroimage*, 36(3):979–992, 2007.

[34] Martin M Monti, Audrey Vanhaudenhuyse, Martin R Coleman, Melanie Boly, John D Pickard, Luaba Tshibanda, Adrian M Owen, and Steven Laureys. Willful modulation of brain activity in disorders of consciousness. *New England journal of medicine*, 362(7):579–589, 2010.

[35] Marie-Aurelie Bruno, Olivia Gosseries, Didier Ledoux, Roland Hustinx, and Steven Laureys. Assessment of consciousness with electrophysiological and neurological imaging techniques. *Current opinion in critical care*, 17(2):146–151, 2011.

[36] Damian Cruse, Srivas Chennu, Camille Chatelle, Tristan A Bekinschtein, Davinia Fernández-Espejo, John D Pickard, Steven Laureys, and Adrian M Owen. Bedside detection of awareness in the vegetative state: a cohort study. *The Lancet*, 378(9809):2088–2094, 2011.

[37] Srivas Chennu, Jitka Annen, Sarah Wannez, Aurore Thibaut, Camille Chatelle, Helena Cassol, Géraldine Martens, Caroline Schnakers, Olivia Gosseries, David Menon, et al. Brain networks predict metabolism, diagnosis and prognosis at the bedside in disorders of consciousness. *Brain*, 140(8):2120–2132, 2017.

[38] Brian L Edlow, Camille Chatelle, Camille A Spencer, Catherine J Chu, Yelena G Bodien, Kathryn L O'Connor, Ronald E Hirschberg, Leigh R Hochberg, Joseph T Giacino, Eric S Rosenthal, et al. Early detection of consciousness in patients with acute severe traumatic brain injury. *Brain*, 140(9):2399–2414, 2017.

[39] B Wilhelm, M Jordan, and N Birbaumer. Communication in locked-in syndrome: effects of imagery on salivary ph. *Neurology*, 67(3):534–535, 2006.

[40] Josef Stoll, Camille Chatelle, Olivia Carter, Christof Koch, Steven Laureys, and Wolfgang Einhäuser. Pupil responses allow communication in locked-in syndrome patients. *Current biology*, 23(15):R647–R648, 2013.

[41] Benjamin Blankertz, Ryota Tomioka, Steven Lemm, Motoaki Kawanabe, and Klaus-Robert Muller. Optimizing spatial filters for robust eeg single-trial analysis. *IEEE Signal processing magazine*, 25(1):41–56, 2007.

[42] Alexandre Barachant, Stéphane Bonnet, Marco Congedo, and Christian Jutten. Riemannian geometry applied to bci classification. In *International conference on latent variable analysis and signal separation*, pages 629–636. Springer, 2010.

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[44] Stuart Geman, Elie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1–58, 1992.

[45] Russell A Poldrack, Chris I Baker, Joke Durnez, Krzysztof J Gorgolewski, Paul M Matthews, Marcus R Munafò, Thomas E Nichols, Jean-Baptiste Poline, Edward Vul, and Tal Yarkoni. Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nature reviews neuroscience*, 18(2):115–126, 2017.

[46] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[47] Ernst Niedermeyer and FH Lopes da Silva. *Electroencephalography: basic principles, clinical applications, and related fields*. Lippincott Williams & Wilkins, 2005.

[48] Benjamin Blankertz, Michael Tangermann, Carmen Vidaurre, Siamac Fazli, Claudia Sannelli, Stefan Haufe, Cecilia Maeder, Lenny E Ramsey, Irene Sturm, Gabriel Curio, et al. The berlin brain–computer interface: non-medical uses of bci technology. *Frontiers in neuroscience*, 4:2050, 2010.

[49] Demetres Kostas and Frank Rudzicz. Thinker invariance: enabling deep neural networks for bci across more people. *Journal of Neural Engineering*, 17(5):056008, 2020.

[50] Matthias Feurer and Frank Hutter. Hyperparameter optimization. In *Automated machine learning: Methods, systems, challenges*, pages 3–33. Springer International Publishing Cham, 2019.

[51] He He and Dongrui Wu. Transfer learning for brain–computer interfaces: A euclidean space data alignment approach. *IEEE Transactions on Biomedical Engineering*, 67(2):399–410, 2019.

[52] Jan Claassen, Kevin Doyle, Adu Matory, Caroline Couch, Kelly M Burger, Angela Velazquez, Joshua U Okonkwo, Jean-Rémi King, Soojin Park, Sachin Agarwal, et al. Detection of brain activation in unresponsive patients with acute brain injury. *New England Journal of Medicine*, 380(26):2497–2505, 2019.

[53] Jacobo Diego Sitt, Jean-Remi King, Imen El Karoui, Benjamin Rohaut, Frederic Faugeras, Alexandre Gramfort, Laurent Cohen, Mariano Sigman, Stanislas Dehaene,

and Lionel Naccache. Large scale screening of neural signatures of consciousness in patients in a vegetative or minimally conscious state. *Brain*, 137(8):2258–2270, 2014.

[54] Andrea Kübler and Niels Birbaumer. Brain–computer interfaces and communication in paralysis: Extinction of goal directed thinking in completely paralysed patients? *Clinical neurophysiology*, 119(11):2658–2666, 2008.

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

[56] Ronald T Seel, Mark Sherer, John Whyte, Douglas I Katz, Joseph T Giacino, Amy M Rosenbaum, Flora M Hammond, Kathleen Kalmar, Theresa Louise-Bender Pape, Ross Zafonte, et al. Assessment scales for disorders of consciousness: evidence-based recommendations for clinical practice and research. *Archives of physical medicine and rehabilitation*, 91(12):1795–1813, 2010.

[57] Paul Gerrard, Ross Zafonte, and Joseph T Giacino. Coma recovery scale–revised: evidentiary support for hierarchical grading of level of consciousness. *Archives of physical medicine and rehabilitation*, 95(12):2335–2341, 2014.

[58] Joseph T Giacino, Caroline Schnakers, Diana Rodriguez-Moreno, Kathy Kalmar, Nicholas Schiff, and Joy Hirsch. Behavioral assessment in patients with disorders of consciousness: gold standard or fool's gold? *Progress in brain research*, 177:33–48, 2009.

[59] Jean-Paul Noel, Camille Chatelle, Serafeim Perdikis, Jane Jöhr, Marina Lopes Da Silva, Philippe Ryvlin, Marzia De Lucia, José del R Millán, Karin Diserens, and Andrea Serino. Peri-personal space encoding in patients with disorders of consciousness and cognitive-motor dissociation. *NeuroImage: Clinical*, 24:101940, 2019.

