Artificial intelligence techniques for studying neural functions in coma and sleep disorders

Inaugural dissertation
of the Faculty of Science,
University of Bern

presented by

Florence Marcelle Aellen

from Saanen, Bern

Supervisor

Ass. Prof. Dr. Athina Tzovara

Institute of Computer Science

Artificial intelligence techniques for studying neural functions in coma and sleep disorders

Inaugural dissertation
of the Faculty of Science,
University of Bern

presented by

Florence Marcelle Aellen

from Saanen, Bern

Supervisor

Ass. Prof. Dr. Athina Tzovara

Institute of Computer Science

Accepted by the Faculty of Science

Bern, 31.03.2023

The Dean
Prof. Dr. Marco Herwegh

Abstract

The use of artificial intelligence in computational neuroscience has increased within the last years. In the field of electroencephalography (EEG) research machine and deep learning, models show huge potential. EEG data is high dimensional, and complex models are well suited for their analysis. However, the use of artificial intelligence in EEG research and clinical applications is not yet established, and multiple challenges remain to be addressed. This thesis is focused on analyzing neurological EEG signals for clinical applications with artificial intelligence and is split into three sub-projects.

The first project is a methodological contribution, presenting a proof of concept that deep learning on EEG signals can be used as a multivariate pattern analysis technique for research. Even though the field of deep learning for EEG has produced many publications, the use of these algorithms in research for the analysis of EEG signals is not established. Therefore for my first project, I developed an analysis pipeline based on a deep learning architecture, data augmentation techniques, and feature extraction method that is class and trial-specific. In summary, I present a novel multivariate pattern analysis pipeline for EEG data based on deep learning that can extract in a data-driven way trial-by-trial discriminant activity.

In the second part of this thesis, I present a clinical application of predicting the outcome of comatose patients after cardiac arrest. Outcome prediction of patients in a coma is today still an open challenge, that depends on subjective clinical evaluations. Importantly, current clinical markers can leave up to a third of patients without a clear prognosis. To address this challenge, I trained a convolutional neural network on EEG signals of coma patients that were exposed to standardized auditory stimulations. This work showed a high predictive power of the trained deep learning model, also on patients that were without a established prognosis based on existing clinical criteria. These results emphasize the potential of deep learning models for predicting outcome of coma and assisting clinicians.

In the last part of my thesis, I focused on sleep-wake disorders and studied whether unsupervised machine learning techniques could improve diagnosis. The field of sleep-wake disorders is convoluted, as they can cooccur within patients, and only a few disorders have clear diagnostic biomarkers. Thus I developed a pipeline based on an unsupervised clustering algorithm to disentangle the full landscape of sleep-wake disorders. First I reproduced previous results in a sub-cohort of patients with central disorders of hypersomnolence. The verified pipeline was then used on the full landscape of sleep-wake disorders, where I identified clear clusters of disorders with clear diagnostic biomarkers. My results call for new biomarkers, to improve patient phenotyping.

Acknowledgements

First I would like to thank my Ph.D. supervisor Athina Tzovara. I am very grateful for all the opportunities that she provided, as well as the support and advice I have received over the last four years. I have learned so much under your supervision and will profit from the acquired skills for years to come.

A very big thank you also goes to all members of our group: Sigurd Alnes, Pinar Göktepe-Kavis, Ruxandra Tivadar, Riccardo Cusinato, Thomas Rusterholz, Camille Mignardot. All of you made these four years so much more interesting. It was always a pleasure to get interrupted and start a conversation to have some fun, besides all the hard work we all put in. Additionally, I want to thank all the master's and bachelor's students, interns, and lab assistants that have made the environment at the lab more interactive.

This thank-you also goes to all the members of the Zen. I thank the PIs for creating an amazing environment for research in Bern. These years were full of collaborations and shared knowledge. I greatly appreciate all of the discussions we had and your feedback on my projects.

Many thanks also to the interfaculty research cooperation (IRC) - decoding sleep, for making this project possible and all of the interactions I shared with the members during this Ph.D.

I would like to thank all of my collaborators at the University of Bern, Inselspital, Chuv, and Retinai. Thank you for your collaboration, your input, and for allowing me to analyze your priceless datasets.

Many thanks also to Dragana Heinzen who made me feel very welcome when I started and for taking care of all the administrative tasks behind the scene. I am also very grateful for Paolo Favaro and his group. To all of you that welcomed me and helped me get started at the beginning of my Ph.D.

None of this would have been possible without the support of my family and friends. Thank you to my parents, my sister, and my brother for all of the time and the memories we have shared thus far. I am very grateful to always be able to find support with you and forget about the stress of work.

Last but not least, I would like to thank my partner Reto. You have supported me so much through this time and I will be forever grateful for all of it.

And lastly not forgetting our two cats, K & M, for sitting with me when everyone else was either offline or already gone to bed.

List of Abbreviations

Adam Adaptive Moment Estimation. 43, 69

AUC Area under the Receiver Operator Characteristic Curve. 13, 14, 17, 43, 44, 46–51, 69, 70, 72, 75, 78, 80, 135–137, 141–144

BCI Brain-Computer-Interfaces. 8, 35, 37, 58, 60, 99, 109

CDH Central Disorder of Hypersomnolence. i, 22–24, 26–29, 33, 83–85, 87, 88, 91–93, 96, 106–108, 146, 147, 161

CNN Convolutional Neural Network. 7, 8, 13, 31, 35–38, 40, 43–53, 56–61, 63, 65, 66, 68–70, 73, 75, 78, 79, 81, 100, 101, 103, 104, 135, 136, 139

CPC Cerebral Performance Category. 11–14, 17, 67, 69, 71, 72, 77, 78, 80, 104, 105

CSA Central Sleep Apnea. 24, 25, 32, 33, 88, 92–94, 96, 97, 106, 107, 113, 152–158, 161–163

CSP Common Spatial Patterns. 37, 59

EEG Electroencephalography. i, 1–19, 21, 29, 31, 32, 35–48, 51, 52, 55–61, 63–73, 75–82, 99–106, 108–111, 113, 134, 135, 137, 138

EMG Electromyography. 21, 108

EOG Electrooculography. 21, 108

ERP Event-Related Potential. 3–5, 15, 36, 45, 57, 64, 65, 76, 78, 81

ESS Epworth Sleepiness Scale. 22, 93

fMRI Functional Magnetic Resonance Imaging. 3, 5, 36

FSS Fatigue Severity Scale. 93–95

GCS Glasgow Coma Scale. 11, 12, 14, 66, 71

Grad-CAM Gradient-Weighted Class Activation Mapping. 10, 13, 101

HT Hypothermia. 11, 13, 17, 18, 64, 66, 71–75, 141

IH Idiopathic Hypersomnia. 22–24, 26, 28, 29, 33, 83–85, 88, 91, 96, 106–108, 146, 161

LSTM Long Short Term Memory. 14, 101

LZ Lempel-Ziv. 18, 71, 75–77, 104

MMN Mismatch Negativity. 15, 16

MSLT Multiple Sleep Latency Test. 21, 22, 24, 26, 85, 91, 108

MVPA Multivariate Pattern Analysis. 5–7, 9–11, 31, 35–38, 40, 44, 45, 49, 50, 56–59, 61, 99, 100, 102, 105, 113

MWT Maintenance of Wakefulness Test. 20, 85

NPV Negative Predictive Value. 70, 72, 75, 141, 143

NREM Non-Rapid Eye Movement. 21, 30, 94, 146

NT Normothermia. 11, 13, 17, 64, 66, 72–76, 81, 104, 141

NT1 Narcolepsy Type 1. 20, 22–24, 27–29, 32, 33, 83–85, 88, 91, 93–96, 106, 107, 113, 152–158, 161

NT2 Narcolepsy Type 2. 22–24, 26, 28, 29, 33, 83–85, 88, 91, 96, 97, 106, 107, 146, 161

OSA Obstructive Sleep Apnea. 20, 21, 24–26, 28, 29, 32, 33, 84, 86, 88–90, 92–94, 96, 97, 106, 107, 113, 152–158, 160–163

PLMS Periodic Limb Movement. 26, 90, 93, 163

PLV Phase-Locking Value. 18, 71, 75–77, 81, 103

PPV Positive Predictive Value. 12, 13, 16–18, 70, 72, 74, 75, 78, 80, 82, 141–143

PSG Polysomnography. 20, 21, 25, 30, 85, 108

ReLu Rectified Linear Unit. 40, 69

REM Rapid Eye Movement. 21, 22, 30, 91, 107, 161

ResNet50 Residual Neural Network. 31, 40, 41, 46, 48, 51, 52, 57, 61, 110, 136

RLS Restless Leg Syndrome. 22, 25, 89, 94, 96, 97, 107, 114, 146, 165

RNN Recurrent Neural Network. 7, 101

ROSC Return of Spontaneous Circulation. 71, 77, 78

SBD Sleep-Related Breathing Disorder. 24, 88, 92, 106, 146

SVM Support Vector Machine. 45, 46, 49–51

SWD Sleep-Wake Disorder. i, 1, 2, 19–22, 24–26, 28, 30–33, 83–85, 87–90, 92–97, 99, 106–108, 113, 114, 146–148, 150, 151

Contents

1 Introduction

The use of algorithms of artificial intelligence, especially machine, and deep learning models, has increased in the field of computational neuroscience in the last few years. These techniques are well-suited for analyzing big databases and complex data. In the field of neurology neural signals in the form of electroencephalography (EEG) are regularly collected and analyzed. They are high dimensional and contain complicated patterns. The analysis of these signals aims at discovering functions of the brain, and also has clinical applications, such as helping to diagnose patients with neurological disorders. The analysis of EEG data of patients in the clinical setting often requires clinical expertise and can be very time-consuming. For example, in sleep scoring, an overnight recording of physiological signals is by clinicians split into stages, making up the architecture of sleep. This information is then used for diagnosing sleep-wake disorders (Bargiotas et al., 2019; Schmidt et al., 2021), but is also crucial for questions in research topics studying sleep (Miskovic et al., 2018; Züst et al., 2019). Sleep scoring is a complex and time-consuming task that requires advanced clinical expertise and training. Another example is the assessment of EEG signals of patients in a comatose state. The manual examination of EEG signals by clinicians is tedious but essential to optimize patient care and to predict their outcome (Rossetti et al., 2016). A last example of the manual analysis of EEG signals in the clinics is the assessment of epileptic activity and seizure detection in epileptic patients. Seizure detection is essential for diagnosing epilepsy, and clinicians need to assess hours of recorded EEG signals (Cho and Jang, 2020). Additionally, there are many benefits of automated seizure detection, as they could, in the future, be used to predict seizures to warn patients (Burrello et al., 2020). The above examples from the clinics rely today on visual scoring by experts and subjective assessment of clinical data. They could benefit from automated algorithms, such as artificial neural networks, that could assist clinicians.

Artificial intelligence has revolutionized many fields, most related to imaging applications and speech or text generation. In the medical field, deep learning has found several applications for medical imaging data. For example, deep learning can accurately detect skin cancers from images (Brinker et al., 2019) or from pathological slides (Parwani, 2019). Deep learning can also be used to accurately segment three-dimensional scans of the retina (Fauw et al., 2018). However, for signals reflecting neural activity, artificial intelligence is, in comparison, underexplored. The modality of the data is different compared to images, and the application to EEG signals is not straightforward. Nevertheless, in recent years, publications on machine and deep learning for EEG signals have shown great potential (Roy et al., 2019; Craik et al., 2019; Zhang et al., 2019).

Several obstacles still need to be overcome before these algorithms can be used to analyze

EEG signals in research and clinical applications. To train deep artificial neural networks, tremendous amounts of data are required. Nowadays, big datasets of EEG data are still not always publicly available, except for data in clinical applications such as sleep scoring or seizure detection. First data collection of EEG data is time-consuming. Second, labeling large amounts of clinical data depends on clinical expertise and is tedious. Therefore big clinical datasets are expensive to build up. The risk with small datasets is that complex models will overfit on a limited number of patients, therefore generalizing poorly and potentially introducing bias. The field of computer vision offers multiple solutions to train artificial neural networks with small datasets (Goodfellow et al., 2016). However, an open challenge is integrating these solutions and models into neuroscience and using them to answer standing clinical questions, such as assisting clinical decision-making.

This introduces the first question this thesis addresses, namely assessing whether artificial deep neural networks can be used to analyze EEG signals for research and if features extracted from a trained model can be used to evaluate changes in EEG activity over the course of an experiment.

The second topic this thesis focuses on is a clinical application where deep learning on EEG signals can assist in predicting outcome of comatose patients after a cardiac arrest. Additionally, the focus is set on patients without a clear prognosis, according to current clinical makers, and it is explored whether deep learning models could give additional predictive information for this cohort.

The last topic of this thesis presents a machine learning algorithm applied to a second clinical application, namely on the diagnosis of sleep-wake disorders. In this case, it is investigated whether unsupervised machine learning techniques could assist in refining phenotypes of patients with sleep-wake disorders.

The introduction of this thesis is split into three parts, each introducing the relevant literature for each of the three topics. Chapter 1.1 introduces how neural signals in the form of EEG are measured in humans, and the use of machine and deep learning techniques for their analysis. Chapter 1.2 presents coma after cardiac arrest and the topic of outcome prediction, with a focus on deep learning models. Lastly, in chapter 1.3, the focus is set on sleep-wake disorders, the challenges with current diagnostic criteria, and how unsupervised machine learning algorithms can assist in identifying unique patient profiles.

1.1 Electroencephalography (EEG) and deep learning

This first section of the introduction outlines how neural signals of the human brain are generated and measured. The later sections describe how artificial neural networks can be used to analyze this data and establishes previous literature on this topic.

1.1.1 Measurement of EGG data

The brain processes information continuously and integrates environmental stimuli from all senses. Information is exchanged between neurons in the brain via electrical currents. Namely a neuron receives input from connected neurons via ions. It is polarized by aggregating these ions, and if the membrane potential rises above a threshold, an action potential, an electrical pulse of 100 mV and 1-2 seconds in duration, is generated, and the neuron is polarised (Bear, 2016). In the post-firing phase, the neuron is in a refractory period for around 10 ms, making it impossible for the neuron to fire again (Hari and Puce, 2017). With electroencephalography (EEG), the postsynaptic potentials (input into the soma and dendrites of the neuron) can be measured from neurons located in the cortex. EEG is a non-invasive recording technique using electrodes placed on the scalp (Bear, 2016). The cortex has a laminar structure of multiple layers, folded into itself to increase the surface, building gyri. The pyramidal neurons are in the cortex perpendicular to the surface, with the somas (neuronal bodies) in deeper layers of the cortex (Hari and Puce, 2017). Therefore the current of large neuronal ensembles firing simultaneously can be measured from the scalp with electrodes. The signal measured on the scalp with the EEG electrodes is a combination of currents from different origins. The biggest contribution comes from the neurons perpendicular to the scalp located directly below the skull, but tangential-oriented neurons and strong deeper currents can also contribute (Bear, 2016). One limitation of EEG recordings is that the measured currents originate from multiple sources. This makes it hard to localize the source of an EEG signal measured on the scalp, as it might have multiple contributions (Hari and Puce, 2017). The biggest limitation of EEG is its relatively poor spatial resolution compared to functional magnetic resonance imaging (fMRI) and other commonly used neuroimaging techniques. The poor spatial resolution is caused by the inhomogeneities of the skull and scalp that distort the signal and result in widespread patterns. The advantage of EEG is its high temporal resolution, which is needed when analyzing fast processes, such as responses to external stimulations, e.g., of auditory or visual nature. (Hari and Puce, 2017)

EEG measures an electrical potential difference between two points. For scalp EEG, one reference electrode is selected, such as the vertex (Cz) electrode or the mastoid. EEG signals in all the remaining electrodes are recorded with respect to this reference electrode. The data is usually re-referenced during preprocessing, most commonly to an average reference.

EEG is especially well suited for analyzing neural responses to external stimulations, commonly referred to as evoked responses. However, the low signal-to-noise ratio of the EEG signal makes the analysis of evoked responses to single external stimuli practically impossible. Therefore, usually, numerous single evoked responses are averaged to compute event-related potentials (ERPs) (Figure 1.1 for an exemplar auditory ERP of 64 elec-

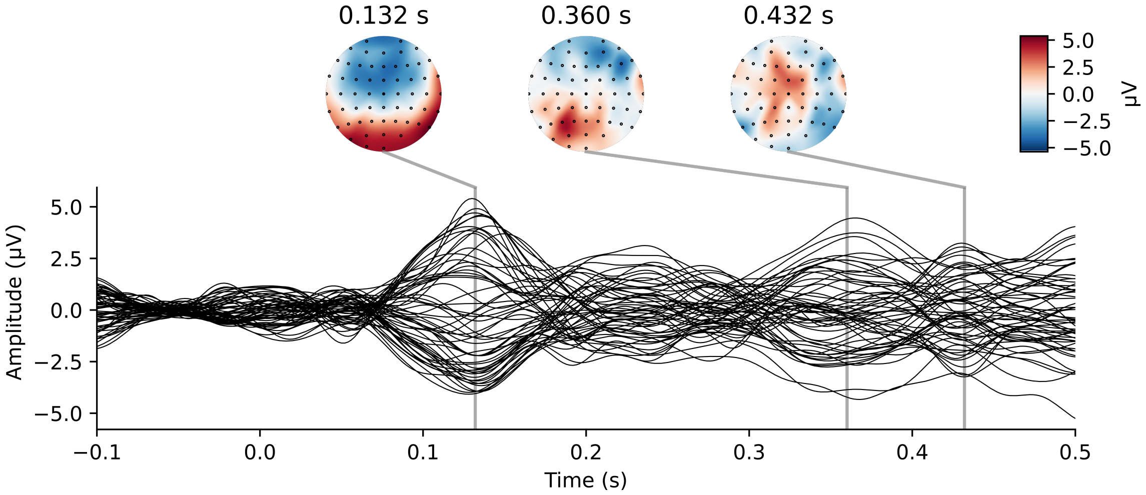


Figure 1.1: Auditory ERP in response to pure tones, as well as topographic maps for local maxima, from 100 ms to 500 ms after stimulus onset. The data was recorded with 64 electrodes and represented here as an average over 127 trials.

trodes). The analysis of ERPs is restricted to an interval around the presented stimulus, for example, 100 ms pre and 500 ms post-stimulus.

EEG signals are measured all over the scalp, with different numbers of electrodes, also called channels. There are some standard numbers of channels and montages, from 16 to 256 electrodes, that are most commonly used. EEG montages for clinical applications are usually restricted to a low number of electrodes. However, a larger number of channels is preferred for research purposes, as they can provide a better estimation of neural activity and its underlying sources (Michel et al., 2004). The recorded EEG data can also be represented as a topographic map, showing the distribution of electrical activity across the head, either at single time-points or averaged over short time intervals (see Figure 1.1 for an example).

In summary, the EEG data that will be of importance in this thesis is recorded with respect to a stimulation, primarily auditory but also visual. The data is epoched with respect to the stimulation, for example, 100 ms pre and 500 ms post-stimulus.

1.1.2 Analysis of EEG data

A mean auditory ERP displays characteristic components, such as a negative deflection at around 100 ms after the onset of an auditory stimulus (N100) or a positive peak at 300 ms (P300) over central and frontal electrodes, following the presentation of an unexpected

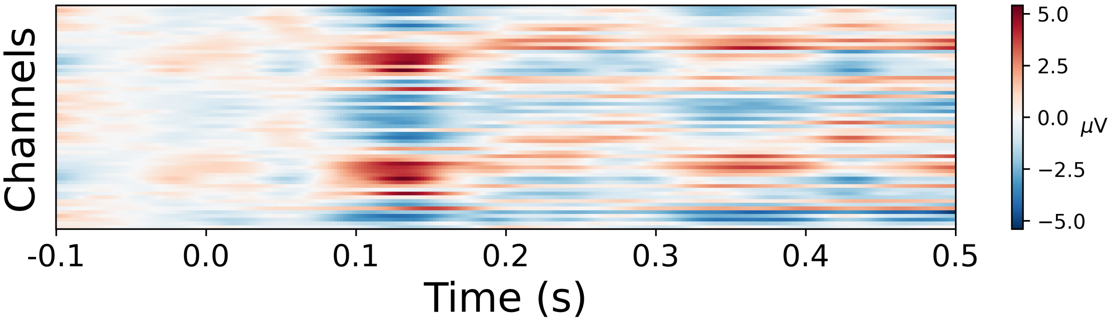


Figure 1.2: Auditory ERP in response to pure tones, represented here as a heatmap, from 100 ms to 500 ms after stimulus onset. The data was recorded with 64 electrodes and represented here as an average over 127 trials.

stimulus (Fonken et al., 2020). Traditional analysis of EEG data focuses on measuring response amplitudes or latencies and compares differences in activations between experimental conditions for single electrodes. For example, (Fischer et al., 1999) showed that patients in a comatose state were more likely to awake if they had a N100 response to auditory stimulations in the Fz electrode, compared to patients that did not have a clear N100 response. However, these analyses are restricted, as they focus on single electrodes and fixed latencies, which are defined based on findings in healthy controls, and may not reflect the characteristics of clinical populations.

1.1.3 Multivariate pattern analysis for EEG data

To address the limitations of analyzing average EEG responses in single electrodes, the field of neuroscience has introduced the use of machine learning techniques (Grootswagers et al., 2017; Kurth-Nelson et al., 2015). (Haynes, 2015) was among the first to propose this new set of analyses, referred to as multivariate pattern analysis (MVPA), initially developed for fMRI data. These analyses focus simultaneously on multiple electrodes and have great advantages compared to univariate analysis (Grootswagers et al., 2017). The integration of information from multiple electrodes increases the sensitivity of identifying differences between experimental conditions (for example response to man-made versus natural stimuli). Additionally, MVPA follow a data-driven approach, which allows for the detection of differences between experimental conditions with minimal a priori assumptions. The most common way of applying MVPA on EEG data consists of training a machine learning classifier to discriminate different conditions at different points in time. As training data, the EEG responses at a given time-point, as a multivariate observation over all recorded channels is used. Only a percentage of all the EEG data is used to train the classifiers, while

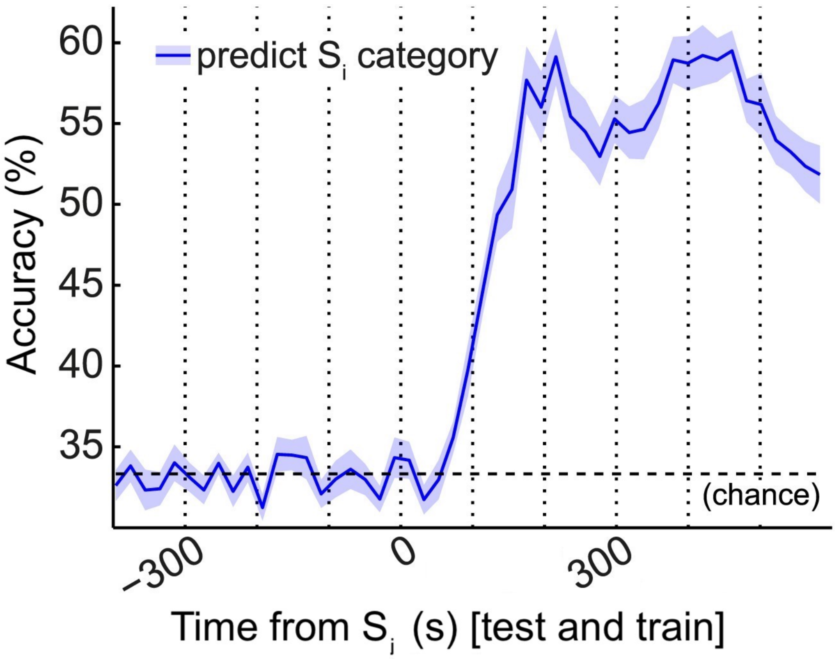


Figure 1.3: Decoding performance, a time course of accuracy. For each time-point, a classifier was trained and tested on an independent test set of EEG data at that time-point. The dashed line indicates the chance level. Adapted from (Kurth-Nelson et al., 2015), published under a "CC BY 4.0 DEED Attribution 4.0 International" licence.

the rest is used to validate and test the model to avoid overfitting. A performance score is then computed, to quantify the percentage of the held-out test trials that are correctly classified, per time-point. This results in a time course of performance scores (see Figure 1.3). Therefore, the analysis of EEG data with MVPA is data-driven, and discriminative information is automatically extracted. A classification score significantly above chance level suggests that the EEG data contained discriminative patterns of activity between the experimental conditions investigated. Identifying time-points with significant decoding performance allows the extraction of latencies of differential EEG responses between experimental conditions, in a data-driven way.

A trained classifier can also be used for further analysis to gain additional insights into the data. For example, the trained classifier from an early temporal latency may be used to test if similar patterns of neural activation can be found at different latencies (temporal generalization; King and Dehaene, 2014). This analysis can identify whether late patterns of sequences are already represented in earlier stages or if the early patterns are reactivated again later. Alternatively, the weights of a trained model together with the EEG data at the electrode level can be used in order to identify which brain regions mostly contribute to an above-chance classification for source localization (van de Nieuwenhuijzen et al., 2013).

However, the majority of existing MVPA algorithms for EEG rely on machine learning classifiers. These have the limitation that they only consider patterns at single latencies. Additionally, it is impossible to extract class or trial-specific information from a trained MVPA model. Most commonly, MVPA classifiers are trained per subject, where the train and test data are from the same subject but distinct sets containing different epochs. Therefore all extracted patterns are subject-specific and not extracted from whole group analysis. These limitations can be addressed with the use of more complex models, such as artificial neural networks.

1.1.4 Deep learning for EEG data

In the field of computer vision, state-of-the-art algorithms for analyzing images are based on deep learning models. Tasks for image analysis, such as image recognition, image segmentation, etc., no longer rely on linear classifiers but use deep multilayer models to achieve outstanding performances (Goodfellow et al., 2016). Following the success of deep learning models in computer vision, they have been used in numerous fields, among others, for medical applications. However, in the field of EEG analysis and especially MVPA, only sparse literature is available.

Similar to MVPA, deep learning models can also be used to analyze EEG signals in the context of answering research questions. Traditional MVPA algorithms suffer from the limitation that they focus on single latencies and are usually less suited for analysis performed on a group level, as inter-subject differences for most research-related tasks may be too big. However, deep learning models, especially convolutional neural networks (CNNs), can consider multiple latencies simultaneously and therefore capture inter-subject differences, such as delayed neural responses. Deep learning models have more trainable parameters than machine learning models and can extract more complex patterns and subtler differences between experimental conditions. Additionally, they are able to detect space-unlocked patterns, because of their translational equivariance property (Goodfellow et al., 2016). This allows for the integration of EEG activity from multiple channels and time-points across the whole trial.

Deep learning can be applied in multiple ways to EEG data. Even though EEG responses have a temporal component, one can also use CNNs with architectures similar to the ones used for image applications. The literature on deep learning for EEG data uses different types of network architectures, such as CNNs, recurrent neural networks (RNNs), autoencoders, or similar (Roy et al., 2019; Craik et al., 2019). RNNs exploit the temporal dimension of the data explicitly. Moreover, CNN architectures have been developed that explicitly use the temporality of the data, by convolutions along the temporal axis before integrating information from the spatial dimension (Lawhern et al., 2018; Schirrmeister et al., 2017). There are also different approaches to how the data is represented for training the networks. For example, the raw EEG signal (channels x time), frequency transforms (frequency x time), or pre-selected features can be used as input for the artificial neural networks (Roy et al., 2019). Each type of architecture and data representation has its advantages and limitations. This thesis focuses on CNNs trained on minimally processed EEG signals, with the shape of (channels x time).

The currently available literature on EEG data and deep learning focuses mainly on brain-computer-interfaces (BCI) (Tang et al., 2016; An et al., 2014), sleep scoring (Fiorillo et al., 2019; Vallat and Walker, 2021; Stephansen et al., 2018), seizure detection (Cho and Jang, 2020; Burrello et al., 2020) or other clinical applications ((Roy et al., 2019; Zhang et al., 2019; Craik et al., 2019) for reviews). Most deep learning studies for EEG that try to automate clinical tasks focus on classification performance. They have the goal to simplify time-consuming and tedious clinical work. However, in applications of deep learning for research purposes, the focus is slightly different, and the initial goal is to test whether classification performance is above chance. The magnitude and level of performance are usually of lesser importance, and the focus is more on feature interpretability.

1.1.5 Data augmentations for EEG data

Deep learning models can efficiently process high-dimensional data and extract more detailed patterns than machine learning methods. They are, therefore, very well suited to analyze EEG data, as that data has high sample frequency, with many temporal points and a large number of recorded electrodes. The bigger or deeper an artificial neural network is, the more data is required to train it, without facing the problem of overfitting. The greatest limitation of deep learning for EEG signals is still the limited amount of available data for training. For clinical applications, the amount of data is limited, because of two main reasons. On the one hand, the labeling of clinical datasets is highly time-consuming. For example, clinicians spend hours on the annotation of EEG recordings, to mark seizures. On the other hand, patient confidentiality often forbids the publication of these datasets. Therefore the acquisition of big publicly available clinical datasets is hard. For research-related purposes, the collection of task-specific datasets is also time and resource expensive.

To address the problems of limited data availability the field of computer vision proposes multiple solutions, one of which is data augmentations. The rationale behind data augmentations is to slightly and randomly distort each training sample and increase the available amount of training data. In computer vision, this would, for example, entail flipping an image vertically as the objects on pictures usually are physically meaningful in both orientations (see (Shorten and Khoshgoftaar, 2019) for a review of data augmentations for computer vision). However, the data augmentation techniques must fit the data and the selected model. For raw EEG signals, with a shape of (channels x times), a flipping of neither the time nor the channel dimension has physical meaning. Yet, flipping the channels from the left to right hemispheres might be more meaningful if the task does not involve hemisphere-specific activations (such as right versus left-hand movements). For the classification of continuous data, such as sleep staging or seizure detection, a sliding window over the data as an augmentation technique might greatly improve classification performance. However, this is not meaningful for epoched EEG signals in response to external stimulation.

Although data augmentations are commonly used for training artificial neural networks for images, their use in the field of EEG research is not established. (Lashgari et al., 2020) presents a review of data augmentation techniques for EEG data, and (Rommel et al., 2022) tested 13 different techniques previously introduced on two datasets (and deep learning models) to compare the improvement in the performance of these methods. The data augmentations in (Rommel et al., 2022) were split into time, frequency, and spatial augmentations. Time augmentations included: Gaussian noise (Wang et al., 2018), the replacement of an arbitrary interval of the signal with zero for all electrodes (Mohsenvand et al., 2020), sign flip (Rommel et al., 2021) and time reversal (Rommel et al., 2021). The frequency augmentations the authors selected were frequency shift (Rommel et al., 2021), random shifts in the phase of the EEG signals (Schwabedal et al., 2018), and bandstop filtering (Cheng et al., 2020; Mohsenvand et al., 2020). Last, the spatial augmentations entailed the switching of hemispheres (Deiss et al., 2018), channel dropout (Saeed et al., 2021), channel shuffling (Saeed et al., 2021), and sensor rotation (Krell and Kim, 2017) (rotating the sensor positions around one of the axis, to simulate a rotation of the EEG cap during data collection). (Rommel et al., 2022) has shown that the random shift of phases in the Fourier space had the biggest improvement in the classification performance of at least 45% on small datasets. For a dataset with two electrodes, the spatial augmentation techniques resulted in meager improvements and no improvement was observed with the sensor rotation technique. This highlights that the augmentation technique needs to match the data, network architecture, and task.

Despite the vast potential of data augmentations for EEG signals, the limited amount of available studies focuses on clinical applications. These techniques remain under-explored for use in research as MVPA decoding algorithms.

1.1.6 Feature visualisations for EEG data

The use of MVPA in research aims, on the one hand, to show an above-chance performance of a classifier, indicating a difference between the experimental conditions. But on the other hand, a trained model has the big advantage that the weights from the classifier can be extracted, giving information on which channels were most informative for the model's decision for different time-points. Unfortunately, this information is neither class nor trial specific. A big advantage of deep learning models is that the prediction of the network can be backpropagated through the network to give trial-specific information about which parts of the input were most important for the model's decision. Aggregated information over multiple trials can then give class-specific features. Feature visualizations of neural networks are important to gain insight into the decision of the models and can increase trust. This is especially important for clinical applications, and additional features can guide clinicians towards important information, for example, as in (Jonas et al., 2019). However, the use of features is also hugely beneficial for research. They allow the extraction of information about which channels or time ranges are most discriminant between conditions. For example, the activations of a trained artificial neural network can expose the most important brain regions for the processes under exploration.

The most common feature visualization methods for deep learning used in computer vision are saliency maps (Simonyan et al., 2013), gradient-weighted class activation mapping (Grad-CAM) (Selvaraju et al., 2019) or kernel visualizations (Goodfellow et al., 2016) and variations thereof. The first two rely on backpropagation, and the latter depends on the forward pass only and involves visualizing kernels of convolutional layers. These techniques have been previously used in the field of EEG. For example (Farahat et al., 2019; Vahid et al., 2020) used saliency maps, (Ghosh et al., 2018; Jonas et al., 2019) a variation of Grad-CAM, and kernel visualisations were used by (Nurse et al., 2016; Schirrmeister et al., 2017; Lawhern et al., 2018). However (Schirrmeister et al., 2017; Ghosh et al., 2018; Lawhern et al., 2018; Farahat et al., 2019; Vahid et al., 2020) presented features on an average level and (Nurse et al., 2016; Lawhern et al., 2018; Jonas et al., 2019) only displayed exemplar single trials. Previous literature has not yet explored how representative features are with respect to the entire dataset or how they change over time.

In summary, deep learning for EEG data shows enormous potential for applications in research as an MVPA method. The more complex models can outperform traditional MVPA techniques. This thesis addresses several open questions in chapter 2. As a proof of concept, this chapter first shows that deep learning can be used as an MVPA method and outperforms traditional MVPA algorithms. Second, three different types of data augmentation techniques are investigated, to overcome the limited data available in the datasets used. Third, the presented MVPA pipeline is validated on two different datasets, in the auditory and visual domains respectively, to discriminate between standard and deviant stimuli.

The stimulations for this task were presented to participants as a deviance paradigm (Section 1.2.3.1). Forth, saliency maps (Simonyan et al., 2013) are used to extract trial-specific information and aggregate them to show class-specific differences. Last, a potential application of MVPA techniques is presented, that can be used to test theories of learning over the course of an experiment purely based on EEG data.

1.2 Coma after cardiac arrest

In the first clinical application of this thesis, I focus on postanoxic patients in a comatose state. Based on deep learning models, the outcome of these patients is predicted. This chapter introduces coma after cardiac arrest and outcome prediction based on currently used clinical markers and novel deep learning models.

Coma after cardiac arrest is a prevalent medical problem today, and within recent years cases have been increasing steadily (Holmberg et al., 2019). As the medical field has advanced and new technologies and medications have emerged, patients are more likely to survive a cardiac arrest. They are therefore more often admitted to intensive care units of hospitals (Rossetti et al., 2016). It has thus become increasingly important to make precise and informed decisions about treatment and maintaining live support for these patients. One of the recent advancements increasing patients' chances of survival is the implementation of targeted temperature management (Rossetti et al., 2016). This protocol decreases the body temperature down to either 33°C, called hypothermia (HT), or 36°C, called normothermia (NT), within the first 24 hours. A decreased body temperature reduces the metabolic rate, and this protocol has the advantage of preserving brain functions (Holzer, 2010). After the first 24 hours of post-anoxic coma, the targeted temperature management is stopped, and patients are rewarmed to their normal body temperature. The standard temperature treatment used to be hypothermia between 32°C and 34°C. However, this was adjusted to 32 °C - 36 °C in 2016 (Donnino et al., 2016). The different temperature treatments each influence brain signals recorded for the patients during clinical evaluations differently (Madhok et al., 2012; Choi et al., 2012). Therefore different temperature treatments might influence clinical assessments and outcome predictions.

The Glasgow Coma Scale (GCS) measures the depth of the coma/unresponsiveness of patients and is evaluated on different scales for eye-opening, motor, and verbal response (Teasdale and Jennett, 1974). Patients are considered to be in a comatose state if their GCS measure is lower than eight.

One of the ways to evaluate outcome of coma is measured by the Cerebral Performance Category (CPC) (Booth et al., 2004), assessed some months after cardiac arrest. A CPC of 1 indicates that a patient has fully recovered and their health is back to the same level as before the hospitalization. A patient with a CPC of 2 shows some moderate impairments, whereas one with a CPC of 3 has severe impairments and requires considerable assistance in daily life. A CPC of 4 indicates a vegetative state, which might be more prevalent in certain countries than others. For example, it is more common in the US to keep patients in a vegetative state alive for years, whereas, in Switzerland, this rarely happens (Stretti et al., 2018; Hirsch, 2005). CPC 5 is a patient who deceased naturally or by stopping life support.

1.2.1 Outcome prediction based on current clinical markers

Outcome prediction has become increasingly important in the intensive care unit for optimal treatment as well as maintenance of live support. It is also important for families of patients to get precise information. Currently used predictors in clinics are based on a multi-modal assessments. This includes the absence of motor response in reaction to pain, which is characterized by a score of three or less on the Glasgow motor scale (a subcategory of the GCS). The absence of a motor response is a reliable indicator of bad outcome at 72 hours (FPRs 24%) (Rossetti et al., 2016). Second, the absence of a pupil or corneal reflexes (short brainstem reflexes) in response to light indicates a bad outcome at 72 hours after coma onset (FPR 0.5%). However, the presence of brainstem reflexes does not necessarily predict good outcome, and the positive predictive value for this marker is especially bad at 72 hours (positive predictive value (PPV) 61%) (Rossetti et al., 2016). The next types of predictors are all based on the EEG signal recorded at the bedside of comatose patients. A third predictor, called discontinuous or suppressed EEG background, describes intervals of attenuated or suppressed EEG during the absence of any stimulation, which is also a predictor for bad outcome at 24 hours (FPRs 0%). However, a continuous background as early as 12 hours is associated with a good outcome (PPV 92%) (Rossetti et al., 2016). An unreactive EEG in response to auditory stimulation, such as a clap, measured by the lack of change in amplitude or frequency, is a predictor for poor outcome (FPRs 7%) (Rossetti et al., 2016). A reactive EEG is, in this case, associated with a positive outcome (PPV 86% - 78%, depending on the temperature treatment). In summary, several clinical predictors exist, but all of these rely on clinical expertise, can be time consuming to evaluate, and are most predictive of non survival (Rossetti et al., 2016; Benghanem et al., 2022).

An open challenge of multi-model approaches based on existing markers is that they can leave patients without a clear prognosis. Based on four clinical markers, brainstem reflexes, motor response, unreactive EEG background, and discontinuous EEG background, patients with an inconclusive prognosis can be defined (Rossetti et al., 2016; Perkins et al., 2021; Westhall et al., 2015). If the outcome prediction of these four markers does not agree one can describe a 'grey zone' of patients without a clear prognosis. For example, a patient with a motor response and brainstem reflex, but a discontinuous and unreactive EEG background, would be in an uncertain state. This criterion can leave up to a third of patients without a clear prognosis. There is, therefore, a demand for new markers that would help predictions for patients in an uncertain state.

1.2.2 Deep learning for prognostication of coma outcome

Prediction of coma outcome is a manual, tedious, and time consuming task that requires a great deal of clinical expertise. Therefore new techniques based on machine or deep learning have been proposed to automate and simplify this work. Multiple deep learning models have been proposed for predicting the outcome of coma within recent years (Jonas et al., 2019; Tjepkema-Cloostermans et al., 2019; Zheng et al., 2021; Altıntop et al., 2022). Most of these publications focus on resting state EEG recordings, i.e., recordings in absence of stimulations.

In one of the first attempts to predict the outcome of postanoxic coma, (Jonas et al., 2019) attempted a classification of 10-second signals recorded from a 5-minute recording without any epileptic artifacts and in the absence of external stimulations. Patients in their cohort were treated with hypothermia and normothermia, and outcome were defined as favorable (CPC 1 and 2) and unfavorable (CPC 3-5). The neural network architecture was a CNN architecture previously used for images but viewed the signal as one-dimensional. However, it allowed for a mixture of channel information already within the first layer. (Jonas et al., 2019) reached a final area under the Receiver Operator Characteristic curve (AUC) score of 0.89 on a test set of 54 patients. They also provided a feature extraction approach by producing feature maps, based on Grad-CAM (Selvaraju et al., 2017) indicating for 26-time-points in each 10 s epoch and each class the strength of the activation of the neural network. Implementing this in the clinics would have the advantage of displaying epochs of data that indicate good or bad outcome, drawing the clinician's attention to essential data segments. However, this method has the weakness that it requires a clinician to select data segments without epileptic artifacts, therefore it is not possible to implement this algorithm in a fully automated pipeline.

In another attempt to use deep learning for outcome prediction, (Tjepkema-Cloostermans et al., 2019) built their model based on a CNN architecture previously used for images and predicted good (CPC 1 and 2) versus bad (CPC 3-5) outcome, both at 12 and 24 hours after cardiac arrest. Similar to (Jonas et al., 2019) they selected a 5-minute artifact-free interval within either 12 or 24 hours and used a 10-second epoch of this data to predict patients' outcome. They reached an AUC score of 0.92 at 12 and 0.88 at 24 hours after a cardiac arrest on an external validation set of 234 patients. At 12 hours good outcome was predicted most accurately with a PPV of 89% and at 24 hours with 65% (Tjepkema-Cloostermans et al., 2019). This study also faces the limitation of not having a fully automated data selection pipeline.

(Zheng et al., 2021) evaluated the classification performance of an artificial neural network from 12 to 96 hours after cardiac arrest to predict the outcome. The novelty of this study was the continuous integration of more information as is standard in clinical evaluations. Their cohort spanned a total of 1'038 patients over seven different centers. This study defined good outcome as CPC 1 and 2, and CPC 3-5 as an unfavorable outcome, which includes patients with major impairments in daily life (CPC 3), patients in a vegetative state (CPC 4), and patients that deceased (CPC 5). From their recordings, for every 5 minutes time window, nine clinically interpretable EEG features were extracted, such as burst suppression ratio, Shannon entropy (Shannon, 1948), or delta (0.1 - 3.5 Hz) band power. These were then used to train the neural network. The best performing architecture was based on bidirectional long short term memory (LSTM) layers and reached an AUC score of 0.78 at 12 hours and a maximum of 0.88 at 66 hours after cardiac arrest (Zheng et al., 2021). With the preselection of interpretable features, this study already restricted the amount of information the neural network could learn from the data. Additionally, their finding of a maximally predictive score of 0.88 at 66 hours after cardiac arrest contradicts the findings from (Jonas et al., 2019) and (Tjepkema-Cloostermans et al., 2019), reaching AUC scores of above 0.90 within 24 hours after cardiac arrest. (Tjepkema-Cloostermans et al., 2019) even found a better predictive power at 12 hours compared to 24 hours.

(Altıntop et al., 2022) attempted a different approach and tried to predict the levels of consciousness based on the patient's GCS score, divided into two levels: low consciousness (GCS 3-5) and high consciousness (GCS 6-8). The dataset contained 39 comatose patients recorded in different states, during resting state (3 times 5 minutes), family interaction (5 minutes), and tactile stimulation (hand touching) by a nurse (5 minutes) for a total of 35 minutes. Each of the 5-minute segments and four of the signals (recording electrodes, with bipolar reference) recorded were used as input to their model, which had a one dimensional convolutional architecture with convolutional and fully connected layers. The study reached an F1-score of 0.85, and an accuracy of 0.83 (Altıntop et al., 2022). The extensive recording protocol makes a possible implementation in a clinical routine difficult. Additionally, while the prediction of the level of consciousness is interesting, it is non-trivial to link the level of consciousness to the outcome of coma.

In summary, existing deep learning techniques reach excellent performances in predicting outcome of coma. However, the majority of existing approaches require either a manual selection of an appropriate time interval for the analysis (Jonas et al., 2019; Tjepkema-Cloostermans et al., 2019) or are based on complicated paradigms where implementation within the clinics would be more difficult, such as interactions with family or nurses (Altıntop et al., 2022). Additionally, the studies actually predicting outcome are based on resting state EEG recordings (Jonas et al., 2019; Tjepkema-Cloostermans et al., 2019; Zheng et al., 2021). The information these studies use is the same that current clinical markers are built on, EEG of patients in the absence of stimulations. Therefore these algorithms will not give additional insight for patients without a clear prognosis based on available clinical markers. However, none of these studies has tested the performance of their models on the subset of patients without a clear prognosis. To address the challenges of predicting outcome of patients in a 'grey zone', novel paradigms are needed to gain additional insight into the neural processes of the brain in a comatose state.

1.2.3 Auditory processing in coma

1.2.3.1 Deviance paradigm

An alternative way to trigger neural functions in coma is through the auditory pathway, by measuring EEG responses to auditory stimuli. The deviance paradigm is a commonly used paradigm in EEG research in particular in coma patients, where a sequence consisting primarily of a single stimulus, for example, of auditory or tactile nature, is presented. This stimulus is, for instance, a pure tone at a specific frequency and is often called a standard stimulus. Once in a while, this repeated presentation is broken, and the standard stimulus is replaced with a deviant stimulus. An example based on auditory stimulations can be found in (Fischer et al., 1999) where participants and coma patients were exposed to a sequence of pure tones at 800 Hz and duration of 75 ms (standard tones), interlaced with 30 ms long deviant tones. During the stimulation protocol, as the regularity in presentation is broken, and a new stimulus is presented, the brain produces an error signal. Comparing the EEG signal of a standard versus a deviant sound reveals a significant difference in the ERP. The error signal calculated as the difference between the response to standard and deviant stimulus is called mismatch negativity (MMN), peaking at around 100-250 ms after stimulus onset (Tivadar et al., 2021), see Figure 1.4 for a mean MMN over a group of healthy participants.

The MMN can also be detected in the absence of attention (Hari and Puce, 2017), which makes the deviance paradigm ideal for studying different levels of consciousness. It has been shown that the MMN response exists in sleep (Nir et al., 2015), anesthesia (Nourski et al., 2018), and coma (Fischer et al., 1999). (An et al., 2021) also demonstrated that the MMN response was present for different types of deviant stimuli and that the topography of MMN responses differed depending on the acoustic feature.

1.2.3.2 Outcome prediction based on mismatch negativity

Responses to auditory stimulations and especially the MMN of a deviance paradigm can be used to assess the integrity of neural functions in coma (Kane et al., 2000; Daltrozzo et al., 2007; Morlet and Fischer, 2014). Among the first to use the MMN as a predictor of the outcome of coma was (Kane et al., 1993) and (Fischer et al., 1999). The later

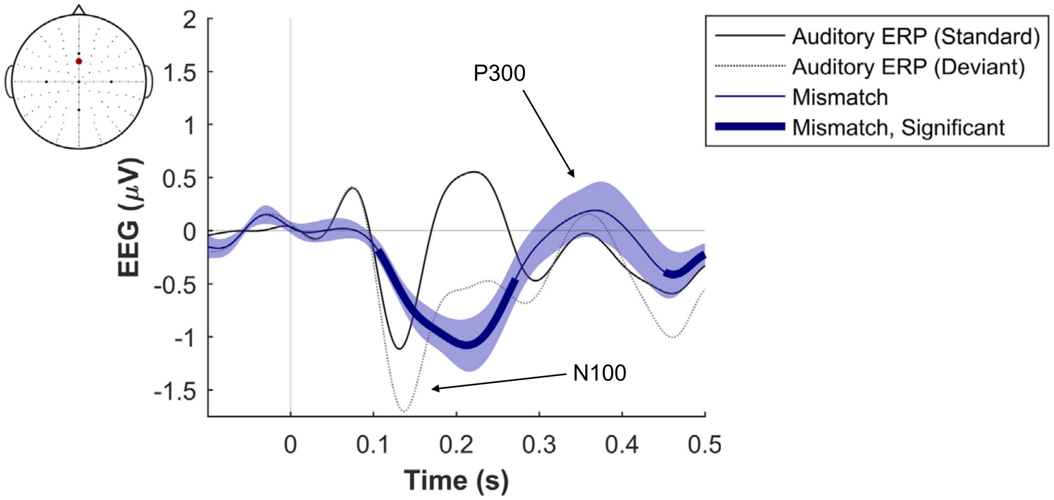


Figure 1.4: Mean auditory response to standard (continuous line) and deviant (dotted line) across a group of 19 healthy participants. Mean MMN in blue with the standard error as the shaded area, dark blue shows a significant difference between standard and deviant responses. Adapted from (Iyer et al., 2017), published under a "CC BY 4.0 DEED Attribution 4.0 International" licence.

study evaluated a cohort of patients several days after the onset of coma. Patients that showed a N100 response to auditory stimulation were then analyzed for an MMN response. The authors found a high positive predictive value of 91% for predicting awakening, the overall accuracy, however, was relatively low. One major drawback of this study is the exclusion of patients without a N100, leaving around a third of patients without a prognosis. The results of (Fischer et al., 1999) were confirmed in many studies, such as (Naccache et al., 2005; Luauté et al., 2005; Fischer et al., 2006), showing the high positive predictive power of the MMN for patients in a comatose state (Kane et al., 2000; Daltrozzo et al., 2007; Morlet and Fischer, 2014). A limitation of these studies is the focus on single electrodes, a univariate EEG analysis. This leaves these studies at a disadvantage over more recent, multivariate studies, which integrate information from multiple EEG channels simultaneously. Additionally, these analyses focused on the later stages of coma, usually multiple days up to months after the onset of coma. (Fischer et al., 2010) also showed that very few patients in a vegetative state, showed an MMN response, months after the onset of coma.

However, as clinical standards are not evaluating the EEG in response to standardized auditory stimuli (Rossetti et al., 2016), the deviance paradigm might introduce additional information that is not yet contained within currently used clinical markers. Predictors based on a deviance paradigm might introduce complementary information for patients that are currently in a 'grey zone'. This is an advantage over the previously introduced deep learning based models. These techniques only use resting state EEG activity, which is already well integrated into the clinical practice, as the same data is used for standard clinical markers.

1.2.3.3 Machine learning for outcome prediction of standardized auditory stimulation

To overcome the disadvantage of the above mentioned analysis focusing on single electrodes, several studies applied multivariate decoding algorithms to analyze EEG data of comatose patients stimulated with an auditory deviance paradigm (Tzovara et al., 2013, 2016; Juan et al., 2016; Pfeiffer et al., 2017, 2018). These studies have focused on training machine learning algorithms on discriminating standard and deviant sounds presented to patients in a deviant paradigm. The first study (Tzovara et al., 2013) showed that the decoding performance of data collected within the first 24 hours after the onset of coma was high for survivors and non survivors. The performance was not predictive for patients' outcome. However, an increase in decoding performance from the first to the second day of coma was predictive of outcome with an accuracy of 70% and a positive predictive value of 100% on a validation dataset of 18 patients. Patients in this study were treated exclusively with therapeutic hypothermia at 33 °C within the first 24 hours. However, they reached normal body temperature again on the second day of recordings. These results were confirmed in (Tzovara et al., 2016) on a bigger cohort of 94 patients reaching a positive predictive value of 82% and an overall accuracy of 63%. For survivors, the difference in AUC score of performance between day one and two correlated to the CPC at three months as well as cognitive scores (Juan et al., 2016). These results indicate that for survivors of coma auditory processing during the early stages of coma provides information about the later functional status after awakening. The results of (Tzovara et al., 2013) were also validated for a cohort of 84 patients treated with controlled normothermia, however, with a lower positive predictive value (Pfeiffer et al., 2017). While these studies focused on multivariate EEG analysis and present a standardized stimulation protocol, they require recordings on two consecutive days.

1.2.3.4 Neural synchrony and neural complexity in coma

The previous paragraph introduced multiple studies using machine learning to predict the outcome of coma. These studies mainly used the multivariate EEG response to sounds, as recorded across electrodes on the scalp, after minimal preprocessing. However, an alternative approach is to model specific features of EEG signals in coma and use those for predicting outcome. The field of consciousness research hypothesizes that the information content of neural activity may reflect levels of consciousness (Casali et al., 2013; Tononi et al., 2016). The information content of EEG signals can be measured via neural synchrony or neural complexity. This is why the concepts of neural synchrony and neural complexity can give insight into the properties of the auditory responses and the apparent levels of consciousness (Alnes et al., 2021). Similar to the deviance paradigm mentioned above, these concepts can be used to distinguish patients' conscious state. Neural synchrony is lower for patients with disorders of consciousness (Carrasco-Gómez et al., 2021a; Binder et al., 2017; Lechinger et al., 2016) and is predictive of the outcome of comatose patients, as survivors show lower neural synchrony compared to non survivors (Carrasco-Gómez et al., 2021a; Zubler et al., 2017). Previous studies have shown that neural complexity was reduced during anesthesia (Sarasso et al., 2015) or sleep (Miskovic et al., 2018).

Neural synchrony can be quantified with the phase-locking value (PLV) (Lachaux et al., 1999), which calculates the Hilbert transform of a signal and extracts the phase lag between pairs of electrodes. (Alnes et al., 2021) investigated the PLV of comatose patients to discover differences in levels of the PLV between survivors versus non survivors of the coma in their response to standardized auditory stimulations. This study calculated the PLV in the alpha frequency band (8-12 Hz), as it had been previously linked to the outcome of coma in patients (Kustermann et al., 2019). The authors of (Alnes et al., 2021) calculated one PLV value per patient, showing that this value was predictive of outcome. They reached a PPV of 0.87 and an accuracy of 0.85 for a cohort of 67 patients treated with therapeutic hypothermia. Survival was predicted with a PLV of higher than 0.69. The study also calculated the PLV of 13 health controls during rest and showed that their PLV was at similar levels between survivors and control participants. This showed that the neural synchrony of surviving comatose patients was more similar to fully conscious healthy controls than non surviving comatose patients.

Neural complexity is described in the literature with the aid of measures of complexity or entropy (Alnes et al., 2021; Sarasso et al., 2015; Miskovic et al., 2018). (Alnes et al., 2021) used Lempel-Ziv (LZ) complexity (Lempel and Ziv, 1976) to investigate the difference levels of neural complexity in comatose patients. Lempel-Ziv complexity was calculated on a binarized version of the EEG signal by identifying the number of unique sub-sequences. Comparing the LZ values of survivors and non survivors (Alnes et al., 2021) found no link between LZ complexity and outcome in comatose patients treated with therapeutic hypothermia. They showed that comatose patients had significantly lower complexity compared to healthy controls. This effect was driven by survivors, as a post hoc test revealed a significant difference between controls and survivors but not between non survivors and controls. In summary, the neural complexity of auditory responses was not linked to the outcome of coma. Nevertheless, it revealed differences between comatose patients and healthy controls, suggesting that it might be related to levels of consciousness.

In summary, this chapter introduced post-anoxic coma, the importance of outcome prediction, and gave an overview of the available literature. Outcome prediction from coma is a tedious task requiring clinical expertise. There have been previous attempts to automate the process with machine or deep learning (Tzovara et al., 2013, 2016; Pfeiffer et al., 2017; Jonas et al., 2019; Tjepkema-Cloostermans et al., 2019; Zheng et al., 2021). However, previous literature has the drawback that the presented models rely on EEG data in the absence of any stimulation and are therefore limited to the same information available in standard evaluations of clinical EEG that are currently performed in the clinics (Jonas et al., 2019; Tjepkema-Cloostermans et al., 2019; Zheng et al., 2021). Other studies that use EEG signals in response to standardized auditory stimulation, therefore integrating complementary information to the clinical practice, either require recordings on two consecutive days (Tzovara et al., 2013, 2016; Pfeiffer et al., 2018) or base the analysis on pre-selected features (Alnes et al., 2021). In chapter 3 of my thesis, I present a pipeline using minimally processed auditory EEG responses of comatose patients to predict their outcome with deep learning models. This work uses EEG data in response to standardized auditory stimulations, providing a novel way to assess the integrity of neural functions in coma. The proposed pipeline is also evaluated on a subgroup of patients currently without a clear prognosis, showing the importance of studying functions of the auditory pathway in patients with decreased levels of consciousness. In a second step, in order to gain insight into which parts of EEG signals are mostly contributing to the network's decisions, the network's prediction was linked to electrophysiological features quantifying neural synchrony and complexity and established clinical markers.

1.3 Sleep-wake disorders

This section introduces sleep-wake disorders, linked to a second clinical application of my thesis, where an unsupervised clustering algorithm is used to disentangle the full landscape of sleep-wake disorders.

There is a great variety of sleep-wake disorders (SWDs) with a collection of different symptoms and causes. Around a third of the population is affected by sleep problems at least once in their life (Kerkhof, 2017). Sleep-wake disorders can greatly disrupt patients' schedules and affect their quality of life (Edinger et al., 2004; Gersh, 2004). Sleep problems usually stem from two different types of symptoms: daytime sleepiness or experiencing trouble falling asleep at night. These symptoms are intermixed, as a person that has trouble falling asleep is also prone to experience daytime sleepiness. Multiple SWDs can therefore cause similar symptoms. Additionally, patients might not always be aware of all their symptoms. For example, someone that snores might not know that they do so, and very similarly they might not realize that they have incidents of interrupted breathing throughout the night, which would indicate sleep apnea. However, without this information diagnosis might be very challenging. A formal diagnosis, done by physicians, might require overnight surveillance of these patients. Furthermore, multiple SWDs can, on the one hand, occur simultaneously (Cappuccio et al., 2010). For example, 28% of patients with narcolepsy also have obstructive sleep apnea (Pataka et al., 2012). On the other hand, SWDs co-occur with other non sleep-wake-related disorders, such as psychiatric disorders, diabetes, or cardinal disorders. All of these factors influence one another and can make a diagnosis challenging. SWDs also vary in their severity. As an example, excessive daytime sleepiness, indicating hypersomnia can be caused by self-inflicted sleep deprivation. But the same symptoms of daytime sleepiness could also be caused by narcolepsy type 1, which is a severe disorder, which might in fact be an autoimmune disease (Mahlios et al., 2013; Barateau et al., 2017).

In summary, the landscape of sleep-wake disorders is very heterogeneous, as they are challenging to diagnose, greatly effect a patient's quality of life, and are quite prevalent in the population. The following sections of this chapter first introduce essential tests used by clinicians for the diagnosis of SWDs, followed by the introduction of the most important SWDs that were studied in the context of this thesis.

1.3.1 Clinical evaluations and variables

Sleep physicians have access to several different tests for the diagnosis of SWDs. Some of these are extensive clinical tests, such as the overnight polysomnography recording, but also include self-evaluations of the patients based on questionnaires. The main tests and evaluations used in this thesis are described below.

1.3.1.1 Clinical tests

The main clinical tests to diagnose SWDs are the overnight polysomnography and multiple sleep latency tests. Other examinations are for example the maintenance of wakefulness test (MWT) or lab evaluations, such as the measurement of the number of hypocretin neurons of the hypothalamus, which is a precise diagnostic criterion for narcolepsy type 1 (Sakurai et al., 1998).

Polysomnography (PSG)

One of the key clinical tools for studying sleep-wake disorders is an overnight polysomnography (PSG) recording. It is a fundamental test, as many measures for sleep assessment and diagnostic markers are extracted from it. Many of the extracted markers are related

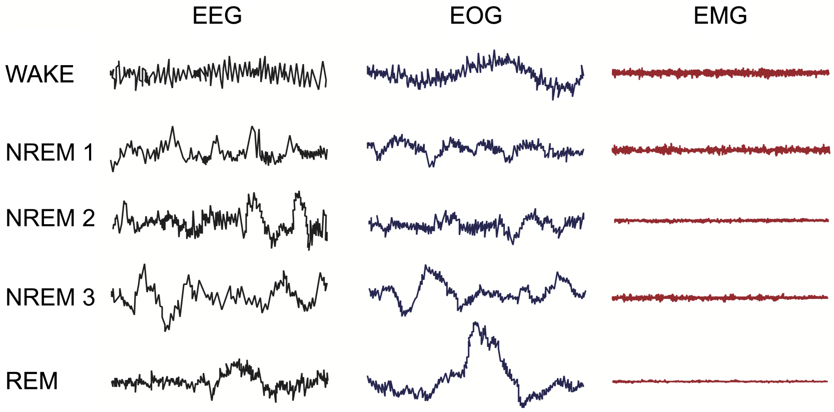


Figure 1.5: Electroencephalography (EEG), Electromyography (EMG), Electrooculography (EOG) signals during the sleep stages wake, NREM 1, NREM 2, NREM 3, and REM. These signals are used for sleep staging. Adapted from (Zhang et al., 2014), published under a "CC BY 3.0 DEED Attribution 3.0 Unported" licence.

to the different sleep stages, for example, how much time is spent in rapid eye movement (REM) or non-rapid eye movement (NREM) sleep. The overnight recording is therefore first sleep staged by a clinician in order to extract these markers. For the test, patients spend their night at the hospital sleeping, while their electroencephalography (EEG), electromyography (EMG), electrooculography (EOG), electrocardiogram is being recorded. In addition, respiratory airflow and effort as well as pulse oximetry are measured and used to identify respiratory-related disorders, such as obstructive sleep apnea. As a first step of evaluation of the PSG recording, the night is scored into the different sleep stages, wake, NREM 1, NREM 2, NREM 3, and REM, with the aid of the collected EEG, EMG, and EOG data (Figure 1.5). Out of the full night PSG recording, many markers can be extracted, like the total amount of sleep, percentage of REM sleep, apnea-hypopnea index (counting the number of episodes of partial or complete blockage of the upper airway), etc. These markers are then used to diagnose sleep-wake disorders by a clinician.

Multiple sleep latency test (MSLT)

The multiple sleep latency test (MSLT) is used to test daytime sleepiness and requires patients to lie down. They then try to fall asleep during the day for 20 minutes at a time, while their EEG, EMG, and EOG are measured. The test is repeated five times and many measures used to diagnose SWDs are extracted. For example, the mean sleep latency i.e. the average time it took a patient to fall asleep, or the mean REM latency, i.e. the average time it took for a patient to reach REM the first time, are measured with this test (Carskadon and Dement, 1982).

1.3.1.2 Self evaluations

Most sleep-wake disorders do not have objective diagnostic biomarkers. However, clinicians' evaluations for sleep-wake disorders are still based on questionnaires, as they give additional insight. There are some disorders, such as parasomnias or restless leg syndrome, which rely almost exclusively on patients' self-reports (ICSD-3, 2014). Additionally, questions related to the daily lives of patients are very important for a diagnosis. A standard evaluation is the Epworth Sleepiness Scale (ESS), measuring daytime sleepiness, for example, how likely someone is to fall asleep during activities, such as talking or riding a car during the day (Johns, 1991). Other questions are related to the subjective amount of sleep patients had during overnight recordings or the subjective feeling of having a burning or tingling sensation in one's legs, for the diagnosis of restless leg syndrome.

1.3.2 Types of sleep-wake disorders

According to the international classification of sleep disorders third edition (ICSD-3, 2014) there are more than 80 sleep-wake disorders. There are six classes of disorders, each containing multiple subclasses. The most important classes of disorders, that are relevant for this thesis, are introduced below, each with their most prevalent subclasses.

1.3.2.1 Central disorders of hypersomnolence (CDH)

Excessive daytime sleepiness for at least three months is the main diagnostic criteria for central disorders of hypersomnolence (CDHs) (ICSD-3, 2014). This can be measured via the ESS (Section 1.3.1.2). As a more objective measure of sleepiness is the mean sleep latency of the MSLT (Section 1.3.1.1, MSLT). An mean sleep latency of below 5 minutes is considered an indication of sleepiness and above 10 minutes for normal alertness (ICSD-3, 2014). The most severe subtypes of CDHs are narcolepsy type 1, narcolepsy type 2, and idiopathic hypersomnia. The most important clinical measures for the distinction between these subtypes are the mean sleep latency and the number of sleep onset REM periods. Sleep onset REM periods are episodes of REM sleep within 15 minutes of falling asleep. As a comparison, the mean latency to a period of REM sleep in the healthy population is 109 ± 65 minutes (Singh et al., 2006).

There are other subtypes of CDHs, such as hypersomnia due to a medical disorder, hypersomnia due to a medication or substance, hypersomnia associated with a psychiatric disorder. (ICSD-3, 2014).

Narcolepsy type 1 (NT1)

Narcolepsy type 1 (NT1), formerly called narcolepsy with cataplexy, is the most severe sub-disorder of CDHs. (Mahlios et al., 2013; Barateau et al., 2017) even suggested that NT1 might be an autoimmune disease. Cataplexy describes the symptom of complete or partial loss of muscle tone without losing consciousness, usually triggered by positive emotions and laughter. A diagnosis of NT1 can be made if a patient reports excessive daytime sleepiness, cataplexy, and mean sleep latency less than eight minutes and at least two sleep onset REM periods (ICSD-3, 2014).

With the discovery of the sleep/wake regulating hypocretin neurons located in the hypothalamus and their link to narcolepsy, a second criterion for the diagnosis of NT1 was introduced (Mignot et al., 2002). Patients with low levels of hypocretin neurons and reports of excessive daytime sleepiness can be diagnosed with NT1 (ICSD-3, 2014). Hypocretin neurons are located in the hypothalamus and are used for regulating sleep/wake states, by producing the neuropeptide orexin (Sakurai et al., 1998). Around 90% to 95% of NT1 patients have low levels of hypocretin neurons (Mignot et al., 2002; Luca et al., 2013).

Narcolepsy type 2 (NT2)

Narcolepsy type 2 (NT2), formerly called narcolepsy without cataplexy, has very similar diagnostic criteria to NT1 (excessive daytime sleepiness, mean sleep latency, sleep onset REM periods), however, patients with NT2 can not show cataplexy or low levels of hypocretin neurons (ICSD-3, 2014). NT2 is hypothesized to be a heterogeneous condition and there currently exists no clear biomarker for an unambiguous diagnosis, similar to the hypocretin deficiency for NT1 (Baumann et al., 2014; Fronczek et al., 2020).

Idiopathic hypersomnia (IH)

Idiopathic hypersomnia (IH) can be challenging to distinguish from NT2 (Trotti et al., 2013; Lopez et al., 2017). The criteria for IH differs from NT2 regarding only the number of sleep onset REM periods, which for NT2 is lower than 2 and for IH strictly higher than 2. Similar to NT2, IH is a very heterogeneous condition and currently, there is no pathophysiological explanation for its origin (Lammers et al., 2020; Fronczek et al., 2020).

Challenges with diagnostic of central disorders of hypersomnia

Diagnosing CDHs can be challenging. The differentiation between NT1 and NT2 is quite clear, as NT1 patients either express cataplexy or have a low number of hypocretin neurons in the hypothalamus (Mignot et al., 2002). However, the rest of the diagnostic criteria are identical. The lab evaluation for levels of hypocretin neurons is not always conducted and was only added as a diagnostic criterion in 2014 (ICSD-3, 2014). Therefore patients that don't express cataplexy are often diagnosed with NT2. However, it has been shown that up to 24% of NT1 patients with low levels of hypocretin neurons do not express cataplexy (Mignot et al., 2002; Andlauer et al., 2012). (Baumann et al., 2014; Fronczek et al., 2020) also highlight that some patients with low or intermediate hypocretin levels but not cataplexy at their first evaluation would develop it many years later. This evidence points towards misclassifications of NT2 patients in the initial absence of cataplexy.

The differentiation between NT2 and IH is even more challenging. The distinction between these two disorders is based on the results of the MSLT, specifically the number of sleep onset REM periods, but it has been shown that the MSLT has a very poor agreement between multiple evaluations (Trotti et al., 2013; Lopez et al., 2017). The former study evaluated a cohort of CDHs patients with two MSLT testings. The re-testing confirmed the diagnosis for only 47% of CDHs patients, more precisely for 33% NT2 patients and 57% of IH patients (Trotti et al., 2013). (Lopez et al., 2017) performed a similar study in 2017. For 81% of NT1 patients, the diagnosis was confirmed. For the non-cataplectic disorders of central hypersomnolence, the diagnosis was confirmed in 39% of patients. 47% of NT2 patients were rediagnosed with NT2, however for 47% of NT2 patients, a diagnosis with a CDH was not confirmed. For IH the percentage of reclassification was lower, at 20% (Lopez et al., 2017). While these studies showed relatively high stability of the diagnosis of NT1, the retest stability for non-cataplectic disorders of central hypersomnolence was below 50%. These studies were of great importance for showing the low test-retest reliability of the MSLT.

(Fronczek et al., 2020) have shown that many of the measures, such as sleep paralysis, hallucinations, and disrupted night sleep, for patients with NT2 lie between values found in NT1 and IH. Therefore some opinions have emerged that NT2 might not exist as a distinct sleep-wake disorder, but only at the border between NT1 and IH. Multiple studies, therefore, call for a redefinition of NT2 and IH on a spectrum (Baumann et al., 2014; Fronczek et al., 2020) and for new useful biomarkers for the distinction of NT2 and IH (Baumann et al., 2014; Lammers et al., 2020).

In conclusion, this paragraph highlights the challenges with the diagnostic criteria for CDHs. There have been multiple suggestions for redefinition of NT1, NT2 and IH (Baumann et al., 2014; Fronczek et al., 2020; Lammers et al., 2020) and calls for new biomarkers (Baumann et al., 2014; Lammers et al., 2020; Dietmann et al., 2021).

1.3.2.2 Sleep-related breathing disorders (SBD)

The category of sleep-related breathing disorders (SBDs) spans a great variety of disorders. The most common is obstructive sleep apnea (OSA), affecting around 2-4% of the population (Young et al., 2002), followed by central sleep apnea (CSA), which can be found in 0.9% of the general population (Donovan and Kapur, 2016). Breathing disorders describe events of loss of breathing during sleep leading to potential awakenings. Patients often do not realize their loss of breath and might not remember the awakening that followed. However, patients generally feel more tired and sleepy during the day (ICSD-3, 2014). OSA and CSA have different pathologies. OSA is caused by an obstruction of the upper airways, which results in patients temporarily receiving too little air. In contrast, CSA is caused by a dysfunction of the central ventilatory control centers, as they fail to initiate ventilatory efforts. Obstructive or central respiratory events are scored in the PSG and patients need to have at least five events per hour (Apnea-Hypopnea Index) for a diagnosis of OSA or CSA respectively (ICSD-3, 2014). The apnea-hypopnea index describes an objective biomarker that can be used to diagnose OSA and CSA reliably. However, the computation of this marker requires the scoring of the overnight PSG by a physician.

1.3.2.3 Insomnia

Patients with insomnia suffer from difficulties in initiating and maintaining sleep. They, therefore, have shorter sleep duration and worse sleep quality, despite good opportunities for sleep. Insomnia is diagnosed mainly based on self-reported markers related to initiating and maintaining sleep, as well as daytime impairments such as fatigue or mood disturbances. Another key criterion for the diagnosis of insomnia is that another disorder should not better explain the reported symptoms (ICSD-3, 2014; Edinger et al., 2004). This suggests that insomnia may be difficult to discriminate from other disorders because its symptoms co-occur in many other sleep-wake disorders (Edinger, 2011).

1.3.2.4 Parasomnias and sleep-related movement disorders

Parasomnias are described as unpleasant physical experiences, such as sleepwalking or sleep terrors. For example, events of parasomnia can entail incomplete awakenings, or movements during REM sleep. Clinicians base the diagnosis of Parasomnia on questionnaires. The evaluated questions are related to dreams, or vocalizations (ICSD-3, 2014).

Simple, stereotypical, and repeated movements that disturb both sleep onset and sleep are some of the criteria for diagnosing sleep-related movement disorders. The only exception is restless leg syndrome (RLS), as the diagnostic criteria differ slightly. RLS patients experience uncomfortable sensations (such as tingling or burning) in the legs that only decrease when moving the legs. For diagnosing RLS, clinicians rely mostly on questionnaires, asking for a tingling or burning sensation in the legs (ICSD-3, 2014).

Parasomnias and movement disorders both express movements during the night, however, the ones observed in parasomnias are more complex and with purpose. In an overnight PSG recording, both disorders would show movements and differentiation based on objective markers, such as the periodic limb movement (PLMS) index, might present a challenge. Additionally, both disorders rely heavily on questionnaires for their diagnosis and do not have objective diagnostic biomarkers.

In summary, the full landscape of SWDs is extremely heterogeneous and entangled. There are multiple problems with the current diagnostic criteria and potential difficulties with diagnostics. For central disorders of hypersomnolence, the poor retest reliability of the MSLT and diagnosis of NT2 versus IH have been mentioned (Trotti et al., 2013; Lopez et al., 2017). Therefore, multiple studies call for new biomarkers (Lopez et al., 2017; Trotti et al., 2013; Baumann et al., 2014; Lammers et al., 2020) or reevaluations of current diagnostic criteria (Baumann et al., 2014; Fronczek et al., 2020; Lammers et al., 2020) within central disorders of hypersomnolence. These limitations are not restricted to CDHs, as biomarkers exist for only a few other disorders. For example, insomnia or sleep-related movement disorders don't have precise biomarkers. Additionally, there are several disorders that heavily rely on self-reported metrics, which are not always reliable, such as parasomnias or restless leg syndrome.

1.3.3 Clustering

Within the heterogeneous and entangled landscape of SWDs the question arises of whether the currently collected diagnostic markers can disentangle the full landscape of SWDs, especially since only very few SWDs have objective biomarkers. Additionally, for prevalent and heterogeneous sleep-wake disorders, such as insomnia or OSA, an open question remains if potential subtypes could be identified, that would characterize patient groups better.

One possible way to address these questions is with unsupervised machine learning techniques, such as clustering, developed in the field of computer vision. Unsupervised clustering is a technique used to discover subgroups of data points with the closest distance in a high dimensional space (Duda et al., 2001). For SWDs this technique can be used to identify new subgroups of patients that share the biggest similarity among the collected clinical markers. This information can then assist in redefining certain SWDs, based on currently used markers. Clustering has been previously used in SWDs, for example in CDHs (Šonka et al., 2015; Cook et al., 2019; Gool et al., 2022), insomnia (McCloskey et al., 2019 11; Miller et al., 2016; Kao et al., 2021) or OSA (Joosten et al., 2011; Bailly et al., 2016; An et al., 2019; Venkatnarayan et al., 2022). However, there have been no previous attempts at clustering the full landscape of sleep-wake disorders.

1.3.3.1 Technical aspects of clustering

There are many different hyperparameters, that need to be considered when building a clustering pipeline. This section covers some of these parameters, with a specific focus on retrospective clinical datasets. First, the type of clustering algorithm used is important. The field of computer vision has proposed many different clustering algorithms, for example, hierarchical or centroid-based algorithms (Duda et al., 2001). Second, a distance metric between data points needs to be defined. Exemplar metrics are, the Euclidean or L_1 distance (Duda et al., 2001). Last, the number of clusters the algorithm should find must be pre-defined or post hoc determined. For centroid-based clusterings the number of clusters is pre-selected. Compared to hierarchical clusterings where a threshold above which clusters are no longer merged is selected. Post selections are usually based on some quality metrics such as explained variance, inter or intra-cluster distance, and for example the elbow method. In all of these cases, the number of clusters, distance threshold, or the used quality metric is a hyperparameter of the clustering pipeline. All of the above-mentioned choices can impact the results.

Applying clustering algorithms developed for computer vision to retrospective clinical data is non trivial. Clinical datasets are oftentimes incomplete, as either patients did not perform all tests, or retrospective data was not properly digitized. There are three options for dealing with missing data for clustering.

First, one can focus on a subset of patients and markers where all data is available. This however may majorly reduce the number of patients available and is not ideal when machine learning algorithms are used, as the risk of overfitting is increased.

Second, there is the possibility to select a distance metric that can handle missing values when comparing data points. One such example is Gower's distance (Gower, 1971), which ignores any dimension where one of the data points is missing and is otherwise equivalent to the L_1 metric. Such a metric has the limitation, that data points that don't have many values in the same dimension are spatially quite close with respect to this metric since only a few entries overall are considered for this distance, compared to two data points with many values given. This could be a disadvantage for clinical datasets where specific tests are only performed if clinicians already suspect one type of disorder and patients with entirely different disorders don't have many values in overlapping dimensions. For example, variables related to cataplexy in NT1 are usually only collected for patients with a disorder of hypersomnia but not for patients with parasomnia. Clinicians are less likely to ask patients with a CDH about night terrors, an essential question for diagnosing parasomnias. Therefore NT1 and parasomnia patients could have a small distance between them, even though their phenotype is completely different.

The third option is to use an imputation technique, for example, replacing missing values with the mean or median of the dataset. There are also machine learning algorithm-based algorithms for imputation, such as Random Forest (Stekhoven and Bühlmann, 2011) or K Nearest Neighbour (Bishop, 2006). The drawback of this option is that markers with only a few values given are not reliable after imputation.

In summary, many different choices need to be made in selecting a clustering pipeline. The decisions are related to the selection of algorithms, metrics, or approaches in dealing with missing data, which are particularly relevant for clustering retrospective clinical data.

1.3.3.2 Clustering for sleep-wake disorders

Despite the methodological challenges, clustering algorithms have a strong potential to identify sub-populations within patient cohorts, without an a priori hypothesis. Therefore, they have been used to address the central open questions in the field of SWDs, namely to describe sub-cohorts of heterogenous and prevalent sleep-wake disorders. In its majority, previous literature has used clustering algorithms for identifying unique phenotypes in CDHs, OSA, or insomnia.

Clustering of central disorders of hypersomnolence

A first attempt in (Šonka et al., 2015) considered a cohort of 96 patients with CDHs, equivalent to what is currently known as NT1, NT2, and IH patients. The analysis was based on a hierarchical clustering with a euclidean distance based on seven clinical variables. The authors found three major clusters within CDHs and a minor cluster only containing two patients with NT2. The study, therefore, concluded that within CDH, three clusters of sub-disorders exist. The first cluster found, contained patients with NT2 and IH. This cluster had the highest number of patients with irresistible unwanted naps. A second cluster had the highest number of patients with great difficulty waking up from naps and contained almost exclusively IH patients. The third cluster contained only patients with NT1 and, as expected, had the highest number of patients with cataplexy.

In an attempt to identify subtypes of hypersomnolence disorder (Cook et al., 2019) used a hierarchical clustering and identified two clusters splitting a cohort of 62 patients into a more severe and milder sub-type of CDHs. The two clusters were identified based on self-reported questionnaires, reflecting total sleep time, time in bed, and sleep inertia. This study suffers however from some major limitations. First, the clustering variables were all self-reported and not objective markers. Second, the study did not clearly define the disorders of the patients. It is unclear whether patients with NT2 were included in the study cohort. And last, the clusters found were never compared to patients' diagnoses according to ICSD diagnostic criteria (ICSD-3, 2014).

The studies mentioned above considered relatively small cohorts of CDH patients. As both studies used unsupervised machine learning algorithms, it is unclear if the algorithms overfit their data, and results might not generalize to larger cohorts. A recent study addressed this limitation by analyzing the European Narcolepsy Network database (Gool et al., 2022). The study comprised a total of 1078 patients and 97 clinical markers. As the database had quite some missing values, an agglomerative clustering with a Gower's distance was performed, which can manage missing values (Section 1.3.3.1). Four of the seven identified clusters contained mostly patients with NT1, describing potentially different NT1 sub-types. All clusters had a high prevalence for cataplexy, a high number of sleep onset REM periods, and low levels of hypocretin neurons, characteristic of NT1 (Section 1.3.2.1). Two clusters contained the majority of the patients without cataplexy; these patients were a mixture of mostly NT2 and IH patients. The difference within these clusters was primarily reported in the markers of sleep drunkenness and difficulty in awakening, among others. These two clusters exhibited no cataplexy and higher numbers of hypocretin neurons compared to the first four clusters, consistent with the diagnostic criteria of NT2 and IH. The last cluster was small and included patients with NT1, NT2, and IH. (Gool et al., 2022) show with their findings that IH and NT2 still lack precise biomarkers and no clear clusters were identified for these disorders.

In summary, (Šonka et al., 2015) and (Gool et al., 2022) found clear clusters of NT1 patients and, in each case, two clusters mixed between patients with NT2 and IH. These studies showed that existing clinical markers can distinguish NT1 from NT2 and IH. However, they discovered a clear need for new biomarkers for NT2 and IH or a redefinition of the clinical criteria. (Šonka et al., 2015) and (Gool et al., 2022) identified some possible markers to distinguish subgroups of non-cataplectic CDHs. Further studies are required to confirm markers for subgroups of CDHs.

Clustering of OSA and Insomnia

OSA and insomnia are prevalent and heterogeneous disorders. Quite some studies have tried to identify subtypes of OSA (Joosten et al., 2011; Bailly et al., 2016; An et al., 2019; Venkatnarayan et al., 2022) or insomnia (Miller et al., 2016; McCloskey et al., 2019 11; Kao et al., 2021) with the aid of unsupervised clustering techniques.

(Joosten et al., 2011; Bailly et al., 2016; Venkatnarayan et al., 2022) used different clustering algorithms (K-Means, hierarchical or feature trees) with cohorts of OSA patients, of various sizes ranging from 100 to 18263 patients. These studies found different numbers of clusters (three to six), which were either separated along demographic data (age, BMI, sex) or disease severity. This suggests that within the heterogeneous disorder of OSA, distinct subtypes exist. It is, however, still unclear how to best identify subtypes of obstructive sleep apnea.

(Miller et al., 2016; McCloskey et al., 2019 11; Kao et al., 2021) base their clustering on overnight recordings and features extracted from the EEG and sleep stages to identify insomnia subtypes. The studies identified between two and six clusters that differed primarily on sleep time (total, REM, or NREM sleep) and sex. They used either hierarchical clusterings or Bayes-Gaussian mixed models (Bishop, 2006). These studies all used features extracted from an overnight PSG recording. For a diagnosis of insomnia disorder, it is not necessary to perform a PSG, as the diagnosis is based on questionnaires. However, PSG could give further insight into insomnia disorder to identify new biomarkers or describe new subtypes.

1.3.3.3 Clustering of sleep outside of the clinical environment

One attempt to cluster a wide range of sleep patterns considered a big population outside the clinical environment. (Katori et al., 2022) presented a clustering of 91'765 participants collected in the UK-Biobank project. The authors analyzed actigraphy data from a cohort of participants with no record of a SWD. The actigraphy data was used to extract 21 variables, such as mean wake time, mean sleep time, and phase. These variables were then used to identify a total of 16 clusters with a density-based spatial clustering of applications with noise clustering (Ester et al., 1996). The authors found clusters of chronotypes ("night owls" and "early birds"), different week-to-weekend cycles, as well as seven clusters expressing insomnia-like patterns. This study is critical as it shows for the first time the use of actigraphy data at this scale for analyzing sleep-wake patterns, however, in participants without reported sleep disorders.

In summary, the above section presented multiple clusterings on a different category of SWDs at a time. The presented results showed a considerable variation between studies, in terms of the number of identified clusters, as well as the variables distinguishing the clusters. The discrepancy in results can emerge from different sources. First, the algorithms and metrics used varied between studies. Second, different approaches were selected for dealing with missing data. Both of these components can influence the results greatly, and for each class of disorders, it remains unknown if other selections of hyperparameters would replicate the results.

One major omission in the literature is that there are, to date, no attempts to cluster the full landscape of sleep-wake disorders. It remains, therefore, an open question if the currently collected clinical markers are sufficient to fundamentally untangle all SWDs. This may allow for identifying more phenotypes of these disorders. This thesis addresses this omission in chapter 4, where an unsupervised clustering algorithm is used to disentangle the full landscape of SWDs. The study focuses on ten different classes of SWDs from the Bernese Sleep registry and presents an automatic data curation and clustering pipeline.

1.4 Thesis Contributions

This thesis contributes to the topic of machine and deep learning in the field of neuroscience and neurology. It is structured around three primary manuscripts, each targeting distinct facets in this field. The first manuscript, (Aellen et al., 2021), sets the methodological basis, where it is shown that deep learning can substitute MVPA techniques for EEG research. In the second paper (Aellen et al., 2023), the strong potential of deep learning for predicting outcome from coma is presented. A third study (Aellen et al., published) is using unsupervised machine learning to cluster sleep-wake disorders based on currently available markers.

1.4.1 CNNs for decoding EEG responses and visualizing trial by trial changes in discriminant features

MVPA algorithms are often used for the analysis of EEG signals. They can be used to extract discriminative patterns from EEG in response to external stimulations. However, most MVPA algorithms are based on simple machine learning classifiers, and most commonly, are trained on a time-point by time-point basis. They operate under the assumption that discriminant information is time-locked, which is a limitation. However, neural patterns can be distributed across trials or subjects and may change in latency or spatial configuration. Therefore using MVPA as an analysis tool on a group level can be challenging.

Deep learning models can address this challenge by incorporating space and time-unlocked patterns. However, the use of deep learning for basic research on EEG signals is still limited, especially as an MVPA method.

In (Aellen et al., 2021) (chapter 2) it is explored if deep learning can be used as an MVPA technique for decoding EEG signals. To this end, the first hypothesis is that using time and space-unlocked information from EEG signals and methods from computer vision can increase classification accuracy. A second hypothesis is that gradient-based feature visualization tools, i.e. saliency maps, can be used to extract class-specific discriminant features.

This contribution, published in the Journal of Neuroscience Methods in December 2021, uses a CNN architecture, which has been developed for image classification (residual neural network (ResNet50) (He et al., 2016)). The study used two different datasets of EEG data of healthy participants once exposed to auditory and once to visual stimulations, presented in a deviance paradigm. In particular, a CNN was trained to discriminate between standard and deviant trials to validate the presented pipeline in different modalities. To compensate for the small sizes of the two datasets, three different types of data augmentations were used, temporal shift, sub-averaging of single trials, and Gaussian noise. Additionally, saliency maps (Simonyan et al., 2013) were used to extract class and trial-specific features and analyzed changes in these features throughout the experiment.

1.4.2 Auditory stimulation and deep learning predict awakening from coma after cardiac arrest

Outcome prediction is of great importance in post-anoxic coma patients. For clinicians, it is important in order to optimize patients' care, and for patients' families, early assessment carries great value. Current predictors used within the clinics leave around a third of the patients without a proper diagnosis. Additionally, most of these predictors exclusively predict bad outcome and only a few predictors for survival exist.

For this reason, the prediction of the outcome from coma is the focus of (Aellen et al., 2023) (chapter 3). The prediction is based on EEG signals of patients in response to standardized auditory stimulations with deep learning models. The hypothesis is that artificial neural networks can assist in predicting patients' chances of survival by extracting interpretable patterns from EEG signals in response to standardized auditory stimulation within the first 24 hours of coma. Furthermore, a second hypothesis is that the presented outcome prediction would be complementary to clinical markers and would improve outcome prognosis for 'grey zone' patients.

This contribution, published by Brain in February 2023, uses EEG signals of 134 post-anoxic coma patients exposed to standardized auditory stimulations. A deep learning model (Lawhern et al., 2018) is used to predict the outcome of these comatose patients. Additionally, it is shown that our model reaches high performance on patients without a clear prognosis, potentially giving additional information for clinical decisions. Last, the network's output is compared to interpretable features related to neural synchrony, neural complexity, and other clinical markers.

1.4.3 Disentangling the complex landscape of sleep-wake disorders with data-driven phenotyping: A study of the Bernese Center

The third part of this thesis focuses on sleep-wake disorders. The landscape of SWDs is convoluted, as multiple sleep-wake disorders can co-occur in single patients. Additionally, only a few disorders with objective biomarkers exist, namely OSA, CSA, and NT1. All other disorders rely heavily on subjective questionnaires or clinical tests with poor retest reliability. Therefore, many studies call for new biomarkers (Baumann et al., 2014; Lammers et al., 2020; Dietmann et al., 2021), or redefinition of current diagnostic criteria

(Baumann et al., 2014; Fronczek et al., 2020).

Previous work attempted unsupervised clustering of patients with a specific sleep-wake disorder to identify markers that would better distinguish sub-classes of these disorders. One such example is (Gool et al., 2022), focusing on CDHs, where they identified clear clusters of NT1 patients. However, NT2 and IH patients were intermixed. A similar approach could be used on the full landscape of SWDs, to address the open question if current clinical markers are enough to disentangle sleep-wake disorders. Yet the literature completely lacks such studies on all SWDs.

The third contribution of this thesis (Aellen et al., published) (chapter 4) therefore asks if unsupervised clustering algorithms would be able to disentangle the current landscape of sleep-wake disorders based on existing biomarkers. Three different patient cohorts are analyzed. First, this work focuses on a cohort of patients with CDHs to test the hypothesis that NT1 would be distinct, but NT2 and IH would be intermixed. A second cohort contained patients with all SWDs and the hypothesis that patients with OSA or CSA and patients with NT1 would be most distinct, as these disorders have clear objective biomarkers, was tested. By contrast, for all other SWDs, the hypothesis was that it would be more challenging to disentangle them based on currently available markers. The last cohort that was analyzed contained only patients with a single SWD.

The manuscript describing this work was, at the time of writing this thesis under review. As working with retrospective data is challenging, a pipeline for data selection, curation, and analysis, based on an unsupervised clustering algorithm was developed and applied to different cohorts of SWDs. The analysis was done on the Bernese Sleep Registry, a database of SWD patients collected over the past 16 years at the Inselspital, the University Hospital of Bern. The entire cohort contained a total of 6'958 patients and more than 300 possible markers.

2 CNNs for decoding EEG responses and visualizing trial by trial changes in discriminant features

Florence M. Aellen[1], Pinar Göktepe-Kavis[1], Stefanos Apostolopoulos[2], Athina Tzovara[1,3,4]

[1] Institute of Computer Science, University of Bern, Switzerland
[2] RetinAI Medical AG, Switzerland
[3] Sleep Wake Epilepsy Center - NeuroTec, Department of Neurology, Inselspital, Bern University Hospital, University of Bern, Switzerland
[4] Helen Wills Neuroscience Institute, University of California, Berkeley, United States

Aellen, F. M., Göktepe-Kavis, P., Apostolopoulos, S., and Tzovara, A. (2021). Convolutional neural networks for decoding electroencephalography responses and visualizing trial by trial changes in discriminant features. *Journal of Neuroscience Methods*, 364:109367. https://doi.org/10.1016/j.jneumeth.2021.109367, published under a "CC BY 4.0 DEED Attribution 4.0 International" licence.

Background: Deep learning has revolutionized the field of computer vision, where convolutional neural networks (CNNs) extract complex patterns of information from large datasets. The use of deep networks in neuroscience is mainly focused to neuroimaging or brain computer interface -BCI- applications. In electroencephalography (EEG) research, multivariate pattern analysis (MVPA) mainly relies on linear algorithms, which require a homogeneous dataset and assume that discriminant features appear at consistent latencies and electrodes across trials. However, neural responses may shift in time or space during an experiment, resulting in under-estimation of discriminant features. Here, we aimed at using CNNs to classify EEG responses to external stimuli, by taking advantage of time- and space- unlocked neural activity, and at examining how discriminant features change over the course of an experiment, on a trial by trial basis.

New method: We present a novel pipeline, consisting of data augmentation, CNN training, and feature visualization techniques, fine-tuned for MVPA on EEG data.

Results: Our pipeline provides high classification performance and generalizes to new datasets. Additionally, we show that the features identified by the CNN for classification are electrophysiologically interpretable and can be reconstructed at the single-trial level to study trial-by-trial evolution of class-specific discriminant activity.

Comparison with existing techniques: The developed pipeline was compared to commonly used MVPA algorithms like logistic regression and support vector machines, as well as to shallow and deep convolutional neural networks. Our approach yielded significantly higher classification performance than existing MVPA techniques (p = 0.006) and comparable results to other CNNs for EEG data.

Conclusion: In summary, we present a novel deep learning pipeline for MVPA of EEG data, that can extract trial-by-trial discriminative activity in a data-driven way.

2.1 Introduction

Multivariate pattern analysis (MVPA) is commonly used in the field of neuroscience to extract discriminative patterns of neural responses to external stimuli (Haynes and Rees, 2006). Although initially developed for functional magnetic resonance imaging (fMRI), MVPA techniques have been adapted for the field of magneto- and electro-encephalography (M/EEG) (Grootswagers et al., 2017). These are most commonly based on linear classifiers, which are applied on sensor-level topographic data, either aggregated across time (Tzovara et al., 2012) or on a time-point by time-point basis (King and Dehaene, 2014). This latter approach is most commonly implemented by training and testing one classifier at a given time-point within a trial (Castegnetti et al., 2020; Demarchi et al., 2019) and identifying time-points for which classification is above chance levels. However, this approach suffers from several drawbacks, as it only allows detecting a fixed time-period of discriminant activity for all experimental conditions and trials. Most MVPA approaches are based on single-trial information, and are thought to be more sensitive than 'classical' event-related potential (ERP) analyses. However, training and testing a classifier at single time-points makes the assumption that discriminant information appears at the same latency and electrode locations across trials, in a time- and space- locked way.

In the past few years, the field of computer vision has gained a tremendous momentum with the introduction of deep learning algorithms (Goodfellow et al., 2016). Deep neural networks, typically relying on convolutional operations (convolutional neural networks -CNNs-) are commonly used to classify different types of images, ranging from everyday objects (He et al., 2016), to challenging medical images (Suzuki, 2017). Because the kernels of the convolutional layers share weights for the whole image, CNNs have the advantage that they are able to detect space unlocked patterns, a property called translational equivariance (Goodfellow et al., 2016) (Section 9.2). The position of the discriminant pattern across observations is therefore irrelevant, which often results in CNNs outperforming 'traditional' machine learning algorithms.

CNNs have been increasingly applied to new fields. In the field of EEG research, deep learning algorithms have been introduced for clinical applications such as detection of epileptic seizures (Cho and Jang, 2020; Burrello et al., 2020), automating sleep scoring (Fiorillo et al., 2019), or predicting outcome of coma patients (Jonas et al., 2019). Applications in basic research mainly focus on brain-computer-interfacess, oftentimes on paradigms based on motor imagery (Schirrmeister et al., 2017; Zhang et al., 2019) and (Roy et al., 2019) for a review on deep learning and EEG. Apart from BCI applications, deep learning techniques for basic EEG research such as MVPA have been introduced but are not widely used in basic research yet. Deep neural networks predominantly profit from extracting features from minimally processed data, yet several algorithms for EEG are based on hand crafted features, such as classification of time frequency (Ghosh et al., 2018; An et al., 2014), or frequency transforms of EEG data within different frequency bands (Kuanar et al., 2018; Bashivan et al., 2016; Tan et al., 2018) or differential entropy (Wang et al., 2018). While these hand crafted features often times have a physiological meaning, such as representing the energy spectrum in a given frequency band, they require making strong assumptions about the underlying task, are not easily translatable across experimental setups and might not fully exploit the features which are most discriminant across experimental conditions. Other deep learning algorithms for EEG use the raw EEG or a minimally processed signal (Schirrmeister et al., 2017; Lawhern et al., 2018; Tang et al., 2016; Nurse et al., 2016; Hajinoroozi et al., 2017), allowing the network to fine-tune its parameters and identify the most discriminant features in the data, by maximizing separability between conditions of interest. The learnt features here usually have not a physiological meaning and their interpretation is oftentimes very complex.