[60] Valentina Galiotta, Ilaria Quattrociocchi, Mariagrazia D'Ippolito, Francesca Schettini, Pietro Aricò, Stefano Sdoia, Rita Formisano, Febo Cincotti, Donatella Mattia, and Angela Riccio. Eeg-based brain-computer interfaces for people with disorders of consciousness: features and applications. a systematic review. *Frontiers in Human Neuroscience*, 16:1040816, 2022.

[61] Jean-Michel Pignat, Etienne Mauron, Jane Jöhr, Charlotte Gilart de Keranflec'h, Dimitri Van De Ville, Maria Giulia Preti, Djalel E Meskaldji, Volker Hömberg, Steven Laureys, Bogdan Draganski, et al. Outcome prediction of consciousness disorders in the acute stage based on a complementary motor behavioural tool. *PloS one*, 11(6):e0156882, 2016.

[62] Caroline Schnakers, Audrey Vanhaudenhuyse, Joseph Giacino, Manfredi Ventura, Melanie Boly, Steve Majerus, Gustave Moonen, and Steven Laureys. Diagnostic accuracy of the vegetative and minimally conscious state: clinical consensus versus standardized neurobehavioral assessment. *BMC neurology*, 9:1–5, 2009.

[63] Derick T Wade. How often is the diagnosis of the permanent vegetative state incorrect? a review of the evidence. *European Journal of Neurology*, 25(4):619–625, 2018.

[64] Niels Birbaumer, Nimr Ghanayim, Thilo Hinterberger, Iver Iversen, Boris Kotchoubey, Andrea Kübler, Juri Perelmouter, Edward Taub, and Herta Flor. A spelling device for the paralysed. *Nature*, 398(6725):297–298, 1999.

[65] Steven Laureys and Nicholas D Schiff. Coma and consciousness: paradigms (re) framed by neuroimaging. *Neuroimage*, 61(2):478–491, 2012.

[66] Ujwal Chaudhary, Ioannis Vlachos, Jonas B Zimmermann, Arnau Espinosa, Alessandro Tonin, Andres Jaramillo-Gonzalez, Majid Khalili-Ardali, Helge Topka, Jens

Lehmberg, Gerhard M Friehs, et al. Spelling interface using intracortical signals in a completely locked-in patient enabled via auditory neurofeedback training. *Nature communications*, 13(1):1236, 2022.

[67] Xiang Zhang, Lina Yao, Xianzhi Wang, Jessica Monaghan, David Mcalpine, and Yu Zhang. A survey on deep learning-based non-invasive brain signals: recent advances and new frontiers. *Journal of neural engineering*, 18(3):031002, 2021.

[68] Saeid Sanei and Jonathon A Chambers. *EEG signal processing*. John Wiley & Sons, 2013.

[69] Eric R Kandel, James H Schwartz, Thomas M Jessell, Steven Siegelbaum, A James Hudspeth, Sarah Mack, et al. *Principles of neural science*, volume 4. McGraw-hill New York, 2000.

[70] Arthur D Craig. How do you feel? interoception: the sense of the physiological condition of the body. *Nature reviews neuroscience*, 3(8):655–666, 2002.

[71] Harold L Atwood and William A MacKay. Essentials of neurophysiology. *(No Title)*, 1989.

[72] Mark Bear, Barry Connors, and Michael A Paradiso. *Neuroscience: Exploring the brain, enhanced edition: Exploring the brain*. Jones & Bartlett Learning, 2020.

[73] Alan L Hodgkin and Andrew F Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of physiology*, 117(4):500, 1952.

[74] György Buzsáki, Costas A Anastassiou, and Christof Koch. The origin of extracellular fields and currents—eeg, ecog, lfp and spikes. *Nature reviews neuroscience*, 13(6):407–420, 2012.

[75] Carl E Stafstrom and Lionel Carmant. Seizures and epilepsy: an overview for neuroscientists. *Cold Spring Harbor perspectives in medicine*, 5(6):a022426, 2015.

[76] Andrea Biasiucci, Robert Leeb, Iñaki Iturrate, Serafeim Perdikis, Abdul Al-Khodairy, Tiffany Corbet, Armin Schnider, Thomas Schmidlin, Huaijian Zhang, Manuela Bassolino, et al. Brain-actuated functional electrical stimulation elicits lasting arm motor recovery after stroke. *Nature communications*, 9(1):2421, 2018.

[77] Harry Nyquist. Certain topics in telegraph transmission theory. *Transactions of the American Institute of Electrical Engineers*, 47(2):617–644, 2009.

[78] Claude E Shannon. Communication in the presence of noise. *Proceedings of the IRE*, 37(1):10–21, 2006.

[79] Ian Daly, Reinhold Scherer, Martin Billinger, and Gernot Müller-Putz. Force: Fully online and automated artifact removal for brain-computer interfacing. *IEEE transactions on neural systems and rehabilitation engineering*, 23(5):725–736, 2014.

[80] Jose Antonio Urigüen and Begoña Garcia-Zapirain. Eeg artifact removal—state-of-the-art and guidelines. *Journal of neural engineering*, 12(3):031001, 2015.

[81] Christoph M Michel and Micah M Murray. Towards the utilization of eeg as a brain imaging tool. *Neuroimage*, 61(2):371–385, 2012.

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

[83] Michal Teplan et al. Fundamentals of eeg measurement. *Measurement science review*, 2(2):1–11, 2002.

[84] Hans Berger. Über das elektroenkephalogramm des menschen. *Archiv für psychiatrie und nervenkrankheiten*, 87(1):527–570, 1929.

[85] Lindsay F Haas. Hans berger (1873–1941), richard caton (1842–1926), and electroencephalography. *Journal of Neurology, Neurosurgery & Psychiatry*, 74(1):9–9, 2003.

[86] Gert Pfurtscheller and FH Lopes Da Silva. Event-related eeg/meg synchronization and desynchronization: basic principles. *Clinical neurophysiology*, 110(11):1842–1857, 1999.

[87] Emma M Whitham, Kenneth J Pope, Sean P Fitzgibbon, Trent Lewis, C Richard Clark, Stephen Loveless, Marita Broberg, Angus Wallace, Dylan DeLosAngeles, Peter Lillie, et al. Scalp electrical recording during paralysis: quantitative evidence that eeg frequencies above 20 hz are contaminated by emg. *Clinical neurophysiology*, 118(8):1877–1888, 2007.

[88] Douglas J Gould, James D Fix, and Jennifer K Brueckner. *High-yield neuroanatomy*. Wolters Kluwer, 2016.