One important aspect common to all MVPA approaches, for basic research and also for clinical or BCI applications, is that of obtaining interpretable features that have an electrophysiological meaning (Haufe et al., 2014). Existing techniques for interpreting results of decoding approaches for M/EEG data provide information about sensor-locked activity, such as the weights or activations of single electrodes or sensors (Haufe et al., 2014). These techniques provide information about the sensors that mostly contribute to an accurate classification, but are not informative about which of the experimental conditions are driving this classification. Other techniques for feature extraction consist of separating the EEG signal in subcomponents that can be then used to visualize condition-specific patterns, like common spatial patterns (CSP) (Koles et al., 1990), but have the limitation of poor temporal resolution, and of poor generalization over multiple participants (Lotte, 2014).

In the case of CNNs, feature interpretability is its own subfield of research. One possibility for interpreting features is visualizing which dimensions of the input data are contributing to the final prediction of the network or what information the weights of the trained kernels contain (Zeiler and Fergus, 2013). This approach has the drawback that the extracted features are not trial nor condition specific. By contrast, gradient based methods, such as saliency maps, can detect discriminant patterns of activity in each individual data sample (Simonyan et al., 2013).

Here, we introduce a novel approach for decoding EEG responses to external stimuli based on CNNs. We present an MVPA pipeline, relying on a deep CNN that extracts time- and space- unlocked patterns of EEG activity; can be generalized to different datasets with minimal assumptions; and has interpretable features in terms of spatio-temporal clusters that drive an accurate classification. We explore this pipeline using two different datasets: first, a dataset consisting of EEG responses to *Repeated* (pure tones) and *Novel* (naturalistic) sounds, with clear and sustained differences in EEG responses. Second, we used a more challenging dataset, consisting of EEG responses to *Repeated* and *Novel* images, whose presentation was mixed across participants, resulting in more subtle condition differences. Our goal is to use this pipeline in order to extract in a data-driven way trial by trial spatio-temporal patterns of discriminant electrophysiological responses, and explore how these change over the course of an experiment.

2.2 Materials and methods

2.2.1 Data

We used two different EEG datasets to (a) build our pipeline, and (b) evaluate whether it generalizes across experimental settings. The first dataset (termed 'Auditory') was used to develop the presented algorithm and fine tune its individual steps. The second dataset (termed 'Visual') was in turn used to examine whether the developed pipeline can also be used on new data and experimental conditions. Both datasets are openly available (Cavanagh et al., 2018; van Peer et al., 2017).

2.2.1.1 Auditory dataset

The first dataset was an auditory oddball paradigm, consisting of repeated presentations of *Standard*, *Target* and *Novel* sounds. The *Standard* and *Target* sounds were sinusoidals at different frequencies, while the *Novel* ones were naturalistic sounds, varying with each presentation. Here, we considered the EEG data of 17 participants from the control group of this dataset, disregarding participants with persistent artifacts or noise in their recordings. For simplicity, we focused on a 2-class classification problem, and considered trials were participants were presented with either a *Standard* or *Novel* sound (Figure 2.1 a and b for mean responses across participants, and Figure A.1 for a topographic representation). The data was recorded with 64 electrodes in a standard 10/20 configuration at a sampling

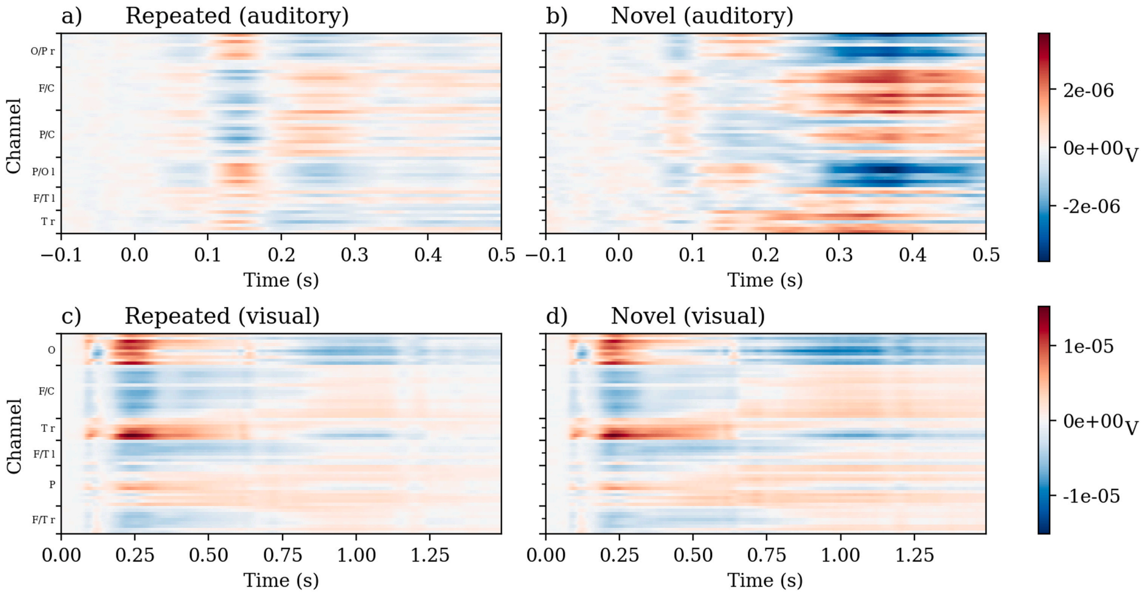


Figure 2.1: Mean evoked responses for the auditory (top) and visual (bottom) dataset, represented as time by electrodes. Panels a and c show the mean of all *Repeated* trials and panels b and d the mean of all *Novel* trials. The y-axis displays the recorded EEG channels, grouped in regions of interest for illustration purposes.

frequency of 500 Hz, initially referenced to the CPz electrode. Four temporal electrodes were removed, as in the original publication of the data (Cavanagh et al., 2018), and the remaining electrodes were re-referenced to a common average reference. Additionally eye blinks were removed with independent component analysis and single trials were extracted on a time window of 0.6 s (−0.1 to 0.5 s relative to stimuli onset). We additionally filtered the data between 0.1 and 20 Hz. This first dataset contained a mean of 129.5 ± 2.7 *Standard* (mean ± standard error reported here and in the following) and 28.5 ± 1 *Novel* trials per participant.

2.2.1.2 Visual dataset

The second dataset was a visual oddball, consisting of a repeated presentation of different sets of images. Similar as in the auditory dataset, we considered two classes of *Familiar* and *Novel* images (Figure 2.1 c and d for mean responses across participants, and Figure A.1 for a topographic representation). We extracted data from 20 participants in total, disregarding participants with prominent artifacts or noise in their recordings. EEG data were recorded at 512 Hz (later down-sampled to 256 Hz) with 64 electrodes in a standard

10/20 montage, referenced to an active common mode sense. EEG data were filtered between 0.1 and 30 Hz and re-referenced to a common average reference. Eye blinks and movement were removed according to (Gratton et al., 1983). Single trials were extracted on a time window of 1.5 s ($0 - 1.5$ s relative to stimuli onset). For the visual dataset we did not include any baseline, as the data that were publicly released were already epoched, without any baseline (van Peer et al., 2017). We additionally inspected single trials visually for artifacts. Noisy trials containing eye blinks or muscle activity were rejected. This resulted in 506 ± 140 *Familiar* trials and 146 ± 40 *Novel* trials per participant.

In the following, for reasons of consistency, we refer to the two classes of both datasets as *Repeated* (replacing *Standard* and *Familiar* from the first and second dataset respectively) and *Novel*. We represented the EEG data as a 2D signal throughout most of the rest of the paper, where the first dimension were the channels and the second the time (Figure 2.1).

2.2.1.3 Train, validation and test sets

Each dataset (Auditory and Visual) was split into a train, validation and test set in a 10-fold procedure. We used these splits to train the neural network 10 times, in a cross validation way. The validation set was used for optimizing the network's hyperparameters and identifying the best fold. The test set was left aside until the very end, and was never used for tuning the neural network or its hyperparameters. The test set was only used to evaluate, in an unbiased way, the network's performance. The available data were split into 81% train, 9% validation and 10% test trials. For the auditory dataset this resulted in 2176 trials for the train, 242 for the validation and 269 for the test dataset. As for the visual dataset there were 10570 trials in the train, 1306 in the validation and 1175 in the test set.

2.2.2 Network architecture

We built our MVPA pipeline around a residual neural network with 50 layers (ResNet50) (He et al., 2016) (Figure 2.2, red box on the left side). This network consists of a convolutional layer, batch normalization, rectified linear unit (ReLu) activation and a MaxPooling layer, followed by four segments of 3/4/6/4 convolutional blocks each. Each convolutional block has three convolutional layers followed by a batch normalization and ReLu activation layer. After the last batch normalization layer, the original input to the convolutional block is added to the output from the batch normalization layer. This residual skip connection is the main novelty of the ResNet50 architectures compared to most convolutional neural networks. The skip connections allow for deeper networks, which can extract more

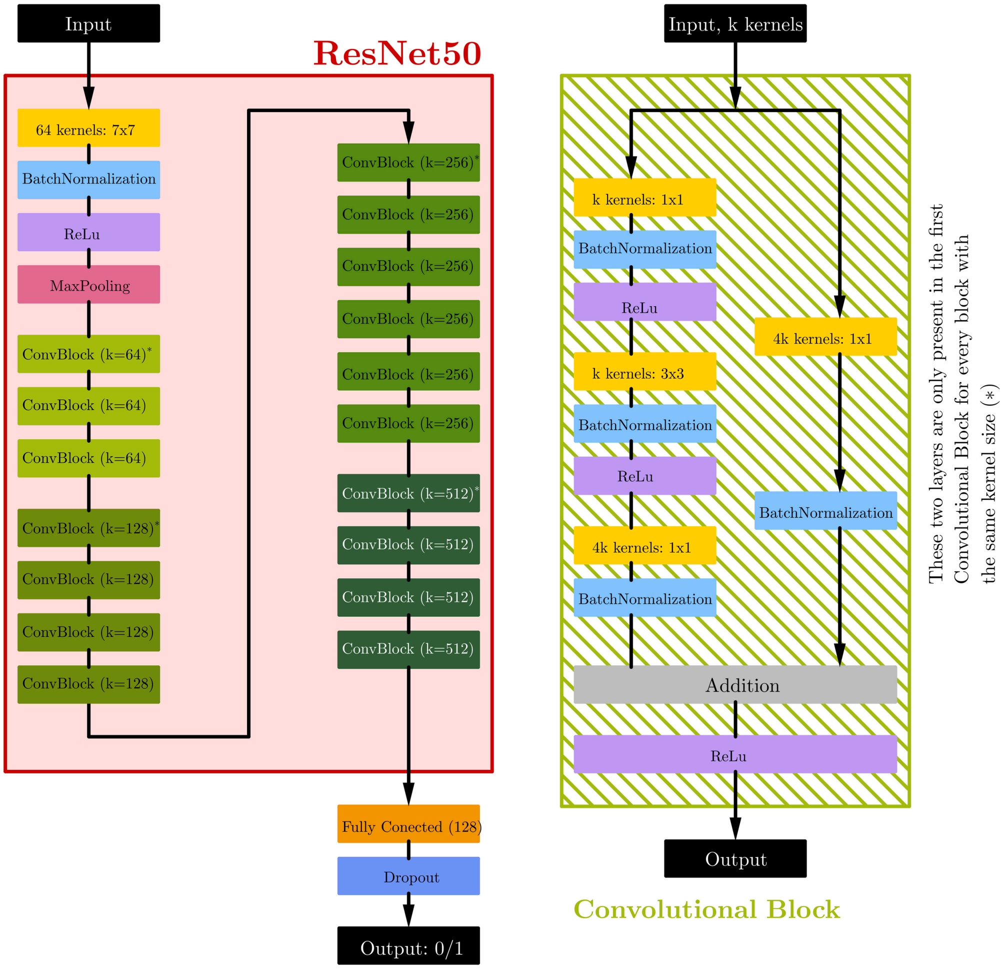


Figure 2.2: Schematic representation of the architecture of the neural network. The main network has four sections of each 3/4/6/4 convolutional blocks. Each convolutional block then contains three convolutional layers. We added a fully connected and dropout layer on top of the ResNet50 architecture. The network outputs either 0 or 1, for the labels *Repeated* or *Novel*.

complex structures from the input data. For the first convolutional block of all the four segments there is an additional convolutional and batch normalization layer on the skip connection due to otherwise mismatched dimensions. All further technical information, such as kernel sizes and padding information can be found in (He et al., 2016). In addition to the standard architecture of ResNet50, we additionally included a fully connected layer with 128 nodes and a dropout layer (with 50% probability), to further restrict overfitting. EEG trials, represented as (Channels) × (Time) were given as input to the network. The

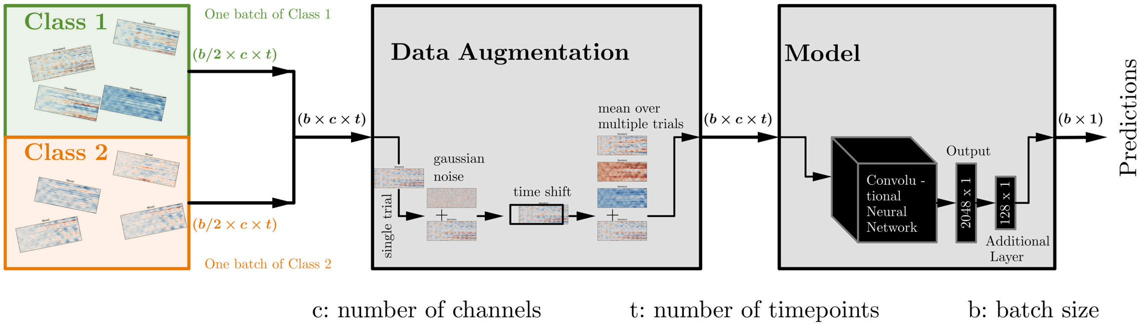


Figure 2.3: Pipeline for data augmentation and training. The pipeline starts with selection of samples in the current batch (left panel), proceeds with data augmentation (central panel) and last inputs the trials into the network for training.

network's output was a probability value per trial, ranging between zero and one, describing the probability of this trial to belong to each of the two classes (the *Repeated* class was assigned the label 0).

2.2.3 Data augmentation

Data augmentation techniques are commonly used in the field of computer vision, to artificially increase the size of an existing dataset and avoid overfitting (see (Shorten and Khoshgoftaar, 2019) for an overview of data augmentations used in computer vision). These techniques essentially distort parts of the input data in a minor but meaningful way before training a network. For example, in the field of computer vision, commonly used data augmentation techniques consist of flipping an input image horizontally, which is ecologically valid, as it is possible to observe object rotations in nature. In the case of EEG data, which are time series, flipping the time dimension would not make sense.

Here, taking into account the nature of EEG, we augmented the available trials in three different ways: (a) time shifts; (b) Gaussian noise and (c) sub-averaging single trials (Figure 2.3). First, in order to account for inter-individual differences in the timing of EEG responses, the available single trials were shifted in time with a random interval of up to 5 time-points in either the positive or negative direction. Second, to account for different levels of noise across participants, we additionally augmented the data by adding Gaussian noise, with a mean of zero and a random standard deviation of 0.1, 0.2 or 0.3 per trial. Third, considering the noisy nature of single-trial scalp EEG responses, we averaged the input data over multiple trials. More specifically, per trial we chose a random number n_k ($k \in (1, b)$) (where b is the batch size) from a triangular distribution (centered around 1) between 1 and 21 (which corresponds to 1/3 of the trials with the same labels in the current batch). For each trial we then chose $n_k - 1$ samples from that batch with the same labels and took the mean over the original and the additional trials. This last technique of averaging single trials is commonly used in classification of EEG responses, in order to improve signal-to-noise ratio (Tzovara et al., 2012).

2.2.4 Optimization and training

In the two datasets used here, the number of trials in the two classes was imbalanced, with a ratio of *Repeated* to *Novel* trials of roughly 4–1. To account for this imbalance, we over-sampled the underrepresented (*Novel*) class during training, by drawing an equal number ($b/2$) of trials from two pots containing all trials from the training set of each of the two classes.

In the training pipeline, a batch of size b of data, containing ($b/2$) trials from each condition went through the data augmentation step before training. To account for the imbalance during validation and test, we measured the area under the Receiver Operator Characteristic curve (AUC) (Macmillan and Creelman, 2004), which consists of the true positive versus false positive rate with respect to multiple thresholds. The network was optimized with an Adaptive Moment Estimation (Adam) optimizer, using the standard parameters proposed in its original implementation (Kingma and Ba, 2014), to minimize the binary cross-entropy loss (Eq. 2.1) between the real labels y of the data and the network's predictions $\hat{y}$.

$$L = -(y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})) \tag{2.1}$$

During training we employed early stopping, so that if the validation loss would not further decrease within ten training epochs the training would stop (Goodfellow et al., 2016) (Section 7.8). The network was trained for maximally 50 epochs. As a final step, we retained the network with the smallest validation loss. In the Results Section 2.3.1 we report the mean train, validation and test AUC score, accuracy and binary cross-entropy loss at that best epoch for all trained networks. The network layers were initialised with imagenet weights (Fchollet, 2016) and the batch size b was chosen to be of 64 samples, because of GPU size limitations. We used the python library tensorflow (2.1.0) with cuda (10.1.168), cudnn (5.1.5) and python (3.6.8). The full training pipeline with the two classes, data augmentation and the CNN is illustrated in Figure 2.3.

2.2.5 Estimation of network performance

To evaluate the network's performance, we computed chance levels in a data-driven way, through random permutations. We randomly shuffled the labels of the training dataset 50 times. For each random shuffle we retrained the networks in the 10-folds of the cross-validation, resulting in 500 'random' networks. Similar as for the networks trained on true labels, we retained the test scores at the epochs of smallest validation loss. We then used these 'random' networks to classify trials from the Test and Validation dataset. Each of the permutations resulted in one chance level classification performance. The actual performance of the network, trained with true labels, was compared to the distribution of AUC values obtained with random permutations with a Wilcoxon signed-rank test. Due to the heavy computational cost of training CNNs, and due to the overwhelmingly low chance-level classification results that we obtained for the first dataset when permuting the true labels, we only computed chance levels via random permutations for the first dataset. For the second dataset (Visual) we compared instead the performance of the CNN with a 'classical' MVPA approach (see Section 2.2.8.1).

2.2.6 Discriminant features

To visualize features from the EEG data that mostly contribute to the network's output, we used a gradient-based technique termed saliency maps (Simonyan et al., 2013). This technique consists of backpropagating the input label of a given trial through the network, to obtain the gradient, i.e. a value per time-point and electrode marking the strength of the contribution of that input point to the decision of the network. For a given input I_0, output class c and a score function $S_c(I_0)$ (here binary cross entropy loss), the gradient w of S_c with respect to an input I at the point I_0 is given as

$$w = \frac{\partial S_c}{\partial I}\bigg|_{I_0}. \tag{2.2}$$

The absolute value of w then gives the saliency map. To calculate saliency maps (in the following called activation maps), we used the implementation from (Kotikalapudi and contributors, 2017). For each fold of the 10-fold cross validation, we trained four networks, resulting in a total of 40 networks. This follows feature visualization approaches that are commonly used in the biomedical field, where physiological data are more noisy and complex compared to natural images (Fauw et al., 2018). Typically, multiple networks are trained per fold and their outputs are averaged to obtain stable features that are consistently identified by all networks (Fauw et al., 2018; Mehrer et al., 2020). Here, we calculated activation maps for each of the 40 networks separately, and then averaged the mean of the obtained maps. This resulted in one activation map per input data point, which were either average ERPs for the two experimental conditions (Section 2.3.4) or single trial ERPs (Section 2.3.6).

2.2.7 Trial by trial representation of discriminant features

As activation maps can be computed at the single-trial level, they can be used to study changes in trial-by-trial neural responses throughout an experiment. Instead of considering all single-trial EEG responses, as in most ERP or MVPA analyses, with single-trial activation maps it is possible to retain the temporal order of trials and compare their evolution from the first, the second, up to the last presentation of the stimuli. Such an approach could allow to assess for instance effects of learning, where neural responses to a given stimulus change from one trial to the next as a function of presentation.

Here, to explore the potential of single-trial feature extraction, as an exemplar test case, we extracted a sequence of single-trials keeping the order of exposure of the participants to the stimuli of each experimental condition. We then averaged single-trial responses over participants for each consecutive stimulus presentation, and calculated the activation maps for each of these responses. This resulted in a sequence of activation maps, which reflect patterns of discriminant EEG activity across consecutive presentations of the experimental stimuli. To quantify changes in activation maps over the course of the experiment, at every trial and time-point we summed up the values of activation maps over all electrodes (Figure A.4 in the Appendix), resulting in one value per trial and time-point. We then fitted a linear regression at every time-point to test whether the overall discriminant EEG activity changed significantly from zero as a function of trial repetitions throughout the experiment. To correct for multiple comparisons across time, 1000 cluster-based permutations were used (Maris and Oostenveld, 2007). With this analysis we identified time-points with a significant change in the activation of the network over the course of the experiment.

As a control, we performed the same analysis on the EEG data. Instead of the activation maps we used EEG activity at every time-point recorded across electrodes, and tested whether there were any latencies where changes in EEG responses during the experiment were significantly different from zero as a function of trial repetitions.

2.2.8 Comparison with existing techniques

2.2.8.1 Comparison to logistic regression and support vector machines

The performance of the CNN was compared to two baseline algorithms, using exemplar MVPA techniques. For this comparison, we chose logistic regression and support vector machine (SVM) (with 'rbf' kernel), as they are commonly used in the field of M/EEG research (Tomioka et al., 2006; Castegnetti et al., 2020; Philiastides et al., 2010). For this comparison, we kept the same splits of train/test/validation and the same cross-validation procedure as for training and testing the CNN. To estimate the hyperparameters of the logistic regression, we pooled all observations from the training set together, and optimized the parameters of penalty (l1 or l2 norm) and inverse regularizer (0.01–100, with logarithmic spacing). The optimized parameters were then used to train and test one classifier for every time-point, resulting in one classification score per time-point. This resulted in one time course of training and test AUC values averaged over the 10-fold cross-validation. For SVM, we use the same approach for optimizing the hyperparameters gamma ('scale' or 'auto') and regularization parameter (0.01–100 with logarithmic spacing). To compare the performance of the logistic regression and SVM with the CNN, we retained the best performance of these two algorithms, at the point of the maximum validation AUC score, and contrasted this with the overall performance of the CNN. Same as for the CNN, chance levels for logistic regression were evaluated by shuffling the labels of the training dataset, and by retraining the classifier of each time-point 50 times. The performance of the real classifier was compared with the distribution of the performance of chance classifiers using Wilcoxon signed-rank tests and was cluster-based corrected for multiple comparisons over time (Maris and Oostenveld, 2007).

2.2.8.2 Comparison with other CNN architectures

We also compared the performance of the ResNet50-based CNN to two other CNNs that are commonly used for decoding EEG signals (Zhang et al., 2019; Ghosh et al., 2018; Williams et al., 2020; Jonas et al., 2019). We used a Shallow and Deep CNN, first introduced in (Schirrmeister et al., 2017). These consist of 2 and 5 convolutional layers, for the Shallow and Deep networks respectively, with additional batch normalization, activation, pooling and dropout layers in between. They don't have any residual skip connections and are therefore shallower than the 50 layered ResNet50. For a fair comparison across all CNN architectures (ResNet50, Shallow and Deep networks), we always used the same training and data augmentation pipeline as described in Section 2.2.4.

Additionally, we evaluated the effect of some of the choices made in the training pipeline introduced here in classification performance (A.2.1, A.2.2, A.2.3). More specifically, we compared a CNN trained with filtered versus. with unfiltered data, a CNN where the underrepresented class was oversampled to a CNN where a weighted binary crossentropy loss was used, and lastly a CNN where we omitted the fully connected layer with 128 nodes before the dropout layer. Details for these control analyses can be found in the Appendix.

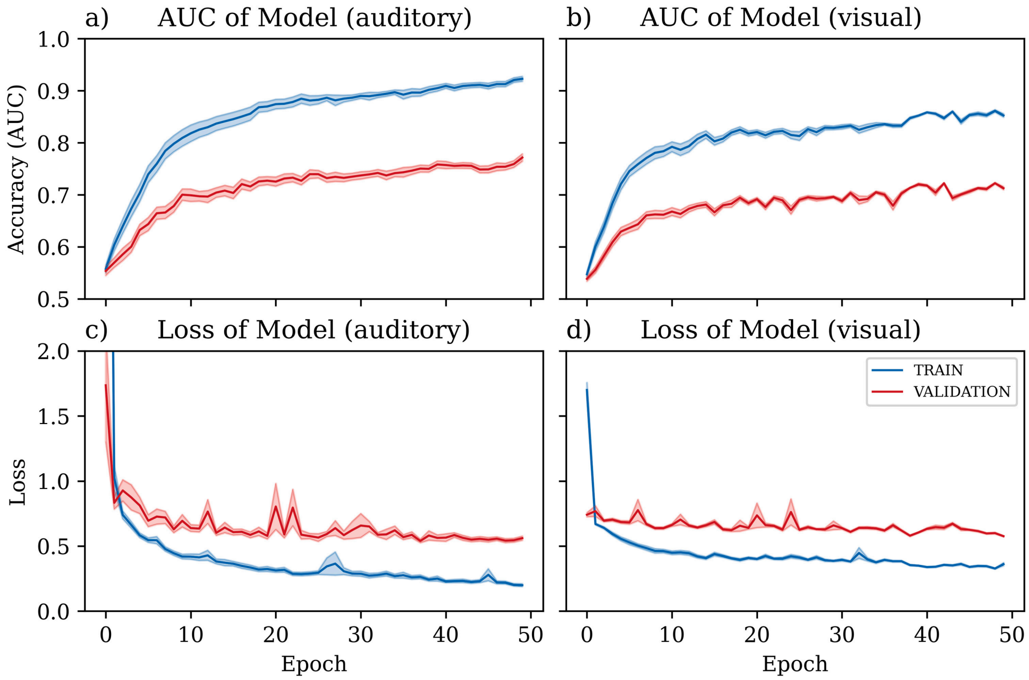


Figure 2.4: Training performance of the CNN on the two datasets. In each plot, the bold line illustrates the mean AUC (panels a and b) or binary cross-entropy loss (panels c and d) over the 40 trained networks and the shaded area the standard error. Blue lines correspond to the scores of the train and red lines to the scores of the validation set, respectively. Panels a and b show the results for the auditory dataset and panels c and d for the visual dataset.

2.3 Results

2.3.1 Training and classification performance of the CNN pipeline

First, we trained CNNs to classify EEG responses to *Repeated* versus *Novel* stimuli, using the training, validation and test folds as described in Section 2.2.1. For the auditory dataset, decoding performance, measured through the AUC, increased for both train and validation sets already within the first 10 epochs of training, and reached a plateau approximately from epoch 20 on (Figure 2.4, panel a). At the same time, the binary cross-entropy loss decreased and reached a plateau already after the first 10 epochs of training (Figure 2.4 panel c). We report an AUC score of 0.89 ± 0.04 on the train, 0.75 ± 0.04 on the validation and 0.72 ± 0.04 on the test set (see Table 2.1). On average across folds, these scores were reached on epoch 28.7 ± 13.7. The high classification performance in the test set

Table 2.1: Results of training the Shallow, Deep convolutional neural network and the ResNet50 for both datasets. We report the binary cross-entropy loss, the AUC score and the accuracy for the train, validation and test sets. The reported results correspond to the mean scores ± standard error at the epoch with the lowest binary cross-entropy loss on the validation set.

	Binary cross-entropy loss			AUC-score			Accuracy		
	Train	Validation	Test	Train	Validation	Test	Train	Validation	Test
Auditory dataset									
ResNet	0.26±0.08	0.43±0.05	1.16±2.33	0.89±0.04	0.75±0.04	0.72±0.04	0.82±0.12	0.79±0.11	0.77±0.10
Shallow Net	0.38±0.01	0.37±0.03	0.44±0.04	0.83±0.01	0.78±0.04	0.75±0.03	0.86±0.01	0.85±0.02	0.82±0.02
Deep Net	0.39±0.03	0.379±0.03	0.47±0.04	0.83±0.01	0.77±0.03	0.73±0.03	0.85±0.01	0.85±0.02	0.82±0.03
Visual dataset									
ResNet	0.40±0.06	0.54±0.03	0.61±0.06	2.82±0.03	0.69±0.02	0.67±0.04	0.76±0.04	0.75±0.03	0.75±0.04
Shallow Net	0.52±0.01	0.48±0.01	0.52±0.04	0.74±0.01	0.67±0.03	0.68±0.02	0.79±0.02	0.78±0.01	0.78±0.02
Deep Net	0.54±0.02	0.49±0.01	0.58±0.06	0.73±0.01	0.63±0.02	0.68±0.01	0.76±0.02	0.76±0.02	0.76±0.02

suggests that the trained networks could extract discriminant features of EEG responses to *Repeated* versus *Novel* sounds, and generalize to new, previously unseen trials.

As a next step, we evaluated whether the training pipeline also generalized to a different dataset and modality. The network training for the visual dataset proceeded in a very similar way as for the auditory. The AUC and the binary cross-entropy loss reached a plateau at around epoch 20 (Figure 2.4, panels b and d). The maximum AUC score was 0.82 ± 0.03 on the train, 0.69 ± 0.02 on the validation and 0.67 ± 0.04 on the test set (see Table 2.1), and corresponded to epoch 17.1 ± 10.4. This result suggests that the training pipeline, even though developed for the auditory dataset, can be applied to a different dataset.

2.3.2 Comparison of the network performance to chance levels

The network's performance was contrasted to chance levels, computed by re-training the CNN on data with randomly shuffled labels (Figure 2.5). During training, the AUC score in the training set slightly increased with training. However, the AUC in the validation set remained around the baseline values of 0.5 (Figure 2.5, panel a). A similar tendency was observed for the binary cross-entropy loss, which decreased over the first few epochs of training, but remained around 0.69 (log(2), corresponding to theoretical chance levels) for all validation epochs (Figure 2.5, panel b). Importantly, networks trained on real data achieved a consistently higher AUC compared to networks trained on data with shuffled labels, across all folds of the cross-validation (Figure 2.5, panel c) (Wilcoxon signed-rank test, $p = 1.2 \times 10^{-83}$).

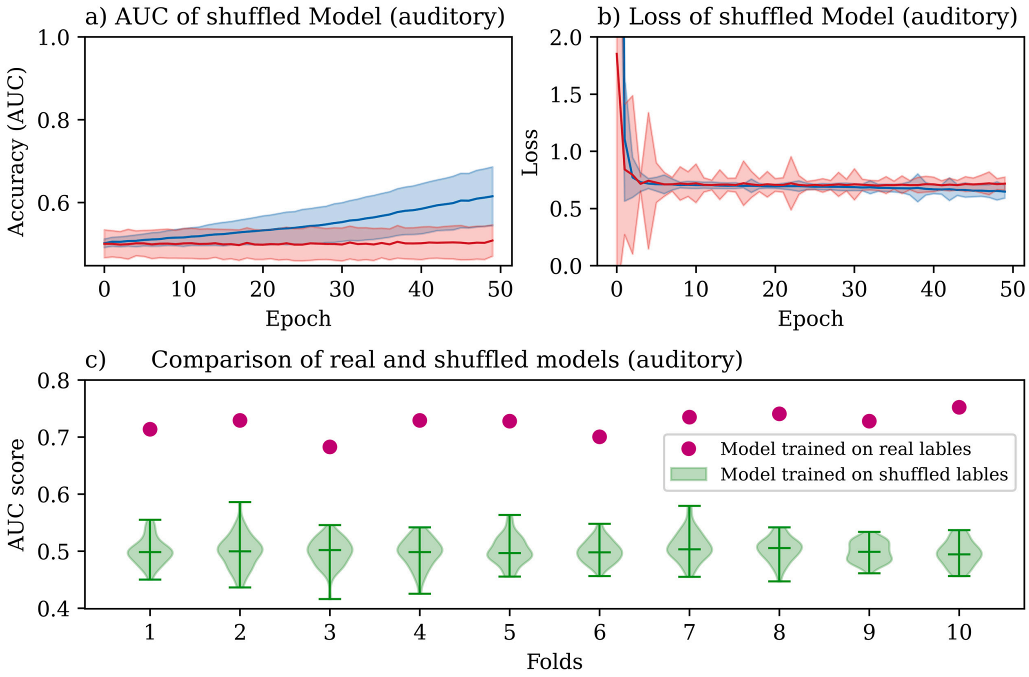


Figure 2.5: Comparison of the CNN classification performance with chance level results of the shuffled CNNs on the auditory dataset. The bold lines illustrate the mean AUC score (panel a) or loss (panel b) over the networks trained with random permutations, and the shaded area the standard error. The blue line shows the scores of the train and the red line the scores of the validation set. Panel c shows the comparison of true and chance level classification performance on the test set. For each fold we show the mean performance of the real network (magenta) versus the distribution of the performance of the 500 shuffled networks (green violin plots).

2.3.3 Comparison with existing techniques

2.3.3.1 Logistic regression and support vector machine

To compare the results obtained with the CNN with existing techniques, we trained linear and non linear 'classical' MVPA algorithms to discriminate *Repeated* versus *Novel* stimuli in both datasets. For every time-point, we trained and tested classifiers based on logistic regression (linear classifier), and SVM with 'rbf' kernel (for a non linear classifier), which resulted in a time course of AUC values, computed on a train and test set (Figure 2.6 panel a for the auditory and b for the visual datasets).

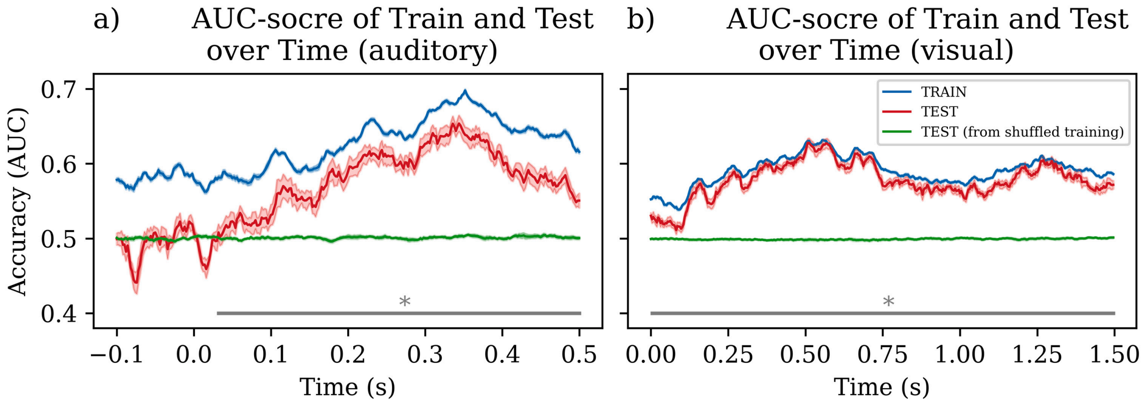


Figure 2.6: Classification performance for a logistic regression classifier trained and tested at every time-point separately. We report the AUC scores of the train (blue) and test set (red), averaged over 10-folds of cross-validation, and the scores of the test set, in the case where the classifier was trained on shuffled data (green). Bold lines show the mean scores over the 10-fold validation and the faded colored regions show the standard error. Horizontal gray lines show the time-periods where classification was significantly above chance levels in the two datasets.

Using logistic regression in the auditory dataset, the classification score of the test set was around 0.5 during the 0.1 s before stimulus onset (Figure 2.6 panel a −0.1 to 0.0 s), and increased after the stimulus onset, reaching a maximum AUC score of 0.63 ± 0.01 on the test set, across folds, at 0.323 ± 0.01 s post-stimulus onset.

In the visual dataset, there was no baseline in the available data (van Peer et al., 2017). Decoding performance started to increase around 0.1 s post-stimulus onset (Figure 2.6 panel b). The maximal decoding performance on the test set was 0.62 ± 0.01, and was reached at 0.70 ± 0.10 s post stimulus onset.

Chance levels, estimated through random permutations, were on average 0.5 throughout the entire trial interval (Figure 2.6, panels a and b, green lines) for both datasets. Decoding performance was significantly above chance levels from 0.032 to 0.5 s post stimulus onset for the auditory, and from 0.0 to 1.5 s post stimulus onset for the visual datasets (Figure 2.6, panels a and b, marked in gray lines).

We also trained and tested a SVM, with the same procedure as for logistic regression. The maximum AUC score reached on the test set was on average 0.58 ± 0.01 at 0.352 ± 0.02 s post-stimulus onset for the auditory and 0.60 ± 0.004 at 0.59 ± 0.06 s post-stimulus onset for the visual dataset.

Next, we contrasted the performance of the CNN with the two baseline MVPA algo-

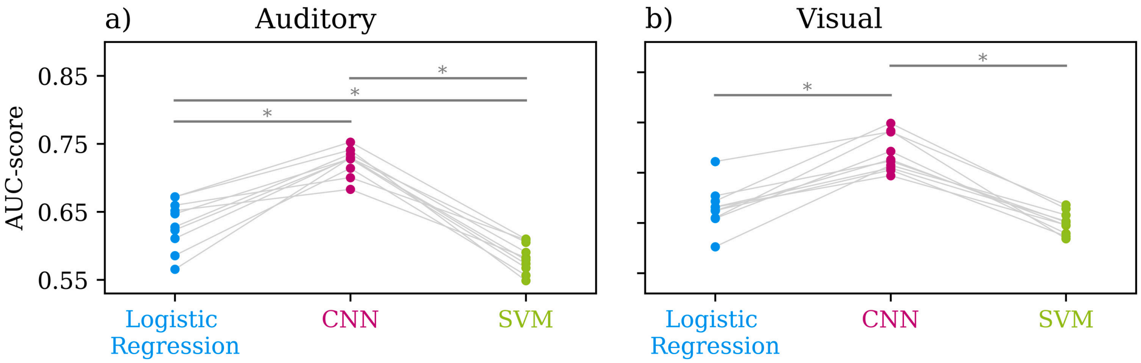


Figure 2.7: Comparison of the performance of logistic regression versus CNN versus SVM for the auditory dataset (panel a) and comparison of logistic regression versus CNN versus SVM for the visual dataset (panel b). Each dot corresponds to the test sets of one of the 10 folds of the cross validation. For the CNN we report the mean AUC score over the 4 trained networks per fold.

rithms. For this comparison, we contrasted the maximum classification performance obtained across time with logistic regression and SVM, to the overall performance obtained with the CNN. This approach is rather conservative, and might penalize the CNN. Nevertheless, in each of the 10 folds of the cross validation, the CNN provided higher classification performance than both the logistic regression and the SVM (Figure 2.7). For both the auditory and visual datasets, the AUC of the CNN was significantly higher than the AUC of logistic regression (Wilcoxon signed-rank test, $p = 0.006$ for both datasets, corrected for multiple comparisons), and than the AUC of SVM (Wilcoxon signed-rank test, $p = 0.006$ for both datasets, corrected for multiple comparisons, Figure 2.7).

2.3.3.2 Comparing different CNN architectures

Additionally, we compared the ResNet50-based pipeline that we developed to other existing CNNs that have been previously used on EEG data, including Shallow and Deep CNNs. The AUC score obtained on the test set with the Shallow CNN was 0.75 ± 0.03 for the auditory and 0.68 ± 0.02 for the visual dataset. The Deep CNN resulted in an AUC of 0.73 ± 0.03 and 0.68 ± 0.01 for the auditory and visual datasets, respectively (Table 2.1 and Figure 2.8). There was no significant difference in AUC values for ResNet50 versus Shallow, ResNet50 versus Deep, or Shallow versus Deep networks (Wilcoxon signed-rank test, $p > 0.08$, corrected for multiple comparisons). These results suggest that the developed pipeline can classify EEG data under different network architectures.

Last, we evaluated the robustness of the developed pipeline under slight modifications

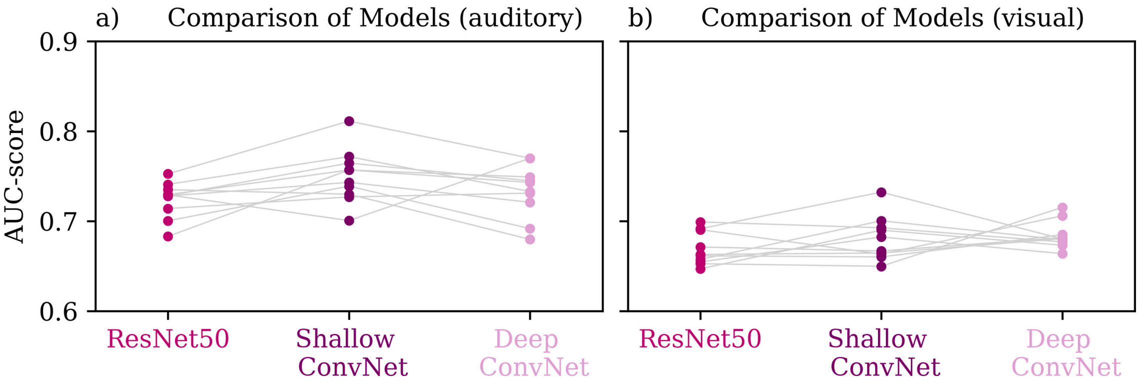


Figure 2.8: Comparison of the performance of the Shallow, Deep convolutional neural networks and ResNet50. The difference in performance was not significant for all comparisons after correcting for multiple comparisons (Wilcoxon signed-rank test, p > 0.08, corrected for multiple comparisons).

in the pipeline architecture or input data (Appendix A.2). Notably, the classification performance remained at similar levels for filtered and unfiltered data (A.2.1), or when omitting the final fully connected layer with 128 nodes before the dropout layer of the CNN (A.2.3). Oversampling of the underrepresented class yielded a higher classification performance than using a weighted binary crossentropy loss (A.2.2).

2.3.4 Extraction of discriminant features

After establishing that the networks can accurately classify *Repeated* from *Novel* stimuli, we next visualized the discriminant features that were driving this classification. Figure 2.9 shows the activation maps of the mean EEG responses to *Repeated* and *Novel* auditory (panels a and b) and visual (panels c and d) stimuli (Figure 2.1). In the representation of activation maps, stronger colors denote that a given electrode and time interval were more relevant in the network's output than lighter ones. For the auditory dataset (Figure 2.9, panels a and b), almost all of the non zero activations appeared after stimulus onset. The highest activation values occurred at different latencies for each experimental condition, ranging from 0.2 to 0.3 s for the *Repeated* and from 0.3 to 0.5 s for the *Novel* trials. For the visual dataset (Figure 2.9, panels c and d), the two experimental conditions (*Repeated* and *Novel*) had a more similar distribution of activations. Most of the non-zero activations for the visual dataset occurred between 0.25 and 0.7 s, at similar latencies for both experimental conditions.

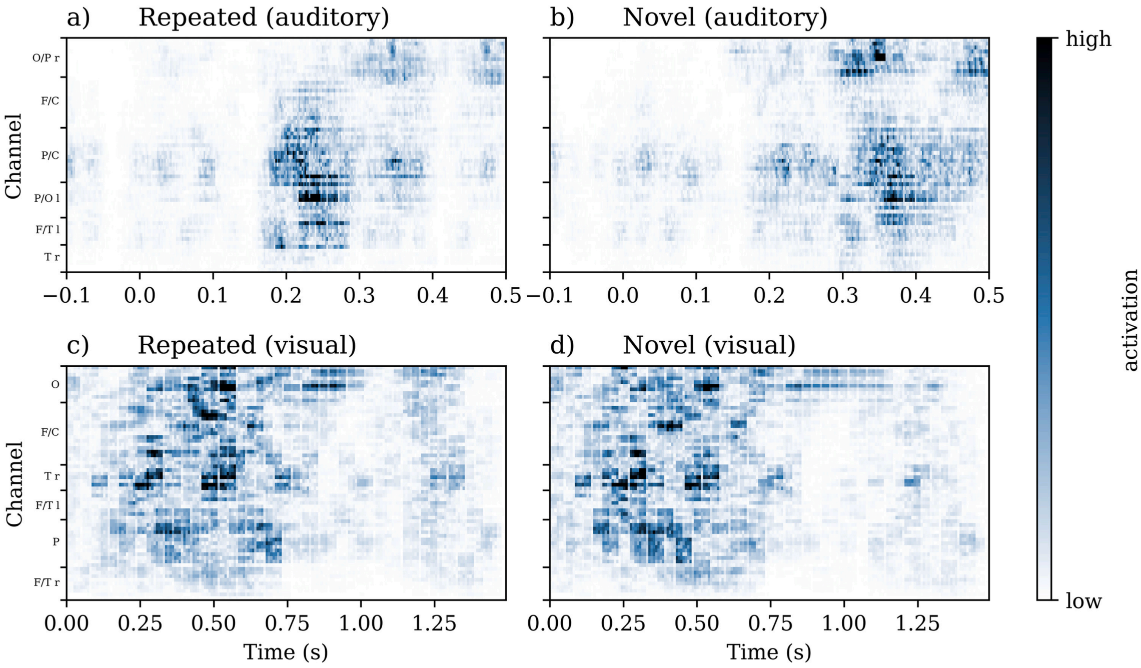


Figure 2.9: The activation maps of the mean trials of the auditory (top) and visual (bottom) dataset for *Repeated* (panels a and c) and *Novel* (panels b and d) trials.

2.3.5 Topographic representation of discriminant features

As an alternative representation, the activations of the CNN were additionally visualized as topographic maps, by reassigning the electrodes to their original location on the scalp. Figure 2.10 shows the activation maps for each dataset and condition as a topographic map, to give a more interpretable visualization of the networks' features. For reasons of consistency, Figure 2.10 provides similar latencies for both datasets. Each topographic map displayed the sum of activations in steps of 0.1 s, as described above the map (Figure 2.10). This topographic representation revealed that the discriminant information for the *Repeated* auditory condition started appearing at $0.2 - 0.3$ s post stimulus onset, mainly at fronto-occipital electrodes (Figure 2.10, panel a). By contrast, for the *Novel* auditory condition, discriminant information was more strongly appearing at centro-parietal electrodes, between 0.3 and 0.4 s post-stimulus onset (Figure 2.10, panel b), matching closely the actual topographic maps of the data (Figure A.1). For the visual dataset, activations were stronger at similar latencies for *Repeated* and *Novel* conditions, starting mainly after 0.2 s post-stimulus onset, and occurring predominantely at central and occipital electrodes (Figure 2.10, panels c and d), closely following the topographic maps of the average data (Figure A.1).

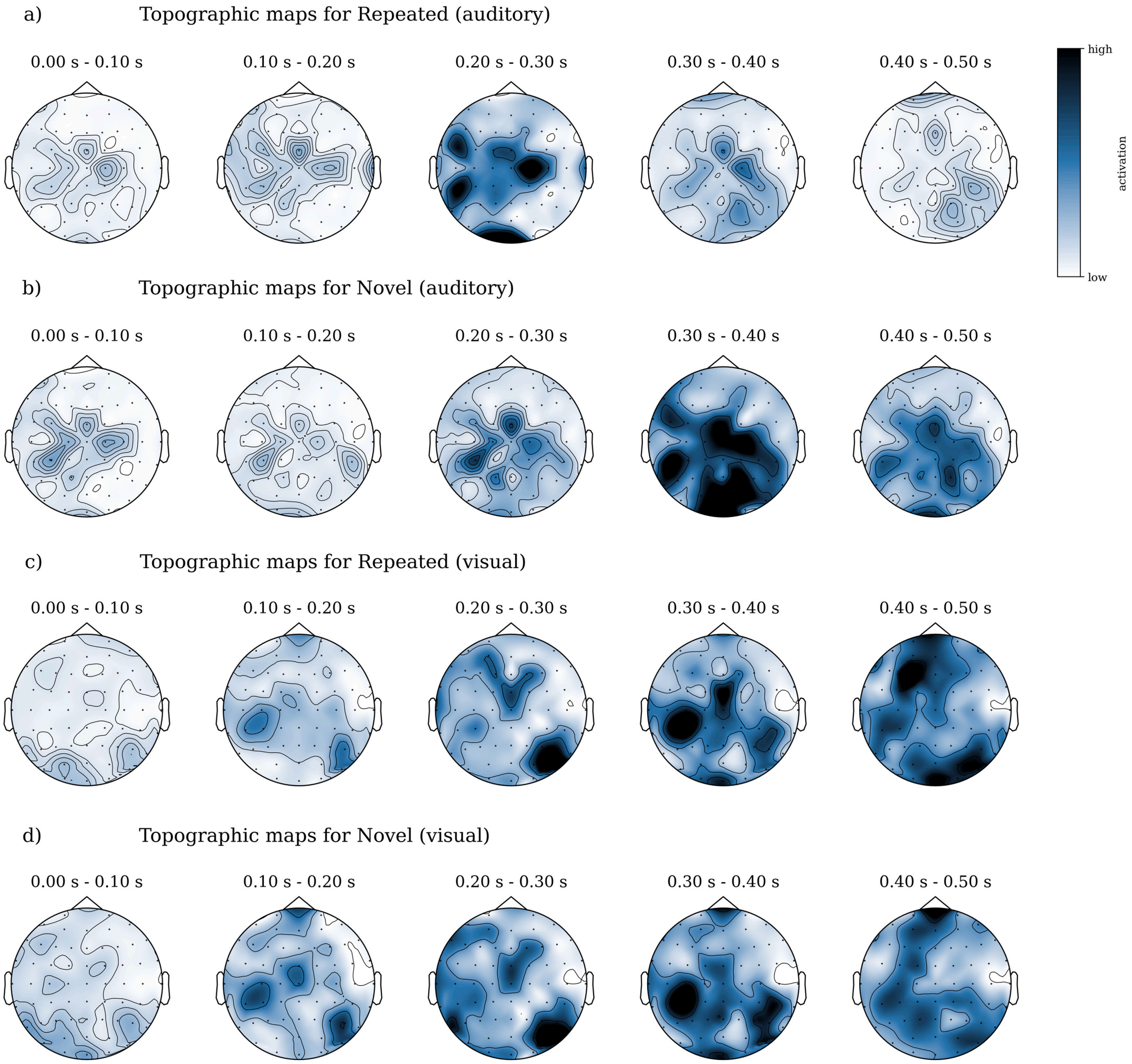


Figure 2.10: Activation maps from Figure 2.9 represented as topographic maps, across datasets and experimental conditions. For reasons of consistency, the topographic activation maps are displayed at the subset of commonly available latencies across the two datasets (i.e. $0 - 0.5$ s post-stimulus onset). Every map corresponds to the sum of activations over intervals of 0.1 s.

2.3.6 Trial by trial changes in discriminant features

As the features of activation maps can be computed at the single-trial level, we performed an exploratory analysis, evaluating trial by trial changes in the activation maps through-

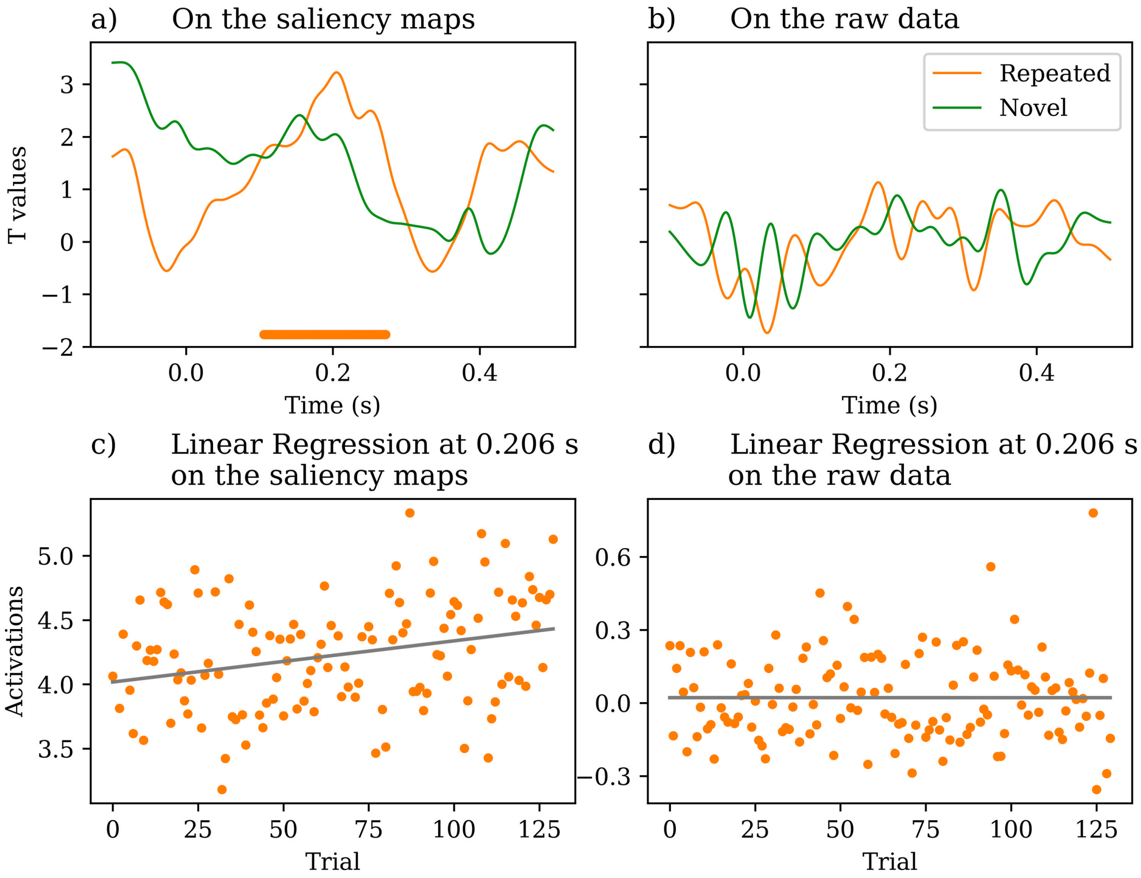


Figure 2.11: Linear regression results quantifying trial by trial changes in the activation maps (panels a and c) and the raw data (panels b and d). The plotted lines correspond to the t-values testing whether the slope of a linear regression was significantly different from zero, for *Repeated* (orange) and *Novel* (green) stimuli. Horizontal lines show periods of significant difference. Panels c and d show trial by trial activations (panel c) or EEG responses (panel d) and regression fit, for the time-point (0.206 s) with the minimal p-value (0.002) for the *Repeated* condition.

out the experiment. Figure 2.11 illustrates the time-course of a linear regression analysis, quantifying trial by trial changes on the activation maps (panel a). The slope of the linear regression was significantly different from zero from 0.106 to 0.272 s for the *Repeated* condition ($p < 0.05$, corrected with cluster-based permutations) (Figure 2.11, orange horizontal line). For the *Novel* condition, there was no period of significant change in the slope of the linear regression throughout the experiment (Figure 2.11, green line).

The same analysis was performed on the EEG data (Section 2.2.7). For the EEG data, the slope of the linear regression was close to zero for both conditions throughout the entire temporal interval (Figure 2.11 panel b), and did not have any periods of significant difference. As an illustration, the trial by trial activations extracted with activation maps and with the raw EEG data at the point of maximum regression (0.206 s post stimulus onset), are displayed in Figure 2.11, panels c and d respectively.

2.4 Discussion

We presented a novel MVPA pipeline for decoding single trial EEG responses to external stimuli and used this pipeline to extract discriminant features at the single trial level. We showed, in two different datasets, that the developed pipeline significantly outperformed commonly used existing MVPA techniques, and that it could detect class-specific discriminant features that are readily interpretable. Our approach resulted in an accurate decoding performance, demonstrated in several ways: (a) generalization of classification to data that the network has not seen during training (test set), (b) generalization of the training pipeline to a different dataset (visual dataset), (c) significantly better classification performance for the original data versus data with shuffled labels, (d) significantly higher classification performance for the network compared to exemplar baseline machine learning algorithms, and (e) comparable decoding performance to existing CNN-based algorithms for EEG data. Additionally, we used feature visualization techniques to characterise the electrodes and time-periods of EEG responses that mostly contribute to an accurate classification. Although several multivariate decoding techniques allow the extraction of discriminant features (Tzovara et al., 2012; Grootswagers et al., 2017), these are typically identified at an across trial level and are shared across experimental conditions. By contrast, our approach allows recovering class- and trial- specific discriminant features, which are informative of the distinct contributions of different experimental conditions to the final classification.

2.4.1 CNNs for MVPA on EEG data

MVPA algorithms have been used to extract patterns of discriminant activity at the single trial level (Lemm et al., 2011; Haufe et al., 2014). The vast majority of these algorithms require a homogeneous dataset as they assume that the discriminant EEG responses appear at consistent latencies and electrode locations across trials (Grootswagers et al., 2017). However, often the discriminant information can be found at different temporal and spatial points over a group of participants or it may shift in time or space over the course of an experiment. The rather conservative approach of most multivariate decoding techniques can result in under-estimation of discriminant features, therefore limiting their interpretability. As CNNs convolve the entire input signal with multiple kernels per layer, they can extract patterns of neural activity which are time- and space- unlocked. Here, we chose a ResNet5050 architecture, which is well known for its breakthrough performance in image classification in computer vision (He et al., 2016). The depth of the network allows it to learn features and find structure in bigger patches of the input data than more shallow convolutional networks (Zeiler and Fergus, 2013). Even though the network was originally developed for classifying images, here we adapted it for the specific case of EEG data.

One main concern of implementing deep learning algorithms for the field of EEG research is that of overfitting the data (Williams et al., 2020). Deep learning architectures typically comprise of hundreds to thousands of hyperparameters, which are prone to overfitting (Srivastava et al., 2014). Data augmentation techniques have been widely used in the field of computer vision (Shorten and Khoshgoftaar, 2019) to overcome this problem. This is a major concern for MVPA, as due to the nature of EEG recordings, it is practically impossible to collect the amounts of data that are often available in computer vision. Here, we overcame this limitation by using data augmentation techniques, which artificially augment the available EEG data and at the same time add variance to them, which makes the network less prone to overfitting to the available data samples. We could exclude that the trained networks were overfitted by showing that they generalize to new data (validation and test datasets). By contrast, the networks trained on randomly shuffled labels did not generalize to test and validation datasets.

The neural networks were trained for a maximum of 50 epochs (Figure 2.4) and the training performance reached a plateau around epoch 15 with a mean score of 0.89 and 0.82 for the auditory and visual datasets respectively. The lower classification performance for the visual dataset was likely due to the nature of the visual event-related potential that resulted in subtle differences between the two experimental conditions (van Peer et al., 2017). The *Repeated* and *Novel* visual stimuli were counter-balanced across participants, and therefore resulted in similar visual features at the average level, where the only difference would be a very subtle difference of the effect of repetition (see also Figure 2.1). By contrast, in the auditory dataset (Cavanagh et al., 2018), *Repeated* and *Novel* sounds were always the same across participants and had very different acoustic characteristics (pure tones vs natural sounds). Therefore, for the auditory dataset, EEG responses were well distinct (Figure 2.1), which resulted in a high classification performance.

When randomly shuffling the labels of the data to estimate a data-driven distribution of chance, the network accuracy could not overcome 0.7 on average for the train data and remained at 0.5 for the test set (Figure 2.7). Taken together, these results suggest that the network was able to learn class-specific features at an above-chance level without overfitting the data.

2.4.2 Comparison with existing techniques

Deep learning techniques have been introduced in the field of EEG research since a few years, but existing techniques predominantly focus on clinical or brain-computer-interfaces applications (Roy et al., 2019). Although deep learning techniques for basic EEG research exist (Kuanar et al., 2018; Bashivan et al., 2016; Wang et al., 2018), these are still at a validation stage, and are seldom used to answer basic research questions. Here, we aimed at examining how deep learning architectures perform in classification tasks that are commonly faced in basic neuroscience, i.e. decoding differences between experimental conditions, and evaluating the stability of decoding features over the course of an experiment. Indeed, the techniques that are readily available for classifying EEG data oftentimes give little to no emphasis on feature interpretability, but rather focus on classification performance (Williams et al., 2020). Although optimizing classification performance is certainly beneficial in basic research, clinical and BCI applications of MVPA algorithms can also profit from interpretable features. Importantly, our approach gave comparable results to other CNN-based architectures for classifying EEG data, based on a Shallow and Deep CNN (Schirrmeister et al., 2017).

Previous studies examining features of CNNs have extracted spectral EEG features (Schirrmeister et al., 2017), which do not contain temporal information, but are collapsed over time and trials. Such an approach is particularly suited for the field of brain-computer-interfaces, where emphasis is given on classification performance, but it is limited for MVPA applications, where emphasis is given on identifying spatial and temporal features that drive an accurate classification (Grootswagers et al., 2017). Other attempts to extract class-specific features based on gradient methods have either reported features at an average level (Farahat et al., 2019; Vahid et al., 2020), or for exemplar trials (Lawhern et al., 2018), without examining how these generalize over time, or how representative they are of the entire dataset. In Figure 2.9 we also show the features on an average level, but additionally we examined how these change over the course of an experiment. (Farahat et al., 2019) and (Lawhern et al., 2018) both explore the features in the context of BCI and (Vahid et al., 2020) and (Lawhern et al., 2018) impose to the used network to start by temporal convolutions followed by spatial ones. Here, instead we consider the EEG data as a spatio-temporal continuum, as it is commonly done in the field of EEG research (Maris and Oostenveld, 2007).

2.4.3 Extraction of class-specific discriminant features

In computer vision, there is a dedicated research area focusing on visualizing which features of the input data are learned by a neural network. Some commonly used techniques consist of visualizing the kernels of the network (which provide only general features for the entire dataset that was used for training), or of gradient-based methods, which backpropagate the input signal through the network (which can reveal class-specific features for each data-point). In neuroscience, there have recently been some studies focusing on feature interpretability (Ghosh et al., 2018; Lawhern et al., 2018; Schirrmeister et al., 2017; Zubarev et al., 2019). Here, we visualize discriminant features with a gradient-based method. The mean activation maps across participants (Figs. 2.9 and 2.10) showed similar spatio-temporal patterns of differential activity as the mean event-related potentials (Figs. 2.1 and A.1).

The advantage of the presented pipeline for MVPA applications is that it allows to identify discriminant features at the single class level. Most existing univariate or multivariate analysis techniques can only identify condition differences, and are agnostic to which experimental condition is driving these differences. Other approaches, like common spatial patterns (CSP), allow the extraction of class specific patterns of EEG activity. However, the CSP components are calculated over a time window and therefore have a poor temporal resolution. Additionally, CSP have the limitation of poor generalization over new participants (Reuderink and Poel, 2008).

With our approach, in the auditory dataset, the strongest activation values were most prominent around 0.2 s post-stimulus onset for the *Repeated* condition and at later latencies, after 0.3 s post-stimulus onset for the *Novel* condition (Figure 2.9 panels a and b). This difference in discriminant intervals for the two conditions is justified by the nature of the auditory stimuli, as *Repeated* sounds were pure tones with a sharp onset time, while *Novel* sounds were naturalistic sounds, which typically have a slower onset, and thus are expected to evoke an EEG response at later latencies (Cavanagh et al., 2018). This information cannot be revealed by existing MVPA techniques, which can only identify at which latency multiple conditions differ. Indeed, our analysis on the same data with an exemplar MVPA approach showed that classification was significantly higher than chance starting after 0.1 s post-stimulus onset, with a prominent peak after 0.3 s (Figure 2.6). However, it is impossible to infer which of the two conditions drives this sustained differential activity. For the visual dataset, the peak in discriminant activity between *Repeated* and *Novel* stimuli appeared at latencies which were qualitatively similar between the two experimental conditions (Figure 2.9 panels c and d). Indeed, in this dataset, the *Novel* and *Repeated* stimuli were all naturalistic images, and were counterbalanced across participants, therefore resulting in similar sensory responses (van Peer et al., 2014). In accordance to previous reports using this dataset, we found that the most prominent differences occurred after 0.1 s post-stimulus onset, and were sustained mostly up to 0.75 s, but also throughout the trial (Figure 2.9 panels c and d). Importantly, when visualized in the form of topographic maps, the discriminant features that were identified through the CNNs match previous reports of this dataset, showing that topographic differences in response to novelty are mainly localized in frontal electrodes (van Peer et al., 2014). For the auditory dataset, the most prominent discriminant features at the topographic level were captured between

0.3 and 0.4 s post-stimulus onset for the *Novel* condition (Figure 2.10). This latency and electrode locations are in accordance with previous reports of this dataset and could reflect a P3a component in response to novelty (Cavanagh et al., 2018).

To highlight the importance of studying discriminant features, we show that activation maps significantly change over the course of the experiment (subsection 2.2.7, Figure 2.11). The positive t-values suggest that there was an increase in network activations over the course of the experiment, consistently observed between 0.106 and 0.272 s, suggesting that EEG responses at this latency activate neurons of the CNN more in later trials of the experiment compared to earlier ones. This could not have been caused by a global change in the strength of neural activity itself as our control analyses on raw EEG data did not find any significant trial by trial changes. Instead, our findings suggest that this increase might be caused by changes in the activation patterns in response to *Repeated* stimuli, which become more distinct across conditions as the experiment unfolds. This interpretation is further supported by the fact that activations increase over the course of the experiment, as participants are increasingly exposed to the presented stimuli, and the two classes (*Repeated* versus *Novel*) start acquiring a more distinct neural representation. Indeed, the observed latency for changes in activation maps is in accordance to the latency of a typical N100 response to auditory stimuli, which is known to habituate with *Repeated* stimuli presentations (Rentzsch et al., 2008).

Studying changes of discriminant features can be relevant not only for basic EEG research, but also for BCI or clinical applications. In BCI applications it is highly relevant to investigate feature interpretability and how these features may manifest or change over long experimental sessions, as this may be relevant to participants' capacity to control an external device (Friedrich et al., 2013). Similarly, studying feature interpretability in clinical applications is particularly important, in order to advance our understanding of which features may underlie algorithmic decisions, which in turn can contribute in gaining novel insights into neurological disorders.

2.4.4 Future directions and limitations

Currently, in the field of EEG research there is a lack of data-driven approaches that can provide information about trial-by-trial changes in EEG responses to external stimuli. Previous studies investigating, for example, learning of new sensory rules, have defined a priori a specific electrode locations and latencies of a response of interest and have examined how these change across trials (Lieder et al., 2013). Here, we refer to an alternative approach, that is 'data-driven' in the sense that features are identified automatically from the data, via means of maximizing discrimination between conditions of interest (i.e. supervised learning), as opposed to a hand-crafted feature selection that relies on a priori hypotheses (Haynes and Rees, 2006). Although the latter approach has been widely used in the literature of basic EEG research, it assumes that the response of interest stays at the same electrode location and latency across all trials, which in cases of changing processes, may not be true. Here, we propose a data-driven approach for identifying discriminant patterns of activity at the single-trial level and show that it is more sensitive than considering raw EEG activity (Figure 2.11). Future studies can apply this approach in experiments involving learning in order to couple changes in discriminant EEG activity with participants' behavior and test learning theories.

One main limitation in the extraction of discriminant features is that they only reveal changes in network activations, but not the cause of these changes. Future experiments could evaluate changes of the network activation within spatial clusters (as identified in Figure A.3), possibly combining those with inverse solutions (He et al., 2006), in order to evaluate the contribution of specific brain regions in significant network activation changes over the course of an experiment (Ieracitano et al., 2021).

Another future direction of research is that of choosing a network architecture. Here, we chose ResNet5050 as an exemplar residual CNN, which has been widely tested in the field of computer vision (He et al., 2016) and biomedical data analysis (Guo and Yang, 2018). This choice is meant as a proof of principle, that demonstrates the feasibility of applying CNNs on MVPA applications for EEG data. Indeed, when changing the network architecture with other CNNs that have been developed for EEG research, classification performance remained at similar levels. Future studies can test different network implementations, to optimize the architecture and range of parameters for specific experimental setups. Additionally, here we chose to focus on a binary classification problem, for reasons of simplicity. Our present pipeline, as well as future attempts, could be easily expanded for multiple classes.

2.5 Conclusion

In summary, we used deep learning techniques to develop a novel MVPA pipeline for EEG data. We showed, in two different datasets, that our pipeline can accurately classify single-trial EEG responses, outperforming existing MVPA approaches, and performing at comparable levels with other deep learning approaches for EEG. Moreover, the neural networks can detect class specific information and discriminant features at the single trial level, a direction that can be used in the future to test theories of learning in a data driven way.

3 Auditory stimulation and deep learning predict awakening from coma after cardiac arrest

Florence M. Aellen[1,2], Sigurd L. Alnes[1,2], Fabian Loosli[1], Andrea O. Rossetti[3], Frédéric Zubler[4], Marzia De Lucia[5], Athina Tzovara[1,2,6,7]

[1] Institute of Computer Science, University of Bern, Bern, Switzerland
[2] Zentrum für Experimentelle Neurologie, Department of Neurology, Inselspital, Bern University Hospital, University of Bern, Bern, Switzerland
[3] Neurology Service, Department of Clinical Neurosciences, Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland
[4] Sleep-Wake-Epilepsy-Center, Department of Neurology, Inselspital, Bern University Hospital, University of Bern, Bern, Switzerland
[5] Laboratory for Research in Neuroimaging (LREN), Department of Clinical Neurosciences, Centre Hospitalier Universitaire Vaudois (CHUV), Lausanne, Switzerland
[6] Sleep Wake Epilepsy Center—NeuroTec, Department of Neurology, Inselspital, Bern University Hospital, University of Bern, Bern, Switzerland
[7] Helen Wills Neuroscience Institute, University of California Berkeley, Berkeley, CA, USA

Aellen, F. M., Alnes, S. L., Loosli, F., Rossetti, A. O., Zubler, F., De Lucia, M., and Tzovara, A. (2023). Auditory stimulation and deep learning predict awakening from coma after cardiac arrest. *Brain.* 146(2):778–788. https://doi.org/10.1093/brain/awac340

Assessing the integrity of neural functions in coma after cardiac arrest remains an open challenge. Prognostication of coma outcome relies mainly on visual expert scoring of physiological signals, which is prone to subjectivity and leaves a considerable number of patients in a ‘grey zone’, with uncertain prognosis. Quantitative analysis of EEG responses to auditory stimuli can provide a window into neural functions in coma and information about patients’ chances of awakening. However, responses to standardized auditory stimulation are far from being used in a clinical routine due to heterogeneous and cumbersome protocols. Here, we hypothesize that convolutional neural networks can assist in extracting interpretable patterns of EEG responses to auditory stimuli during the first day of coma that are predictive of patients’ chances of awakening and survival at 3 months. We used convolutional neural networks (CNNs) to model single-trial EEG responses to auditory

stimuli in the first day of coma, under standardized sedation and targeted temperature management, in a multicentre and multiprotocol patient cohort and predict outcome at 3 months. The use of CNNs resulted in a positive predictive power for predicting awakening of 0.83 ± 0.04 and 0.81 ± 0.06 and an area under the curve in predicting outcome of 0.69 ± 0.05 and 0.70 ± 0.05, for patients undergoing therapeutic hypothermia and normothermia, respectively. These results also persisted in a subset of patients that were in a clinical 'grey zone'. The network's confidence in predicting outcome was based on interpretable features: it strongly correlated to the neural synchrony and complexity of EEG responses and was modulated by independent clinical evaluations, such as the EEG reactivity, background burst-suppression or motor responses. Our results highlight the strong potential of interpretable deep learning algorithms in combination with auditory stimulation to improve prognostication of coma outcome.

3.1 Introduction

Most survivors of cardiac arrest are initially in a coma. Outcome prognostication has become an integral part of post-resuscitation care (Rossetti et al., 2016; Perkins et al., 2021). Currently used outcome prediction techniques mainly rely on expert multi-modal assessments of clinical variables and physiological signals (Rossetti et al., 2016) like electroencephalography (EEG), which is routinely used to evaluate the integrity of neural functions at the patients' bedside. EEG evaluations consist of visual assessments which can be time consuming and prone to subjectivity (Westhall et al., 2015). In addition, current clinical markers for outcome prognostication are unable to provide a clear prognosis for a considerable proportion of patients, classifying them as indeterminate, or part of a 'gray-zone' (Perkins et al., 2021), and highlighting a clear need for developing novel markers of outcome.

A putative marker for assessing the integrity of neural functions in coma patients are EEG responses to auditory stimulation (Fischer et al., 2004; Daltrozzo et al., 2007; Morlet and Fischer, 2014; Liu et al., 2021). Auditory event-related potentials (ERPs) have been previously linked to chances of awakening from coma (Fischer et al., 2004; Daltrozzo et al., 2007; Morlet and Fischer, 2014; Liu et al., 2021), but standardized ERPs, assessed in a quantitative way, are not routinely used for outcome prognosis (Nolan et al., 2021). Typically, auditory ERPs are evaluated by averaging hundreds of EEG responses to the same standardized auditory stimuli, and extracting aggregate characteristics, like the presence or absence of characteristic ERP deflections, like the N100 (Fischer et al., 1999, 2008), their amplitude or latency (Fischer et al., 1999), or differential responses to sequences of standard and deviant sounds (Fischer et al., 1999; Kane et al., 1993; Naccache et al., 2005; Luauté et al., 2005). These approaches have the disadvantage that the features used to predict coma outcome are selected a priori at an average ERP level, overlooking the richness of EEG responses, and neglecting potentially important characteristics, thus likely leading to unreliable prognosis.

More recent attempts to explore the prognostic value of auditory stimulation consist of modeling single-trial EEG responses to sounds, with the use of machine learning techniques to extract patient-specific EEG patterns that quantify auditory discrimination (Tzovara et al., 2013, 2016). The progression of auditory discrimination from first to second day of coma is informative of patients' chances of awakening (Tzovara et al., 2016; Pfeiffer et al., 2017, 2018). Moreover, the neural synchrony across voltage measurements of auditory EEG responses in the first day of coma is also predictive of awakening, further corroborating the early prognostic value of auditory ERPs (Alnes et al., 2021). However, despite clear links between auditory processing in coma and patients' outcome, shown over multiple studies and approaches, standardized auditory stimulation is currently not used in the clinical routine as a prognostic marker. A major limitation for this discrepancy is that existing studies report findings either in small patient cohorts with highly curated features (i.e. average EEG responses over pre-defined time windows and specific electrodes) and limited predictive power (Morlet and Fischer, 2014; Liu et al., 2021; Fischer et al., 1999; Naccache et al., 2005; Luauté et al., 2005), or require two EEG recordings over two consecutive days (Tzovara et al., 2013, 2016; Pfeiffer et al., 2017, 2018). In order to fully exploit the multidimensional ERP features and their relevance to coma outcome, there is a critical need for assessing EEG responses to auditory stimulation in a more robust and straightforward way.

In recent years, advances in the field of machine learning have given rise to powerful tools for modeling brain signals (Craik et al., 2019; Roy et al., 2019). Convolutional neural networks (CNNs) are particularly promising in extracting in a data-driven way rich features of EEG data, and have been shown to outperform traditional techniques (Craik et al., 2019; Roy et al., 2019; Lawhern et al., 2018; Aellen et al., 2021). Despite their huge potential, the use of CNNs in acute neuro-critical prognostication remains limited. One challenge in using CNNs in medical applications is that it is difficult to trace which features of the EEG data are relevant for the decisions that CNNs are making (Aellen et al., 2021). The very few studies that have used CNNs to predict outcome from coma rely on the same continuous EEG recordings of resting state activity that are used in the clinics via visual evaluations (Tjepkema-Cloostermans et al., 2019; Jonas et al., 2019; Zheng et al., 2021; Altıntop et al., 2022), and these networks have shown a remarkable precision in discriminating patients who later survive from those who do not. It remains unknown whether CNNs can be applied on EEG responses to standardized auditory stimuli to assist in outcome prognosis, and to provide additional insights for those patients for whom existing clinical assessments do not result in a conclusive prognostication.

Here, we made the hypothesis that CNNs would be able to extract patterns of EEG responses to standardized auditory stimuli that relate to patients' chances of awakening from coma and survival at three months. We additionally hypothesized that the outcome prediction of CNNs would be complementary to currently used clinical variables for prognostication, and would have the potential to improve prognosis for patients with indeterminate prognosis. Last, we performed exploratory analyses to identify which features of the EEG data are relevant for the outcome prediction provided by the CNNs. We extracted measures of confidence of the network's decisions, and linked those to (a) clinical variables currently used for outcome prognosis (Westhall et al., 2016), and (b) features of EEG responses to sounds that quantify neural synchrony, which have been recently shown to be informative of patients' outcome (Rossetti et al., 2016; Alnes et al., 2021; Westhall et al., 2016). To this aim, we analyzed EEG responses to auditory stimulation during the first day of coma, recorded in a multi-center and multi-protocol cohort of coma patients following cardiac arrest at four different hospitals (Nielsen et al., 2013).

3.2 Materials and methods

3.2.1 Patients and procedure

We recorded data from a cohort of 145 comatose patients following cardiac arrest (33 female, 63.3 ± 1.2 years old, mean ± standard error)), admitted to the intensive care units of the University Hospitals Lausanne (121 patients), Bern (18 patients), Sion (4 patients), and Fribourg (2 patients) between December 2009 and April 2017. Patients have been previously described in (Pfeiffer et al., 2017, 2018; Alnes et al., 2021). Informed written consent was obtained prior to EEG recordings from a family member, legal representative, or treating clinician not involved in this study. The ethical committees of the Cantons of Bern, Fribourg, Valais and Vaud, Switzerland approved the experimental protocol.

Upon admission to the hospital, patients were in an acute coma, i.e. a score of less than 6 on the Glasgow Coma Scale and treated with targeted temperature management for 24 hours; 79 patients were treated with targeted temperature management at 33°C (therapeutic hypothermia; HT) and 55 patients at 36°C (normothermia; NT). Controlled temperature treatment was based on ice packs or intravenous ice-cold fluids together with a feedback controlled cooling device (Arctic Sun System, Medivance, Louisville or Thermogard XP; ZOLL Medical, Zug, Switzerland), for 24 hours after cardiac arrest and was subsequently removed. Propofol (2–3 mg/kg/h), Midazolam (0.1 mg/kg/h) and Fentanyl (1.5 lg/kg/h) were administered for analgesia and sedation, while Vecuronium, Rocuronium, or Atracurium for controlling shivering.

Decisions to withdraw clinical care were based on a multidisciplinary approach (Rossetti et al., 2016). Namely, care was withdrawn if two or more of the following criteria were present at 72 hours or more, after sedation weaning (Rossetti et al., 2010; Tsetsou et al., 2018): unreactive background EEG; epileptiform EEG and/or myoclonous that was resistant to treatment; incomplete return of brainstem reflexes; bilateral absence of N20 in somatosensory evoked potential; the concomitant presence of major hypoxic/ischemic lesions in structural magnetic resonance imaging and neuron-specific enolase levels more than twice above 75 μgl were additionally considered (Nolan et al., 2015). Importantly, all clinical decisions were blinded to the output of the neural networks.

Patients' outcome was defined at three months after cardiac arrest via a semi-structured phone interview via the Cerebral Performance Category (CPC) (Booth et al., 2004). A CPC of 1 indicates full recovery; of 2 return of consciousness with moderate disability; a CPC of 3 consciousness with severe disability; while a CPC of 4 coma or persistent vegetative state, and a CPC of 5 death. For our analyses we considered patients with CPC 1–3 at 3 months after coma onset as Survivors (N=79). Patients with a CPC of 5 were considered as patients with poor outcome (Non Survivors, N = 55). In our cohort, no patient was classified with a CPC of 4, possibly due to the clinical practice in the participating hospitals, where the decision to withdraw life sustaining treatment for patients who fail to regain consciousness is regularly reassessed even after the acute phase.

In accordance with previous investigations focusing on prediction of outcome in coma patients following cardiac arrest (Pfeiffer et al., 2018), we did not analyze patients that regained consciousness during their stay at the hospital but later died, for example because of other comorbidities. This resulted in the exclusion of 11 patients, resulting in a cohort of 134 patients in total.

3.2.2 Auditory stimulation protocol

EEG recordings were conducted at the patients' bedside, within 24 hours after cardiac arrest while all patients were in a comatose state. Patients were presented with a series of pure tones as previously described (Tzovara et al., 2016; Pfeiffer et al., 2017, 2018). Tones consisted of 16-bit stereo sounds, sampled at 44.1 kHz, with a 10 ms linear amplitude envelope applied at stimulus onset and offset to avoid clicks. Between each sound, there was a 700 ms interstimulus interval. Standard sounds were presented in 70% of the trials and had a pitch of 1000 Hz and duration of 100 ms. Deviant sounds differed from the standards in duration (150 ms), interaural time difference (left ear leading with 700 μs), or pitch (1200 Hz). Stimuli were presented in a pseudo-randomized order, in a way that at least one standard sound was presented between two deviant stimuli. The auditory stimulation protocol consisted in total of 1500 tones, split into three blocks, each lasting ~7 min. Similar to a previous study investigating neural properties of auditory processing in coma (Alnes et al., 2021), in the present study we focused on responses to standard and duration deviant sounds, as they have been previously shown to be highly informative of coma outcome (Tzovara et al., 2016; Alnes et al., 2021). A detailed evaluation of the predictive value of all EEG responses to all sound types is provided in the Appendix, Section B.4.

3.2.3 Recording setup and preprocessing

As the data of this study was multi-center and multi-protocol, EEG was recorded at the patient's bedside with 19 or 62 electrodes, depending on the original study design (Tzovara et al., 2016; Pfeiffer et al., 2017, 2018; Alnes et al., 2021), positioned according to the international 10-20 system. The data were collected with a sampling frequency of either 1000 or 1200 Hz. For consistency, data recorded with 1200 Hz were down-sampled to 1000 Hz and data recorded with 62 electrodes were reduced to the overlapping set of 19 electrodes. Across all channels the impedance was kept below 10 kΩ. The online reference for the electrodes was set as Fpz and in the course of preprocessing they were re-referenced to a common average reference. Epochs were extracted from 50 ms before stimulus onset to 500 ms after. Artifacts were rejected with a criterion of ± 100 μV on all electrodes. Noisy electrodes were interpolated using three dimensional splines (Perrin et al., 1987). After re-referencing, the data were filtered from 0.1 to 40 Hz, and in a control analysis, to ensure that our findings are not driven by muscle activity, from 0.1 to 20 Hz. Additionally, the EEG epochs were visually inspected and noisy epochs were manually removed. After preprocessing, we obtained 347.22 ± 9.23 trials (mean ± standard error) per patient, 204.76 ± 8.66 in response to standard sounds and 142.46 ± 1.42 in response to duration deviants.

3.2.4 Training of CNN

For training the neural network we adopted a 10-fold cross validation procedure, i.e. we split the patients ten times into three different sets of train, validation and test patients, in a way that data from each patient was included in only one of the three sets. The train set had 60% of patients (80 patients), the validation set 20% (27 patients) and the test set 20% (27 patients). The single trials of all patients from the train set were used to train the neural networks and the trials from the validation set were used to evaluate any hyperparameters (e.g. learning rate, early stopping, etc.). The test set was used only after training and optimizing the network, for an objective evaluation of the model's performance on unseen patients.

We trained a convolutional neural network called EEGNet (Lawhern et al., 2018) to predict patients' outcomes, which has been designed specifically for EEG data. The overall network size is moderate, giving smaller training times while still achieving robust results. The original network architecture was changed slightly for the purpose of our study. First,

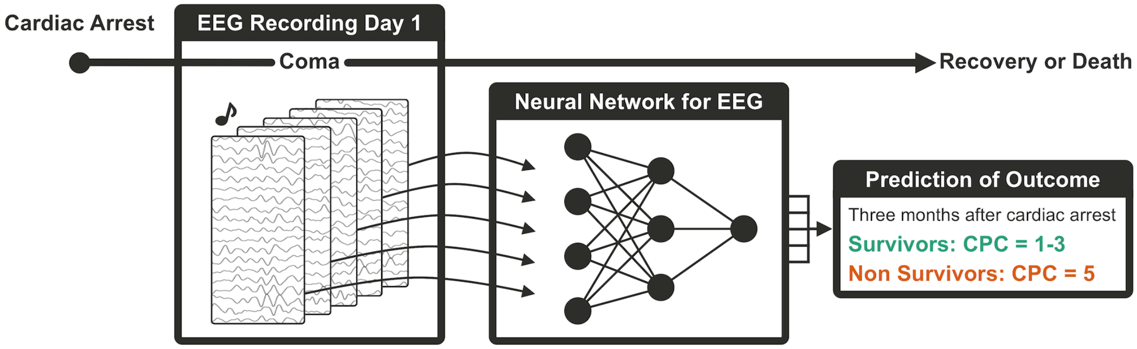


Figure 3.1: Schematic description of the deep learning algorithm for predicting outcome from coma. Auditory stimulation and EEG recordings are performed on the first day of coma, shortly after a cardiac arrest. Single-trial EEG responses are then given as input to a convolutional neural network, which classifies them as belonging to a survivor versus non survivor. The average of all single-trial predictions provides the network's confidence of predicting survival. Survival is defined at three months via the CPC score.

according to the recommendations as in the original study (Lawhern et al., 2018) the filter length of the first convolutional layer was changed to 512, instead of the original 64. Consequently, the number of temporal, spatial and pointwise filters were increased to 16, 4 and 64 respectively, while the activation function was set to ReLu. These changes were made as preliminary analysis on a small subset of patients showed a more stable training.

For optimizing the model we used the binary cross-entropy loss function and the Adaptive Moment Estimation (Adam) (Kingma and Ba, 2014), with a learning rate of 5×10^{-6}, all other parameters were unchanged from the default suggestions. We trained the network for a maximum of 100 epochs, but employed early stopping after 30 epochs if the validation loss of two consecutive epochs was smaller than a threshold, as it is standard practice in the field (Prechelt, 1998). To evaluate the network performance we additionally used the area under the Receiver Operator Characteristic curve (AUC) (Macmillan and Creelman, 2004).

3.2.5 Outcome prediction based on network's output

As the neural network was trained on single-trial EEG responses to sounds (Figure 3.1), it classified single-trial EEG responses as belonging to a survivor versus non survivor. To link the network's output to coma outcome at the single patient level, for each patient we computed the mean classification performance across all trials, which we term "confidence of predicting survival". The confidence score ranged between zero and one and indicated the confidence of the model's prediction of coma outcome, as the mean label that the model assigned across single-trial EEG responses to sounds of a given patient. If that score was above 0.5, then a patient was classified as a survivor, while below 0.5 as a non survivor, as per convention based on a sigmoid function (Figure 3.1).

For evaluating the network's performance, for all metrics we report mean ± standard error values obtained across the 10-folds of cross validation. We additionally plot the network's output in relation to patients' outcome for the best model, defined as the one with the highest AUC score of the validation set. The reported final results remain objective, as they concern the test set of patients which was not used to train, optimize, or select the best model. We additionally evaluated the positive predictive value (PPV) and negative predictive v (NPV), as the ratio of correctly predicted survivors among all predicted survivors (PPV), or correctly predicted non survivors among all predicted non survivors (negative predictive v (NPV)).

3.2.6 Outcome prediction in patients with indeterminate prognosis

To assess the additional utility of our method for current clinical practice, we evaluated the network's predictions for a subset of coma patients whose outcome prognosis was inconclusive based on existing clinical tests. Based on previous literature and recommendations, we considered the motor response, EEG reactivity, EEG continuity and brainstem reflexes (Rossetti et al., 2016; Perkins et al., 2021). If the above-mentioned variables showed a discrepancy (e.g., present motor response and brainstem reflexes, but a discontinuous and irritative EEG), a patient was defined as being in a 'gray zone'. For this definition, we did not include patients where only brainstem reflex was present and all other variables predicted a negative outcome, due to the low positive predictive power of brainstem reflex for good outcome (Rossetti et al., 2016). This resulted in 48 patients fulfilling the criteria for a clinical 'gray zone' (32 survivors, 16 non survivors). Due to their relatively low number, for this analysis we merged patients from train, validation and test sets, and examined the overall distribution of the outcome prediction resulting from the CNN. In the Appendix in the Section (B.1) we show two additional analyses on patients in an uncertain state. The first one was done on a test set containing only patients belonging to the 'gray zone' and for the second network all 'gray zone' patients were part of the test set.

3.2.7 Exploring links between network's output and electrophysiological features of EEG responses

To explore the features of EEG responses related to the network's decisions, we explored (a) measures of neural synchrony and complexity, previously shown to relate to patients' outcome and presence of consciousness, respectively (Alnes et al., 2021) and (b) well established clinical variables currently used for outcome prognosis (Rossetti et al., 2016; Perkins et al., 2021).

We first computed the phase-locking value (PLV), which quantifies the synchrony of EEG responses to auditory stimuli (Lachaux et al., 1999). The PLV was computed for electrode pairs in the alpha range, which has been recently shown to be predictive of patients' outcome in two different cohorts of patients undergoing therapeutic hypothermia (Alnes et al., 2021). Here, we calculated the mean PLV per patient across electrode pairs, applying the same procedure as previously reported, based on a subset of patients included in the present study (Alnes et al., 2021). The neural complexity was quantified via the Lempel-Ziv (LZ) complexity, which measures the number of unique patterns present in a signal (Lempel and Ziv, 1976). Both PLV and LZ were calculated on single-trial EEG responses, similar to the neural network, to capture trial-by-trial characteristics of EEG responses to the auditory stimuli. Here, we expanded the analysis that was recently reported for both measures, on a larger patient cohort. Our goal on the one hand was to link these measures to the network's output (via Pearson correlation coefficient, $p_{\mathrm{corr}} < 0.01$, Bonferoni corrected), and on the other, to validate these measures in patients treated with targeted temperature management at 36°C, as they were previously only reported for patients treated at 33°C. We excluded 2 patients that were part of a previous study (Alnes et al., 2021), because they awoke in the hospital but subsequently died before the CPC assessment at 3 months (see Patients and procedure).

Second, we compared predictions of the network in subgroups of patients defined according to the following binary markers: presence of brainstem reflexes (pupil and corneal reflexes present), presence of motor response (motor GCS ≥ 4), reactive EEG background (change in amplitude and/or frequency after stimulus, judged visually), discontinuous or suppressed EEG background (Hirsch et al., 2021) and irritative EEG (presence of electrographic seizures or status epilepticus, sporadic epileptiform discharges, spiky or sharp periodic discharges or rhythmic spike waves), using Mann-Withney tests ($p_{\mathrm{corr}} < 0.01$), and Bonferroni correction for multiple comparisons. We additionally explored correlations between the network's confidence and time to return of spontaneous circulation (ROSC), with a Pearson correlation coefficient, as well as links between confidence and hospital site, or CPC at three months with Kruskal-Wallis H-tests.

Table 3.1: Prediction of outcome of the trained neural networks. We report the mean ± standard error over all ten trained folds, as well as the performance of the best fold. We report the AUC, PPV and NPV, with respect to survival, of the train, validation and test sets, for all patients and separately the sub-cohorts of patients treated with hypothermia and normothermia. The performance of the best fold is also visualized in (Figure 3.2)

	AUC-score			PPV-score			NPV-score		
	Train	Validation	Test	Train	Validation	Test	Train	Validation	Test
Mean over 10-folds									
All	0.81±0.00	0.75±0.03	0.70±0.04	0.90±0.01	0.85±0.02	0.83±0.03	0.70±0.01	0.66±0.04	0.57±0.04
Hypothermia	0.81±0.01	0.72±0.04	0.69±0.05	0.92±0.02	0.86±0.02	0.83±0.04	0.65±0.02	0.56±0.08	0.53±0.05
Normothermia	0.80±0.01	0.72±0.05	0.70±0.05	0.87±0.01	0.71±0.10	0.81±0.06	0.70±0.02	0.66±0.06	0.57±0.05
Best fold									
All	0.79	0.83	0.83	0.86	0.83	0.92	0.71	0.89	0.71
Hypothermia	0.72	0.93	0.91	0.82	0.90	1.00	0.60	1.00	0.67
Normothermia	0.88	0.68	0.73	0.93	0.75	0.75	0.83	0.67	0.75

3.3 Results

3.3.1 Outcome of coma patients

Out of the 134 patients analyzed, 79 (59 %) survived, where survival was defined as CPC 1 (46 patients), 2 (22 patients) and 3 (11 patients) at three months. 55 out of 134 patients (41 %) had a poor outcome, corresponding to a CPC of 5, and no patient was in a vegetative state (CPC of 4).

3.3.2 Prediction of outcome based on the neural network

The neural networks trained to discriminate auditory EEG responses of patients that later survived from those who did not reach a mean AUC score of 0.81 ± 0.00 on the train, 0.75 ± 0.03 on the validation and 0.70 ± 0.04 on the test sets (Figure 3.2, Table 3.1). On the test set we obtained a PPV of 0.83 ± 0.03 and a NPV of 0.57 ± 0.04 (Table 3.1). For patients treated with targeted temperature management at 33°C, the average PPV was 0.83 ± 0.04 and for those treated at 36°C 0.81 ± 0.06. The difference in PPV over the two treatments over the ten folds was not significant ($p = 0.67$, Wilcoxon signed rank test), implying that the networks performed at similar levels for patients treated with targeted temperature management at 33 and 36°C. The AUC scores were also replicated with a control analysis where the neural networks were trained with EEG data filtered between 0.1 and 20 Hz, resulting in a mean AUC score of 0.74 ± 0.03, PPV of 0.86 ± 0.02 and NPV of 0.62 ± 0.03 on the test set.