[89] Herbert H Jasper. Ten-twenty electrode system of the international federation. *Electroencephalogr Clin Neurophysiol*, 10:371–375, 1958.

[90] Rodney J Croft and Robert J Barry. Removal of ocular artifact from the eeg: a review. *Neurophysiologie Clinique/Clinical Neurophysiology*, 30(1):5–19, 2000.

[91] Irina I Goncharova, Dennis J McFarland, Theresa M Vaughan, and Jonathan R Wolpaw. Emg contamination of eeg: spectral and topographical characteristics. *Clinical neurophysiology*, 114(9):1580–1593, 2003.

[92] Xiao Jiang, Gui-Bin Bian, and Zean Tian. Removal of artifacts from eeg signals: a review. *Sensors*, 19(5):987, 2019.

[93] Dante Mantini, Mauro Gianni Perrucci, Simone Cugini, Antonio Ferretti, Gian Luca Romani, and Cosimo Del Gratta. Complete artifact removal for eeg recorded during

continuous fmri using independent component analysis. *Neuroimage*, 34(2):598–607, 2007.

[94] Rolf Verleger, Theo Gasser, and Joachim Möcks. Correction of eog artifacts in event-related potentials of the eeg: Aspects of reliability and validity. *Psychophysiology*, 19(4):472–480, 1982.

[95] Gabriele Gratton, Michael GH Coles, and Emanuel Donchin. A new method for off-line removal of ocular artifact. *Electroencephalography and clinical neurophysiology*, 55(4):468–484, 1983.

[96] Alan V Oppenheim. *Discrete-time signal processing*. Pearson Education India, 1999.

[97] Todd C Handy. *Event-related potentials: A methods handbook*. MIT press, 2005.

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

[99] Fabien Lotte and Marco Congedo. Eeg feature extraction. *Brain–Computer Interfaces 1: Foundations and Methods*, pages 127–143, 2016.

[100] Steven M Kay. *Modern spectral estimation*. Pearson Education India, 1988.

[101] Olaf Sporns, Giulio Tononi, and Rolf Kötter. The human connectome: a structural description of the human brain. *PLoS computational biology*, 1(4):e42, 2005.

[102] Susumu Mori and Peter CM Van Zijl. Fiber tracking: principles and strategies–a technical review. *NMR in Biomedicine: An International Journal Devoted to the Development and Application of Magnetic Resonance In Vivo*, 15(7-8):468–480, 2002.

[103] Karl J Friston. Functional and effective connectivity: a review. *Brain connectivity*, 1(1):13–36, 2011.

[104] Paul L Nunez, Ramesh Srinivasan, Andrew F Westdorp, Ranjith S Wijesinghe, Don M Tucker, Richard B Silberstein, and Peter J Cadusch. Eeg coherency: I: statistics, reference electrode, volume conduction, laplacians, cortical imaging, and interpretation at multiple scales. *Electroencephalography and clinical neurophysiology*, 103(5):499–515, 1997.

[105] Bharat Biswal, F Zerrin Yetkin, Victor M Haughton, and James S Hyde. Functional connectivity in the motor cortex of resting human brain using echo-planar mri. *Magnetic resonance in medicine*, 34(4):537–541, 1995.

[106] John Geweke. Measurement of linear dependence and feedback between multiple time series. *Journal of the American statistical association*, 77(378):304–313, 1982.

[107] Karl J Friston, Lee Harrison, and Will Penny. Dynamic causal modelling. *Neuroimage*, 19(4):1273–1302, 2003.

[108] George L Gerstein and Donald H Perkel. Simultaneously recorded trains of action potentials: analysis and functional interpretation. *Science*, 164(3881):828–830, 1969.

[109] Karl J Friston, Chris D Frith, Peter F Liddle, and Richard SJ Frackowiak. Functional connectivity: the principal-component analysis of large (pet) data sets. *Journal of Cerebral Blood Flow & Metabolism*, 13(1):5–14, 1993.

[110] Arnold Neumaier and Tapio Schneider. Estimation of parameters and eigenmodes of multivariate autoregressive models. *ACM Transactions on Mathematical Software (TOMS)*, 27(1):27–57, 2001.

[111] Alexander Y Kaplan, Sergei L Shishkin, Ilya P Ganin, Ivan A Basyul, and Alexander Y Zhigalov. Adapting the p300-based brain–computer interface for gaming: a review. *IEEE Transactions on Computational Intelligence and AI in Games*, 5(2):141–149, 2013.

[112] David Marshall, Damien Coyle, Shane Wilson, and Michael Callaghan. Games, gameplay, and bci: the state of the art. *IEEE Transactions on Computational Intelligence and AI in Games*, 5(2):82–99, 2013.

[113] Laurent Bonnet, Fabien Lotte, and Anatole Lécuyer. Two brains, one game: design and evaluation of a multiuser bci video game based on motor imagery. *IEEE Transactions on Computational Intelligence and AI in games*, 5(2):185–198, 2013.

[114] Serafeim Perdikis, Luca Tonin, Sareh Saeedi, Christoph Schneider, and José del R. Millán. The cybathlon bci race: Successful longitudinal mutual learning with two tetraplegic users. *PLOS Biology*, 16(5):e2003787, May 2018.

[115] Ryan Beveridge, Shane Wilson, and Damien Coyle. Can teenagers control a 3d racing game using motion-onset visual evoked potentials? *Brain-Computer Interfaces*, 4(1-2):102–113, 2017.

[116] Serafeim Perdikis, Luca Tonin, and Jose del R Millan. Brain racers. *IEEE Spectrum*, 54(9):44–51, 2017.

[117] Janggil Oh and Jongheon Kim. Military application study of bci technology using brain waves in republic of korea army: Focusing on personal firearms. *Journal of Advances in Military Studies*, 5(1):35–48, 2022.

[118] Margaret Kosal and Joy Putney. Neurotechnology and international security: Predicting commercial and military adoption of brain-computer interfaces (bcis) in the united states and china. *Politics and the Life Sciences*, 42(1):81–103, 2023.

[119] Niels Birbaumer. Breaking the silence: brain–computer interfaces (bci) for communication and motor control. *Psychophysiology*, 43(6):517–532, 2006.

[120] Serafeim Perdikis, Robert Leeb, John Williamson, Amy Ramsay, Michele Tavella, Lorenzo Desideri, Evert-Jan Hoogerwerf, Abdul Al-Khodairy, Roderick Murray-

Smith, and J d R Millán. Clinical evaluation of braintree, a motor imagery hybrid bci speller. *Journal of neural engineering*, 11(3):036003, 2014.