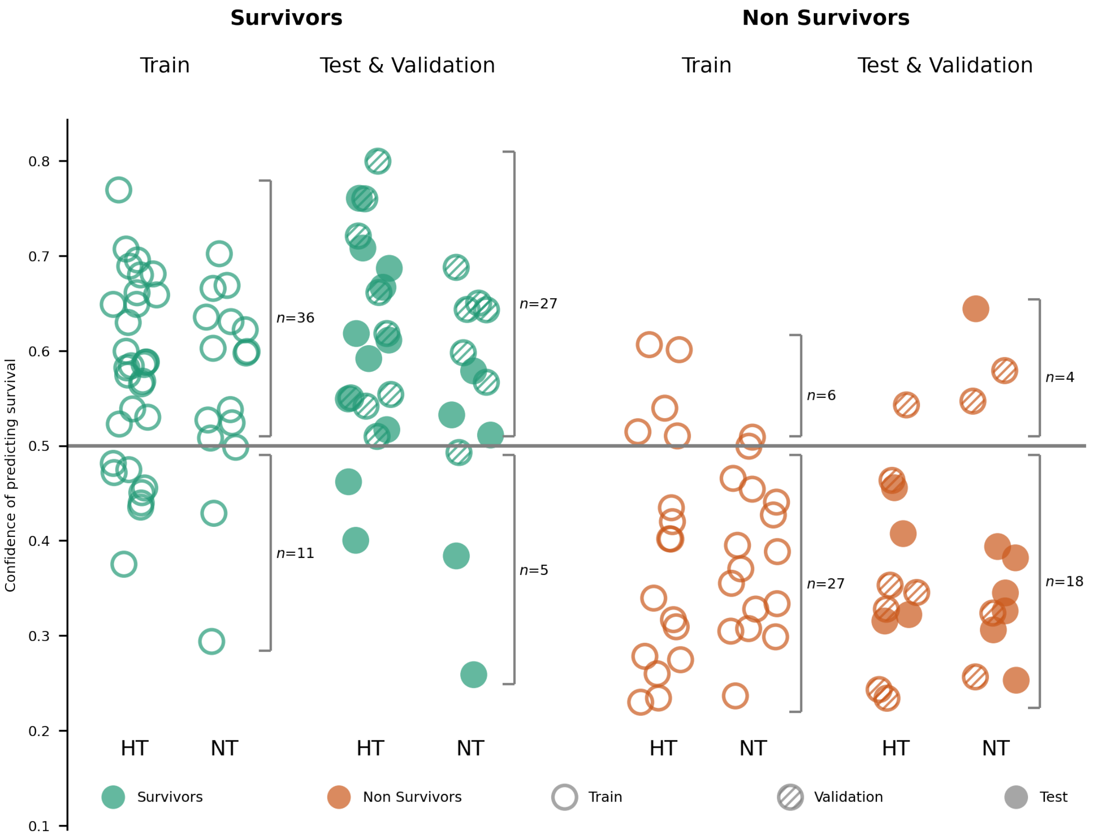


Figure 3.2: Prediction of outcome based on the convolutional neural network. Confidence scores were computed for each patient by averaging the network's outcome prediction for all single-trial EEG responses to sounds of a given patient. A patient was predicted to be a survivor if the confidence score was above 0.5, otherwise a non-survivor. Data from patients of the train set (empty circles) were used to train the network, while data from patients in the validation set (shaded circles) were used to evaluate hyperparameters. The predictive value of the network was evaluated on a separate test set of patients (full circles), whose data were never used to train or optimize the network. For each of the sets, we split patients into HT (hypothermia, targeted temperature management at 33°C) and NT (normothermia, patients with targeted temperature management at 36°C). For the numerical performance scores see Table 3.1.

We next focused on one single fold of the cross-validation, and evaluated the confidence scores assigned by the network to individual patients (Figure 3.2), as our goal was to investigate the neural properties of EEG signals that may mediate the network's outcome

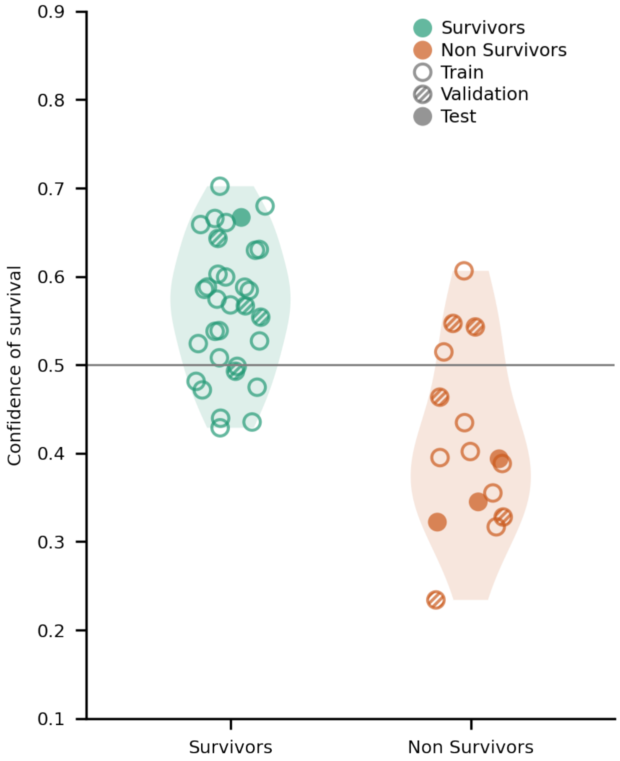


Figure 3.3: Confidence of survival assigned by the network for patients with uncertain outcome prognosis based on existing outcome predictors. Confidence scores for patients in this subset followed the distribution of confidence scores in the entire patient cohort (Figure 3.2). The empty circles show patients in the train set, hatched circles patients in the validation and full circles patients in the test set. The green circles represent survivors and the orange ones non survivors.

prediction. In the validation and test sets, we obtained a sensitivity of 84% for survivors and a specificity of 82% for non survivors. Out of all the patients classified as survivors in the train set (N = 42 patients), 36 awoke from coma, resulting in a PPV of 86%, while in the validation/test sets 27 out of 31 patients that were predicted as survivors awoke, resulting in a PPV of 87%. These results were largely similar for patients treated with different temperature treatments (Figure 3.2 HT, NT and Table 3.1).

3.3.3 Outcome prediction for patients in a 'gray zone'

We next focused on patients who, from a clinical viewpoint, were part of a 'gray zone', i.e. cases where currently used outcome predictors indicated indeterminate outcome. 48 patients fulfilled these criteria (N = 32 survivors and N = 16 non survivors, Figure 3.3). The distribution of confidence values assigned by the network in this subset of patients followed the distribution of the full cohort (Figure 3.2 and 3.3). For this subset of patients we obtained a PPV of 0.86, NPV of 0.60 and AUC score of 0.75, based on the CNN. These scores were at similar levels as the ones obtained with the full cohort, suggesting that although this group of patients had indeterminate prognosis based on existing clinical tests, they were not ‘peculiar’ cases for the neural network.

3.3.4 Investigating the interpretability of the neural network's outcome prediction

We next evaluated the electrophysiological properties of EEG signals that the network's output may be reflecting while providing a prediction about patients' outcome. We first evaluated the PLV of EEG responses to sounds, previously shown to be predictive of patients' outcome when treated with hypothermia (Alnes et al., 2021). Here, we first replicated these results for patients in normothermia. The mean PLV for survivors treated with hypothermia was 0.75 ± 0.01 and for survivors treated with normothermia 0.72 ± 0.02. For non survivors, we found a mean PLV of 0.55 ± 0.02 and 0.57 ± 0.03 for patients treated with hypothermia and normothermia, respectively. When statistically tested, we found a significant main effect of outcome on PLV ($F = 101.68$; $p_{corr} < 0.01$), while nor the main effect of temperature treatment ($F = 4.07$; $p_{corr} = 0.046$), neither the outcome by temperature interaction were significant ($F = 2.17$; $p = 0.14$). Next, we explored the predictive power of the PLV in patients treated with normothermia, as done previously for the subset of patients treated with hypothermia (partially overlapping with those included in our previous study (Alnes et al., 2021)). For patients in hypothermia, the PLV provided a PPV of 0.85 and NPV of 0.83, as previously reported (Alnes et al., 2021). For patients in normothermia, the PLV resulted in a PPV of 0.77 and NPV of 0.80, (Figure 3.4A).

Importantly, the confidence of survival assigned by the network to each patient, strongly correlated with the mean PLV across electrodes ($r = 0.76$; $p_{corr} < 0.01$) (Figure 3.4B). This correlation was not trivially driven by the fact that both measures predict outcome, as it remained significant when tested for survivors ($r = 0.49$; $p_{corr} < 0.01$) and non survivors ($r = 0.65$; $p_{corr} < 0.01$) separately.

Next, we computed the Lempel-Ziv complexity of EEG responses to sounds, which by itself has been shown to not be predictive of outcome when only including patients treated at 33°C (Alnes et al., 2021). We confirmed this previous finding for patients of the present study who were treated at 36°C, whose distribution of LZ values was similar to the distribution of LZ values for patients in treated at 33°C (Figure 3.4C).

Interestingly, although LZ complexity by itself was not predictive of outcome, it showed a significant negative correlation with the network's confidence ($r = -0.57$; $p_{corr} < 0.01$; Figure 3.4D), so that the higher the network's confidence in predicting survival, the lower the complexity of EEG responses to sounds. A strong correlation between network's con-

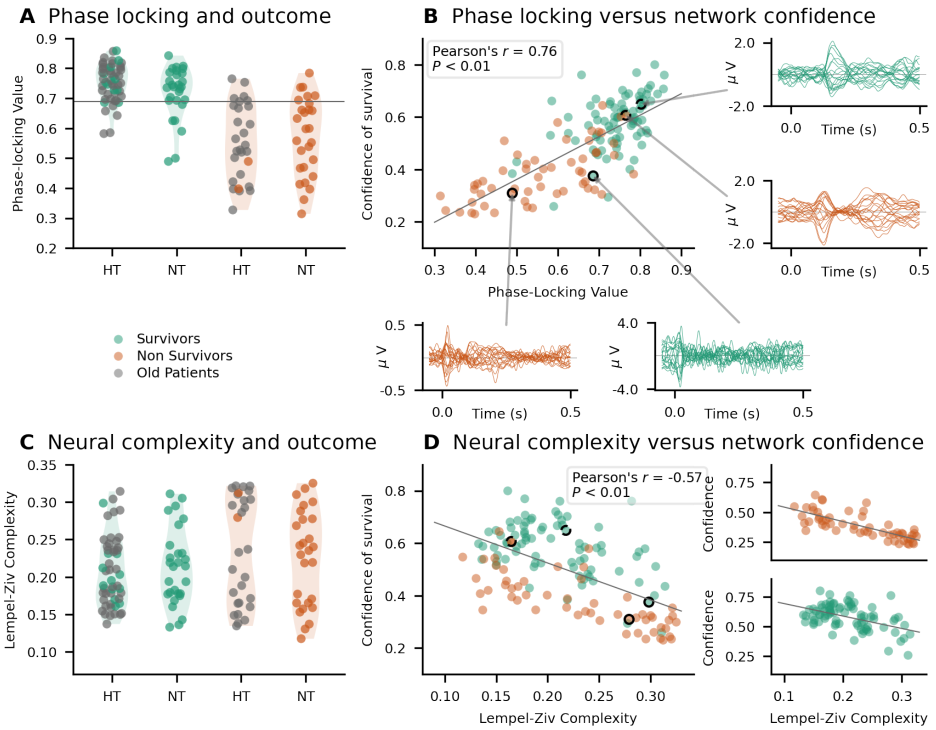


Figure 3.4: Investigating the interpretability of the neural network's output. **(A)** phase-locking value (PLV) for survivors (left) and non-survivors (right). The horizontal line corresponds to a threshold in PLV of 0.69, identified and already evaluated in a subset of patients (gray circles), previously reported in (Alnes et al., 2021). The previously reported distribution of PLV values was replicated in a new cohort, predominantly treated with NT. **(B)** Correlation of network confidence to PLV. The network's confidence of survival was strongly correlated to the average PLV (Pearson's r = 0.76; $P_{corr} < 0.01$). Exemplar EEG responses to sounds have been plotted for two survivors and two non-survivors, one correctly and one incorrectly classified based on the neural network. **(C)** LZ complexity of auditory ERPs for each patient. As previously reported in a subset of patients (Alnes et al., 2021), LZ complexity was not informative of patients' outcome. **(D)** Correlation between LZ complexity and the network's confidence of survival for the entire patient cohort (r $= -0.53$; $P_{corr} < 0.01$). The correlation values are also plotted separately for survivors (bottom plot) and non-survivors (top plot). The highlighted circles with borders mark the patients whose exemplar EEG responses are shown in B.

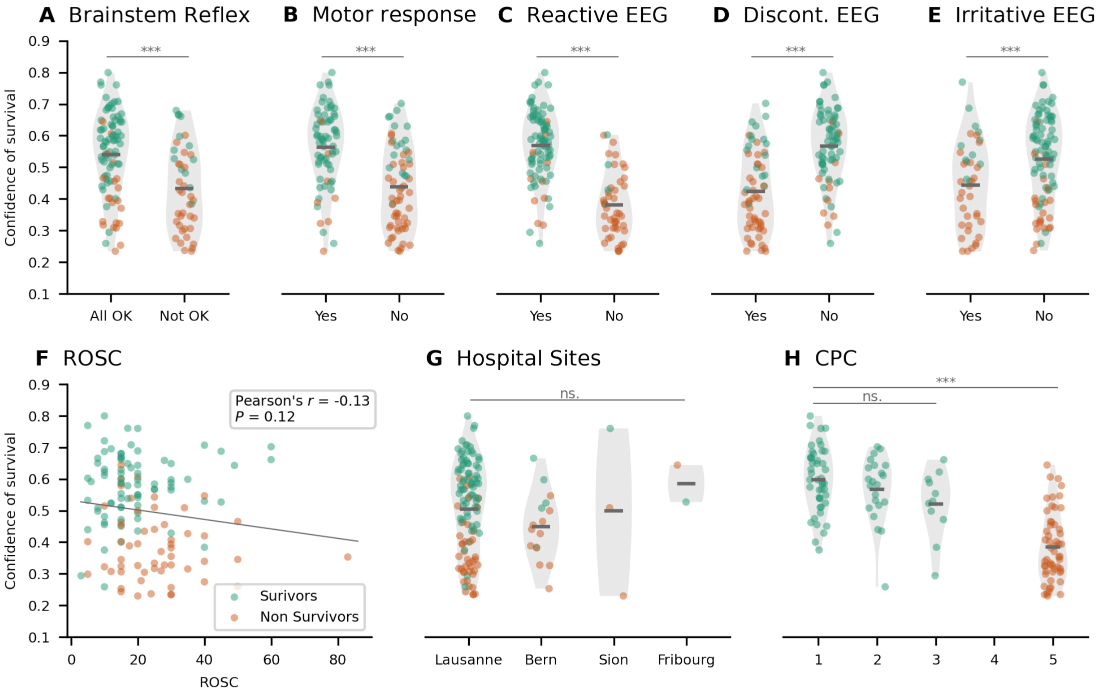


Figure 3.5: Links between network's confidence in predicting survival and clinical variables currently used for outcome prognosis. The network's confidence scores were statistically compared for patients with and without: **(A)** Brainstem Reflex ($p_{corr} < 0.01$), **(B)** Motor response ($p_{corr} < 0.01$), **(C)** Reactive EEG ($p_{corr} < 0.01$), **(D)** Discontinuous EEG ($p_{corr} < 0.01$) and **(E)** Irritative EEG ($p_{corr} < 0.01$). **(F)** Correlation of the network's confidence of survival and time to return of spontaneous circulation (ROSC, $r = -0.13$; $p = 0.12$). **(G)** Absence of link between hospital sites and the network's confidence scores ($H = 3.48$; $p = 0.32$). **(H)** Differences in the network's confidence scores across CPC outcomes. The main effect of CPC was significant when considering all outcomes (CPC 1,2,3,5) ($H = 61.90$; $p_{corr} < 0.01$), but not within the group of survivors (CPC 1−3, $H = 4.36$; $p = 0.11$).

fidence and EEG complexity was also observed for survivors ($r = -0.48$; $p_{corr} < 0.01$) and non survivors ($r = -0.76$; $p_{corr} < 0.01$) separately. PLV and LZ complexity were also negatively correlated, albeit with a weaker correlation than each of these measures did with the confidence of the neural network (Pearson's $r = -0.39$; $p_{corr} < 0.01$).

Overall, these results suggest that the network assigned higher confidence scores for awakening for patients with high PLV and low complexity, or in other words, with stronger neural synchrony and temporal structure in their EEG responses. Exemplar EEG traces of these responses for a correctly classified survivor, and a misclassified non-survivor show rather smooth EEG responses to the sounds, compared to exemplar traces of a correctly classified non survivor and misclassified survivor, where the EEG responses are more ‘stochastic’ (Figure 3.4B).

3.3.5 Comparison with clinical variables

Lastly, we compared the confidence of the network’s prediction with clinical variables currently used for outcome prognosis. We found a significant difference ($p_{corr} < 0.01$; Mann-Withney U-test) in the network’s confidence values between patients with and without presence of brainstem reflex, motor response, reactive EEG background, discontinuous or suppressed EEG background and irritative EEG (Figure 3.5A to E). Patients with brainstem reflexes, motor responses, or reactive EEG, all of which are considered indicators of good outcome had significantly higher confidence scores compared to patients without (Figure 3.5A-C). The opposite was observed for patients with discontinuous or irritative EEG, which are considered indicators of poor outcome (Figure 3.5D and E). Interestingly, EEG reactivity, which has a prognostic value for good outcome (Westhall et al., 2016), provided similar levels of predicting awakening as the neural network (PPV = 0.88). It is worth noting however, that the prognostic performance of EEG reactivity is likely biased, as this score is used in the clinical interventions to influence outcome.

Last, the network’s confidence did not correlate with ROSC ($r = -0.13$; $p = 0.12$, Figure 3.5F), while no significant difference was found in confidence scores across the four hospital sites (Kruskal-Wallis, $H = 3.48$; $p = 0.32$, Figure 3.5G). As expected based on the outcome prediction results, there was a main effect of CPC on network’s confidence when testing for CPC 1−5 (Kruskal-Wallis, $H = 61.91$; $p_{corr} < 0.01$) (Figure 3.5H). However, there was no significant difference of confidence within the group survivors, for CPC 1, 2 and 3 (Kruskal-Wallis, $H = 4.35$; $p = 0.11$).

3.4 Discussion

We studied the prognostic value of EEG responses to auditory stimulation, combined with deep learning in predicting coma outcome after cardiac arrest. We showed that convolutional neural networks are powerful in extracting single-trial information from auditory ERPs on the first day of coma, and at predicting survival three months later, with a positive predictive power of 0.83 ± 0.03, negative predictive power of 0.57 ± 0.04 and an overall AUC of 0.70 ± 0.04. These results were not available to clinicians treating the patients, and did not influence patients’ outcome. Predicting patients’ chances of awakening was at similar levels for patients receiving targeted temperature management at 33 and 36°C. The performance of the neural network was separately evaluated on patients in a “gray-zone”, where clinical variables gave inconclusive results, reaching a positive predictive value of 0.86, suggesting that it might have the potential to assist in prognostication in these currently indeterminate cases. Lastly, we showed that the confidence scores of the neural network in predicting survival were strongly correlated to the phase locking and complexity of auditory EEG responses to the auditory stimuli, so that patients that were confidently characterized as survivors had high synchrony and low complexity in their neural responses.

3.4.1 Auditory stimulation for predicting outcome from coma

The main novelty and advantages of our approach of combining auditory stimulation with deep learning to predict coma outcome are threefold: (a) it is semi-automatic, based on a single EEG recording performed within 24h after cardiac arrest, and, if confirmed in another dataset, can objectively be used to assist in predicting patients’ chances of awakening from coma; (b) it relies on the auditory pathway which is currently not actively used in the clinical routine for outcome prediction, and can therefore provide additional clinical insights for patients in a ‘gray zone’, whose outcome is indeterminate based on existing techniques; (c) the output of the neural network is not a simple binary prediction of outcome, but it exploits a continuum of confidence values, which are then directly linked and strongly correlated to interpretable features of EEG responses.

Our work follows a large body of literature showing links between neural responses to auditory stimulation in comatose or unresponsive patients and patients’ outcome (Fischer et al., 2004; Daltrozzo et al., 2007; Morlet and Fischer, 2014; Liu et al., 2021). Standardized auditory stimulation in particular has been proposed to be informative of patients’ chances of awakening (Fischer et al., 1999, 2008; Kane et al., 1993; Naccache et al., 2005; Luauté et al., 2005). However, to date it is not regularly used in the clinical routine, as the majority of existing techniques are not robust enough, rendering a clinical implementation challenging (Tzovara et al., 2016; Pfeiffer et al., 2017, 2018). Here, we overcame this challenge by focusing on one single EEG recording of 20 min, and analyzed EEG responses to auditory stimulation with convolutional neural networks. These networks have the strong advantage that they can detect discriminative patterns even in heterogeneous datasets, with minimal a priori assumptions and preprocessing (Craik et al., 2019; Roy et al., 2019; Aellen et al., 2021; Tjepkema-Cloostermans et al., 2019; Jonas et al., 2019; Altıntop et al., 2022). Importantly, all steps of the presented method are automatic, apart from visual inspection of the data and manual rejection of artifacts, which were done to ensure high quality data. Future studies can investigate whether this step can also be automated and replaced with existing algorithms for EEG data cleaning (Jiang et al., 2019).

3.4.2 Neural networks assisting prognostication of coma outcome

Neural networks have been used in several fields showing an astonishing potential to automate and improve prognostication of various neurological disorders (Craik et al., 2019; Roy et al., 2019) but their use in neuro-critical care and coma outcome prognosis remains limited. The few existing studies using neural networks to predict coma outcome are based on EEG recordings in the absence of external stimulation. These have shown a remarkable performance (AUC = 0.89 (Jonas et al., 2019); AUC at 24h = 0.88 (Tjepkema-Cloostermans et al., 2019); best AUC at 24 h = 0.89 (Zheng et al., 2021)), in predicting awakening. In our study, we obtained a mean AUC score of 0.70 ± 0.04 and of 0.83 for the best fold (Table 3.1), and a mean PPV of 0.83 ± 0.03 and 0.92 for the best fold (Table 3.1). The AUC values we obtained are slightly lower than previous studies. However, it is worth noting that our study relies on a complementary piece of information to what is currently used in the clinical routine, that is standardized auditory stimulation. Our approach is therefore predictive of awakening, and not of overall outcome, compared to previous studies using neural networks, based on resting state EEG. Therefore suggesting that the most informative measure is the PPV, which is comparable to the AUC scores reported in the literature. As such, we could show that neural networks have the potential to provide concrete diagnostic information for patients in a clinical 'gray zone', for whom currently available clinical tests are inconclusive. Future studies, with larger patient cohorts can evaluate the clinical applicability of these networks, and confirm these results.

One main limitation of our approach, and also of most previous studies (with few exceptions like (Juan et al., 2016)), is that we only predict a binary outcome of survival versus non survival. Survivors are defined as patients with a CPC 1−3, which corresponds to varying levels of autonomy and quality of life (Booth et al., 2004; Juan et al., 2018). Although a CPC of 1−2 generally represents a satisfactory quality of life, a CPC of 3 can be a heterogeneous class. Because we trained the network to specifically discriminate between patients with CPC 1−3 versus patients with a CPC of 5 (death), we could not find a more fine-grained link between the network's output and the state of survival at 3 months. Future studies, expanding on larger patient cohorts, could use the CPC score already during training the network to evaluate whether auditory responses in the acute coma phase can be informative not only of survival, but also of its quality.

3.4.3 Electrophysiological features contributing to the network's confidence in outcome prediction

One major concern for the use of neural networks in a clinical environment is that of interpretability, i.e. tracing features in the data that were important for decisions made by a network. Here, we addressed this concern by showing that the output of the neural networks was strongly correlated to features of their EEG activity like the phase locking, previously shown to reflect coma severity (Alnes et al., 2021; Zubler et al., 2017; Carrasco-Gómez et al., 2021b) and also to clinical evaluations (Westhall et al., 2016; Rossetti et al., 2017). The phase locking of EEG responses to sounds, recently shown to be predictive of patients' chances of awakening from coma (Alnes et al., 2021), was strongly correlated to the networks' confidence in predicting survival, such that the higher the phase locking, the stronger the network's confidence. Crucially, this link was observed not only across survivors and non survivors which might be considered trivial as both measures predict outcome, but also within the group of survivors and non survivors separately. This suggests that the decisions made by the neural network for assessing the confidence of survival are strongly linked to the strength of neural synchrony across EEG electrodes. Importantly, not only did we observe a strong correlation between PLV and network's output, but we also replicated previous findings about the predictive value of PLV (Alnes et al., 2021) in a new patient cohort, treated with targeted temperature management at 36°C (NT).

The advantage of using convolutional neural networks over hand crafted and pre-selected features, such as the PLV, is that the network automatically extracts multivariate features of the ERP response, which are most discriminant between patient outcomes. This approach is fully data-driven, and the input signal is minimally pre-processed, as opposed to computing the PLV, where one needs to have strong a priori assumptions about the electrode-pairs, frequency bands, and specific measure to be used (see SI of (Alnes et al., 2021)).

More surprisingly, we also observed a strong negative correlation between the network's confidence of predicting survival and the complexity of EEG responses to sounds. Neural complexity *per se*, at the single-patient level, was not predictive of patients' outcome. This finding has been previously reported for patients treated with targeted temperature management at 33°C (Alnes et al., 2021), and here we replicated it for patients treated at 36°C. Neural complexity can be considered a proxy to neuronal noise, of how structured EEG responses are across time (Luppi et al., 2019). Interestingly, here we found that the higher the network's confidence for a patient to survive the coma, the lower the complexity of the patients' EEG responses to auditory stimulation. This effect was particularly strong for non survivors. This finding suggests that although complexity *per se* is not informative of coma outcome, the network is extracting some features in EEG responses that relate to the levels of neuronal noise or structure in the EEG signal, which are in turn informative of coma outcome.

Finally, we also found significantly higher confidence in the network's predictions of survival assigned to patients that have intact brainstem reflexes, motor responses, and reactive EEG, compared to those who do not (Figure 3.5A-C). All these features are considered to be markers of preserved neural functioning, and to indicate good outcome (Rossetti et al., 2016). Interestingly, when focusing on reactive EEG, which has a prognostic value for good outcome (Westhall et al., 2016), we found that it provided similar levels of predicting awakening as the neural network (PPV = 0.88). Notably, unlike the EEG reactivity, the neural network was not used to inform clinical decisions about these patients. By contrast, the network's confidence of survival was significantly lower for patients with discontinuous EEG and irritative EEG than patients without (Figure 3.5D and E), which are considered signs of poor outcome (Westhall et al., 2016).

In summary, the strong links between the network's output and clinical or electrophysiological features across patients strengthen the view that (a) the neural network's decisions are based on clinically relevant features and (b) can provide similar levels of performance as currently used techniques (Westhall et al., 2016), but with higher level of automation and objectivity (Caroyer et al., 2021) and minimal preprocessing in the EEG data.

3.5 Conclusion

In summary, we show, for the first time, the strong potential of standardized auditory stimulation in combination with deep learning to predict awakening from coma. Our approach provides an objective and semi-automatic way to quantify a patient's chances of awakening and surviving at 3 months, already from the first few hours after coma onset. For two different patient cohorts, treated with controlled temperature management at 33 and 36°C, and recorded across four different hospitals, the neural network provides a high positive predictive value of awakening, of 0.83 ± 0.04 and 0.81 ± 0.06 respectively. This finding was also confirmed for a sub-group of patients whose outcome prognosis was indeterminate with currently used diagnostic criteria. Importantly, we could show strong links between the output of the neural network and electrophysiological characteristics of the EEG responses to sounds reflecting neural synchrony and complexity, which have been previously associated with the presence of consciousness from a theoretical point of view. Our work calls for a more systematic use of standardized auditory stimulation in the clinical routine, in combination with state-of-the-art deep learning algorithms, to assist and improve early prognostication of coma outcome.

4 Disentangling the complex landscape of sleep-wake disorders with data-driven phenotyping: A study of the Bernese Center

Florence M. Aellen[1,2], Julia Van der Meer[3], Anelia Dietmann[3], Markus Schmidt[2], Claudio L. A. Bassetti[2], Athina Tzovara[1,2,4]

[1] Institute of Computer Science, University of Bern, Bern, Switzerland
[2] Center for Experimental Neurology, Department of Neurology, Inselspital, Bern University Hospital and University of Bern, Bern, Switzerland
[3] Department of Neurology, Inselspital, Bern University Hospital and University of Bern, Bern, Switzerland
[4] Sleep Wake Epilepsy Center - NeuroTec, Department of Neurology, Inselspital, Bern University Hospital, University of Bern, Bern, Switzerland

Aellen, F. M., Van der Meer, J., Dietmann, A., Schmidt, M., Bassetti, C. L. A., and Tzovara, A. (in review). Disentangling the complex landscape of sleep-wake disorders with data-driven phenotyping: A study of the Bernese Center.

Background and Objective: The diagnosis of sleep-wake disorders (SWDs) is challenging because of the existence of only few accurate biomarkers, the frequent co-existence of multiple SWDs and/or comorbidities. The aim of this study was to assess in a large cohort of well characterized SWDs patients the potential of a data-driven approach in the identification of single SWD.

Methods: The data set analyzed contained 6'958 patients from the Bernese Sleep registry and a total of 300 variables/biomarkers including questionnaires, results of polysomnography/vigilance tests, and final clinical diagnoses. A pipeline, based on unsupervised machine learning, was created to extract and cluster the clinical data. Our analysis was performed on three cohorts: patients with central disorders of hypersomnolence (CDHs), a full cohort of SWDs spanning and a clean cohort without coexisting SWDs.

Results: A first analysis focused on the cohort of patients with CDHs and revealed four patient clusters: two clusters for narcolepsy type 1 but not for narcolepsy type 2 or idiopathic hypersomnia. In the full cohort of SWDs nine clusters were found: four contained patients with obstructive and central sleep apnea, one with NT1 and four with intermixed SWDs. In the cohort of patients without coexisting SWDs an additional cluster of patients with chronic insomnia was identified.

Discussion: This study confirms the existence of clear clusters of NT1 in CDHs, but mainly intermixed groups in the full spectrum of SWDs, with the exception of sleep apneas and NT1. New biomarkers are needed for a better phenotyping and diagnosis of SWDs.

4.1 Introduction

Sleep-wake disorders (SWDs) are a significant and growing public health concern: they affect over one third of the human population (Kerkhof, 2017) and severely reduce life quality and life expectancy. Patients with SWDs are at higher risk for cardiometabolic, brain and mental disease as well as accidents (Grandner, 2017). Yet, despite these major implications the diagnosis of SWDs is challenging due to the rarity of specific disease markers, the coexistence of multiple SWDs and frequent comorbidities (Cappuccio et al., 2010).

In central disorders of hypersomnolence (CDHs) narcolepsy type 1 (NT1) is associated with reliable biomarkers such as cataplexy and a hypocretin deficiency, the other CDHs (including narcolepsy type 2 (NT2) and idiopathic hypersomnia (IH)) lack selective and specific disease markers (ICSD-3, 2014; Trotti et al., 2013; Cairns et al., 2019; Zhang et al., 2022).

Similarly, the need for new biomarkers and better phenotypical characterization of sleep apnea (Zinchuk and Yaggi, 2020) and parasomnias (Erickson and Vaughn, 2019), has been seen as necessary for personalized management of these SWDs. Overall, a large body of literature suggests that existing diagnostic criteria for SWDs may be insufficient to categorize large groups of SWDs patients (Grandner, 2017; Lammers et al., 2020; Dietmann et al., 2021).

Machine learning algorithms can assist in identifying groups of patients that share common clinical characteristics, based on large amounts of clinical data collected in routine clinical examinations, moving towards a data-driven identification of SWD phenotypes. Unsupervised learning algorithms have already been used on data of patients with CDHs (Šonka et al., 2015; Cook et al., 2019; Gool et al., 2022). A recent study has shown that clustering algorithms can easily identify patients with NT1, and additional clusters of patients with or without sleep drunkenness, that had been formally diagnosed with NT2 and IH (Gool et al., 2022). Other attempts to cluster patients with SWDs have focused on heterogeneous disorders, such as obstructive sleep apnea (OSA) or insomnias to identify distinct subtypes

(Bailly et al., 2016; Venkatnarayan et al., 2022; Joosten et al., 2011; An et al., 2019).

Considering the lack of studies covering the full spectrum of SWDs, the aim of this study is to evaluate the extent to which individual SWD can be disentangled on the basis of existing biomarkers. To this aim, we focused on the Bernese Sleep Registry, a comprehensive database of patients with sleep disorders collected over the past 16 years at the Sleep Wake Epilepsy Center of Inselspital, the University Hospital of Bern, Switzerland. As a proof of concept, we first focused on a cohort of patients with CDHs. We hypothesized that based on clinical markers, we would be able to identify clear clusters of patients with NT1, mainly characterized by the presence of cataplexy, and mixed clusters for patients with IH and NT2, replicating previous findings (Gool et al., 2022). On the full spectrum of SWDs, we made the hypothesis that patients with obstructive or central sleep apnea, and patients with NT1 would be mostly distinct, because of the existence of multiple clear diagnostic markers. By contrast, we hypothesized that distinguishing other SWDs like insomnias or parasomnias would be more challenging based on existing clinical markers.

4.2 Materials and methods

4.2.1 Study population

The cohort of studied patients was evaluated in the multidisciplinary sleep-wake laboratory of the Inselspital, University Hospital of Bern between 2000 and 2016. Patients' evaluation typically included a clinical assessment, standardized questionnaires (including the Bernese Sleep Questionnaire, composed of 150 questions) and when needed ancillary tests such as a full night polysomnography (PSG), multiple sleep latency test (MSLT), maintenance of wakefulness test (MWT), actigraphy or others. Data from the different sources and covering the longitudinal follow-up of patients were consolidated in a REDCap database hosted at the Clinical Trials Unit Bern (Bern Sleep Registry). The cohort contained more than 300 possible markers per patient and a total of 6'958 patients. For the current analysis, the first sleep laboratory (baseline) visit was analyzed. Based on the clinical diagnoses, reports and markers, patient diagnoses were retrospectively re-classified by two independent sleep specialists (Bargiotas et al., 2019) according to the International classification of sleep disorders, 3rd edition (ICSD-3, 2014). Main demographics and main diagnoses are summarized in Table 4.1. Up to five different SWDs were made per patient (Appendix, Figure C.2).

Standard protocol approvals, registrations, and patient consents

The use of the patient data presented in this study was approved by the local ethics committee (Kantonale Ethikkommission Bern (KEK) Nr. 2022-00415). Patients gave written consent for the use of their data in scientific studies (General Consent). In order to cover our patient database since 2000, the data of the "SNS project" (Kantonale Ethikkommission Bern, Nr. 2016-00409), which was ethically approved shortly after the introduction of the General Consent, were included even in cases where no General Consent was available, according to HRO[a], Art. 9, as described in detail in the ethics protocol Nr. 2022-00415.

Table 4.1: Number of patients per diagnosis and main demographics, referring to the cohort included in the present study prior to downsampling OSA patients (corresponding to Figure 4.1B middle panel)

Diagnosis	Female sex (%)	Age (year)	BMI	Waist-Hip ratio
Insomnia (N = 469)	0.48	48.5 ± 14.0	26.9 ± 5.8	0.90 ± 0.08
	Missing: 8	Missing: 0	Missing: 11	Missing: 117
Obstructive Sleep Apnea (N = 2'834)	0.19	53.0 ± 12.7	30.3 ± 6.1	0.95 ± 0.09
	Missing: 60	Missing: 3	Missing: 60	Missing: 688
Central Sleep Apnea (N = 303)	0.12	57.4 ± 14.0	29.8 ± 5.4	0.96 ± 0.11
	Missing: 1	Missing: 0	Missing: 6	Missing: 56
Other Sleep Related Breathing Disorders (N = 27)	0.44	45.2 ± 17.7	30.3 ± 10.0	0.94 ± 0.10
	Missing: 0	Missing: 0	Missing: 1	Missing: 3
Narcolepsy Type I (N = 54)	0.41	36.4 ±17.5	26.3 ± 5.2	0.90 ± 0.08
	Missing: 0	Missing: 0	Missing: 0	Missing: 19
Narcolepsy Type II & Idiopathic Hypersomnia (N = 42)	0.52	26.8 ± 11.8	23.8 ± 3.8	0.85 ± 0.07
	Missing: 0	Missing: 0	Missing: 1	Missing: 11
Other Central Disorders of Hypersomnolence (N = 204)	0.50	35.5 ± 13.5	25.2 ± 4.4	0.88 ± 0.11
	Missing: 1	Missing: 0	Missing: 4	Missing: 26
Parasomnias (N = 144)	0.35	47.1 ± 19.9	24.9 ± 4.9	0.89 ± 0.07
	Missing: 0	Missing: 0	Missing: 3	Missing: 35
Sleep-Related Movement Disorders (N = 211)	0.46	50.6 ± 15.7	26.8 ± 4.9	0.88 ± 0.09
	Missing: 3	Missing: 0	Missing: 5	Missing: 77
Isolated Symptoms and Normal Variants (N = 166)	0.27	52.9 ± 14.4	27.9 ± 6.1	0.92 ± 0.09
	Missing: 7	Missing: 0	Missing: 3	Missing: 60

4.2.2 Processing pipeline

The processing pipeline involved six steps, of data collection and marker extraction, data curation, selection of good quality markers and patients, normalization of data, imputation of missing values (Epsilon-machine, 2022) and clustering, (Figure 4.1), which are described in full detail in Appendix C.

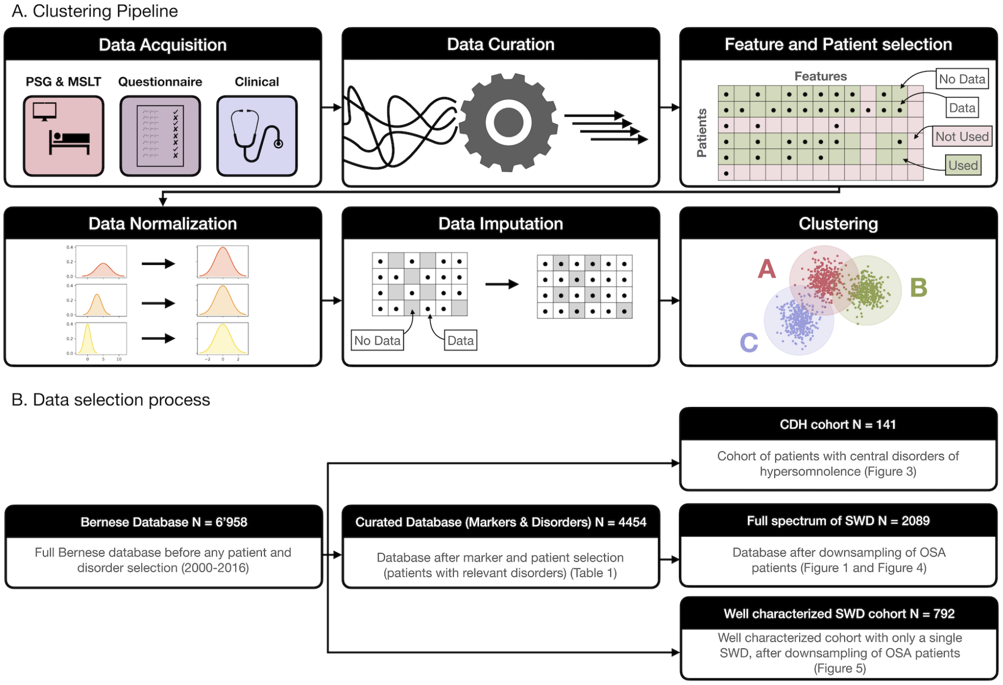

A. Clustering Pipeline
Data Acquisition
PSG & MSLT
Questionnaire
Clinical
Data Curation
Feature and Patient selection
Features
Patients
No Data
Data
Not Used
Used
Data Normalization
Data Imputation
No Data
Data
Clustering
A
B
C
B. Data selection process
CDH cohort N = 141
Cohort of patients with central disorders of hypersomnolence (Figure 3)
Bernese Database N = 6'958
Full Bernese database before any patient and disorder selection (2000-2016)
Curated Database (Markers & Disorders) N = 4454
Database after marker and patient selection (patients with relevant disorders) (Table 1)
Full spectrum of SWD N = 2089
Database after downsampling of OSA patients (Figure 1 and Figure 4)
Well characterized SWD cohort N = 792
Well characterized cohort with only a single SWD, after downsampling of OSA patients (Figure 5)

Figure 4.1: **(A)** Processing Pipeline, starting from data acquisition, curation and selection of markers and patients with high quality data. The selected variables were then normalized and missing entries were imputed. The final step was an unsupervised clustering algorithm. **(B)** Data selection process, describing the sample size of the entire Bern Sleep Registry and each of the three cohorts included in the main analyses, across different steps of the processing pipeline.

4.2.3 Patient cohorts

Three patient cohorts were analyzed, (a) a cohort of patients with CDHs, (b) a cohort of patients covering the full spectrum of SWDs, and (c) a cohort of well-characterised patients with one single SWD. For each of these cohorts, the patient selection started from the Bern Sleep Registry from scratch (Figure 4.1B, left panel).

4.2.3.1 CDH cohort

For this cohort, we used a set of 22 markers selected among the available markers, based on clinical expertise (ICSD-3, 2014). This first cohort included a total of 141 patients, out

Table 4.2: Number of patients for the three cohorts. The last line for each cohort (Total number of patients in the cohort) refers to the total number of patients used for the clustering, so after the "feature and patient" selection steps of the pipeline (Figure 4.1B).

Cohort	Diagnosis		# Patients
CDH Cohort			
	Narcolepsy Type 1	(NT1)	78
	Narcolepsy Type 2	(NT2)	19
	Idiopathic Hypersomnia	(IH)	44
Full Cohort of SWD			
	Insomnia	(Insomnia)	469
	Obstructive Sleep Apnea	(OSA)	469
	Central Sleep Apnea	(CSA)	303
	Other Sleep Related Breathing Disorders	(Other SBD)	27
	Narcolepsy Type I	(NT1)	54
	Narcolepsy Type II & Idiopathic Hypersomnia	(NT2&IH)	42
	Other Central Disorders of Hypersomnolence	(Other CDH)	204
	Parasomnias	(Parasomnia)	144
	Sleep-Related Movement Disorders	(Movement Disorders)	211
	Isolated Symptoms and Normal Variants	(Isolated Symptoms)	166
Full Cohort of SWD with well characterized patients with a single SWD			
	Long-Term Insomnia	(Insomnia)	211
	Obstructive Sleep Apnea	(OSA)	290
	Central Sleep Apnea	(CSA)	168
	Narcolepsy Type I	(NT1)	23
	NREM-Related Parasomnias	(Parasomnia)	42
	Restless-Leg Syndrome	(Movement Disorders)	58

of which 78 suffered from NT1, 19 from NT2 and 44 from IH (see Table 4.2).

4.2.3.2 Full cohort of SWDs

Second, the analysis was expanded to the full spectrum of SWDs, namely on the following diagnosis and sub diagnosis: insomnia, obstructive sleep apnea (OSA), central sleep apnea (CSA), other sleep-related breathing disorders (a grouping of all other sleep-related breathing disorders not stated explicitly here; Other SBD), narcolepsy type 1 (NT1), narcolepsy type 2 and idiopathic hypersomnia (NT2&IH), other central disorders of hypersomnolence (a grouping for all other CDHs not stated explicitly here; Other CDH), parasomnias, sleep-related movement disorders, isolated symptoms and normal variants.

This led to a selection of 36 markers in a data driven way (Appendix, Section C.1), while 15 additional markers were added based on clinical importance, resulting in a dataset of 51 markers in total (Appendix, Section C.3.2) and a total of 4'454 patients. As the vast majority of patients in this dataset were diagnosed with OSA (N = 2'834 patients out of 4'454), these patients were downsampled to 469 patients, the same number as the second largest group of patients, with a primary diagnosis of insomnia. This step was necessary as otherwise all clusters are dominated by patients with OSA (Appendix, Section C.6.1). This resulted in a cohort containing 2'089 patients (Table 4.2, for an overview of the distribution of the diagnoses, and Figure 4.2 for an overview of which secondary SWD patients in this cohort were diagnosed with).

4.2.3.3 Full cohort of SWDs with well characterized patients with a single SWD

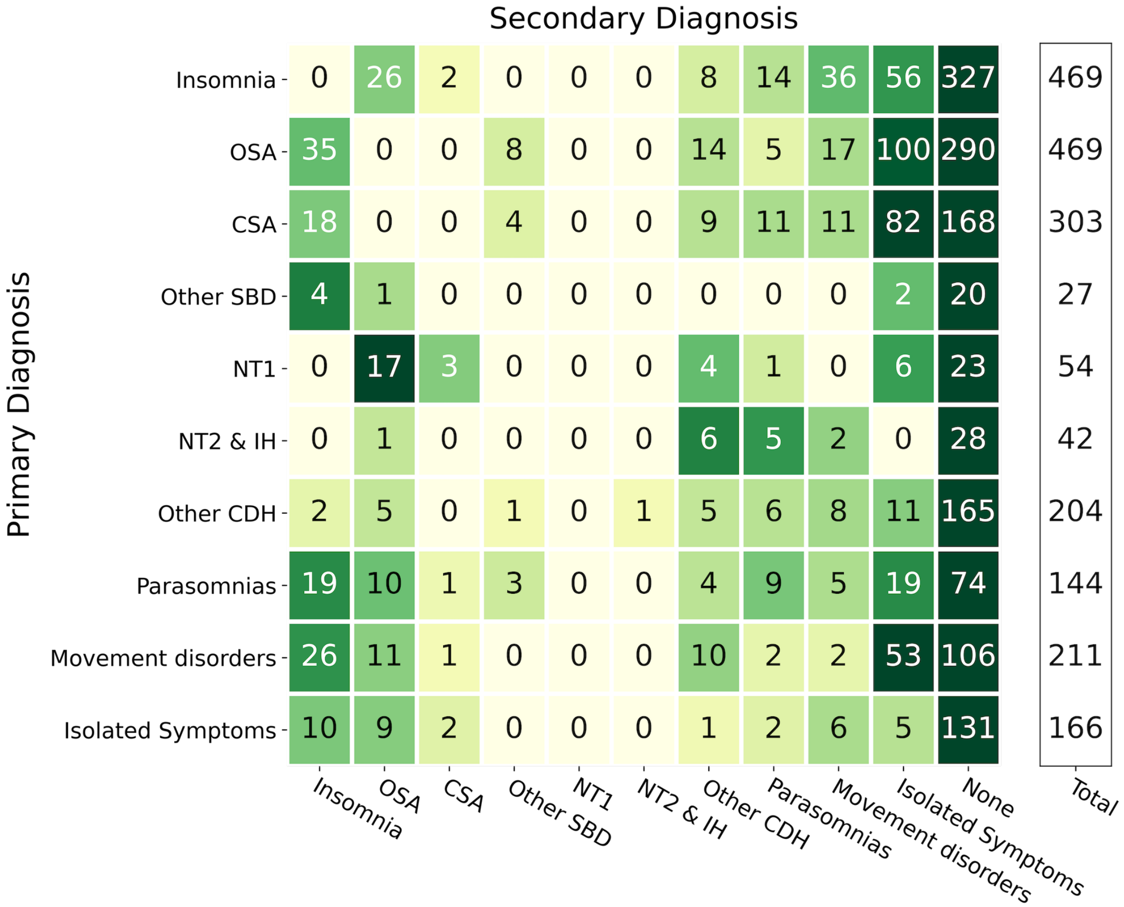
Because the full cohort of patients with SWDs had high levels of coexisting SWDs (Figure 4.2), we selected a control cohort, consisting of well characterized patients, where only those with a primary and no secondary diagnosis were included, and where certain SWDs were divided into subcategories. For example, for insomnia only chronic insomnia patients were considered. See Appendix, Section C.2 for a description of disorders used for this analysis. For the well-characterized sub-cohort, we used the same clinical markers as for the full cohort. However, because of a decreased number of patients compared to the full cohort, 4 markers were not available for a sufficiently large number of patients and were thus excluded. Questions from the custom questionnaire (Bern sleep questionnaire, including questions that are used to differentiate parasomnias from restless leg syndrome (RLS)) were not consistently digitized, and were therefore excluded from our analysis (Appendix, Section C.8). This left a dataset of 47 markers and 792 patients (Table 4.2).