[121] Betts Peters, Brandon Eddy, Deirdre Galvin-McLaughlin, Gail Betz, Barry Oken, and Melanie Fried-Oken. A systematic review of research on augmentative and alternative communication brain-computer interface systems for individuals with disabilities. *Frontiers in human neuroscience*, 16:952380, 2022.

[122] Luca Randazzo, Inaki Iturrate, Serafeim Perdikis, and J d R Millán. mano: A wearable hand exoskeleton for activities of daily living and neurorehabilitation. *IEEE Robotics and Automation Letters*, 3(1):500–507, 2017.

[123] Niels Birbaumer and Leonardo G Cohen. Brain–computer interfaces: communication and restoration of movement in paralysis. *The Journal of physiology*, 579(3):621–636, 2007.

[124] Ravikiran Mane, Tushar Chouhan, and Cuntai Guan. Bci for stroke rehabilitation: motor and beyond. *Journal of neural engineering*, 17(4):041001, 2020.

[125] Haiyun Huang, Jie Chen, Jun Xiao, Di Chen, Jun Zhang, Jiahui Pan, and Yuanqing Li. Real-time attention regulation and cognitive monitoring using a wearable eeg-based bci. *IEEE Transactions on Biomedical Engineering*, 2024.

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

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

[128] J d R Millán, Rüdiger Rupp, Gernot R Müller-Putz, Roderick Murray-Smith, Claudio Giugliemma, Michael Tangermann, Carmen Vidaurre, Febo Cincotti, Andrea Kübler, Robert Leeb, et al. Combining brain–computer interfaces and assistive technologies: state-of-the-art and challenges. *Frontiers in neuroscience*, 4:161, 2010.

[129] Yang Bai, Yajun Lin, and Ulf Ziemann. Managing disorders of consciousness: the role of electroencephalography. *Journal of neurology*, 268:4033–4065, 2021.

[130] Yvonne Höller, Jürgen Bergmann, Aljoscha Thomschewski, Martin Kronbichler, Peter Höller, Julia S Crone, Elisabeth V Schmid, Kevin Butz, Raffaele Nardone, and Eugen Trinka. Comparison of eeg-features and classification methods for motor imagery in patients with disorders of consciousness. *PloS one*, 8(11):e80479, 2013.

[131] Christoph Guger, Brendan Allison, Rossella Spataro, Vincenzo La Bella, Andrea Kammerhofer, Florian Guttmann, Tim J von Oertzen, Jitka Annen, Steven Laureys, Alexander Heilinger, et al. Mindbeagle—a new system for the assessment and communication with patients with disorders of consciousness and complete locked-in syndrom. In *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 3008–3013. IEEE, 2017.

[132] Camille Chatelle, Srivas Chennu, Quentin Noirhomme, Damian Cruse, Adrian M Owen, and Steven Laureys. Brain–computer interfacing in disorders of consciousness. *Brain injury*, 26(12):1510–1522, 2012.

[133] Damien Coyle, Jacqueline Stow, Karl McCreadie, Jacinta McElligott, and Áine Carroll. Sensorimotor modulation assessment and brain-computer interface training in disorders of consciousness. *Archives of physical medicine and rehabilitation*, 96(3):S62–S70, 2015.

[134] Christoph Guger, Rossella Spataro, Brendan Z Allison, Alexander Heilinger, Rupert Ortner, Woosang Cho, and Vincenzo La Bella. Complete locked-in and locked-in

patients: command following assessment and communication with vibro-tactile p300 and motor imagery brain-computer interface tools. *Frontiers in neuroscience*, 11:251, 2017.

[135] Mikhail A Lebedev and Miguel AL Nicolelis. Brain-machine interfaces: from basic science to neuroprostheses and neurorehabilitation. *Physiological reviews*, 97(2):767–837, 2017.

[136] Hongyu Zhang, Le Jiao, Songxiang Yang, Haopeng Li, Xinzhan Jiang, Jing Feng, Shuhuai Zou, Qiang Xu, Jianheng Gu, Xuefeng Wang, et al. Brain–computer interfaces: the innovative key to unlocking neurological conditions. *International Journal of Surgery*, 110(9):5745–5762, 2024.

[137] Meenalosini Vimal Cruz, Suhaima Jamal, and Sibi Chakkaravarthy Sethuraman. A comprehensive survey of brain–computer interface technology in health care: Research perspectives. *Journal of Medical Signals & Sensors*, 15(6):16, 2025.

[138] Santosh Chandrasekaran, Matthew Fifer, Stephan Bickel, Luke Osborn, Jose Herrero, Breanne Christie, Junqian Xu, Rory KJ Murphy, Sandeep Singh, Matthew F Glasser, et al. Historical perspectives, challenges, and future directions of implantable brain-computer interfaces for sensorimotor applications. *Bioelectronic medicine*, 7(1):14, 2021.

[139] Kogulan Paulmurugan, Vimalan Vijayaragavan, Sayantan Ghosh, Parasuraman Padmanabhan, and Balázs Gulyás. Brain–computer interfacing using functional near-infrared spectroscopy (fnirs). *Biosensors*, 11(10):389, 2021.

[140] Rihui Li, Dalin Yang, Feng Fang, Keum-Shik Hong, Allan L Reiss, and Yingchun Zhang. Concurrent fnirs and eeg for brain function investigation: a systematic, methodology-focused review. *Sensors*, 22(15):5865, 2022.

[141] Sasha Burwell, Matthew Sample, and Eric Racine. Ethical aspects of brain computer interfaces: a scoping review. *BMC medical ethics*, 18(1):60, 2017.

[142] Gert Pfurtscheller and Christa Neuper. Motor imagery and direct brain-computer communication. *Proceedings of the IEEE*, 89(7):1123–1134, 2001.

[143] François-Benoît Vialatte, Monique Maurice, Justin Dauwels, and Andrzej Cichocki. Steady-state visually evoked potentials: focus on essential paradigms and future perspectives. *Progress in neurobiology*, 90(4):418–438, 2010.

[144] Lawrence Ashley Farwell and Emanuel Donchin. Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and clinical Neurophysiology*, 70(6):510–523, 1988.

[145] Gert Pfurtscheller, Brendan Z Allison, Günther Bauernfeind, Clemens Brunner, Teodoro Solis Escalante, Reinhold Scherer, Thorsten O Zander, Gernot Mueller-Putz, Christa Neuper, and Niels Birbaumer. The hybrid bci. *Frontiers in neuroscience*, 4:1283, 2010.

[146] John Whyte and Robin Myers. Incidence of clinically significant responses to zolpidem among patients with disorders of consciousness: a preliminary placebo controlled trial. *American journal of physical medicine & rehabilitation*, 88(5):410–418, 2009.

[147] Dorothée Lulé, Quentin Noirhomme, Sonja C Kleih, Camille Chatelle, Sebastian Halder, Athena Demertzi, Marie-Aurélie Bruno, Olivia Gosseries, Audrey Vanhaudenhuyse, Caroline Schnakers, et al. Probing command following in patients with disorders of consciousness using a brain–computer interface. *Clinical Neurophysiology*, 124(1):101–106, 2013.

[148] Jie Cai, Jiawei Luo, Shulin Wang, and Sheng Yang. Feature selection in machine learning: A new perspective. *Neurocomputing*, 300:70–79, 2018.

[149] Pat Langley. *Elements of machine learning*. Morgan Kaufmann, 1996.

[150] Manuela Vasconcelos and Nuno Vasconcelos. Natural image statistics and low-complexity feature selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):228–244, 2008.

[151] Jaume Gibert, Ernest Valveny, and Horst Bunke. Feature selection on node statistics based embedding of graphs. *Pattern Recognition Letters*, 33(15):1980–1990, 2012.

[152] Bo Li, Chun-Hou Zheng, and De-Shuang Huang. Locally linear discriminant embedding: An efficient method for face recognition. *Pattern Recognition*, 41(12):3813–3821, 2008.

[153] Biao Jie, Daoqiang Zhang, Bo Cheng, and Dinggang Shen. Manifold regularized multi-task feature selection for multi-modality classification in alzheimer's disease. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 275–283. Springer, 2013.

[154] François Fleuret. Fast binary feature selection with conditional mutual information. *Journal of Machine learning research*, 5(Nov):1531–1555, 2004.

[155] Kenneth P Burnham and David R Anderson. *Model selection and multimodel inference: a practical information-theoretic approach*. Springer, 2002.

[156] Jue Wang, Kun Guo, and Shouyang Wang. Rough set and tabu search based feature selection for credit scoring. *Procedia Computer Science*, 1(1):2425–2432, 2010.

[157] Roman W Swiniarski and Andrzej Skowron. Rough set methods in feature selection and recognition. *Pattern recognition letters*, 24(6):833–849, 2003.

[158] Bing Liu. Supervised learning. In *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*, pages 63–132. Springer, 2011.

[159] Derek Greene, Pádraig Cunningham, and Rudolf Mayer. Unsupervised learning and clustering. In *Machine learning techniques for multimedia: Case studies on organization and retrieval*, pages 51–90. Springer, 2008.

[160] Aized Amin Soofi and Arshad Awan. Classification techniques in machine learning: applications and issues. *Journal of Basic & Applied Sciences*, 13:459–465, 2017.

[161] Sebastian Halder, Ana Matran-Fernandez, Rab Nawaz, Marina Lopes da Silva, Tommaso Bertoni, Jean-Paul Noel, Jane Jöhr, Andrea Serino, Karin Diserens, Reinhold Scherer, et al. To repeat or not to repeat? erp-based assessment of the level of consciousness-a case study. 2024.

[162] Sarah Wannez, Lizette Heine, Marie Thonnard, Olivia Gosseries, Steven Laureys, and Coma Science Group Collaborators. The repetition of behavioral assessments in diagnosis of disorders of consciousness. *Annals of neurology*, 81(6):883–889, 2017.

[163] Jing Wang, Xiaohua Hu, Zhouyao Hu, Ziwei Sun, Steven Laureys, and Haibo Di. The misdiagnosis of prolonged disorders of consciousness by a clinical consensus compared with repeated coma-recovery scale-revised assessment. *BMC neurology*, 20(1):343, 2020.

[164] Jan Claassen, Daniel Kondziella, Ayham Alkhachroum, Michael Diringer, Brian L Edlow, Joseph J Fins, Olivia Gosseries, Yousef Hannawi, Benjamin Rohaut, Caroline Schnakers, et al. Cognitive motor dissociation: gap analysis and future directions. *Neurocritical care*, 40(1):81–98, 2024.

[165] Anna Estraneo, Vincenzo Loreto, Ivan Guarino, Virginia Boemia, Giuseppe Paone,

Pasquale Moretta, and Luigi Trojano. Standard eeg in diagnostic process of prolonged disorders of consciousness. *Clinical Neurophysiology*, 127(6):2379–2385, 2016.

[166] Davinia Fernández-Espejo and Adrian M Owen. Detecting awareness after severe brain injury. *Nature Reviews Neuroscience*, 14(11):801–809, 2013.

[167] Olivia Gosseries, Haibo Di, Steven Laureys, and Mélanie Boly. Measuring consciousness in severely damaged brains. *Annual Review of Neuroscience*, 37(1):457–478, 2014.

[168] Christoph Schneider, Serafeim Perdikis, Marina Silva, Jane Jöhr, Alexander Pincherle, José del R Millán, and Karin Diserens. Motor attempt eeg paradigm as a diagnostic tool for disorders of consciousness. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 4681–4684. IEEE, 2018.

[169] Ruben GL Real, Sandra Veser, Helena Erlbeck, Monica Risetti, Dominik Vogel, Friedemann Müller, Boris Kotchoubey, Donatella Mattia, and Andrea Kübler. Information processing in patients in vegetative and minimally conscious states. *Clinical Neurophysiology*, 127(2):1395–1402, 2016.

[170] Jun Xiao, Yanbin He, Tianyou Yu, Jiahui Pan, Qiuyou Xie, Caiyun Cao, Heyi Zheng, Weitian Huang, Zhenghui Gu, Zhuliang Yu, et al. Toward assessment of sound localization in disorders of consciousness using a hybrid audiovisual brain–computer interface. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 30:1422–1432, 2022.

[171] Luca Tonin, Serafeim Perdikis, Taylan Deniz Kuzu, Jorge Pardo, Bastien Orset, Kyuhwa Lee, Mirko Aach, Thomas Armin Schildhauer, Ramón Martínez-Olivera, and José del R. Millán. Learning to control a bmi-driven wheelchair for people with severe tetraplegia. *iScience*, 25(12):105418, December 2022.

[172] Mushfika Sultana and Serafeim Perdikis. Automatic feature selection for sensorimotor rhythms brain-computer interface fusing expert and data-driven knowledge. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 32:3422–3431, 2024.

[173] Andrey Eliseyev, Ian Jerome Gonzales, Anh Le, Kevin Doyle, Jennifer Egbebike, Angela Velazquez, Sachin Agarwal, David Roh, Soojin Park, E Sander Connolly, et al. Development of a brain-computer interface for patients in the critical care setting. *PLoS One*, 16(1):e0245540, 2021.