4.2.4 Clustering

For our main analysis, we used a K-Means clustering algorithm implemented in the python package sklearn (Virtanen et al., 2020) with euclidean distance. The clustering was calculated for multiple numbers of cluster centers, for each of which five different metrics were evaluated. The final number of clusters was selected based on the explained variance, similar to previous publications (Salmanpour et al., 2022; Hassan et al., 2021) (Appendix, Section C.7). Because K-means clustering is sensitive to its initialization, we also repeated the clustering with multiple initializations, all of which showed similar results as the ones presented here.

4.2.4.1 Control analysis

We additionally performed a comprehensive set of control analyses to ensure that our results hold across multiple implementations of the K-Means algorithm, with agglomerative clustering and across multiple clinical control analyses, where markers used for diagnosis of



	Insomnia	OSA	CSA	Other SBD	NT1	NT2 & IH	Other CDH	Parasomnias	Movement disorders	Isolated Symptoms	None	Total
Insomnia	0	26	2	0	0	0	8	14	36	56	327	469
OSA	35	0	0	8	0	0	14	5	17	100	290	469
CSA	18	0	0	4	0	0	9	11	11	82	168	303
Other SBD	4	1	0	0	0	0	0	0	0	2	20	27
NT1	0	17	3	0	0	0	4	1	0	6	23	54
NT2 & IH	0	1	0	0	0	0	6	5	2	0	28	42
Other CDH	2	5	0	1	0	1	5	6	8	11	165	204
Parasomnias	19	10	1	3	0	0	4	9	5	19	74	144
Movement disorders	26	11	1	0	0	0	10	2	2	53	106	211
Isolated Symptoms	10	9	2	0	0	0	1	2	6	5	131	166

Figure 4.2: Primary and secondary sleep-wake diagnosis in 2’089 selected patients from the Bernese Center (2000-2016). The confusion matrix shows on the y-axis the primary and on the x-axis the secondary SWD, if present, illustrating the complicated landscape of SWDs (Figure C.2 for tertiary - quinary diagnoses). The number of patients with OSA was downsampled from 2’834 to 469 (see Figure 4.1) while only patients with the diagnosis on the y-axis specified were included in this visualization. The diagnosis “Isolated Symptoms” contains some isolated symptoms and normal variants, such as: “Excessive fragmentary myoclonus,” “Hynagogic foot tremor and Alma”, “Hypnic jerks” and ”PLMS” (see Appendix, Section C.4.1).

disorders were excluded and the number of OSA patients was not downsampled (Appendix, Sections C.5 and C.6).

4.2.4.2 Feature extraction

To assess how the available markers were distributed across diagnoses and patient clusters, we visualized them in the form of a barcode, calculated similar to (Gool et al., 2022). In particular, patients were split according to their primary diagnosis, or according to their assigned cluster, and the distribution of each marker was calculated for each available diagnosis or cluster. Resulting values were normalized with respect to 10'000 random draws of the same number of patients per diagnosis or cluster. The plotted features, in the form of a barcode, show the deviation of the actual mean computed from the data, with respect to the randomly generated distribution, in measures of standard deviations.

4.3 Results

4.3.1 CDH cohort

For the cohort of CDHs, we extracted the available clinical markers across patients in the form of a barcode (Figure 4.3A). In accordance with the existing diagnostic criteria, we found an over-representation of the marker cataplexy in patients with NT1 compared to NT2 and IH. The marker mean REM latency (of the MSLT) on the other hand had a high value for IH, but not for any of the narcolepsy diagnosis (Figure 4.3A).

Next, we performed a clustering analysis to group together patients with similar markers in a data-driven way. The clustering analysis revealed four clusters, with the first two containing almost exclusively patients diagnosed with NT1 (Figure 4.3C, first cluster: 100% NT1 patients, and the second cluster 93% of NT1 patients). A third cluster contained a majority of patients with IH, and the fourth a mixture mainly between NT2 and IH (Figure 4.3C).

The barcodes with the distribution of each marker for the different clusters revealed that the discriminating markers between clusters 1 and 2 and clusters 3 and 4 were the high prevalence of cataplexy, which is a diagnostic criterion for NT1 (ICSD-3, 2014), and the presence of naps. Clusters 1 and 2 were characterized by short mean REM latency and low prevalence for sleep drunkenness. Clusters 1 and 2 were separated from one another mainly due to the presence of hallucinations, sleep paralysis, the ratio of men/women among others. Cluster 3 was separated from cluster 4 mostly by sex, but weaker differences were also found for REM latency, mean REM latency and naps. These results suggest that our approach was able to discriminate, in a data driven way, NT1 patients from patients with NT2 and IH, confirming previous reports based on different cohorts (Gool et al., 2022).

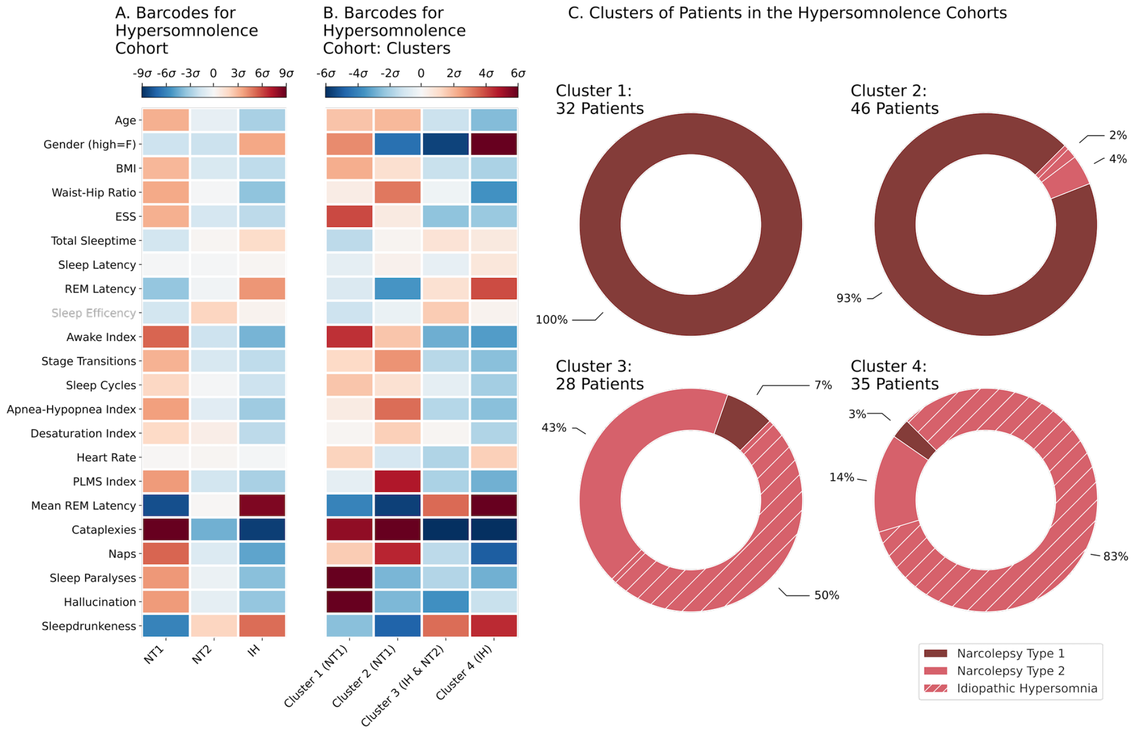


Figure 4.3: Characterisation and clustering for the Hypersomnolence cohort: **(A)** Barcode plot prior to clustering, where patients are split according to their primary diagnosis. Marker names that are written in gray indicate a low number of a given value before imputation, and therefore should be interpreted cautiously. **(B)** Barcode plot where patients are split according to the identified clusters, showing the distribution of the markers per cluster. **(C)** Identified patient clusters, highlighting the distribution of primary diagnoses within each cluster (percentages) and the total number of patients per cluster.

4.3.2 Full cohort of SWDs

After establishing that our clustering and marker selection pipeline reproduced previously reported findings (Gool et al., 2022) we proceeded with a larger patient cohort, including the full spectrum of SWDs. Patients with insomnia had the highest value of percent wakefulness after sleep onset, low sleep efficiency among others, consistent with their clinical phenotype (Figure 4.4A). Patients with OSA and CSA showed high values of apnea-hypopnea index, apnea index, hypopnea index, apnea-hypopnea duration, maximum apnea duration, desaturation index, age, BMI, and waist-hip ratio compared to other disorders. For other SBD we found the highest values of duration O2 desaturation $< 90\%$ (min) and O2 $< 80\%$ (min) and lowest values of O2% mean and O2% minimum. All CDHs contained

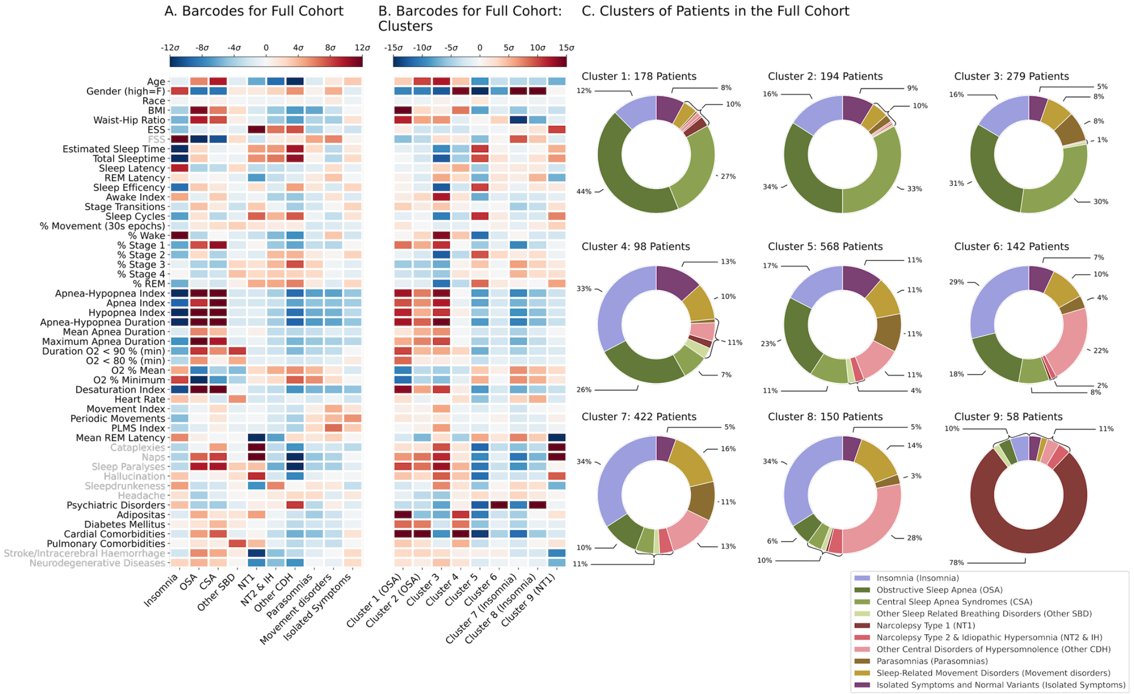


Figure 4.4: Clustering for the full cohort of SWDs: **(A)** Barcode plot where patients are split according to their primary diagnosis. Marker names that are written in gray indicate a low number of a given value before imputation, and therefore should be interpreted cautiously. **(B)** Barcode plot for the identified clusters, showing the distribution of the markers per cluster. **(C)** Clustering results, highlighting the identified 9 clusters for the full cohort. For each cluster we display the distribution of primary disorders of the patients assigned to it in percentage, as well as the total number of patients in the cluster.

young patients, with high ESS, estimated sleep time and total sleep time. Parasomnias showed relatively high values of fatigue (fatigue severity scale (FSS)) and, as expected, movement disorders showed the highest PLMS index.

For the full cohort clustering, we found a total of nine clusters, which were ordered according to the prevalence of OSA: three clusters (clusters 1, 2 and 3, Figure 4.4C) contained more than 60% of patients with either OSA or CSA. Only clusters 1 and 2 contained a relatively well-defined cohort, with at least one third of patients having the same primary diagnosis, namely OSA. The next three clusters (clusters 4, 5 and 6, Figure 4.4C) showed a mixed distribution of different disorders. Clusters 7 and 8 contained slightly more than a third of patients with a primary diagnosis of insomnia. The remaining two thirds of cluster 7 were mixed, while cluster 8 contained 28% of other CDHs patients. Cluster 9 contained 78% of patients with NT1, and contained almost all NT1 patients that were part of our cohort, making it a clear cluster of NT1 (Figure 4.4C).

We next evaluated the most prevalent markers for each cluster (Figure 4.4B). The first three clusters, primarily containing patients with OSA and CSA, had as expected higher values for markers related to breathing which are associated with OSA and CSA (ICSD-3, 2014). Cluster 1 showed the highest BMI, hip-waist ratio, obesity frequency, and oxygen desaturation index. Clusters 1 and 2 showed a high prevalence of diabetes and cardiac comorbidities compared to cluster 3. Cluster 3 showed a distinct pattern with short sleep time and low sleep efficiency as compared to the first two clusters, pointing to sleep-disordered with insomnia-like symptoms. Cluster 4 was characterized by a high prevalence of women with high sleep latency, overweight-associated comorbidities, and inconspicuous classical breathing disorder markers. Cluster 5 showed a high prevalence of young men with long sleep time and efficiency without any comorbidities in terms of psychiatric, pulmonary or cardiac disorders, without narcolepsy-specific markers, and normal values for markers related to apnea. Patients assigned in clusters 6 and 8 had particularly high prevalence of psychiatric disorders, with a main difference of high prevalence of men in cluster 6 as compared to women in cluster 8. Similar to cluster 8, cluster 7 showed a high prevalence of women, in combination with a low hip-waist ratio and an exceptionally high fatigue (FSS score). Last, cluster 9, which predominantly contained patients with NT1, showed extreme values in all markers expected for patients with NT1. In summary, for the full cohort we could obtain some well characterized patient groups, primarily consisting of patients with breathing-related sleep disorders and NT1, but for the remaining disorders no clear clusters could be identified.

4.3.3 Full cohort of SWDs with well characterized patients with a single SWD

We next evaluated whether the lack of clearly defined clusters for a great number of SWDs was due to the high prevalence of secondary diagnoses. To this aim, we focused on a subset of the full cohort of SWDs, which only contained patients with one single well-defined primary diagnosis.

In this cohort, the first four clusters (Clusters 1,2,3 and 4; Figure 4.5B) contained predominantly patients diagnosed with OSA and CSA. Clusters 5 and 6 were rather mixed, both with more than a third of OSA patients, and 30% and 25% insomnia patients, respectively. Clusters 7 and 9 were both clusters of insomnia and cluster 8 contained 65% of patients diagnosed with NT1. RLS and NREM-related parasomnias both were mostly represented in clusters 6 and 9. The barcodes for this well-defined cohort (Figure 4.5A) showed that clusters 2 and 3 contained the highest values for markers related to breathing disorders. Clusters 1 and 4 showed less extreme, but still high values for these markers. Clusters 4 and 5 contained most of the OSA and CSA patients with diabetes and/or cardinal comor-

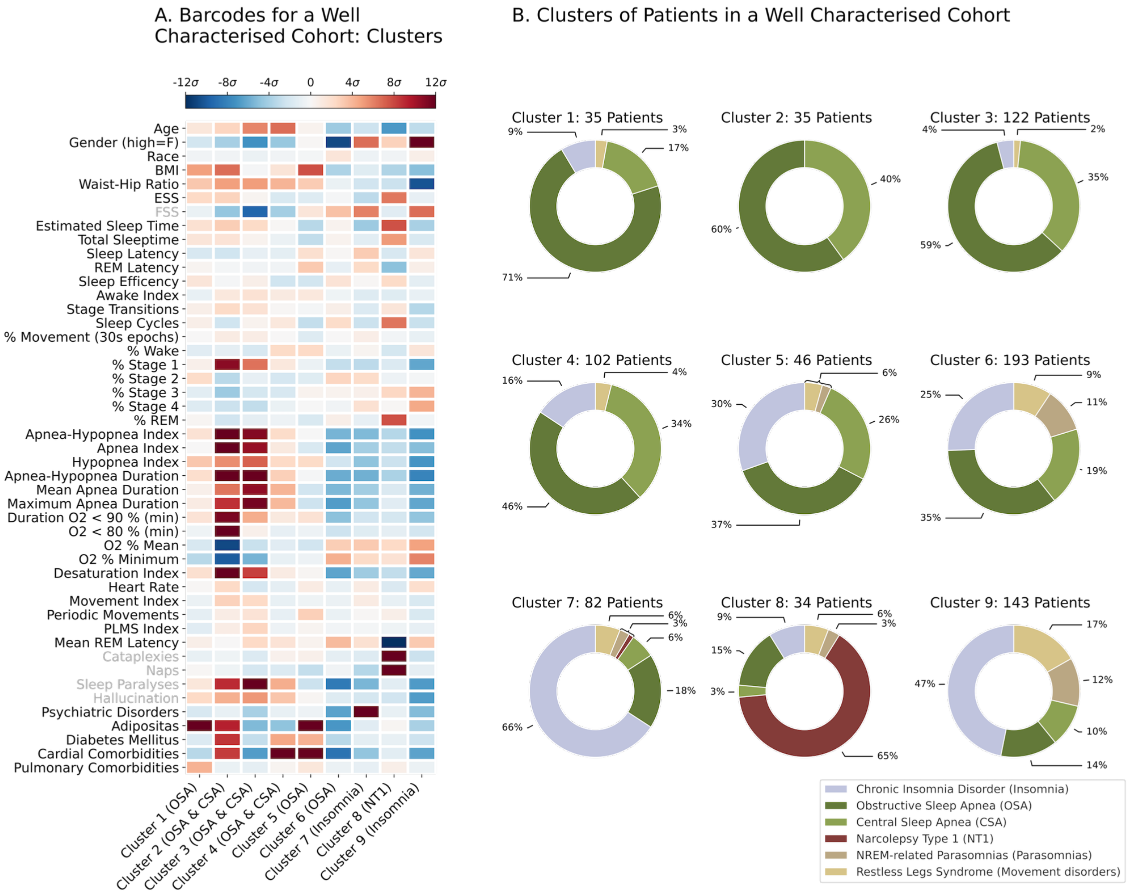


Figure 4.5: Clustering for the full cohort of SWDs, as the cohort is similar to the full cohort presented in Figure 4.4, the barcode for the diagnoses is not shown: **(A)** Barcode plot for the identified clusters, showing the distribution of the markers per cluster. Marker names that are written in gray indicate a low number of a given value before imputation, and therefore should be interpreted cautiously. **(B)** Distribution of identified disorders within the obtained clusters.

bidities. Cluster 6 mainly contained young men with normal values for markers related to apnea and without non-sleep related comorbidities, such as psychiatric, pulmonary or cardiac disorders. In cluster 7 we found mostly women with high values for psychiatric disorders and high FSS. Cluster 8 showed a typical marker pattern of patients with NT1. Cluster 9 contained mostly women, with low BMI and high FSS values.

4.4 Discussion

We show, for the first time, a data-driven phenotyping of patients covering the full spectrum of SWDs. In a cohort of CDHs we identified two clear clusters of patients with NT1 and two mixed clusters with NT2 and IH patients, mirroring previous findings on different patients (Gool et al., 2022). In the cohort covering the full spectrum of SWDs, we obtained clear clusters for patients with breathing disorders (OSA/CSA) and NT1, while in a sub-cohort of patients with a single diagnosis, clusters with patients with insomnia could also be identified. All other SWDs were rather intermixed, highlighting, for the first time in a data-driven way, the lack of objective, digitized markers for diagnosing the full spectrum of SWDs that represents the clinical reality. These results were confirmed in multiple of control analyses.

4.4.1 Data-driven phenotyping in the CDH cohort

In the CDH cohort, we confirmed our hypothesis that NT1 was well distinct from NT2 and IH. Notably, two of the four identified clusters contained nearly exclusively NT1 patients. The two non-NT1 clusters were distinct from the NT1 clusters with respect to well known NT1 markers and the presence of sleep drunkenness, a marker of idiopathic hypersomnolence (Trotti et al., 2013; Kretzschmar et al., 2016). A recent study showed similar results (Gool et al., 2022), performing an unsupervised clustering on a different cohort of patients with CDHs. This study included a much larger number of patients, likely reflecting higher variability, and found seven clusters, four of which were clusters of NT1. This first part of our analyses confirmed these previous findings, corroborating the need for new markers for re-classifying CDHs (Dietmann et al., 2021). The accordance with previous results validated our analysis pipeline, which we additionally applied in the full cohort of SWDs.

4.4.2 Redefining the landscape of SWD

In the full spectrum of SWDs, we confirmed our hypothesis that some disorders, like OSA, CSA, or NT1, were separated in distinct clusters. Indeed, these disorders are characterized by relatively clear clinical criteria and therefore have distinct markers, such as multiple breathing-related markers the apnea-hypopnea index for the case of OSA/CSA, or cataplexy in the case of NT1 (ICSD-3, 2014).

By contrast, for other SWDs like e.g. parasomnia and RLS, our clustering algorithm was unable to identify distinct clusters based on currently available markers. Although these disorders have very distinct symptoms (ICSD-3, 2014), they do not have specific markers differentiating them from other SWDs. Thus, in the clinical setting, for example RLS can be reliably discriminated from parasomnias on the basis of questionnaires Appendix, Section C.8). The occurrence of insomnia in all clusters supports the notion that insomnia is often diagnosed in the presence of other SWDs (Edinger, 2011). Our results thus suggest that the available clinical markers can distinguish certain SWDs, but fail for others, highlighting the complexity of the full spectrum of SWDs.

4.4.3 Stability of clustering

By performing multiple control analysis we confirmed our main results, that patients with OSA/CSA and NT2 were well distinct, however the remaining SWDs were intermixed. None of the additional analysis was able to clearly distinguish the full spectrum SWDs, indicating that our results were not solely driven by the selected pipeline.

These results suggest that the clusters we found adequately represent the landscape of SWDs, based on available clinical markers. Indeed, in the clinical practice misdiagnoses are very common for several of the clinical tests (Baumann et al., 2014; Fronczek et al., 2020; Lammers et al., 2020), and test-retest reliability is surprisingly low (Trotti et al., 2013; Lopez et al., 2017). Here, we confirm the low reliability in diagnosing SWDs (a) by showing high comorbidities in the existing diagnoses, and (b) by showing that robust machine learning techniques failed to distinguish all SWDs based on digitized markers, calling for additional markers (Dietmann et al., 2021) and possible re-definitions of diagnostic criteria.

4.4.4 Limitations

Even though we have performed numerous control analyses to ensure stability of clustering results, we cannot exclude the possibility that an alternative pipeline may better capture the true distribution of the data. However, our clustering pipeline was able to reproduce previously reported results (Gool et al., 2022) on a different group of patients, and all of our control analyses validated our main results. This leads us to believe that the identified patient clusters are reliable.

One limitation for the full spectrum of SWDs is that the diagnosis of some of these disorders is based on self-reports which are collected in the form of questionnaires. In our cohort, these were collected on paper and their digitization could not always be reliably performed. We therefore did not include some of this clinically relevant information. For example, including questionnaires about self-reported symptoms might have improved the discrimination of RLS from parasomnias (Appendix, Section C.8). Future studies can address this limitation with the collection of prospective datasets, employing digitized questionnaires.

As a last limitation, in the present study we had to exclude patients with circadian-sleep-wake disorders, as there were not enough patients with high quality data available. Future studies should replicate our findings on bigger cohorts in terms of markers and patients.

5 Discussion

My Ph.D. project aimed at analyzing neurological data for clinical applications and was split into three different parts. The goal of the first part was to analyze EEG signals in order to answer a methodological question if as a proof of concept deep learning could be used as an MVPA pipeline. The second part was a clinical application to predict outcome of coma with deep learning models based on EEG data. The last part focused on a second clinical application to analyze sleep-wake disorders with unsupervised clustering algorithms to disentangle the full landscape of SWDs. In the following chapter I will discuss each of the projects separately, as well as three general topics, namely deep learning for EEG signals, pre-selected features versus data-driven approaches for classification, and artificial intelligence for medical applications.

5.1 CNNs for decoding EEG responses and visualizing trial by trial changes in discriminant features

5.1.1 Summary

Deep learning has been increasingly used in neuroscience in the last few years as an analysis tool for EEG signals. However, most studies using artificial neural networks on EEG focus on BCI applications, seizure detection, sleep staging, or similar. It remains an open question if artificial neural networks could be used as an MVPA algorithm for research purposes. Currently used MVPA algorithms are mostly based on classical machine learning algorithms, such as logistic regression or support vector machines, but EEG signals are high dimensional and complex. The currently used MVPA algorithms are not powerful enough to capture patterns of neural activity accurately, especially at the group level. Additionally, these methods make the assumption that responses are located at the same latencies across trials, an assumption that is only met in very homogeneous datasets. Data collection for research is expensive and time-consuming, which results in datasets with a limited number of participants and trials. The field of deep learning has multiple solutions for dealing with small datasets, but they need to be translated into the new context of EEG signals.

In the first part of my thesis (Aellen et al., 2021), I showed that deep learning could be used as an MVPA algorithm by presenting a complete analysis pipeline suggesting appropriate data augmentation techniques for EEG signals. To detect which parts of the EEG signal contributed predominately to the network's decision, saliency maps, a gradient-based technique was used. This allowed for the extraction of class and trial-specific information. I showed that the extracted features were meaningful, as the features mirrored the EEG data on the group level per experimental condition (Figures 2.1 & 2.9, Chapter 2). Additionally, the class and trial-specific information were used to show an exemplar analysis technique for the evolution of the network activation over the course of the experiment, which can be used to test theories of learning. This work presented a novel pipeline based on a convolutional neural network for analyzing EEG data showing great potential to be used as an MVPA for research purposes. The developed pipeline contained data augmentation, network training, and feature visualization techniques.

5.1.2 Deep learning for basic EEG research

Artificial neural networks have been previously applied to analyze EEG data (Roy et al., 2019; Craik et al., 2019). Most applications of deep learning models focus on clinical applications, such as sleep staging (Fiorillo et al., 2019; Vallat and Walker, 2021; Stephansen et al., 2018), outcome prediction of coma (Jonas et al., 2019; Aellen et al., 2023) or detection of seizures (Burrello et al., 2020; Cho and Jang, 2020).
Although deep learning has been used for basic research (Kuanar et al., 2018; Bashivan et al., 2016; Wang et al., 2018), the applications remain limited and these techniques are not standardly used today. The pipeline developed in this work focused on common tasks in basic neuroscience, such as decoding the difference between experimental conditions and the extraction of stable features on a single trial level. The performance scores that the presented pipeline reached were comparable to previously proposed network architectures (Shallow and Deep neural network by (Schirrmeister et al., 2017), Figure 2.8). The network architecture that was selected had initially been proposed for the classification of images (He et al., 2016) and did not explicitly exploit the temporal dimensionality of the data. Previous literature on deep learning and EEG (Schirrmeister et al., 2017; Lawhern et al., 2018; Kuanar et al., 2018; Bashivan et al., 2016; Wang et al., 2018) has suggested architectures that considered the temporal dimension before the spatial one, thus giving greater importance to temporal patterns in the data. Here however the EEG data was studied as a spatial-temporal continuum, as it is commonly done in the field of EEG research (Maris and Oostenveld, 2007).

The extraction of EEG features from convolutional neural networks has been explored previously. For example, (Schirrmeister et al., 2017) computed spectral EEG features, based on their trained networks. However, their results did not contain temporal information, as they were collapsed over trials and time. Other studies reported gradient-based methods, however, they presented the features on an average level (Farahat et al., 2019; Vahid et al., 2020) or on exemplar trials (Lawhern et al., 2018) only. Compared to these studies the work presented in this thesis also analyzed a generalization of features over time and how they were represented over the full dataset.

5.1.3 Limitations

The main limitation of this work is the small number of hyperparameters that were explored. Only minimal variations of the network architecture were tested (Appendix, Section A.2.3). Additionally, I investigated the use of unfiltered versus filtered EEG data and used the proposed pipeline to oversample the underrepresented class versus a weighted binary cross-entropy loss. However, there would have been many more hyperparameters to explore. For example, the network architecture was initially developed for the classification of images in the field of computer vision in 2016 (He et al., 2016). There are more recent and more powerful architectures that could improve decoding performance. The used model also did not explicitly take into account the temporal dimension of the EEG data, which could be exploited with an RNN (Delvigne et al., 2022), LSTM (Bashivan et al., 2016; Kuanar et al., 2018; Zheng et al., 2021) architecture or a CNN model developed for EEG, such as EEGNet (Lawhern et al., 2018). Another hyperparameter that was under-explored was the feature extraction method, as a Grad-CAM algorithm (Selvaraju et al., 2017) could have also been used.

A second limitation of the study is the limited number of participants available. A bigger number of participants adds more variation in the EEG data and, therefore, neural patterns captured by the algorithm. The trained model would thus generalize better to new participants. I overcame this limitation by employing data augmentation techniques. However, the question remains whether their added value is as good as a more extensive dataset with more natural variation.

5.1.4 Future directions in deep learning for EEG signals

One of the biggest questions emerging from this study is exploring different network architectures, feature extraction techniques, and how the architecture influences the extracted features. Previous literature in computer vision (Zeiler and Fergus, 2013) has shown that features visualized with the aid of the kernels of convolutional layers are dependent on the used layer. Features visualized from higher layers show more coarse patterns associated with high-level features compared to kernels from lower layers that show high-level and detailed features (Zeiler and Fergus, 2013). The depth of the architecture, therefore, influences the extracted features. Additionally, the features extracted with gradient-based extraction techniques, such as saliency maps (Simonyan et al., 2013) or Grad-CAM (Selvaraju et al., 2019) depend on the algorithms. The features calculated in this work were gradient-based and the used extraction technique could thus influence the features.

Additionally, the depth of the artificial neural network could influence the resolution of the saliency maps. Similarly, the shape of kernels could influence patterns of the discriminant features. Yet it remains unexplored how the network architecture or feature extraction technique could influence the extracted features for EEG signals.

In a different direction, considering deep learning as an MVPA technique for EEG analysis, an open question remains if these models could be suited for source reconstruction. A major advantage of MVPA algorithms is that the weights of a trained classifier can be used for source localization to reveal which region differed between two conditions, e.g., stimuli related to animate versus inanimate objects (van de Nieuwenhuijzen et al., 2013). Also unexplored is whether discriminant features extracted from a trained deep neural network could be used to infer sources of activation with the aid of source localization. Additionally, such an algorithm could be used to analyze changes in sources throughout the experiment.

5.2 Auditory stimulation and deep learning predict awakening from coma after cardiac arrest

5.2.1 Summary

Outcome prediction for post-cardiac arrest patients in a comatose state is essential for clinicians to optimize patient care as well as the close family of the patient. Currently used clinical markers have the limitation that they leave up to a third of patients without a clear prognosis, and most of these markers are sensitive to predict non survival (Rossetti et al., 2016). Only a few clinical markers with high positive predictive power for survival exist.

The second part of my thesis used a dataset of EEG recordings of coma patients that were exposed to standardized auditory stimulations presented in a deviance paradigm, which has been shown previously (Fischer et al., 1999) to be predictive of positive outcome. Brain responses to standardized auditory stimulations are not currently used in the clinical routine and could provide additional insight for patients without a clear prognosis. The cohort of patients in this work was recorded in four different hospitals in Switzerland and under two different temperature treatments. The selected network architecture (EEGNet (Lawhern et al., 2018)), explicitly exploits the temporal dimension of the EEG data and reached a high performance in predicting survival. I also analyzed the EEG of patients without a clear prognosis and found for this cohort high positive predictive power. The network's performance was however slightly lower than the full cohort's performance. Additionally, to gain insight into the network's decision, the output of the network, measuring its confidence in predicting survival, was compared to measures related to neural synchrony and complexity, showing a correlation for both. Neural synchrony was positively correlated with the network's confidence in predicting survival, thus patients with high neural synchrony were predicted by the network be to more likely to survive. The neural complexity values of patients were negatively correlated with the network's output, i.e. patients with high complexity in their auditory EEG responses were predicted by the network to be less likely to survive. This indicates that the network extracted information related to these measures. For the first time ever this work has shown that artificial neural networks can be used for coma outcome prediction based on EEG data in response to standardized auditory stimulations. Additionally, my pipeline showed good performance on a subcohort of patients without a clear prognosis, based on current clinical markers.

5.2.2 Predicting coma outcome with deep learning

Previous literature has explored the use of convolutional neural networks for the prediction of outcome from coma (Jonas et al., 2019; Tjepkema-Cloostermans et al., 2019; Zheng et al., 2021). However, the use of such algorithms is far from being used as standard tools within clinical practices. This is because of lacking trust in these algorithms and because a standard implementation is not straightforward. (Jonas et al., 2019) and (Tjepkema-Cloostermans et al., 2019) base their prediction pipeline on intervals of EEG recordings without epileptic activity. The selection of such a recording interval requires manual work performed by a clinician. Nevertheless, these studies show impressive scores for predicting outcome (AUC = 0.89 (Jonas et al., 2019) AUC at 24 hours = 0.88 (Tjepkema-Cloostermans et al., 2019), best AUC at 15 hours = 0.89 (Zheng et al., 2021)). This thesis showed a mean AUC score of 0.70 ± 0.04 and 0.83 for the best fold, and a mean PPV of 0.83 ± 0.03 and 0.92 for the best fold. These performances are slightly lower than the current literature. However, the presented pipeline integrates complementary information to current clinical markers and thus great performance for patients in a 'grey zone'.

5.2.3 Electrophysiological features for comatose patients and networks confidence

In chapter 3 a strong correlation between the network's confidence in predicting survival and the PLV values of comatose patients was found. A high PLV was correlated with the network's confidence in predicting survival. These results could be considered trivial, as the PLV, as well as the network, predict outcome of coma. However, this correlation was also found for survivors and non survivors separately. This suggests that the network's output and therefore learned features are strongly connected to the PLV.

Similarly, a correlation between the network's confidence in predicting survival and the LZ complexity was found, even though this measure alone was shown not to be predictive of outcome of comatose patients. Also for this measure, a separate correlation between survivors and non survivors and the network's confidence could be observed.

The CNN automatically extracted features from the EEG data of the comatose patients related to the neural complexity and neural synchrony. This leaves the network with an advantage over had-crafted and interpretable features, as this extraction was fully data-driven, without an initial hypothesis about the data.

(Alnes et al., 2021) showed that the Lempel-Ziv complexity was significantly different between comatose patients and healthy controls. This effect was driven by survivors. Non survivors, however, did not show a significant difference from controls with respect to the LZ values. Some patients even had complexity values at the same level as awake controls. These results were confirmed in chapter 3, also with patients that were treated with therapeutic normothermia. The complexity values of survivors and non survivors did not show a significant difference between the two temperature treatments. Therefore a subgroup of non survivors had high values of complexity, similar to the awake controls. These results are contradictory to current literature, which found lower values of LZ for decreased levels of consciousness (Luppi et al., 2019; Miskovic et al., 2018). The relatively high complexity values that can be observed in Aellen et al. (2023) and chapter 3 of this thesis may be driven by the fact that all patients were recorded within the first 24 hours of coma. During this time, a patient's EEG and metabolism undergo drastic changes. It is, however, unclear if this effect would hold in later stages of coma and how the LZ values would change on a second day, and if they could then be predictive of the outcome.

5.2.4 Limitations

In my work, I only performed limited hyperparameter or network architecture tuning. The used architecture was relatively shallow, with only three convolutional layers (Lawhern et al., 2018). It was selected because of the only 19 available EEG channels in some of the recordings. The hyperparameter tuning that was performed for this work was limited to a subset of the cohort. It remains unexplored if a deeper network architecture would increase the performance or result in overfitting.

My study predicts outcome from coma as survivors (CPC 1-3) versus non survivors (CPC 5), compared to other deep learning based studies that predict favorable (CPC 1-2) versus unfavorable (CPC 3-5) outcome (Jonas et al., 2019; Tjepkema-Cloostermans et al., 2019; Zheng et al., 2021). (Juan et al., 2016) showed that the CPC correlated with the measures of outcome found in (Tzovara et al., 2013), where the outcome prediction was based on machine learning techniques. No significant difference in the network's predicted score of confidence of survival and the CPC of patients was found for the subgroup of survivors. But it remains unknown if an artificial neural network could distinguish the varying levels of CPC. Such a prediction of the CPC, could provide information not just about awakening, but also on the quality of life patients have after three months.

The cohort of patient data analysed in this work was collected over many years, and some of the data collected more than ten years ago were recorded with only 19 electrodes. Current research usually uses more channels, and 19 channels are considered a relatively limited number. Although I obtained a high performance in predicting awakening, I cannot exclude that higher dimensional data might capture more details in neural responses and improve classification performance further.

5.2.5 Future directions for deep networks in outcome prediction of coma

From a methodological standpoint, this work did not explore many different hyperparameters and network architectures. Future studies could therefore explore deeper and more complex architectures to evaluate if a trained model's generalization would improve.

The prediction of outcome of patients for this work was based on a deviance protocol of standardized auditory stimulations, with a duration of around 20 minutes. The relatively long recording time and the large amount of data collected for the train set are of great value to capture the variation within the signals. It however remains unexplored how much data the prediction pipeline requires for an accurate prediction of patients' outcome in the test set. (Gómez-Tapia et al., 2022) found that for fingerprinting based on EEG signals, i.e., identifying an individual based on their EEG signal, up to three seconds of recording data is enough. An open question, that future studies could explore is how many auditory stimulations and thus how long recordings per patient are required for an accurate prediction of outcome with the prediction pipeline presented in this thesis. A reduced recording length could be of value for a possible use of such a model in a clinical setting.

The sound stimulations were presented as a deviance paradigm with three different deviant sounds, as previous studies have shown the relevance of deviance paradigms for predicting coma outcome (Fischer et al., 1999; Kane et al., 1993; Naccache et al., 2005; Luauté et al., 2005). However, analyzing deviance responses was not the main focus of the study. A supplemental analysis showed that the network's confidence of survival did not depend on the type of auditory stimulation (Appendix, Section B.4). (An et al., 2021) were able to differentiate the neural response to different types of deviant sounds in wake participants. Previous work in comatose patients (Tzovara et al., 2013, 2016) trained MVPA classifiers to differentiate standard and one deviant sound and was able to do so on a patient by patient level. Future studies could thus explore with deep learning models if, on a group level in comatose patients, the different sound types could be disentangled based on the EEG responses. This possibility to distinguish different types of deviant sounds could be related to survival, and potentially improve outcome prognosis.

Previous studies on outcome prediction of coma with deep learning reported contradictory results on the optimal time after the onset of coma for outcome prediction. (Zheng et al., 2021) reported that the best outcome prediction can be expected at 66 hours, compared to (Tjepkema-Cloostermans et al., 2019) which reported 12 hours. EEG data from a later recording of the cohort used in this study could be analyzed with the use of deep learning, to explore the optimal latency for predicting survival, to reinforce one of these results.

5.3 Disentangling the complex landscape of sleep-wake disorders with data-driven phenotyping: A study of the Bernese Center

5.3.1 Summary

The landscape of sleep-wake disorders is complex, as different SWDs can co-occur within one patient. Additionally, not many disorders have objective diagnostic biomarkers. Already within central disorders of hypersomnolence, the different disorders are convoluted because the distinction between narcolepsy type 2 and idiopathic hypersomnia is based on a clinical test with poor retest reliability (Trotti et al., 2013; Lopez et al., 2017).

In an attempt to disentangle the full landscape of SWDs, in the third part of this thesis, a data processing pipeline based on an unsupervised clustering algorithm was built. My first results focus on CDHs, where I confirm previous results that NT1 is clearly distinguishable but not NT2 and IH. On the full cohort of SWDs, I found for the sleep-related breathing disorders (OSA and CSA) and NT1 clear clusters; all other disorders were intermixed. As a third analysis, I only considered patients with a well-defined, single SWD and found additional clusters of insomnia. These results were also confirmed with multiple control analyses. My findings support the need for new markers on the full landscape of SWDs. This work showed for the first time an unsupervised clustering on the full landscape of sleep-wake disorders and that currently collected clinical markers were not able to fully discriminate the full landscape of SWDs, with the exception of OSA, CSA and NT1.

5.3.2 Unsupervised clustering can assist in phenotyping patients with SWDs

Previous attempts to find distinct clusters within CDHs have revealed different results (Gool et al., 2022; Šonka et al., 2015). Both of these studies found clear clusters of NT1, highlighting that this disorder has clear biomarkers. For NT2 and IH however, the literature found intermixed results. (Gool et al., 2022) identified two clusters of NT2 and IH, that were most distinct in variables related to sleep drunkenness and awakenings. (Šonka et al., 2015) found unwanted naps and awakenings from naps as the biggest difference between clusters of NT2 and IH. In the third part of this thesis, the difference in clusters between NT2 and IH was mainly driven by patients' sex, REM latency, mean REM latency, and the number of naps. Additionally, for NT1 clear distinct clusters were identified, confirming previous findings.

For the first time ever this work used an unsupervised clustering algorithm on the full cohort of sleep-wake disorders and found clear clusters of disorders with clear biomarkers, such as OSA, CSA and NT1. However, no clear clusters were found for RLS or parasomnias, even though they have clearly distinct symptoms. Additionally, the results were consistent with previous literature, showing that insomnia was rather intermixed with other disorders (Edinger, 2011), and could only be distinguished when a clean cohort was considered, containing only patients with single SWDs.

5.3.3 Limitations

I performed numerous control analyses and tested different clustering algorithms, metrics, imputation techniques, and other steps of the pipeline. On the one hand, I proved the stability of my clustering, showing that most choices of hyperparameters resulted in similar clusters. On the other hand, I showed that no other choices of hyperparameters would reveal clearer clusters of SWDs. This implies that my choices of hyperparameters were justified. However, it is always possible that other unexplored clustering algorithms or metrics exist that can identify additional pathologies within the landscape of SWDs.

Another drawback of this study is the data collection step, as much of the data based on questionnaires was collected on paper and only later digitized. This could lead to possible errors and loss of data. A lot of the questionnaire data was excluded because only insufficient good-quality data was available. Similarly, not all sleep-wake disorders have the same frequency of occurrence in the population (Ohayon, 2011). The maximum possible number of patients was included here, to represent all SWDs. This, however, lead to an imbalance in the data and potentially introduced bias. This was addressed in the case of OSA patients that made up almost two third of the original cohort, by downsampling patients with this disorder. Because of the very low number of available samples, circadian sleep-wake disorders were excluded altogether.

5.3.4 Future directions for phenotyping of SWDs with artificial intelligence

In this study, it was impossible to disentangle the full landscape of sleep-wake disorders based on the available and currently used markers. Many of the markers used in the clustering were extracted from the overnight PSG or MSLT recording. This raises the question if possible new markers could be extracted from these recordings for disorders that do not have clear biomarkers, such as restless leg syndrome. (Gholami et al., 2022) were able to distinguish sleep-related breathing disorders from healthy controls, by using EEG channels from the overnight PSG recording, extracting features, and training machine learning algorithms. A future project could therefore attempt to classify all sleep-wake disorders from the raw data of the PSG with deep learning models. In such an attempt, it would be possible to consider different sources, for example EEG, EMG, EOG. With such rich data and deep learning models, a better distinction of SWDs could be possible. Deep learning models could extract macro or micro patterns from PSG recordings, related for example to sleep stages or power spectra that could best distinguish different SWDs.

A different approach could focus on patients' treatments instead of their diagnosed disorders. In CDHs, recommendations of treatment depend on the symptoms shown by patients and not their diagnosis (Khan and Trotti, 2015). Therefore patients with different disorders might get the same treatments. It could be meaningful for future studies to focus on treatments, especially successful treatments, instead of patients' diagnoses. However, this information is rarely collected and would require longitudinal data collection. Future studies could therefore gather reports about treatments and the impact of medication for prospective cohort studies.

Wearable devices and home monitoring have become increasingly used in research and for clinical purposes (Schmidt et al., 2021; Dietmann et al., 2021; Katori et al., 2022; Ancona et al., 2022). The diagnosis of IH now takes into consideration actigraphy data collected over two weeks (ICSD-3, 2014). The use of actigraphy data provides several advantages, such as that patients are in their own homes and beds during recording (Katori et al., 2022). Additionally, it also allows for the collection of data over multiple weeks. Future studies, recording and analyzing actigraphy data could make diagnosis more precise, as the results of the clinical tests would not only depend on a single night when patients are in an unfamiliar environment.

5.4 Deep learning for EEG signals in neuroscience

The literature presents a number of studies using deep learning to analyze EEG signals, in order to automate and simplify the diagnosis of neurological disorders or for BCI applications (Roy et al., 2019; Craik et al., 2019; Zhang et al., 2019; Lawhern et al., 2018; Schirrmeister et al., 2017; Jonas et al., 2019; Fiorillo et al., 2019; Stephansen et al., 2018; Cho and Jang, 2020; Burrello et al., 2020; Aellen et al., 2021, 2023; Heilmeyer et al., 2018). This literature explores many different applications, network architectures, and data representations. Previous work on deep learning and EEG signals are therefore very diverse. However, these applications or analysis methods are not yet established within clinics or research practices. There are numerous problems that need to be addressed before these methods can be applied in neurological practices or neuroscience. One such problem is the limited quantity of available EEG recordings. This can be addressed with solutions, such as data augmentations, that were discussed above, in regards to the contribution of chapter 2. This section covers two additional problems for using deep learning techniques for EEG signals.

5.4.1 Architectures of deep learning models on EEG data

Previous literature has introduced many new network architectures that can be applied to EEG signals. The two contributions on deep learning and EEG data in this thesis ((Aellen et al., 2021), in chapter 2 and (Aellen et al., 2023), in chapter 3) have the limitation that not many architectures and hyperparameters were explored. Most publications introduce a novel network architecture for a specific clinical problem, and there are rarely any ablation studies showing why the choice of parameters was optimal. Other publications introducing a new architecture for general use, such as (Lawhern et al., 2018), compare their network on a few datasets with different architectures. Another example is (Schirrmeister et al., 2017), who performed hyperparameter tuning on the proposed architecture, e.g., with/without dropout, batch normalization, etc. (Heilmeyer et al., 2018) tested four different network architectures over six different BCI datasets on single patient learning and single folds, comparing training scores across a total of 100 subjects. However, to this day there exists no large-scale testing of many different architectures and multiple datasets, spanning many different applications of EEG data. Such a study could for example span different datasets from BCI, seizure predictions, sleep staging, prognostication of coma, and basic research. Chapter 2 of this thesis addressed such a research direction on a very small scale. Three different network architectures were tested on two different datasets from the field of basic research, showing no significant difference between performances over 10 folds (Figure 2.8, in chapter 2). Future work could expand this investigation on more architectures, features, and datasets from different modalities, such as sleep-scoring, BCI, outcome prediction from coma, and EEG research datasets.

5.4.2 Variance of deep neural network performance for EEG signals

This thesis included two chapters exploring deep learning for EEG signals. Chapter 2, showed three different network architectures (ResNet5050 (He et al., 2016), a shallow and deep convolutional network (Schirrmeister et al., 2017)), that were each trained and tested on two different datasets, one based on auditory and the other on visual stimulations. The AUC scores for all network architectures and datasets were spread around ten percentage points, over all the 10 folds trained. A similar spread of AUC scores could be observed for predicting outcome from coma in chapter 3. These examples, taken together, spanned four different network architectures and three different datasets, each showing that AUC scores can have a spread of around 10 percentage points around the mean values, trained over 10 folds. These big differences in the performance of the same network over different folds are typical for the literature on EEG and deep learning and can be observed in many other studies (Schirrmeister et al., 2017; Heilmeyer et al., 2018). These big differences in performance and their implications for clinical or basic research, are currently underexplored and ignored in the field of EEG research based on artificial neural networks.