[174] Yelena G Bodien, Judith Allanson, Paolo Cardone, Arthur Bonhomme, Jerina Carmona, Camille Chatelle, Srivas Chennu, Mary Conte, Stanislas Dehaene, Paola Finoia, et al. Cognitive motor dissociation in disorders of consciousness. *New England Journal of Medicine*, 391(7):598–608, 2024.

[175] Andrew M Goldfine, Jonathan D Victor, Mary M Conte, Jonathan C Bardin, and Nicholas D Schiff. Determination of awareness in patients with severe brain injury using eeg power spectral analysis. *Clinical Neurophysiology*, 122(11):2157–2168, 2011.

[176] Jitka Annen, Steven Laureys, and Olivia Gosseries. Brain-computer interfaces for consciousness assessment and communication in severely brain-injured patients. *Handbook of clinical neurology*, 168:137–152, 2020.

[177] Caroline Schnakers, Fabien Perrin, Manuel Schabus, Roland Hustinx, Steve Majerus, Gustave Moonen, Melanie Boly, Audrey Vanhaudenhuyse, Marie-Aurelie Bruno, and Steven Laureys. Detecting consciousness in a total locked-in syndrome: an active event-related paradigm. *Neurocase*, 15(4):271–277, 2009.

[178] Sylvain Baillet, John C Mosher, and Richard M Leahy. Electromagnetic brain mapping. *IEEE Signal processing magazine*, 18(6):14–30, 2002.

[179] Scott Makeig, Stefan Debener, Julie Onton, and Arnaud Delorme. Mining event-related brain dynamics. *Trends in cognitive sciences*, 8(5):204–210, 2004.

[180] P. Welch. The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Transactions on Audio and Electroacoustics*, 15(2):70–73, June 1967.

[181] Mélanie Boly, Marcello Massimini, Marta Isabel Garrido, Olivia Gosseries, Quentin Noirhomme, Steven Laureys, and Andrea Soddu. Brain connectivity in disorders of consciousness. *Brain connectivity*, 2(1):1–10, 2012.

[182] Audrey Vanhaudenhuyse, Quentin Noirhomme, Luaba J-F Tshibanda, Marie-Aurelie Bruno, Pierre Boveroux, Caroline Schnakers, Andrea Soddu, Vincent Perlbarg, Didier Ledoux, Jean-François Brichant, et al. Default network connectivity reflects the level of consciousness in non-communicative brain-damaged patients. *Brain*, 133(1):161–171, 2010.

[183] Bolin Cao, Yan Chen, Ronghao Yu, Lixiang Chen, Ping Chen, Yihe Weng, Qinyuan Chen, Jie Song, Qiuyou Xie, and Ruiwang Huang. Abnormal dynamic properties of functional connectivity in disorders of consciousness. *Neuroimage: clinical*, 24:102071, 2019.

[184] Jean-Rémi King and Stanislas Dehaene. Characterizing the dynamics of mental representations: the temporal generalization method. *Trends in cognitive sciences*, 18(4):203–210, 2014.

[185] Moti Salti, Simo Monto, Lucie Charles, Jean-Remi King, Lauri Parkkonen, and Stanislas Dehaene. Distinct cortical codes and temporal dynamics for conscious and unconscious percepts. *elife*, 4:e05652, 2015.

[186] Aaron Schurger, Ioannis Sarigiannidis, Lionel Naccache, Jacobo D Sitt, and Stanislas Dehaene. Cortical activity is more stable when sensory stimuli are consciously perceived. *Proceedings of the National Academy of Sciences*, 112(16):E2083–E2092, 2015.

[187] Adenauer G Casali, Olivia Gosseries, Mario Rosanova, Mélanie Boly, Simone Sarasso, Karina R Casali, Silvia Casarotto, Marie-Aurélie Bruno, Steven Laureys, Giulio Tononi, et al. A theoretically based index of consciousness independent of sensory processing and behavior. *Science translational medicine*, 5(198):198ra105–198ra105, 2013.

[188] Michael Schartner, Anil Seth, Quentin Noirhomme, Melanie Boly, Marie-Aurelie Bruno, Steven Laureys, and Adam Barrett. Complexity of multi-dimensional spontaneous eeg decreases during propofol induced general anaesthesia. *PloS one*, 10(8):e0133532, 2015.

[189] Mario Rosanova, M Fecchio, Silvia Casarotto, Simone Sarasso, AG Casali, A Pigorini, A Comanducci, F Seregni, G Devalle, Giuseppe Citerio, et al. Sleep-like cortical off-periods disrupt causality and complexity in the brain of unresponsive wakefulness syndrome patients. *Nature communications*, 9(1):4427, 2018.

[190] Michele Angelo Colombo, Martino Napolitani, Melanie Boly, Olivia Gosseries, Silvia Casarotto, Mario Rosanova, Jean-Francois Brichant, Pierre Boveroux, Steffen Rex, Steven Laureys, et al. The spectral exponent of the resting eeg indexes the presence of consciousness during unresponsiveness induced by propofol, xenon, and ketamine. *NeuroImage*, 189:631–644, 2019.

[191] Janna D Lendner, Randolph F Helfrich, Bryce A Mander, Luis Romundstad, Jack J Lin, Matthew P Walker, Pal G Larsson, and Robert T Knight. An electrophysiological marker of arousal level in humans. *elife*, 9:e55092, 2020.

[192] Ying Zhao, Anqi Wang, Weiqiao Zhao, Nantu Hu, Steven Laureys, and Haibo Di. The neural correlates of consciousness: A spectral exponent approach to diagnosing disorders of consciousness. *Brain Sciences*, 15(4):377, 2025.

[193] William H Curley, Peter B Forgacs, Henning U Voss, Mary M Conte, and Nicholas D Schiff. Characterization of eeg signals revealing covert cognition in the injured brain. *Brain*, 141(5):1404–1421, 2018.

[194] Yelena G Bodien, Camille Chatelle, and Brian L Edlow. Functional networks in disorders of consciousness. In *Seminars in neurology*, volume 37, pages 485–502. Thieme Medical Publishers, 2017.

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

[196] Srivas Chennu, Paola Finoia, Evelyn Kamau, Judith Allanson, Guy B Williams, Martin M Monti, Valdas Noreika, Aurina Arnatkeviciute, Andrés Canales-Johnson, Francisco Olivares, et al. Spectral signatures of reorganised brain networks in disorders of consciousness. *PLoS computational biology*, 10(10):e1003887, 2014.

[197] Huan Yang, Hang Wu, Lingcong Kong, Wen Luo, Qiuyou Xie, Jiahui Pan, Wuxiu Quan, Lianting Hu, Dantong Li, Xuehai Wu, et al. Precise detection of awareness in disorders of consciousness using deep learning framework. *NeuroImage*, 290:120580, 2024.

[198] Zihan Wang, Junqi Yu, Jiahui Gao, Yang Bai, and Zhijiang Wan. Mutapt: A multi-task pre-trained transformer for classifying state of disorders of consciousness using eeg signal. *Brain Sciences*, 14(7):688, 2024.

[199] Charlotte Maschke, Catherine Duclos, Adrian M Owen, Karim Jerbi, and Stefanie Blain-Moraes. Aperiodic brain activity and response to anesthesia vary in disorders of consciousness. *NeuroImage*, 275:120154, 2023.

[200] Yong Wang, Zikang Niu, Xiaoyu Xia, Yang Bai, Zhenhu Liang, Jianghong He, and Xiaoli Li. Application of fast perturbational complexity index to the diagnosis and prognosis for disorders of consciousness. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 30:509–518, 2022.

[201] Charlotte Maschke, Jordan O'Byrne, Michele Angelo Colombo, Melanie Boly, Olivia Gosseries, Steven Laureys, Mario Rosanova, Karim Jerbi, and Stefanie Blain-Moraes. Critical dynamics in spontaneous eeg predict anesthetic-induced loss of consciousness and perturbational complexity. *Communications Biology*, 7(1):946, 2024.

[202] Abraham Lempel and Jacob Ziv. On the complexity of finite sequences. *IEEE Transactions on information theory*, 22(1):75–81, 1976.

[203] Sigurd L Alnes, Marzia De Lucia, Andrea O Rossetti, and Athina Tzovara. Complementary roles of neural synchrony and complexity for indexing consciousness and chances of surviving in acute coma. *NeuroImage*, 245:118638, 2021.

[204] F Kaspar and HG Schuster. Easily calculable measure for the complexity of spatiotemporal patterns. *Physical review A*, 36(2):842, 1987.

[205] Maciej Kamiński, Mingzhou Ding, Wilson A Truccolo, and Steven L Bressler. Evaluating causal relations in neural systems: Granger causality, directed transfer function and statistical assessment of significance. *Biological cybernetics*, 85:145–157, 2001.

[206] Charles W Anderson, Erik A Stolz, and Sanyogita Shamsunder. Multivariate autoregressive models for classification of spontaneous electroencephalographic signals dur-

ing mental tasks. *IEEE Transactions on Biomedical Engineering*, 45(3):277–286, 1998.

[207] Gideon Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.

[208] Tapio Schneider and Arnold Neumaier. Algorithm 808: Arfit—a matlab package for the estimation of parameters and eigenmodes of multivariate autoregressive models. *ACM Transactions on Mathematical Software (TOMS)*, 27(1):58–65, 2001.

[209] G Pfurtscheller and C Neuper. Motor Imagery and Direct Brain-Computer Communication. *Proc. IEEE*, 89:1123–1134, 2001.

[210] Gert Pfurtscheller. Functional brain imaging based on erd/ers. *Vision research*, 41(10-11):1257–1260, 2001.

[211] A Carballedo, M Doyle, G Lavelle, J Gormley, V O'keane, and T Frodl. Affective network hyperconnectivity and hypoconnectivity of cognitive control and ventral attention networks in adults with high neuroticism scores. *European Psychiatry*, 30(S1):1–1, 2015.

[212] Athena Demertzi, Aaron Kucyi, Adrián Ponce-Alvarez, Georgios A Keliris, Susan Whitfield-Gabrieli, and Gustavo Deco. Functional network antagonism and consciousness. *Network Neuroscience*, 6(4):998–1009, 2022.

[213] Yelena G Bodien, Cecilia A Carlowicz, Camille Chatelle, and Joseph T Giacino. Sensitivity and specificity of the coma recovery scale–revised total score in detection of conscious awareness. *Archives of physical medicine and rehabilitation*, 97(3):490–492, 2016.

[214] Robert G Kowalski, Flora M Hammond, Alan H Weintraub, Risa Nakase-Richardson, Ross D Zafonte, John Whyte, and Joseph T Giacino. Recovery of consciousness and

functional outcome in moderate and severe traumatic brain injury. *JAMA neurology*, 78(5):548–557, 2021.

[215] Betty Wutzl, Stefan M Golaszewski, Kenji Leibnitz, Patrick B Langthaler, Alexander B Kunz, Stefan Leis, Kerstin Schwenker, Aljoscha Thomschewski, Juergen Bergmann, and Eugen Trinka. Narrative review: quantitative eeg in disorders of consciousness. *Brain Sciences*, 11(6):697, 2021.

[216] Aurelie Selfslagh, Solaiman Shokur, Debora SF Campos, Ana RC Donati, Sabrina Almeida, Seidi Y Yamauti, Daniel B Coelho, Mohamed Bouri, and Miguel AL Nicolelis. Non-invasive, brain-controlled functional electrical stimulation for locomotion rehabilitation in individuals with paraplegia. *Scientific reports*, 9(1):6782, 2019.

[217] Sebastian Halder, Ana Matran-Fernandez, Rab Nawaz, Marina Lopes da Silva, Tommaso Bertoni, Jean-Paul Noel, Jane Jöhr, Andrea Serino, Karin Diserens, Reinhold Scherer, and Serafeim Perdikis. To repeat or not to repeat? erp-based assessment of the level of consciousness - a case study. 2024.

[218] Martin Spüler. Questioning the evidence for bci-based communication in the complete locked-in state. *PLOS Biology*, 17(4):1–5, 04 2019.

[219] Quentin Noirhomme, Damien Lesenfants, Francisco Gomez, Andrea Soddu, Jessica Schrouff, Gaëtan Garraux, André Luxen, Christophe Phillips, and Steven Laureys. Biased binomial assessment of cross-validated estimation of classification accuracies illustrated in diagnosis predictions. *NeuroImage: Clinical*, 4:687–694, 2014.

[220] Serafeim Perdikis, Fabien Bourban, Vincent Rouanne, José del R. Millán, and Robert Leeb. Effects of data sample dependence on the evaluation of BCI performance. In *7th International BCI Meeting*. International BCI Society, 2018.

[221] Jitka Annen, Maddalena M Filippini, Estelle Bonin, Helena Cassol, Charlène Aubinet, Manon Carrière, Olivia Gosseries, Aurore Thibaut, Alice Barra, Audrey Wolff, et al. Diagnostic accuracy of the crs-r index in patients with disorders of consciousness. *Brain injury*, 33(11):1409–1412, 2019.