One hypothesis is that these spreads could depend on the architecture. For example, one could expect that smaller architectures might be more stable overall. This would be an advantage for the potential implementation of deep learning applications in a clinical setting. An alternative hypothesis is related to the data, as some folds contain by nature a lot more noisy data, and independently of the architecture their generalization to the test set will be poorer. These questions need to be addressed before such algorithms could be implemented within the clinics, otherwise, the performance of such a model would substantially depend on the selected fold. Even a selection based on the best performance on the validation or test sets introduces a bias toward that specific test population, which may happen to contain samples that are not representative of the entire dataset. One way to address this limitation is to train multiple networks on the same folds. For example, in chapter 2 each fold was trained four times, totaling 40 trained networks. The reported AUC scores per fold were then shown as a mean over the performances of these four networks. A similar technique has been used in medical imaging, to analyze stable features (Fauw et al., 2018), but it has not been thoroughly explored yet for EEG research. This leaves the open question of how many different networks would need to be trained for stable performances.

5.5 Data-driven versus feature based algorithms for neural signals

The two chapters (2, 3) that trained deep learning models on EEG signals were based on minimally preprocessed data. However, the use of data-driven versus feature-based models is still an ongoing discussion. EEG data is high dimensional, and the raw data can be rather hard to interpret. Therefore often, some frequency transform or entropy measures are calculated from the raw EEG signals, before using classifiers (Kuanar et al., 2018; Bashivan et al., 2016; Tan et al., 2018; Wang et al., 2018). Using pre-selected features has the advantage that a suitable, trained classifier is interpretable, as one can extract its weights and infer the most essential features for its classification decisions. Feature-based models are beneficial for simpler machine learning techniques with less optimizable parameters. However, a thorough feature selection greatly restricts the power of more sophisticated classification models, such as deep artificial neural networks. Using pre-selected features to train a neural network can limit the network's capacity to infer the most discriminative patterns in the data. Deep artificial neural networks can approximate any function (Zhou, 2020). Therefore if a feature such as a frequency transform would have the biggest difference between conditions, a deep enough neural network could approximate this calculation. This approach allows for a data-driven extraction of features and decoding of EEG signals. However, a limitation of such an approach is that the model becomes intransparent, and learning features are not directly extractable. Algorithms such as saliency maps (Simonyan et al., 2013) can help visualize important patterns in the input data. However, the information extracted might be harder to interpret than features from classifiers trained with pre-selected features.

5.6 Artificial intelligence for neurological signals

Artificial intelligence has a vast potential to aid clinicians in their diagnosis and assessments of EEG signals and to save costs by optimizing care. However, the use of methods based on artificial intelligence in the field of EEG signals is still limited. This is not because of a lack of studies exploring different clinical applications of using artificial intelligence for EEG (Fiorillo et al., 2019; Cho and Jang, 2020; Burrello et al., 2020; Jonas et al., 2019; Aellen et al., 2023). Yet, before clinicians can use these methods daily, several obstacles must be solved. Additionally, efforts must be made to evaluate and reduce bias in the proposed models.

5.6.1 From Research to clinical applications

There are a great number of publications on artificial intelligence for clinical applications, from sleep-staging (Fiorillo et al., 2019; Vallat and Walker, 2021; Stephansen et al., 2018) to seizure detection (Cho and Jang, 2020; Burrello et al., 2020) or prognosis of coma outcome (Tzovara et al., 2013; Jonas et al., 2019; Aellen et al., 2023). Some studies are based on simpler machine learning models and others on very sophisticated deep learning architectures. However, these techniques are not integrated into the clinics, and they still have to address multiple challenges. The reasons are manifold, among others, one point is lacking trust that these methods perform as promised.

This is also highlighted in (Yu et al., 2020), where the authors compare the performance of deep learning models trained on detecting lung cancers for a public challenge. While the models showed promising performances on a public data set, to demonstrate the lack of generalization, this study evaluated the models on an independent test set, that was not published. The performances on the test set were disappointing, as some models even performed at chance level. Additionally, the authors correlated the performance of the public and private sets and found no significant correlation. A better performance on the public data thus did not indicate a better performance on the private set. This challenge provided a training set of 1397 patients, which is not considered small. However, for most clinical applications, there are no public datasets even of this size available, except for tasks like sleep staging or seizure detection that are explored more frequently (Fiorillo et al., 2019; Burrello et al., 2020; Cho and Jang, 2020). The availability of substantial medical datasets is limited, which represents an important drawback for implementing artificial intelligence-based algorithms in clinics. First, because of patient confidentiality, the data is rarely published. Second, collecting and labeling medical data can be costly and time-consuming, as it requires clinical expertise. Therefore, medical applications based on artificial intelligence have limited data and may poorly generalize to new data collected from different hospitals. The open science community is growing and increasingly more clinical data is published online, even though the process is slow, there is more awareness. However, it remains still unexplored how to best apply a previously trained model within a given hospital. One hypothesis is that fine-tuning on a local subset of patient data could improve the performance of pre-trained models.

5.6.2 Bias in artificial intelligence and healthcare

An additional problem for automated algorithms is bias. Before models based on artificial intelligence can be used in a clinical setting, it is essential to identify and reduce sources of bias. In a collaboration that I did during my Ph.D., we identified three primary sources of bias for artificial intelligence and big data in healthcare (Norori et al., 2021). First, we present data-driven bias, which describes the bias found within the used dataset. Most of the publicly available neuroscience datasets were collected within a scientific setting and, therefore, often contain a bias toward a western, educated, industrialized, rich, and democratic population (Henrich et al., 2010). As demographic information is often not included in publicly released EEG data, these forms of bias are challenging to detect. Second, we identified algorithmic biases, for example, unbalanced datasets, non-appropriate evaluation metrics, or incorrect estimations of chance levels. The third source of bias is of human nature and can be very hard to detect as it might be very subtle and emerge from old societal sources (Norori et al., 2021). To bring artificial intelligence to clinical applications there is a strong need to first ensure that biases are eliminated. A recent study showed that correctly trained artificial neural networks would not necessarily suffer from bias and generalize well across subcohorts with different demographic information (Davatzikos, 2023). An additional solution might be the use of data augmentations techniques, such as those presented in chapter 2. These techniques allow to address biasses, as they can generate additional samples of under-represented populations and allow for the integration of more synthetic data points.

5.7 Conclusion

This thesis contributes to the field of artificial intelligence for neurology and clinical applications.

I first presented a methodological proof of concept that deep learning for EEG signals can be implemented as an MVPA algorithm for research purposes. Additionally, an exemplar feature extraction algorithm was presented and an application of trial-by-trial learning over the course of the experiment that could be used to test theories of learning was introduced.

In the second part of my thesis, I focused on post-cardiac arrest patients in a comatose state, deep learning, and EEG responses to sounds showing great potential in predicting survival, especially in a cohort of patients that, according to current clinical markers, are without a precise diagnosis. Further, the network's confidence in predicting survival was correlated to physiological properties of EEG signals, that have been linked to the integrity of auditory functions in coma. Specifically, the output of the network was strongly correlated with neural synchrony and neural complexity EEG responses to standardized auditory stimulations.

In the last contribution, I developed an unsupervised clustering pipeline based on clinical variables that was applied to patients with sleep-wake disorders. For some disorders, namely obstructive sleep apnea, central sleep apnea, and narcolepsy type 1 clear clusters were identified. However, it was not possible to disentangle other disorders, such as parasomnia or restless leg syndrome, based on the currently collected clinical markers. This work highlights that the markers currently collected are not sufficient to extract clusters of sleep-wake disorders in a data-driven way and calls for further studies to identify new biomarkers.

6 Bibliography

Aellen, F. M., Göktepe-Kavis, P., Apostolopoulos, S., and Tzovara, A. (2021). Convolutional neural networks for decoding electroencephalography responses and visualizing trial by trial changes in discriminant features. *Journal of Neuroscience Methods*, 364:109367.

Aellen, F. M., Alnes, S. L., Loosli, F., Rossetti, A. O., Zubler, F., De Lucia, M., and Tzovara, A. (2023). Auditory stimulation and deep learning predict awakening from coma after cardiac arrest. *Brain*, 146(2):778–788.

Aellen, F. M., Van der Meer, J., Dietmann, A., Schmidt, M., Bassetti, C. L. A., and Tzovara, A. (published). Disentangling the complex landscape of sleep-wake disorders with data-driven phenotyping: A study of the bernese center.

Alnes, S. L., Lucia, M. D., Rossetti, A. O., and Tzovara, A. (2021). Complementary roles of neural synchrony and complexity for indexing consciousness and chances of surviving in acute coma. *NeuroImage*, 245:118638.

Altıntop, C. G., Latifoğlu, F., Karayol Akın, A., and Çetin, B. (2022). A novel approach for detection of consciousness level in comatose patients from EEG signals with 1-d convolutional neural network. *Biocybernetics and Biomedical Engineering*, 42(1):16–26.

An, H., Ho Kei, S., Auksztulewicz, R., and Schnupp, J. W. H. (2021). Do auditory mismatch responses differ between acoustic features? *Frontiers in Human Neuroscience*, 15.

An, H.-J., Baek, S.-H., Kim, S.-W., Kim, S.-J., and Park, Y.-G. (2019). Clustering-based characterization of clinical phenotypes in obstructive sleep apnoea using severity, obesity, and craniofacial pattern. *European Journal of Orthodontics*, page cjz041.

An, X., Kuang, D., Guo, X., Zhao, Y., and He, L. (2014). A deep learning method for classification of eeg data based on motor imagery.

Ancona, S., Faraci, F. D., Khatab, E., Fiorillo, L., Gnarra, O., Nef, T., Bassetti, C. L. A., and Bargiotas, P. (2022). Wearables in the home-based assessment of abnormal movements in parkinson's disease: a systematic review of the literature. *Journal of Neurology*, 269(1):100–110.

Andlauer, O., Moore, H., Hong, S.-C., Dauvilliers, Y., Kanbayashi, T., Nishino, S., Han, F., Silber, M. H., Rico, T., Einen, M., Kornum, B. R., Jennum, P., Knudsen, S., Nevsimalova, S., Poli, F., Plazzi, G., and Mignot, E. (2012). Predictors of hypocretin (orexin) deficiency in narcolepsy without cataplexy. *Sleep*, 35(9):1247–1255.

Bailly, S., Destors, M., Grillet, Y., Richard, P., Stach, B., Vivodtzev, I., Timsit, J.-F., Lévy, P., Tamisier, R., Pépin, J.-L., and scientific council and investigators of the French national sleep apnea registry (OSFP) (2016). Obstructive sleep apnea: A cluster analysis at time of diagnosis. *PLOS ONE*, 11(6):e0157318.

Barateau, L., Liblau, R., Peyron, C., and Dauvilliers, Y. (2017). Narcolepsy type 1 as an autoimmune disorder: Evidence, and implications for pharmacological treatment. *CNS Drugs*, 31(10):821–834.

Bargiotas, P., Dietmann, A., Haynes, A. G., Kallweit, U., Calle, M. G., Schmidt, M., Mathis, J., and Bassetti, C. L. (2019). The swiss narcolepsy scale (SNS) and its short form (sSNS) for the discrimination of narcolepsy in patients with hypersomnolence: a cohort study based on the bern sleep–wake database. *Journal of Neurology*, 266(9):2137–2143.

Bashivan, P., Rish, I., Yeasin, M., and Codella, N. (2016). Learning representations from eeg with deep recurrent-convolutional neural networks.

Baumann, C. R., Mignot, E., Lammers, G. J., Overeem, S., Arnulf, I., Rye, D., Dauvilliers, Y., Honda, M., Owens, J. A., Plazzi, G., and Scammell, T. E. (2014). Challenges in diagnosing narcolepsy without cataplexy: A consensus statement. *Sleep*, 37(6):1035–1042.

Bear, author, M. F. (2016). *Neuroscience : exploring the brain.* Fourth edition. Philadelphia : Wolters Kluwer, [2016] ©2016.

Benghanem, S., Pruvost-Robieux, E., Bouchereau, E., Gavaret, M., and Cariou, A. (2022). Prognostication after cardiac arrest: how EEG and evoked potentials may improve the challenge. *Annals of Intensive Care*, 12(1):111.

Binder, M., Górska, U., and Griskova-Bulanova, I. (2017). 40 hz auditory steady-state responses in patients with disorders of consciousness: Correlation between phase-locking index and coma recovery scale-revised score. *Clinical Neurophysiology*, 128(5):799–806.

Bishop, C. M. (2006). *Pattern recognition and machine learning.* Information science and statistics. Springer.

Booth, C. M., Boone, R. H., Tomlinson, G., and Detsky, A. S. (2004). Is this patient dead, vegetative, or severely neurologically impaired?: Assessing outcome for comatose survivors of cardiac arrest. *JAMA*, 291(7):870.

Brinker, T. J., Hekler, A., Enk, A. H., Berking, C., Haferkamp, S., Hauschild, A., Weichenthal, M., Klode, J., Schadendorf, D., Holland-Letz, T., von Kalle, C., Fröhling, S., Schilling, B., and Utikal, J. S. (2019). Deep neural networks are superior to dermatologists in melanoma image classification. *European Journal of Cancer*, 119:11–17.

Burrello, A., Schindler, K., Benini, L., and Rahimi, A. (2020). Hyperdimensional computing with local binary patterns: One-shot learning of seizure onset and identification of ictogenic brain regions using short-time ieeg recordings. *IEEE Transactions on Biomedical Engineering*, 67(2):601–613.

Cairns, A., Trotti, L. M., and Bogan, R. (2019). Demographic and nap-related variance of the MSLT: results from 2,498 suspected hypersomnia patients. *Sleep Medicine*, 55:115–123.

Cappuccio, F. P., Miller, M. A., and Lockley, S. W. (2010). *Sleep, Health and Society: From Aetiology to Public Health.* Oxford University Press.

Caroyer, S., Depondt, C., Rikir, E., Mavroudakis, N., Peluso, L., Silvio Taccone, F., Legros, B., and Gaspard, N. (2021). Assessment of a standardized EEG reactivity protocol after cardiac arrest. *Clinical Neurophysiology*, 132(7):1687–1693.

Carrasco-Gómez, M., Keijzer, H. M., Ruijter, B. J., Bruña, R., Tjepkema-Cloostermans, M. C., Hofmeijer, J., and van Putten, M. J. (2021a). EEG functional connectivity contributes to outcome prediction of postanoxic coma. *Clinical Neurophysiology*, 132(6):1312–1320.

Carrasco-Gómez, M., Keijzer, H. M., Ruijter, B. J., Bruña, R., Tjepkema-Cloostermans, M. C., Hofmeijer, J., and van Putten, M. J. (2021b). EEG functional connectivity contributes to outcome prediction of postanoxic coma. *Clinical Neurophysiology*, 132(6):1312–1320.

Carskadon, M. A. and Dement, W. C. (1982). The multiple sleep latency test: What does it measure. *Sleep*, 5:S67–S72.

Casali, A. G., Gosseries, O., Rosanova, M., Boly, M., Sarasso, S., Casali, K. R., Casarotto, S., Bruno, M.-A., Laureys, S., Tononi, G., and Massimini, M. (2013). A theoretically based index of consciousness independent of sensory processing and behavior. *Science Translational Medicine*, 5(198):198ra105–198ra105.

Castegnetti, G., Tzovara, A., Khemka, S., Melin, F., Barnes, G., Dolan, R., and Bach, D. (2020). Representation of probabilistic outcomes during risky decision-making. *Nature Communications*, 11.

Cavanagh, J. F., Kumar, P., Mueller, A. A., Richardson, S. P., and Mueen, A. (2018). Diminished eeg habituation to novel events effectively classifies parkinson's patients. *Clinical Neurophysiology*, *129*(2):409 – 418.

Cheng, J. Y., Goh, H., Dogrusoz, K., Tuzel, O., and Azemi, E. (2020). Subject-aware contrastive learning for biosignals.

Cho, K.-O. and Jang, H.-J. (2020). Comparison of different input modalities and network structures for deep learning-based seizure detection. *Scientific Reports*, 10.

Choi, H. A., Badjatia, N., and Mayer, S. A. (2012). Hypothermia for acute brain injury—mechanisms and practical aspects. *Nature Reviews Neurology*, 8(4):214–222.

Cook, J., Rumble, M., and Plante, D. (2019). Identifying subtypes of hypersomnolence disorder: a clustering analysis. *Sleep Medicine*, 64:71–76.

Craik, A., He, Y., and Contreras-Vidal, J. L. (2019). Deep learning for electroencephalogram (eeg) classification tasks: a review. *Journal of Neural Engineering*, 16(3):031001.

Daltrozzo, J., Wioland, N., Mutschler, V., and Kotchoubey, B. (2007). Predicting coma and other low responsive patients outcome using event-related brain potentials: A meta-analysis. *Clinical Neurophysiology*, 118(3):606–614.

Davatzikos, R. W. P. C. C. (2023). Bias in machine learning models can be significantly mitigated by careful training: Evidence from neuroimaging studies. *Proceedings of the National Academy of Sciences*, 120(6):e2211613120.

Deiss, O., Biswal, S., Jin, J., Sun, H., Westover, M. B., and Sun, J. (2018). Hamlet: Interpretable human and machine co-learning technique.

Delvigne, V., Wannous, H., Dutoit, T., Ris, L., and Vandeborre, J.-P. (2022). Phydaa: Physiological dataset assessing attention. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(5):2612–2623.

Demarchi, G., Sanchez, G., and Weisz, N. (2019). Automatic and feature-specific prediction-related neural activity in the human auditory system. *Nature Communications*, 10.

Dietmann, A., Wenz, E., Meer, J., Ringli, M., Warncke, J. D., Edwards, E., Schmidt, M. H., Bernasconi, C. A., Nirkko, A., Strub, M., Miano, S., Manconi, M., Acker, J., Manitius, S., Baumann, C. R., Valko, P. O., Yilmaz, B., Brunner, A., Tzovara, A., Zhang, Z., Largiadèr, C. R., Tafti, M., Latorre, D., Sallusto, F., Khatami, R., and Bassetti, C. L. A. (2021). The swiss primary hypersomnolence and narcolepsy cohort study (SPHYNCS): Study protocol for a prospective, multicentre cohort observational study. *Journal of Sleep Research*, 30(5).

Donnino, M. W., Andersen, L. W., Berg, K. M., Reynolds, J. C., Nolan, J. P., Morley, P. T., Lang, E., Cocchi, M. N., Xanthos, T., Callaway, C. W., and Soar, J. (2016). Temperature management after cardiac arrest: An advisory statement by the advanced life support task force of the international liaison committee on resuscitation and the american heart association emergency cardiovascular care committee and the council on cardiopulmonary, critical care, perioperative and resuscitation. *Resuscitation*, 98:97–104.

Donovan, L. M. and Kapur, V. K. (2016). Prevalence and characteristics of central compared to obstructive sleep apnea: Analyses from the sleep heart health study cohort. *Sleep*, 39(7):1353–1359.

Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern classification*. Wiley, 2nd ed edition.

Edinger, J. D., Bonnet, M. H., Bootzin, R. R., Doghramji, K., Dorsey, C. M., Espie, C. A., Jamieson, A. O., McCall, W. V., Morin, C. M., and Stepanski, E. J. (2004). Derivation of Research Diagnostic Criteria for Insomnia: Report of an American Academy of Sleep Medicine Work Group. *Sleep*, 27(8):1567–1596.

Edinger, J. D. (2011). Testing the reliability and validity of DSM-IV-TR and ICSD-2 insomnia diagnoses: Results of a multitrait-multimethod analysis. *Archives of General Psychiatry*, 68(10):992.

Epsilon-machine (2022). Missingpy: Missing data imputation for python. `https://github.com/epsilon-machine/missingpy`. Accessed: 2022-09-06.

Erickson, J. and Vaughn, B. V. (2019). Non-REM parasomnia: The promise of precision medicine. *Sleep Medicine Clinics*, 14(3):363–370.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231. AAAI Press.

Farahat, A., Reichert, C., Sweeney-Reed, C. M., and Hinrichs, H. (2019). Convolutional neural networks for decoding of covert attention focus and saliency maps for eeg feature visualization. *bioRxiv*.

Fauw, J., Ledsam, J., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O'Donoghue, B., Visentin, D., Driessche, G., Lakshminarayanan, B., Meyer, C., Mackinder, F., Bouton, S., Ayoub, K., Chopra, R., King, D., Karthikesalingam, A., and Ronneberger, O. (2018). Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine*, 24.

Fchollet (2016). Github deep learning models. `https://github.com/fchollet/deep-learning-models/releases/download/v0.2/resnet50_weights_tf_dim_ordering_tf_kernels_notop.h5`. Accessed: 2010-09-30.

Fiorillo, L., Puiatti, A., Papandrea, M., Ratti, P.-L., Favaro, P., Roth, C., Bargiotas, P., Bassetti, C. L., and Faraci, F. D. (2019). Automated sleep scoring: A review of the latest approaches. *Sleep Medicine Reviews*, 48:101204.

Fischer, C., Morlet, D., Bouchet, P., Luaute, J., Jourdan, C., and Salord, F. (1999). Mismatch negativity and late auditory evoked potentials in comatose patients. *Clinical Neurophysiology*, 110(9):1601–1610.

Fischer, C., Luaute, J., Adeleine, P., and Morlet, D. (2004). Predictive value of sensory and cognitive evoked potentials for awakening from coma. *Neurology*, 63(4):669–673.

Fischer, C., Luauté, J., Némoz, C., Morlet, D., Kirkorian, G., and Mauguière, F. (2006). Improved prediction of awakening or nonawakening from severe anoxic coma using tree-based classification analysis*:. *Critical Care Medicine*, 34(5):1520–1524.

Fischer, C., Dailler, F., and Morlet, D. (2008). Novelty p3 elicited by the subject's own name in comatose patients. *Clinical Neurophysiology*, 119(10):2224–2230.

Fischer, C., Luaute, J., and Morlet, D. (2010). Event-related potentials (MMN and novelty p3) in permanent vegetative or minimally conscious states. *Clinical Neurophysiology*, 121(7):1032–1042.

Fonken, Y. M., Kam, J. W. Y., and Knight, R. T. (2020). A differential role for human hippocampus in novelty and contextual processing: Implications for p300. *Psychophysiology*, 57(7):e13400.

Friedrich, E. V., Scherer, R., and Neuper, C. (2013). Long-term evaluation of a 4-class imagery-based brain–computer interface. *Clinical Neurophysiology*, 124(5):916–927.

Fronczek, R., Arnulf, I., Baumann, C. R., Maski, K., Pizza, F., and Trotti, L. M. (2020). To split or to lump? classifying the central disorders of hypersomnolence. *Sleep*, 43(8):zsaa044.

Gersh, S. F. Q. B. J. (2004). Cardiovascular consequences of sleep-disordered breathing: Past, present and future. *Circulation*, 109(8):951–957.

Gholami, B., Behboudi, M. H., Khadem, A., Shoeibi, A., and Gorriz, J. M. (2022). Sleep apnea diagnosis using complexity features of eeg signals. In Ferrández Vicente, J. M., Álvarez-Sánchez, J. R., de la Paz López, F., and Adeli, H., editors, *Artificial Intelligence in Neuroscience: Affective Analysis and Health Applications*, pages 74–83, Cham. Springer International Publishing.

Ghosh, A., dal Maso, F., Roig, M., Mitsis, G. D., and Boudrias, M.-H. (2018). Deep semantic architecture with discriminative feature visualization for neuroimage analysis.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Gool, J. K., Zhang, Z., Oei, M. S., Mathias, S., Dauvilliers, Y., Mayer, G., Plazzi, G., del Rio-Villegas, R., Cano, J. S., Šonka, K., Partinen, M., Overeem, S., Peraita-Adrados, R., Heinzer, R., Martins da Silva, A., Högl, B., Wierzbicka, A., Heidbreder, A., Feketeova, E., Manconi, M., Bušková, J., Canellas, F., Bassetti, C. L., Barateau, L., Pizza, F., Schmidt, M. H., Fronczek, R., Khatami, R., and Lammers, G. J. (2022). Data-driven phenotyping of central disorders of hypersomnolence with unsupervised clustering. *Neurology*, 98(23):e2387–e2400.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, 27(4):857.

Grandner, M. A. (2017). Sleep, health, and society. *Sleep Medicine Clinics*, 12(1):1–22.

Gratton, G., Coles, M., and Donchin, E. (1983). A new method for off-line removal of ocular artifact. *Electroencephalography and Clinical Neurophysiology*, 55(4):468 – 484.

Grootswagers, T., Wardle, S. G., and Carlson, T. A. (2017). Decoding dynamic brain patterns from evoked responses: A tutorial on multivariate pattern analysis applied to time series neuroimaging data. *Journal of Cognitive Neuroscience*, 29(4):677–697. PMID: 27779910.

Guo, S. and Yang, Z. (2018). Multi-channel-resnet: An integration framework towards skin lesion analysis. *Informatics in Medicine Unlocked*, 12:67 – 74.

Gómez-Tapia, C., Bozic, B., and Longo, L. (2022). On the minimal amount of eeg data required for learning distinctive human features for task-dependent biometric applications. *Frontiers in Neuroinformatics*, 16.

Hajinoroozi, M., Mao, Z., Lin, Y.-P., and Huang, Y. (2017). Deep transfer learning for cross-subject and cross-experiment prediction of image rapid serial visual presentation events from eeg data.

Hari, R. and Puce, A. (2017). *MEG-EEG Primer*. Oxford University Press.

Hassan, B. A., Rashid, T. A., and Hamarashid, H. K. (2021). A novel cluster detection of COVID-19 patients and medical disease conditions using improved evolutionary clustering algorithm star. *Computers in Biology and Medicine*, 138:104866.

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., and Bießmann, F. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, 87:96 – 110.

Haynes, J.-D. and Rees, G. (2006). Decoding mental states from brain activity in human. *Nature reviews. Neuroscience*, 7:523–34.

Haynes, J.-D. (2015). A primer on pattern-based approaches to fmri: Principles, pitfalls, and perspectives. *Neuron*, 87(2):257–270.

He, B., Hori, J., and Babiloni, F. (2006). *Electroencephalography (EEG): Inverse Problems*, page 1355–1363. American Cancer Society.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Heilmeyer, F. A., Schirrmeister, R. T., Fiederer, L. D. J., Volker, M., Behncke, J., and Ball, T. (2018). A large-scale evaluation framework for eeg deep learning architectures. In *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1039–1045.

Henrich, J., Heine, S. J., and Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2):61–83.

Hirsch, J. (2005). Raising consciousness. *The Journal of Clinical Investigation*, 115(5):1102–1102.

Hirsch, L. J., Fong, M. W., Leitinger, M., LaRoche, S. M., Beniczky, S., Abend, N. S., Lee, J. W., Wusthoff, C. J., Hahn, C. D., Westover, M. B., Gerard, E. E., Herman, S. T., Haider, H. A., Osman, G., Rodriguez-Ruiz, A., Maciel, C. B., Gilmore, E. J., Fernandez, A., Rosenthal, E. S., Claassen, J., Husain, A. M., Yoo, J. Y., So, E. L., Kaplan, P. W., Nuwer, M. R., van Putten, M., Sutter, R., Drislane, F. W., Trinka, E., and Gaspard, N. (2021). American clinical neurophysiology society's standardized critical care EEG terminology: 2021 version. *Journal of Clinical Neurophysiology*, 38(1):1–29.

Holmberg, M. J., Ross, C. E., Fitzmaurice, G. M., Chan, P. S., Duval-Arnould, J., Grossestreuer, A. V., Yankama, T., Donnino, M. W., Andersen, L. W., for the American Heart Association's Get With The Guidelines–Resuscitation Investigators*, Chan, P., Grossestreuer, A. V., Moskowitz, A., Edelson, D., Ornato, J., Berg, K., Peberdy, M. A., Churpek, M., Kurz, M., Starks, M. A., Girotra, S., Perman, S., Goldberger, Z., Guerguerian, A.-M., Atkins, D., Foglia, E., Fink, E., Lasa, J. J., Roberts, J., Bembea, M., Gaies, M., Kleinman, M., Gupta, P., Sutton, R., and Sawyer, T. (2019). Annual incidence of adult and pediatric in-hospital cardiac arrest in the united states. *Circulation: Cardiovascular Quality and Outcomes*, 12(7):e005580.

Holzer, M. (2010). Targeted temperature management for comatose survivors of cardiac arrest. *New England Journal of Medicine*, 363(13):1256–1264. PMID: 20860507.

ICSD-3 (2014). International classification of sleep disorders.

Ieracitano, C., Mammone, N., Hussain, A., and Morabito, F. (2021). A novel explainable machine learning approach for eeg-based brain-computer interface systems. *Neural Computing and Applications*, pages 1–14.

Iyer, P. M., Mohr, K., Broderick, M., Gavin, B., Burke, T., Bede, P., Pinto-Grau, M., Pender, N. P., McLaughlin, R., Vajda, A., Heverin, M., Lalor, E. C., Hardiman, O., and Nasseroleslami, B. (2017). Mismatch negativity as an indicator of cognitive sub-domain dysfunction in amyotrophic lateral sclerosis. *Frontiers in Neurology*, 8.

Jiang, X., Bian, G.-B., and Tian, Z. (2019). Removal of artifacts from eeg signals: A review. *Sensors*, 19(5).

Johns, M. W. (1991). A New Method for Measuring Daytime Sleepiness: The Epworth Sleepiness Scale. *Sleep*, 14(6):540–545.

Jonas, S., Rossetti, A. O., Oddo, M., Jenni, S., Favaro, P., and Zubler, F. (2019). Eeg-based outcome prediction after cardiac arrest with convolutional neural networks: Performance and visualization of discriminative features. *Human Brain Mapping*, 40(16):4606–4617.

Joosten, S. A., Hamza, K., Sands, S., Turton, A., Berger, P., and Hamilton, G. (2011). Phenotypes of patients with mild to moderate obstructive sleep apnoea as confirmed by cluster analysis: Cluster analysis of OSA phenotypes. *Respirology*, 17(1):99–107.

Juan, E., De Lucia, M., Tzovara, A., Beaud, V., Oddo, M., Clarke, S., and Rossetti, A. O. (2016). Prediction of cognitive outcome based on the progression of auditory discrimination during coma. *Resuscitation*, 106:89–95.

Juan, E., De Lucia, M., Beaud, V., Oddo, M., Rusca, M., Viceic, D., Clarke, S., and Rossetti, A. O. (2018). How do you feel? subjective perception of recovery as a reliable surrogate of cognitive and functional outcome in cardiac arrest survivors:. *Critical Care Medicine*, 46(4):e286–e293.

Kane, N., Curry, S., Butler, S., and Cummins, B. (1993). Electrophysiological indicator of awakening from coma. *The Lancet*, 341(8846):688.

Kane, N. M., Butler, S. R., and Simpson, T. (2000). Coma outcome prediction using event-related potentials: P_3 and mismatch negativity. *Audiology and Neurotology*, 5(3):186–191.

Kao, C.-H., D'Rozario, A. L., Lovato, N., Wassing, R., Bartlett, D., Memarian, N., Espinel, P., Kim, J.-W., Grunstein, R. R., and Gordon, C. J. (2021). Insomnia subtypes characterised by objective sleep duration and NREM spectral power and the effect of acute sleep restriction: an exploratory analysis. *Scientific Reports*, 11(1):24331.

Katori, M., Shi, S., Ode, K. L., Tomita, Y., and Ueda, H. R. (2022). The 103,200-arm acceleration dataset in the UK biobank revealed a landscape of human sleep phenotypes. *Proceedings of the National Academy of Sciences*, 119(12):e2116729119.

Kerkhof, G. A. (2017). Epidemiology of sleep and sleep disorders in the netherlands. *Sleep Medicine*, 30:229–239.

Khan, Z. and Trotti, L. M. (2015). Central disorders of hypersomnolence. *Chest*, 148(1):262–273.

King, J.-R. and Dehaene, S. (2014). Characterizing the dynamics of mental representations: the temporal generalization method. *Trends in Cognitive Sciences*, 18(4):203 – 210.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization.

Koles, Z., Lazar, M., and Zhou, S. (1990). Spatial patterns underlying population differences in the background eeg. *Brain topography*, 2(4):275—284.

Kotikalapudi, R. and contributors (2017). keras-vis. `https://github.com/raghakot/keras-vis`.

Krell, M. M. and Kim, S. K. (2017). Rotational data augmentation for electroencephalographic data. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 471–474. IEEE.

Kretzschmar, U., Werth, E., Sturzenegger, C., Khatami, R., Bassetti, C. L., and Baumann, C. R. (2016). Which diagnostic findings in disorders with excessive daytime sleepiness are really helpful? a retrospective study. *Journal of Sleep Research*, 25(3):307–313.

Kuanar, S., Athitsos, V., Pradhan, N., Mishra, A., and Rao, K. R. (2018). Cognitive analysis of working memory load from eeg, by a deep recurrent neural network. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2576–2580.

Kurth-Nelson, Z., Barnes, G., Sejdinovic, D., Dolan, R., and Dayan, P. (2015). Temporal structure in associative retrieval. *eLife*, 4:e04919.

Kustermann, T., Nguepnjo Nguissi, N. A., Pfeiffer, C., Haenggi, M., Kurmann, R., Zubler, F., Oddo, M., Rossetti, A. O., and De Lucia, M. (2019). Electroencephalography-based power spectra allow coma outcome prediction within 24 h of cardiac arrest. *Resuscitation*, 142:162–167.

Lachaux, J.-P., Rodriguez, E., Martinerie, J., and Varela, F. J. (1999). Measuring phase synchrony in brain signals. *Human Brain Mapping*, 8(4):194–208.

Lammers, G. J., Bassetti, C. L., Dolenc-Groselj, L., Jennum, P. J., Kallweit, U., Khatami, R., Lecendreux, M., Manconi, M., Mayer, G., Partinen, M., Plazzi, G., Reading, P. J., Santamaria, J., Sonka, K., and Dauvilliers, Y. (2020). Diagnosis of central disorders of hypersomnolence: A reappraisal by european experts. *Sleep Medicine Reviews*, 52:101306.

Lashgari, E., Liang, D., and Maoz, U. (2020). Data augmentation for deep-learning-based electroencephalography. *Journal of Neuroscience Methods*, 346:108885.

Lawhern, V. J., Solon, A. J., Waytowich, N. R., Gordon, S. M., Hung, C. P., and Lance, B. J. (2018). Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of Neural Engineering*, 15(5):056013.

Lechinger, J., Wielek, T., Blume, C., Pichler, G., Michitsch, G., Donis, J., Gruber, W., and Schabus, M. (2016). Event-related EEG power modulations and phase connectivity indicate the focus of attention in an auditory own name paradigm. *Journal of Neurology*, 263(8):1530–1543.

Lemm, S., Blankertz, B., Dickhaus, T., and Müller, K.-R. (2011). Introduction to machine learning for brain imaging. *NeuroImage*, 56(2):387 – 399. Multivariate Decoding and Brain Reading.

Lempel, A. and Ziv, J. (1976). On the complexity of finite sequences. *IEEE Transactions on Information Theory*, 22(1):75–81.

Lieder, F., Daunizeau, J., Garrido, M. I., Friston, K. J., and Stephan, K. E. (2013). Modelling trial-by-trial changes in the mismatch negativity. *PLOS Computational Biology*, *9*(2):1–16.

Liu, Y., Huang, H., Su, Y., Wang, M., Zhang, Y., Chen, W., Liu, G., and Jiang, M. (2021). The combination of n60 with mismatch negativity improves the prediction of awakening from coma. *Neurocritical Care*.

Lopez, R., Doukkali, A., Barateau, L., Evangelista, E., Chenini, S., Jaussent, I., and Dauvilliers, Y. (2017). Test–retest reliability of the multiple sleep latency test in central disorders of hypersomnolence. *Sleep*, 40(12).

Lotte, F. (2014). *A Tutorial on EEG Signal-processing Techniques for Mental-state Recognition in Brain–Computer Interfaces*, pages 133–161. Springer London, London.

Luauté, J., Fischer, C., Adeleine, P., Morlet, D., Tell, L., and Boisson, D. (2005). Late auditory and event-related potentials can be useful to predict good functional outcome after coma. *Archives of Physical Medicine and Rehabilitation*, 86(5):917–923.

Luca, G., Haba-Rubio, J., Dauvilliers, Y., Lammers, G.-J., Overeem, S., Donjacour, C. E., Mayer, G., Javidi, S., Iranzo, A., Santamaria, J., Peraita-Adrados, R., Hor, H., Kutalik, Z., Plazzi, G., Poli, F., Pizza, F., Arnulf, I., Lecendreux, M., Bassetti, C., Mathis, J., Heinzer, R., Jennum, P., Knudsen, S., Geisler, P., Wierzbicka, A., Feketeova, E., Pfister, C., Khatami, R., Baumann, C., Tafti, M., and (EU-NN), E. N. N. (2013). Clinical, polysomnographic and genome-wide association analyses of narcolepsy with cataplexy: a european narcolepsy network study. *Journal of Sleep Research*, 22(5):482–495.

Luppi, A. I., Craig, M. M., Pappas, I., Finoia, P., Williams, G. B., Allanson, J., Pickard, J. D., Owen, A. M., Naci, L., Menon, D. K., and Stamatakis, E. A. (2019). Consciousness-specific dynamic interactions of brain integration and functional diversity. *Nature Communications*, 10(1):4616.

Macmillan, N. and Creelman, D. (2004). *Detection Theory: A User's Guide*, volume xix. Psychology Press.

Madhok, J., Wu, D., Xiong, W., Geocadin, R. G., and Jia, X. (2012). Hypothermia amplifies somatosensory-evoked potentials in uninjured rats. *Journal of Neurosurgical Anesthesiology*, 24(3):197–202.

Mahlios, J., De la Herrán-Arita, A. K., and Mignot, E. (2013). The autoimmune basis of narcolepsy. *Current Opinion in Neurobiology*, 23(5):767–773. Circadian rhythm and sleep.

Maris, E. and Oostenveld, R. (2007). Nonparametric statistical testing of eeg- and meg-data. *Journal of neuroscience methods*, 164(1):177—190.

McCloskey, S., Jeffries, B., Koprinska, I., Miller, C. B., and Grunstein, R. R. (2019-11). Data-driven cluster analysis of insomnia disorder with physiology-based qEEG variables. *Knowledge-Based Systems*, 183:104863.

Mehrer, J., Spoerer, C. J., Kriegeskorte, N., and Kietzmann, T. C. (2020). Individual differences among deep neural network models. *Nature Communications*, 11(1):5725.

Michel, C. M., Murray, M. M., Lantz, G., Gonzalez, S., Spinelli, L., and Grave de Peralta, R. (2004). Eeg source imaging. *Clinical Neurophysiology*, 115(10):2195–2222.

Mignot, E., Lammers, G. J., Ripley, B., Okun, M., Nevsimalova, S., Overeem, S., Vankova, J., Black, J., Harsh, J., Bassetti, C., Schrader, H., and Nishino, S. (2002). The Role of Cerebrospinal Fluid Hypocretin Measurement in the Diagnosis of Narcolepsy and Other Hypersomnias. *Archives of Neurology*, 59(10):1553–1562.

Miller, C. B., Bartlett, D. J., Mullins, A. E., Dodds, K. L., Gordon, C. J., Kyle, S. D., Kim, J. W., D'Rozario, A. L., Lee, R. S., Comas, M., Marshall, N. S., Yee, B. J., Espie, C. A., and Grunstein, R. R. (2016). Clusters of insomnia disorder: An exploratory cluster analysis of objective sleep parameters reveals differences in neurocognitive functioning, quantitative EEG, and heart rate variability. *Sleep*, 39(11):1993–2004.

Miskovic, V., MacDonald, K. J., Rhodes, L. J., and Cote, K. A. (2018). Changes in EEG multiscale entropy and power-law frequency scaling during the human sleep cycle. *Human Brain Mapping*, 40(2):538–551.

Mohsenvand, M. N., Izadi, M. R., and Maes, P. (2020). Contrastive representation learning for electroencephalogram classification. In Alsentzer, E., McDermott, M. B. A., Falck, F., Sarkar, S. K., Roy, S., and Hyland, S. L., editors, *Proceedings of the Machine Learning for Health NeurIPS Workshop*, volume 136 of *Proceedings of Machine Learning Research*, pages 238–253. PMLR.

Morlet, D. and Fischer, C. (2014). MMN and novelty p3 in coma and other altered states of consciousness: A review. *Brain Topography*, 27(4):467–479.

Naccache, L., Puybasset, L., Gaillard, R., Serve, E., and Willer, J.-C. (2005). Auditory mismatch negativity is a good predictor of awakening in comatose patients: a fast and reliable procedure. *Clinical Neurophysiology*, 116(4):988–989.

Nielsen, N., Wetterslev, J., Cronberg, T., Erlinge, D., Gasche, Y., Hassager, C., Horn, J., Hovdenes, J., Kjaergaard, J., Kuiper, M., Pellis, T., Stammet, P., Wanscher, M., Wise, M. P., Åneman, A., Al-Subaie, N., Boesgaard, S., Bro-Jeppesen, J., Brunetti, I., Bugge, J. F., Hingston, C. D., Juffermans, N. P., Koopmans, M., Køber, L., Langørgen, J., Lilja, G., Møller, J. E., Rundgren, M., Rylander, C., Smid, O., Werer, C., Winkel, P., and Friberg, H. (2013). Targeted temperature management at 33°c versus 36°c after cardiac arrest. *New England Journal of Medicine*, 369(23):2197–2206.

Nir, Y., Vyazovskiy, V. V., Cirelli, C., Banks, M. I., and Tononi, G. (2015). Auditory Responses and Stimulus-Specific Adaptation in Rat Auditory Cortex are Preserved Across NREM and REM Sleep. *Cerebral Cortex*, 25(5):1362–1378.

Nolan, J. P., Soar, J., Cariou, A., Cronberg, T., Moulaert, V. R., Deakin, C. D., Bottiger, B. W., Friberg, H., Sunde, K., and Sandroni, C. (2015). European resuscitation council and european society of intensive care medicine guidelines for post-resuscitation care 2015. *Resuscitation*, 95:202–222.

Nolan, J. P., Sandroni, C., Böttiger, B. W., Cariou, A., Cronberg, T., Friberg, H., Genbrugge, C., Haywood, K., Lilja, G., Moulaert, V. R. M., Nikolaou, N., Olasveengen, T. M., Skrifvars, M. B., Taccone, F., and Soar, J. (2021). European resuscitation council and european society of intensive care medicine guidelines 2021: post-resuscitation care. *Intensive Care Medicine*, 47(4):369–421.

Norori, N., Hu, Q., Aellen, F. M., Faraci, F. D., and Tzovara, A. (2021). Addressing bias in big data and AI for health care: A call for open science. *Patterns*, 2(10):100347.

Nourski, K. V., Steinschneider, M., Rhone, A. E., Kawasaki, H., Howard, M. A., and Banks, M. I. (2018). Auditory predictive coding across awareness states under anesthesia: An intracranial electrophysiology study. *Journal of Neuroscience*.

Nurse, E., Mashford, B. S., Yepes, A. J., Kiral-Kornek, I., Harrer, S., and Freestone, D. R. (2016). Decoding eeg and lfp signals using deep learning: Heading truenorth.

Ohayon, M. M. (2011). Epidemiological overview of sleep disorders in the general population. *Sleep Med Res*, 2(1):1–9.

Parwani, A. V. (2019). Next generation diagnostic pathology: use of digital pathology and artificial intelligence tools to augment a pathological diagnosis. *Diagnostic Pathology*, 14(1):138, s13000–019–0921–2.

Pataka, A. D., Frangulyan, R. R., Mackay, T. W., Douglas, N. J., and Riha, R. L. (2012). Narcolepsy and sleep-disordered breathing. *European Journal of Neurology*, 19(5):696–702.

Perkins, G. D., Callaway, C. W., Haywood, K., Neumar, R. W., Lilja, G., Rowland, M. J., Sawyer, K. N., Skrifvars, M. B., and Nolan, J. P. (2021). Brain injury after cardiac arrest. *The Lancet*, 398(10307):1269–1278.

Perrin, F., Pernier, J., Bertnard, O., Giard, M., and Echallier, J. (1987). Mapping of scalp potentials by surface spline interpolation. *Electroencephalography and Clinical Neurophysiology*, 66(1):75–81.

Pfeiffer, C., Nguissi, N. A. N., Chytiris, M., Bidlingmeyer, P., Haenggi, M., Kurmann, R., Zubler, F., Oddo, M., Rossetti, A. O., and De Lucia, M. (2017). Auditory discrimination improvement predicts awakening of postanoxic comatose patients treated with targeted temperature management at 36 °c. *Resuscitation*, 118:89–95.

Pfeiffer, C., Nguissi, N. A. N., Chytiris, M., Bidlingmeyer, P., Haenggi, M., Kurmann, R., Zubler, F., Accolla, E., Viceic, D., Rusca, M., Oddo, M., Rossetti, A. O., and De Lucia, M. (2018). Somatosensory and auditory deviance detection for outcome prediction during postanoxic coma. *Annals of Clinical and Translational Neurology*, 5(9):1016–1024.

Philiastides, M. G., Biele, G., Vavatzanidis, N., Kazzer, P., and Heekeren, H. R. (2010). Temporal dynamics of prediction error processing during reward-based decision making. *NeuroImage*, 53(1):221 – 232.

Prechelt, L. (1998). *Early Stopping - But When?*, pages 55–69. Springer Berlin Heidelberg, Berlin, Heidelberg.

Rentzsch, J., Jockers-Scherübl, M. C., Boutros, N. N., and Gallinat, J. (2008). Test–retest reliability of p50, n100 and p200 auditory sensory gating in healthy subjects. *International Journal of Psychophysiology*, 67(2):81 – 90.

Reuderink, B. and Poel, M. (2008). Robustness of the common spatial patterns algorithm in the bci-pipeline. *IEEE Transactions on Circuits and Systems I-regular Papers - IEEE TRANS CIRCUIT SYST-I*.

Rommel, C., Moreau, T., Paillard, J., and Gramfort, A. (2021). Cadda: Class-wise automatic differentiable data augmentation for eeg signals.

Rommel, C., Paillard, J., Moreau, T., and Gramfort, A. (2022). Data augmentation for learning predictive models on EEG: a systematic comparison. *Journal of Neural Engineering*, 19(6):066020.

Rossetti, A. O., Oddo, M., Logroscino, G., and Kaplan, P. W. (2010). Prognostication after cardiac arrest and hypothermia: A prospective study. *Annals of Neurology*, pages NA–NA.

Rossetti, A. O., Rabinstein, A. A., and Oddo, M. (2016). Neurological prognostication of outcome in patients in coma after cardiac arrest. *The Lancet Neurology*, 15(6):597–609.

Rossetti, A. O., Tovar Quiroga, D. F., Juan, E., Novy, J., White, R. D., Ben-Hamouda, N., Britton, J. W., Oddo, M., and Rabinstein, A. A. (2017). Electroencephalography predicts poor and good outcomes after cardiac arrest: A two-center study*. *Critical Care Medicine*, 45(7):e674–e682.

Roy, Y., Banville, H., Albuquerque, I., Gramfort, A., Falk, T. H., and Faubert, J. (2019). Deep learning-based electroencephalography analysis: a systematic review.

Saeed, A., Grangier, D., Pietquin, O., and Zeghidour, N. (2021). Learning from heterogeneous eeg signals with differentiable channel reordering. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1255–1259. IEEE.