[222] Alison O'Shea, Rehan Ahmed, Gordon Lightbody, Elena Pavlidis, Rhodri Lloyd, Francesco Pisani, Willian Marnane, Sean Mathieson, Geraldine Boylan, and Andriy Temko. Deep learning for eeg seizure detection in preterm infants. *International journal of neural systems*, 31(08):2150008, 2021.

[223] Weiming Chen, Yiqing Luo, and Jie Wang. Three-branch temporal-spatial convolutional transformer for motor imagery eeg classification. *IEEE Access*, 12:79754–79764, 2024.

[224] Sagnik De, Anurag Singh, and Ashish Kumar Bhandari. A novel vision transformer based multimodal fusion approach for clinical mdd diagnosis using eeg and audio signals. *IEEE Transactions on Computational Biology and Bioinformatics*, 2025.

[225] Mariam Khayretdinova, Alexey Shovkun, Vladislav Degtyarev, Andrey Kiryasov, Polina Pshonkovskaya, and Ilya Zakharov. Predicting age from resting-state scalp eeg signals with deep convolutional neural networks on td-brain dataset. *Frontiers in aging neuroscience*, 14:1019869, 2022.

[226] Federico Del Pup, Andrea Zanola, Louis Fabrice Tshimanga, Alessandra Bertoldo, Livio Finos, and Manfredo Atzori. The role of data partitioning on the performance of eeg-based deep learning models in supervised cross-subject analysis: a preliminary study. *Computers in Biology and Medicine*, 196:110608, 2025.

[227] Yuchao He, Xin Wang, Zijian Yang, Lingbin Xue, Yuming Chen, Junyu Ji, Feng Wan, Subhas Chandra Mukhopadhyay, Lina Men, Michael Chi Fai Tong, et al. Clas-

sification of attention deficit/hyperactivity disorder based on eeg signals using a eeg-transformer model. *Journal of Neural Engineering*, 20(5):056013, 2023.

[228] Banghua Yang, Fenqi Rong, Yunlong Xie, Du Li, Jiayang Zhang, Fu Li, Guangming Shi, and Xiaorong Gao. A multi-day and high-quality eeg dataset for motor imagery brain-computer interface. *Scientific Data*, 12(1):488, 2025.

[229] Hyeonyeong Nam, Jun-Mo Kim, WooHyeok Choi, Soyeon Bak, and Tae-Eui Kam. The effects of layer-wise relevance propagation-based feature selection for eeg classification: a comparative study on multiple datasets. *Frontiers in Human Neuroscience*, 17:1205881, 2023.

[230] Min-jae Kim, Young Chul Youn, and Joonki Paik. Deep learning-based eeg analysis to classify normal, mild cognitive impairment, and dementia: Algorithms and dataset. *NeuroImage*, 272:120054, 2023.

[231] Marleen C Tjepkema-Cloostermans, Catarina da Silva Lourenço, Barry J Ruijter, Selma C Tromp, Gea Drost, Francois HM Kornips, Albertus Beishuizen, Frank H Bosch, Jeannette Hofmeijer, and Michel JAM van Putten. Outcome prediction in postanoxic coma with deep learning. *Critical care medicine*, 47(10):1424–1432, 2019.

[232] Gaowei Xu, Xiaoang Shen, Sirui Chen, Yongshuo Zong, Canyang Zhang, Hongyang Yue, Min Liu, Fei Chen, and Wenliang Che. A deep transfer convolutional neural network framework for eeg signal classification. *IEEe Access*, 7:112767–112776, 2019.

[233] Ali Al-Saegh, Shefa A Dawwd, and Jassim M Abdul-Jabbar. Deep learning for motor imagery eeg-based classification: A review. *Biomedical Signal Processing and Control*, 63:102172, 2021.

[234] Yannick Roy, Hubert Banville, Isabela Albuquerque, Alexandre Gramfort, Tiago H

Falk, and Jocelyn Faubert. Deep learning-based electroencephalography analysis: a systematic review. *Journal of neural engineering*, 16(5):051001, 2019.

[235] Alexander Craik, Yongtian He, and Jose L Contreras-Vidal. Deep learning for electroencephalogram (eeg) classification tasks: a review. *Journal of neural engineering*, 16(3):031001, 2019.

[236] Demetres Kostas, Stephane Aroca-Ouellette, and Frank Rudzicz. Bendr: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of eeg data. *Frontiers in Human Neuroscience*, 15:653659, 2021.

[237] Zied Tayeb, Juri Fedjaev, Nejla Ghaboosi, Christoph Richter, Lukas Everding, Xingwei Qu, Yingyu Wu, Gordon Cheng, and Jörg Conradt. Validating deep neural networks for online decoding of motor imagery movements from eeg signals. *Sensors*, 19(1):210, 2019.

[238] Daniel Freer and Guang-Zhong Yang. Data augmentation for self-paced motor imagery classification with c-lstm. *Journal of neural engineering*, 17(1):016041, 2020.

[239] Elnaz Lashgari, Dehua Liang, and Uri Maoz. Data augmentation for deep-learning-based electroencephalography. *Journal of Neuroscience Methods*, 346:108885, 2020.

[240] Thorir Mar Ingolfsson, Michael Hersche, Xiaying Wang, Nobuaki Kobayashi, Lukas Cavigelli, and Luca Benini. Eeg-tcnet: An accurate temporal convolutional network for embedded motor-imagery brain–machine interfaces. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2958–2965. IEEE, 2020.

[241] Christina Sartzetaki, Panagiotis Antoniadis, Nick Antonopoulos, Ioannis Gkinis, Agamemnon Krasoulis, Serafeim Perdikis, and Vassilis Pitsikalis. Beyond within-subject performance: A multi-dataset study of fine-tuning in the eeg domain. In

2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pages 4429–4435. IEEE, 2023.

[242] Gita Sarafraz, Armin Behnamnia, Mehran Hosseinzadeh, Ali Balapour, Amin Meghrazi, and Hamid R Rabiee. Domain adaptation and generalization of functional medical data: A systematic survey of brain data. *ACM Computing Surveys*, 56(10):1–39, 2024.

[243] Elisa Visani, Gianvittorio Luria, Davide Sattin, Davide Rossi Sebastiano, Stefania Ferraro, Ferruccio Panzica, Matilde Leonardi, and Silvana Franceschetti. Entropy metrics correlating with higher residual functioning in patients with chronic disorders of consciousness. *Brain Sciences*, 12(3):332, 2022.