Sakurai, T., Amemiya, A., Ishii, M., Matsuzaki, I., Chemelli, R. M., Tanaka, H., Williams, S., Richardson, J. A., Kozlowski, G. P., Wilson, S., Arch, J. R., Buckingham, R. E., Haynes, A. C., Carr, S. A., Annan, R. S., McNulty, D. E., Liu, W.-S., Terrett, J. A., Elshourbagy, N. A., Bergsma, D. J., and Yanagisawa, M. (1998). Orexins and orexin receptors: A family of hypothalamic neuropeptides and g protein-coupled receptors that regulate feeding behavior. *Cell*, 92(4):573–585.

Salmanpour, M. R., Shamsaei, M., Hajianfar, G., Soltanian-Zadeh, H., and Rahmim, A. (2022). Longitudinal clustering analysis and prediction of parkinson's disease progression using radiomics and hybrid machine learning. *Quantitative Imaging in Medicine and Surgery*, 12(2):906–919.

Sarasso, S., Boly, M., Napolitani, M., Gosseries, O., Charland-Verville, V., Casarotto, S., Rosanova, M., Casali, A., Brichant, J.-F., Boveroux, P., Rex, S., Tononi, G., Laureys, S., and Massimini, M. (2015). Consciousness and complexity during unresponsiveness induced by propofol, xenon, and ketamine. *Current Biology*, 25(23):3099–3105.

Schirrmeister, R. T., Springenberg, J. T., Fiederer, L. D. J., Glasstetter, M., Eggensperger, K., Tangermann, M., Hutter, F., Burgard, W., and Ball, T. (2017). Deep learning with convolutional neural networks for eeg decoding and visualization. *Human Brain Mapping*, 38(11):5391–5420.

Schmidt, M. H., Dekkers, M. P., Baillieul, S., Jendoubi, J., Wulf, M.-A., Wenz, E., Fregolente, L., Vorster, A., Gnarra, O., and Bassetti, C. L. (2021). Measuring sleep, wakefulness, and circadian functions in neurologic disorders. *Sleep Medicine Clinics*, 16(4):661–671.

Schwabedal, J. T. C., Snyder, J. C., Cakmak, A., Nemati, S., and Clifford, G. D. (2018). Addressing class imbalance in classification problems of noisy signals by using fourier transform surrogates.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2019). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.

Shorten, C. and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, 6:1–48.

Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps.

Singh, M., Drake, C. L., and Roth, T. (2006). The prevalence of multiple sleep-onset REM periods in a population-based sample. *Sleep*, 29(7):890–895.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958.

Stekhoven, D. J. and Bühlmann, P. (2011). MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118.

Stephansen, J. B., Olesen, A. N., Olsen, M., Ambati, A., Leary, E. B., Moore, H. E., Carrillo, O., Lin, L., Han, F., Yan, H., Sun, Y. L., Dauvilliers, Y., Scholz, S., Barateau, L., Hogl, B., Stefani, A., Hong, S. C., Kim, T. W., Pizza, F., Plazzi, G., Vandi, S., Antelmi, E., Perrin, D., Kuna, S. T., Schweitzer, P. K., Kushida, C., Peppard, P. E., Sorensen, H. B. D., Jennum, P., and Mignot, E. (2018). Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. *Nature Communications*, 9(1):5229.

Stretti, F., Klinzing, S., Ehlers, U., Steiger, P., Schuepbach, R., Krones, T., and Brandi, G. (2018). Low level of vegetative state after traumatic brain injury in a swiss academic hospital:. *Anesthesia & Analgesia*, 127(3):698–703.

Suzuki, K. (2017). Overview of deep learning in medical imaging. *Radiological Physics and Technology*, 10.

Tan, C., Sun, F., and Zhang, W. (2018). Deep transfer learning for eeg-based brain computer interface.

Tang, Z., Li, C., and Sun, S. (2016). Single-trial eeg classification of motor imagery using deep convolutional neural networks. *Optik - International Journal for Light and Electron Optics*, 130.

Teasdale, G. and Jennett, B. (1974). Assessment of coma and impaired consciousness: A practical scale. *The Lancet*, 304(7872):81–84. Originally published as Volume 2, Issue 7872.

Tivadar, R. I., Tivadar, R. I., Knight, R., Knight, R. T., and Tzovara, A. (2021). Automatic sensory predictions: A review of predictive mechanisms in the brain and their link to conscious processing. *Frontiers in Human Neuroscience*.

Tjepkema-Cloostermans, M. C., da Silva Lourenço, C., Ruijter, B. J., Tromp, S. C., Drost, G., Kornips, F. H. M., Beishuizen, A., Bosch, F. H., Hofmeijer, J., and van Putten, M. J. A. M. (2019). Outcome prediction in postanoxic coma with deep learning*:. *Critical Care Medicine*, 47(10):1424–1432.

Tomioka, R., Aihara, K., and Müller, K.-R. (2006). Logistic regression for single trial eeg classification. In *Proceedings of the 19th International Conference on Neural Information Processing Systems*, NIPS'06, page 1377–1384, Cambridge, MA, USA. MIT Press.

Tononi, G., Boly, M., Massimini, M., and Koch, C. (2016). Integrated information theory: from consciousness to its physical substrate. *Nature Reviews Neuroscience*, 17(7):450–461.

Trotti, L. M., Staab, B. A., and Rye, D. B. (2013). Test-retest reliability of the multiple sleep latency test in narcolepsy without cataplexy and idiopathic hypersomnia. *Journal of Clinical Sleep Medicine*, 09(8):789–795.

Tsetsou, S., Novy, J., Pfeiffer, C., Oddo, M., and Rossetti, A. O. (2018). Multimodal outcome prognostication after cardiac arrest and targeted temperature management: Analysis at 36 °c. *Neurocritical Care*, 28(1):104–109.

Tzovara, A., Murray, M. M., Plomp, G., Herzog, M. H., Michel, C. M., and Lucia], M. D. (2012). Decoding stimulus-related information from single-trial eeg responses based on voltage topographies. *Pattern Recognition*, *45*(6):2109 – 2122. Brain Decoding.

Tzovara, A., Rossetti, A. O., Spierer, L., Grivel, J., Murray, M. M., Oddo, M., and De Lucia, M. (2013). Progression of auditory discrimination based on neural decoding predicts awakening from coma. *Brain*, 136(1):81–89.

Tzovara, A., Rossetti, A. O., Juan, E., Suys, T., Viceic, D., Rusca, M., Oddo, M., and Lucia, M. D. (2016). Prediction of awakening from hypothermic postanoxic coma based on auditory discrimination: Awakening from postanoxic coma. *Annals of Neurology*, 79(5):748–757.

Vahid, A., Mückschel, M., Stober, S., Stock, A.-K., and Beste, C. (2020). Applying deep learning to single-trial eeg data provides evidence for complementary theories on action control. *Communications Biology*.

Vallat, R. and Walker, M. P. (2021). An open-source, high-performance tool for automated sleep staging. *eLife*, 10:e70092.

van de Nieuwenhuijzen, M., Backus, A., Bahramisharif, A., Doeller, C., Jensen, O., and van Gerven, M. (2013). Meg-based decoding of the spatiotemporal dynamics of visual category perception. *NeuroImage*, 83:1063–1073.

van Peer, J. M., Grandjean, D., and Scherer, K. R. (2014). Sequential unfolding of appraisals: Eeg evidence for the interaction of novelty and pleasantness. *Emotion*, 14(1):51–63.

van Peer, J. M., Coutinho, E., Grandjean, D., and Scherer, K. R. (2017). Emotion-antecedent appraisal checks: Eeg and emg datasets for novelty and pleasantness. `https://doi.org/10.5281/zenodo.197404`.

Venkatnarayan, K., Krishnaswamy, U. M., Rajamuri, N. K. R., Selvam, S., Veluthat, C., Devaraj, U., Ramachandran, P., and D'Souza, G. (2022). Identifying phenotypes of obstructive sleep apnea using cluster analysis. *Sleep and Breathing*.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, I., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., SciPy 1.0 Contributors, Vijaykumar, A., Bardelli, A. P., Rothberg, A., Hilboll, A., Kloeckner, A., Scopatz, A., Lee, A., Rokem, A., Woods, C. N., Fulton, C., Masson, C., Häggström, C., Fitzgerald, C., Nicholson, D. A., Hagen, D. R., Pasechnik, D. V., Olivetti, E., Martin, E., Wieser, E., Silva, F., Lenders, F., Wilhelm, F., Young, G., Price, G. A., Ingold, G.-L., Allen, G. E., Lee, G. R., Audren, H., Probst, I., Dietrich, J. P., Silterra, J., Webber, J. T., Slavič, J., Nothman, J., Buchner, J., Kulick, J., Schönberger, J. L., de Miranda Cardoso, J. V., Reimer, J., Harrington, J., Rodríguez, J. L. C., Nunez-Iglesias, J., Kuczynski, J., Tritz, K., Thoma, M., Newville, M., Kümmerer, M., Bolingbroke, M., Tartre, M., Pak, M., Smith, N. J., Nowaczyk, N., Shebanov, N., Pavlyk, O., Brodtkorb, P. A., Lee, P., McGibbon, R. T., Feldbauer, R., Lewis, S., Tygier, S., Sievert, S., Vigna, S., Peterson, S., More, S., Pudlik, T., Oshima, T., Pingel, T. J., Robitaille, T. P., Spura, T., Jones, T. R., Cera, T., Leslie, T., Zito, T., Krauss, T., Upadhyay, U., Halchenko, Y. O., and Vázquez-Baeza, Y. (2020). SciPy 1.0: fundamental algorithms for scientific computing in python. *Nature Methods*, 17(3):261–272.

Wang, F., Zhong, S.-h., Peng, J., Jiang, J., and Liu, Y. (2018). Data augmentation for eeg-based emotion recognition with deep convolutional neural networks. In Schoeffmann, K., Chalidabhongse, T. H., Ngo, C. W., Aramvith, S., O'Connor, N. E., Ho, Y.-S., Gabbouj, M., and Elgammal, A., editors, *MultiMedia Modeling*, pages 82–93, Cham. Springer International Publishing.

Westhall, E., Rosén, I., Rossetti, A. O., van Rootselaar, A.-F., Wesenberg Kjaer, T., Friberg, H., Horn, J., Nielsen, N., Ullén, S., and Cronberg, T. (2015). Interrater variability of EEG interpretation in comatose cardiac arrest patients. *Clinical Neurophysiology*, 126(12):2397–2404.

Westhall, E., Rossetti, A. O., van Rootselaar, A.-F., Wesenberg Kjaer, T., Horn, J., Ullén, S., Friberg, H., Nielsen, N., Rosén, I., Åneman, A., Erlinge, D., Gasche, Y., Hassager, C., Hovdenes, J., Kjaergaard, J., Kuiper, M., Pellis, T., Stammet, P., Wanscher, M., Wetterslev, J., and Wise, M. P. (2016). Standardized EEG interpretation accurately predicts prognosis after cardiac arrest. *Neurology*, 86(16):1482–1490.

Williams, J. M., Samal, A., Rao, P. K., and Johnson, M. R. (2020). Paired trial classification: A novel deep learning technique for mvpa. *Frontiers in Neuroscience*, 14:417.

Young, T., Peppard, P. E., and Gottlieb, D. J. (2002). Epidemiology of obstructive sleep apnea: A population health perspective. *American Journal of Respiratory and Critical Care Medicine*, 165(9):1217–1239.

Yu, K.-H., Lee, T.-L. M., Yen, M.-H., Kou, S. C., Rosen, B., Chiang, J.-H., and Kohane, I. S. (2020). Reproducible machine learning methods for lung cancer detection using computed tomography images: Algorithm development and validation. *J Med Internet Res*, 22(8):e16709.

Zeiler, M. D. and Fergus, R. (2013). Visualizing and understanding convolutional networks.

Zhang, X., Yao, L., Wang, X., Monaghan, J., Mcalpine, D., and Zhang, Y. (2019). A survey on deep learning based brain computer interface: Recent advances and new frontiers.

Zhang, Y., Zhang, X., Liu, W., Luo, Y., Yu, E., Zou, K., and Liu, X. (2014). Automatic sleep staging using multi-dimensional feature extraction and multi-kernel fuzzy support vector machine. *Journal of Healthcare Engineering*, 5(4):505–520.

Zhang, Y., Ren, R., Yang, L., Zhang, H., Shi, Y., Vitiello, M. V., Tang, X., and Sanford, L. D. (2022). Comparative polysomnography parameters between narcolepsy type 1/type 2 and idiopathic hypersomnia: A systematic review and meta-analysis. *Sleep Medicine Reviews*, 63:101610.

Zheng, W.-L., Amorim, E., Jing, J., Wu, O., Ghassemi, M., Lee, J. W., Sivaraju, A., Pang, T., Herman, S. T., Gaspard, N., Ruijter, B. J., Tjepkema-Cloostermans, M. C., Hofmeijer, J., Van Putten, M., and Westover, B. (2021). Predicting neurological outcome from electroencephalogram dynamics in comatose patients after cardiac arrest with deep learning. *IEEE Transactions on Biomedical Engineering*, pages 1–1.

Zhou, D.-X. (2020). Universality of deep convolutional neural networks. *Applied and Computational Harmonic Analysis*, 48(2):787–794.

Zinchuk, A. and Yaggi, H. K. (2020). Phenotypic subtypes of OSA: A challenge and opportunity for precision medicine. *Chest*, 157(2):403–420.

Zubarev, I., Zetter, R., Halme, H.-L., and Parkkonen, L. (2019). Adaptive neural network classifier for decoding meg signals. *NeuroImage*, 197:425 – 434.

Zubler, F., Steimer, A., Kurmann, R., Bandarabadi, M., Novy, J., Gast, H., Oddo, M., Schindler, K., and Rossetti, A. O. (2017). EEG synchronization measures are early outcome predictors in comatose patients after cardiac arrest. *Clinical Neurophysiology*, 128(4):635–642.

Züst, M. A., Ruch, S., Wiest, R., and Henke, K. (2019). Implicit vocabulary learning during sleep is bound to slow-wave peaks. *Current Biology*, 29(4):541–553.e7.

Šonka, K., Šusta, M., and Billiard, M. (2015). Narcolepsy with and without cataplexy, idiopathic hypersomnia with and without long sleep time: a cluster analysis. *Sleep Medicine*, 16(2):225–231.

A Appendix for CNNs for decoding EEG responses and visualizing trial by trial changes in discriminant features

Florence M. Aellen[1], Pinar Göktepe-Kavis[1], Stefanos Apostolopoulos[2], Athina Tzovara[1,3,4]

[1] Institute of Computer Science, University of Bern, Switzerland
[2] RetinAI Medical AG, Switzerland
[3] Sleep Wake Epilepsy Center - NeuroTec, Department of Neurology, Inselspital, Bern University Hospital, University of Bern, Switzerland
[4] Helen Wills Neuroscience Institute, University of California, Berkeley, United States

Aellen, F. M., Göktepe-Kavis, P., Apostolopoulos, S., and Tzovara, A. (2021). Convolutional neural networks for decoding electroencephalography responses and visualizing trial by trial changes in discriminant features. *Journal of Neuroscience Methods*, 364:109367. https://doi.org/10.1016/j.jneumeth.2021.109367

A.1 Data in topographic representation

In Figure A.1 we show the mean EEG data in a topographic map representation.

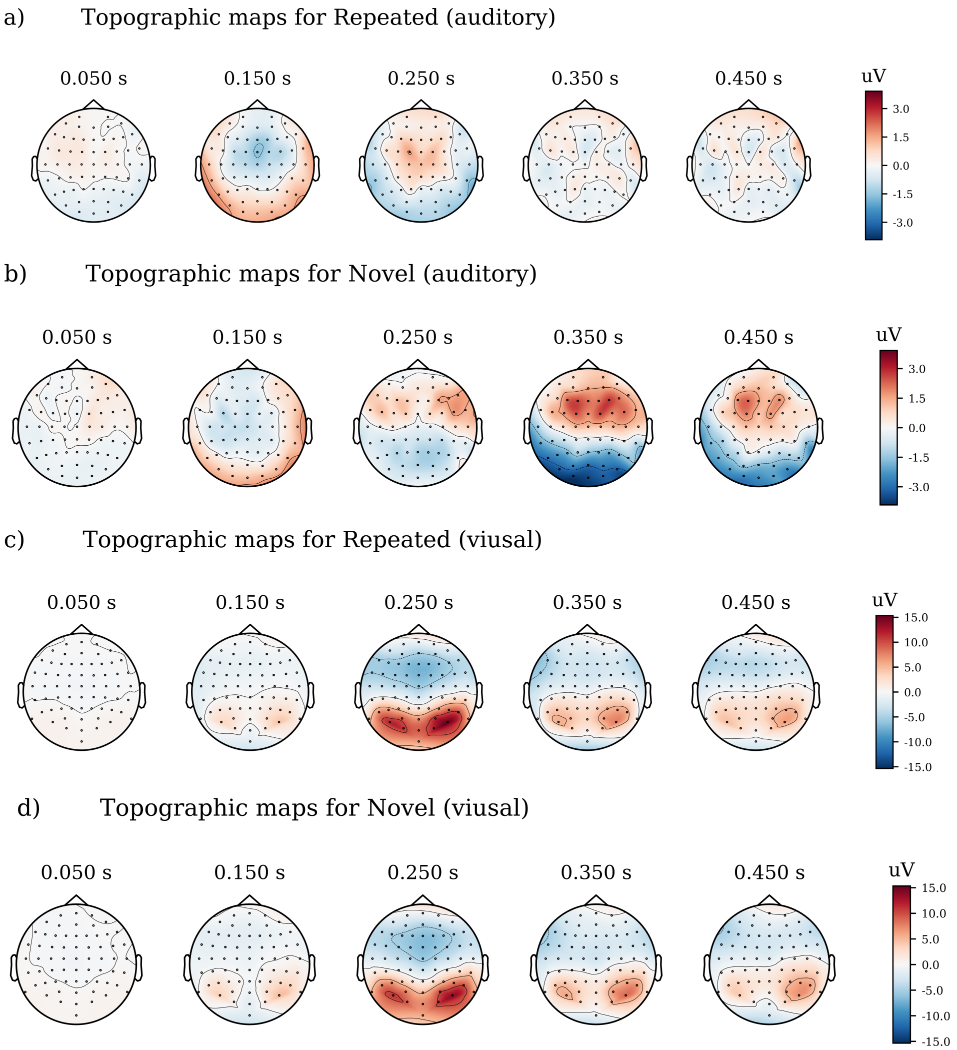


Figure A.1: The mean EEG data (Figure 2.1) for both conditions and both datasets as topographic maps.

A.2 Comparison of different settings

This section covers some comparison between different training options and justifies some of the choices that were made during for the training pipeline.

A.2.1 Training on filtered versus. unfiltered data

We evaluated whether filtering the EEG data would affect the network performance. In our main analyses, EEG data of the auditory dataset were filtered between 1 and 20 Hz. Here, we re-trained the CNNs without filtering the EEG data. Figure A.2, panel a and Table A.1 show the classification performance and loss over 10-folds of the auditory dataset, for filtered and unfiltered data. The network's performance was similar for filtered (0.70 ± 0.05) versus unfiltered (0.70 ± 0.04) data. However, when filtering the data, the network was slightly faster to converge, and needed to be trained on average for 28 ± 16 epochs until early stopping, while the network trained on unfiltered data needed to be trained for 33 ± 13 epochs.

Table A.1: Comparison of Binary cross-entropy loss, AUC-score and Accuracy for the training with filtered and unfiltered data over a 10-fold cross validation.

	Binary cross-entropy loss			AUC-score			Accuracy		
	Train	Validation	Test	Train	Validation	Test	Train	Validation	Test
Auditory dataset									
Filtered Data	0.26±0.08	0.43±0.05	1.16±2.33	0.89±0.04	0.75±0.04	0.72±0.04	0.82±0.12	0.79±0.11	0.77±0.10
Unfiltered Data	0.27±0.06	0.59±0.09	0.42±0.03	0.89±0.02	0.73±0.06	0.70±0.04	0.79±0.07	0.77±0.07	0.75±0.07

A.2.2 Training on oversampling versus training with weighted binary crossentropy loss

Because the datasets used in the present study are imbalanced, with a ratio of approximately 4:1 for *Repeated*:*Novel* trials, for our main analyses we oversampled the under-represented class (*Novel*). Moreover, we evaluated whether the weighted binary cross entropy loss could also account for the imbalanced data. To this aim, we re-trained the CNNs for the auditory dataset without over-sampling the *Novel* class, but using the weighted binary cross entropy loss function. The AUC score on the test dataset was significantly higher when oversampling the *Novel* class (0.72 ± 0.04), compared to when using the weighted binary cross entropy loss (0.62 ± 0.08) (Wilcoxon test, p = 0.002) (Figure A.2, panel b and Table A.2).

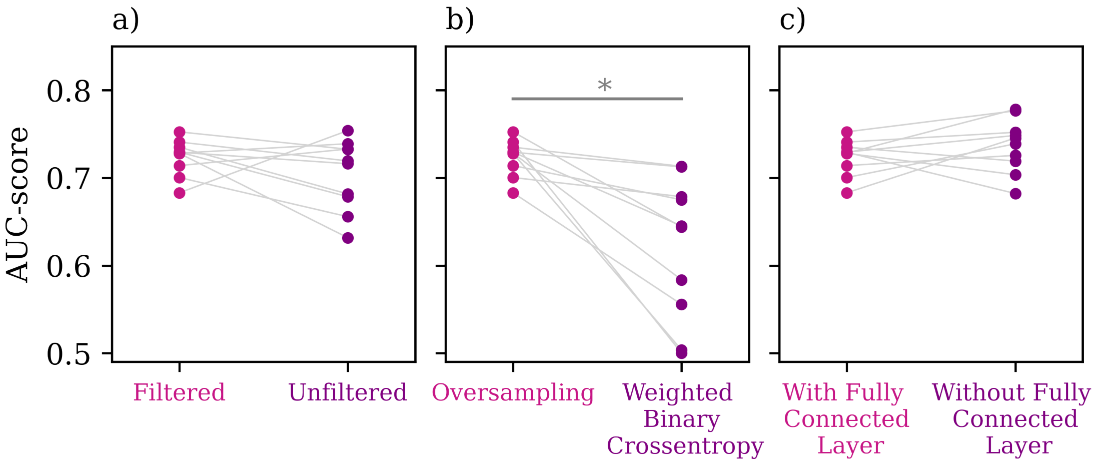


Figure A.2: 10-fold cross validation comparison between test AUC-scores of a training a) of filtered versus unfiltered, b) with oversampling versus a training with weighted binary crossentropy and c) of a training with and without the fully connected 128 node layer after the ResNet5050 architecture on the auditory dataset

A.2.3 Training of the ResNet architecture with and without a fully connected layer

The architecture of trained networks followed a typical ResNet5050 architecture, with the addition of a dense and a dropout layer (see 2.2.2). We tested the effects of adding a fully connected layer before the dropout layer by re-training the CNNs for the auditory dataset with and without the fully connected 128 nodes layer before the dropout layer. Training with the fully connected layer resulted in an AUC of 0.77 ± 0.10, while training without resulted in an AUC of 0.79 ± 0.02 (Figure A.1, panel c and Table A.3). The difference in classification performance with and without the fully connected layer across the 10-folds of cross validation was not significant ($p = 0.05$, with a Wilcoxon signed-rank test).

Table A.2: Comparison of Binary cross-entropy loss, AUC-score and Accuracy for the training with over sampled data and weighted binary crossentropy loss over a 10-fold cross validation.

	Binary cross-entropy loss			AUC-score			Accuracy		
	Train	Validation	Test	Train	Validation	Test	Train	Validation	Test
Auditory dataset									
Oversampled Data	0.26±0.08	0.43±0.05	1.16±2.33	0.89±0.04	0.75±0.04	0.72±0.04	0.82±0.12	0.79±0.11	0.77±0.10
Weighted Binary Corssentropy Loss	0.35±0.16	0.51±0.08	6990.3± 20968.9	0.69±0.14	0.63±0.10	0.62±0.08	0.81±0.03	0.79±0.05	0.77±0.06

Table A.3: Comparison of Binary cross-entropy loss, AUC-score and Accuracy for the training with over sampled data and weighted binary crossentropy loss over a 10-fold cross validation.

	Binary cross-entropy loss			AUC-score			Accuracy		
	Train	Validation	Test	Train	Validation	Test	Train	Validation	Test
Auditory dataset									
With Fully Connected Layer	0.26±0.08	0.43±0.05	1.16±2.33	0.89±0.04	0.75±0.04	0.72±0.04	0.82±0.12	0.79±0.11	0.77±0.10
Without Fully Connected Layer	0.40±0.07	0.46±0.03	0.88±0.62	0.83±0.03	0.76±0.04	0.73±0.03	0.85±0.03	0.81±0.02	0.79±0.02

A.3 Data representation

EEG data were represented as a 2D signal, consisting of channels by time (Figure 2.1). As the channel order of presentation is arbitrary, we homogenized the representation of EEG data by re-ordering the EEG channels with a k-means clustering algorithm. For each dataset, we considered the single-trials of both conditions together and clustered the data dimension (time-points $*$ trials) $\times$ electrodes into six clusters of brain regions with similar activity throughout the experiment. The clusters were then used to reorder the electrodes and get a better representation of the data. In Figure A.3 in the Appendix we show the clusters for both datasets. The total number of clusters was set to six a priori, as the clustering was not an essential step for our analyses, but rather a way to improve data visualization. Figure A.3 shows on panel a the clusters for the auditory and on panel b the clusters for the visual dataset.

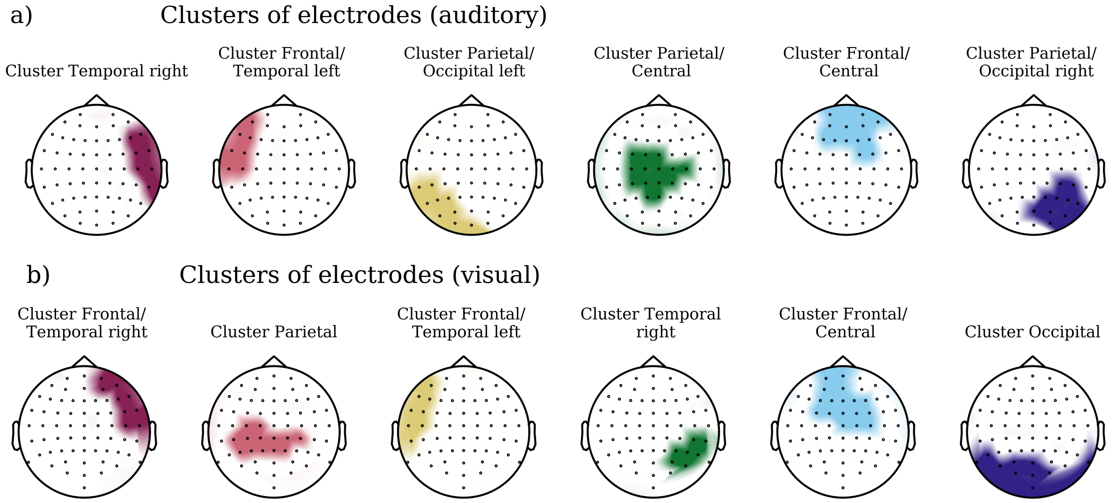

a)
Clusters of electrodes (auditory)
Cluster Temporal right
Cluster Frontal/ Temporal left
Cluster Parietal/ Occipital left
Cluster Parietal/ Central
Cluster Frontal/ Central
Cluster Parietal/ Occipital right
b)
Clusters of electrodes (visual)
Cluster Frontal/ Temporal right
Cluster Parietal
Cluster Frontal/ Temporal left
Cluster Temporal right
Cluster Frontal/ Central
Cluster Occipital

Figure A.3: The clusters of electrodes for both datasets, which were used for improving the visualization of input EEG data seen in Figure 2.1.

A.4 Illustration of trial by trial representation of discriminant features

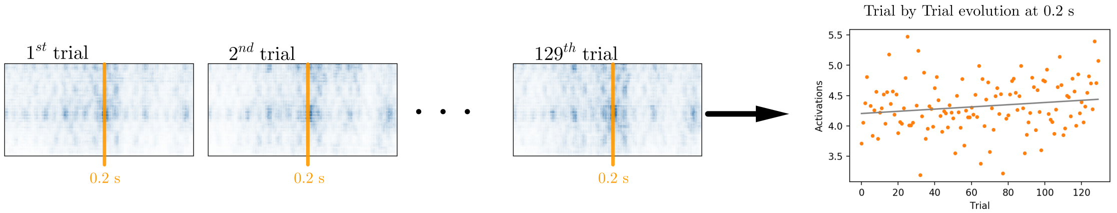


Figure A.4: For every participant and both stimuli we build a sequence in order of which the stimuli was presented to the participants. We take the mean over all participants and calculate the activation maps. Then for every time-point (here 0.2 s) separately we sum up the salient activation and build the sequence seen on the right. We test if the slope of a linear regression through the resulting points is significantly different from zero.

B Appendix for auditory stimulation and deep learning predict awakening from coma after cardiac arrest

Florence M. Aellen[1,2], Sigurd L. Alnes[1,2], Fabian Loosli[1], Andrea O. Rossetti[3], Frédéric Zubler[4], Marzia De Lucia[5], Athina Tzovara[1,2,6,7]

[1] Institute of Computer Science, University of Bern, Bern, Switzerland
[2] Zentrum für Experimentelle Neurologie, Department of Neurology, Inselspital, Bern University Hospital, University of Bern, Bern, Switzerland
[3] Neurology Service, Department of Clinical Neurosciences, Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland
[4] Sleep-Wake-Epilepsy-Center, Department of Neurology, Inselspital, Bern University Hospital, University of Bern, Bern, Switzerland
[5] Laboratory for Research in Neuroimaging (LREN), Department of Clinical Neurosciences, Centre Hospitalier Universitaire Vaudois (CHUV), Lausanne, Switzerland
[6] Sleep Wake Epilepsy Center—NeuroTec, Department of Neurology, Inselspital, Bern University Hospital, University of Bern, Bern, Switzerland
[7] Helen Wills Neuroscience Institute, University of California Berkeley, Berkeley, CA, USA

Aellen, F. M., Alnes, S. L., Loosli, F., Rossetti, A. O., Zubler, F., De Lucia, M., and Tzovara, A. (2023). Auditory stimulation and deep learning predict awakening from coma after cardiac arrest. *Brain.* 146(2):778–788. https://doi.org/10.1093/brain/awac340

B.1 Evaluation of outcome prediction results for patients in a 'gray zone'

In the results presented in the main manuscript we imitate as closely as possible a 'real life' situation where a CNN would be used in a clinical practice. One could envision training the network with all available data, and then using it to predict outcome in new patients, as they arrive in the intensive care unit, without setting explicit ratios of determinate vs.

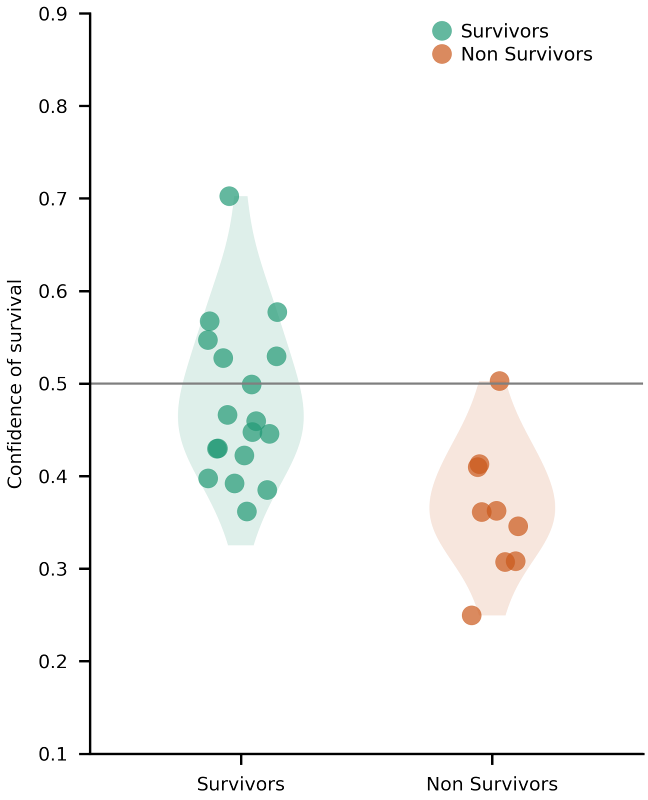


Figure B.1: Confidence of survival assigned by the network for patients in the test set for the median network. Confidence scores of survival for the test set, for a network, where the test set only contained patients in an uncertain state.

'gray zone' patients in the train or test sets. The main manuscript reports aggregate results across train/validation/test sets (Figure 3.3), as the distribution of the confidence scores of the network can still be informative about whether 'gray zone' patients are perceived as particular cases or outliers by the network. Here, we perform control analyses to evaluate systematically the generalizability of the network on 'gray zone' patients:

1. First, instead of randomly splitting patients to train/test/validation, we now curate this split, so that some of the 'gray zone' patients will be part of the 10-fold train/validation datasets, and a fixed (but high) number of 'gray zone' patients will be part of the test set.

2. Second, we trained one network using exclusively patients with determinate outcomes for train/validation, and kept all 'gray zone' patients as our test set, to evaluate generalization of results.

B.2 Training neural networks with a curated patient split resulting in a test set containing only 'gray zone' patients

Here, we trained a neural network where the test set only contained patients from the 'gray zone'. 27 'gray zone' patients were randomly selected for the test set, and the remaining 21 'gray zone' patients were randomly split between the train and validation sets. This resulted in a train set of 80 patients, a validation set of 27 patients and a test set of 27, exclusively 'gray zone' patients. This split was repeated in a cross validation, and contained the same patient numbers per set as the network that is reported in the main text. With this approach, on the test set of 'gray zone' patients, we obtained an AUC score of 0.64 ± 0.01, a PPV of 0.86 ± 0.02 and NPV of 0.43 ± 0.01 (Table B.1 and Figure B.1 for the median fold). These results are very close to the values obtained with the networks presented originally (AUC: 0.70 ± 0.04, PPV: 0.83 ± 0.03 and NPV: 0.57 ± 0.04), and suggest an unbiased and robust outcome prediction on 'gray zone' patients, which were previously unseen by the network.

Table B.1: Scores reached on the networks, where the test set contained only patients in an uncertain state. First, we report the mean ± standard error over all ten trained folds, as well as the performance of the median fold. We also report the AUC, PPV and NPV, with respect to survival, of the train, validation and test sets, for all patients and separately for the sub-cohorts of patients treated with hypothermia and normothermia.

	AUC-score			PPV-score			NPV-score		
	Train	Validation	Test	Train	Validation	Test	Train	Validation	Test
Mean over 10 folds									
All	0.81±0.01	0.77±0.02	0.64±0.01	0.90±0.01	0.90±0.03	0.86±0.02	0.71±0.01	0.67±0.03	0.43±0.01
Hypothermia	0.83±0.01	0.78±0.03	0.60±0.03	0.93±0.01	0.93±0.03	0.79±0.04	0.67±0.02	0.63±0.03	0.41±0.01
Normothermia	0.77±0.02	0.74±0.04	0.72±0.02	0.82±0.03	0.71±0.11	0.96±0.03	0.71±0.02	0.67±0.04	0.43±0.04
Best fold									
All	0.82	0.77	0.61	0.94	1.00	0.86	0.70	0.63	0.40
Hypothermia	0.83	0.75	0.51	0.96	1.00	0.67	0.62	0.62	0.36
Normothermia	0.79	0.80	0.79	0.91	1.00	1.00	0.76	0.67	0.50

B.3 Training of neural networks with all 'gray zone' patients in the test set

Second, we trained one network, where the test set contained all patients in the 'gray zone' (N=48) and the train and validation set contained the rest of the patients (N=86). In this case, we obtained an AUC score of 0.67, PPV of 0.85 and NPV of 0.46 on the

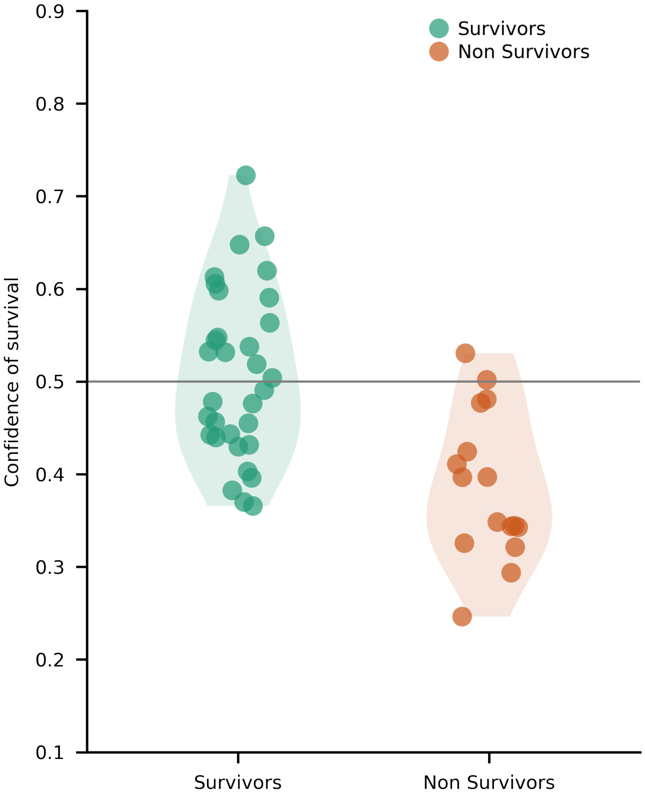


Figure B.2: Confidence of survival assigned by the network for patients in the test set i.e. all patients in the 'gray zone'. Confidence scores of survival for the test set, i.e. all patients in an uncertain state.

test set. The PPV and AUC are lower than the ones obtained with curating the splits of train/validation/test sets, but are well in line with the main results of our manuscript, that our approach is primarily sensitive to predicting survival.

One important caveat of this second approach is that we now have a smaller train set of only 64 patients, compared to the original 80 patients, which can result in less accurate training of the network and therefore less strong outcome prediction. Importantly, this new network was trained without any of the 'gray zone' patients and therefore performs, as expected, worse than a network which was trained with at least some 'gray zone' patients.

For this approach we trained one neural network, with a train set of 64 patients, validation set of 22 patients and test set of 48 patients. We reached the following scores:

Table B.2: Scores reached on the network, where all patients with an uncertain outcome were assigned to the test set. As here only one fold was trained we report the AUC, PPV and NPV scores for the train, validation and test sets.

AUC-score			PPV-score			NPV-score		
Train	Validation	Test	Train	Validation	Test	Train	Validation	Test
0.85	0.82	0.67	0.93	0.83	0.85	0.77	0.80	0.46

B.4 Network performance on different types of auditory stimulations

For all the analysis in the main manuscript we only considered data where patients were exposed to standard or duration deviant sounds. However, for the original paradigm data was also recorded for location and pitch deviant sounds. Here we evaluated the performance of the neural networks on different types of auditory stimulations. We found in our dataset the following number of trials per patient and stimulation type:

Table B.3: Mean number of trials per patient and per stimulation type. The mean ± standard error number of trials for each patient in the dataset.

Standard	Duration	Location	Pitch
204.76±8.66	142.46±1.42	141.84±1.48	141.40±1.48

For all patients in the test we separately test the performance of the network on trials where patients were exposed to standard and duration deviant sounds. And extended the analysis to the previously unseen location and pitch deviant sounds. We found the following performances per stimulation type:

Table B.4: Mean AUC score per stimulation type. The mean ± standard error of the AUC scores per type of auditory stimulation for all patients in the test set. The neural network was only trained on the first two sounds (standard and duration) , the second set of sounds (location and pitch) were only used for this analysis.

Standard	Duration	Location	Pitch
0.698±0.035	0.701±0.035	0.702±0.037	0.709±0.038

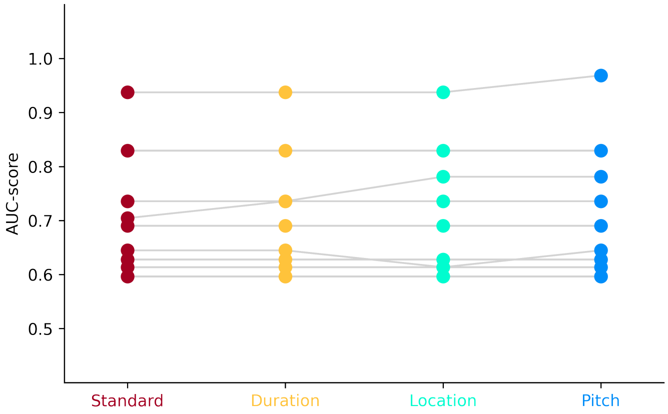


Figure B.3: Comparison of performance per stimulation type. Each dot shows the AUC score of one network on the test set. We show the scores for all ten folds, corresponding folds connected with a gray line.

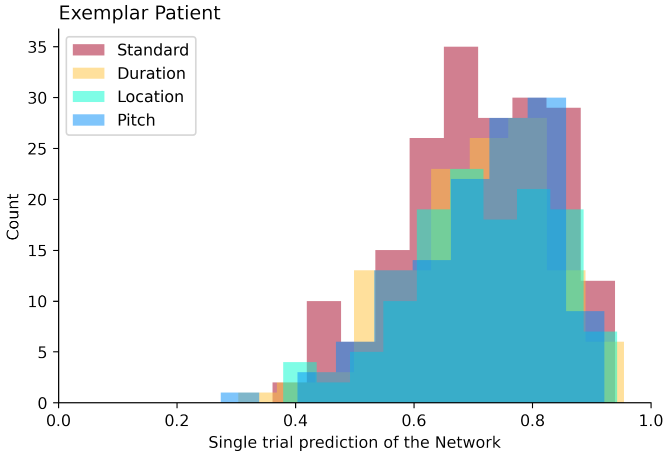


Figure B.4: Exemplar distribution of the predictions per trial for one patient. The network predicts per single trial a value between 0 (non survivor) and 1 (survivor). Here a distribution of these predictions are shown for an exemplar patient, correctly predicted by the network as a survivor.

C Appendix for disentangling the complex landscape of sleep-wake disorders with data-driven phenotyping: A study of the Bernese Center

Florence M. Aellen[1,2], Julia Van der Meer[3], Anelia Dietmann[3], Markus Schmidt[2], Claudio L. A. Bassetti[2], Athina Tzovara[1,2,4]

[1] Institute of Computer Science, University of Bern, Bern, Switzerland
[2] Center for Experimental Neurology, Department of Neurology, Inselspital, Bern University Hospital and University of Bern, Bern, Switzerland
[3] Department of Neurology, Inselspital, Bern University Hospital and University of Bern, Bern, Switzerland
[4] Sleep Wake Epilepsy Center - NeuroTec, Department of Neurology, Inselspital, Bern University Hospital, University of Bern, Bern, Switzerland

Aellen, F. M., Van der Meer, J., Dietmann, A., Schmidt, M., Bassetti, C. L. A., and Tzovara, A. (in review). Disentangling the complex landscape of sleep-wake disorders with data-driven phenotyping: A study of the Bernese Center.

C.1 Processing pipeline

The first step in our processing pipeline includes data collection and marker extraction (Figure 4.1). The extracted markers were in a second step curated. Evident mistakes during manual data collection, such as having letters in fields requiring numbers, were corrected by setting these data as missing. The third step involved selecting good quality markers in a data driven way, such that no more than 50% of the values of a given marker across all patients were missing. Moreover, markers were excluded if more than 90% of the patients had the same value for this marker, as this would indicate that the marker was only collected for a subset of the patients and a standard (null) value was kept for the rest. As an example, patients that indicated the presence of cataplexy were asked if it manifested in the head, chin, etc. For patients that did not report cataplexy, the questions for the body parts were each set to “No”/zero. To this set of markers with high data quality, a set of clinically relevant markers was added (Appendix, Section C.3.2).

All markers were corrected to match meaningful ranges, such as a reported age less than zero and values out of range were set to missing (a list of all ranges is reported in: Appendix, Sections C.3.1 and C.3.2). After this correction patients with more than 1/3 of markers missing were excluded from further analyses. The fourth step of the pipeline involved normalizing all markers to the interval [0,1], as it is common practice prior to using machine learning algorithms. In a fifth step, all missing values were imputed using a Random Forest imputation (Epsilon-machine, 2022). The final step of the pipeline was an unsupervised clustering.

For each of the selected cohorts, the selection of patients was restarted from the full Bern Sleep Registry from the top (Figure 4.1B, left panel). This approach was chosen over pre-selecting one single cohort and then breaking it down to sub-cohorts, to maximize the number of patients that could be included in each of the analyses, as the desired markers for each sub-cohort were slightly different, and were chosen for each cohort considering clinical relevance and markers with non missing values. Patient selection for these cohorts is illustrated in Figure 4.1B.

C.2 Description of disorders in full cohort of SWD with well characterized patients with a single SWD

The full cohort of SWDs contained a lot of patients with heterogeneous disorders, such as insomnia, containing patients with chronic insomnia, short term insomnia and other insomnia disorders. In the control analysis of well characterized patients with single SWD we focused on single sub-diagnoses. For insomnia, we selected “chronic insomnia disorder”, for parasomnia “NREM-related parasomnia” and for movement disorders “restless leg syndrome”. The groups “other sleep-related breathing disorder”, “narcolepsy type 2 & idiopathic hypersomnia” and “other central disorders of hypersomnolence” were excluded, because of either limited number of patients (N <50), or because of comorbidities with other non-SWDs, such as psychiatric disorders or medication. Patients with “isolated symptoms and normal variants” were also excluded, as they only describe single symptoms or variants.

C.3 Markers

C.3.1 Markers for clustering CDH

Table C.1 describes all markers used for clustering for CDHs, as well as the ranges used for correcting these markers. The ranges were used to correct for errors (e.g. age less than 0 years old), which might have occurred while entering retrospective values into the database. If values were out of range they were set to ‘missing’. For the hypersomnolence cohort, the included markers were selected based on clinical expertise.

Table C.1: Markers used for the clustering of the hypersomnolence cohort and valid interval.

Clinical Markers	Min	Max
Age	0	100
Gender (F = 2)	1	2
Waist-Hip Ratio	0.17	4.75
ESS	0	24
Total Sleep Time	0	850
Sleep Latency	0	300
REM Latency	0	610
Sleep Efficiency	0	100
Awake Index	0	100
Stage Transitions	0	500
Sleep Cycles	0	15
Apnea-Hypopnea Index	0	200
Desaturation Index	0	200
Heart Rate	0	200
PLMS Index	0	260
Mean REM Latency	0	35
Cataplexies	0	1
Naps	0	1
Sleep Paralysis	0	1
Hallucination	0	1
Sleep Drunkenness	0	1

C.3.2 Markers for clustering sleep disorders in the full cohort

Table C.2 highlights all markers used for clustering the full cohort of SWDs. For each marker we provide the min-max range used for correcting possible mistakes while entering the retrospective data into the database (e.g. age <0) and whether the marker was included because of clinical criteria (i.e. relevance for diagnosing a specific SWD), or because of data-driven criteria (i.e. presence in more than 50% of patients and no more than 90% of patients with identical values). If any of the available values were out of range they were set to missing.

Table C.2: Clinical markers used for clustering the full cohort of SWDs, their minimal and maximal values and the selection type (either data-driven, clinical or both).

Clinical Markers	Min	Max	Selection Type
Age	0	100	data-driven & clinical
Gender (F = 2)	1	2	data-driven & clinical
Race	1	5	data-driven & clinical
BMI	10	60	data-driven & clinical
Waist-Hip Ratio	0.17	4.75	data-driven & clinical
ESS	0	24	data-driven & clinical
FSS	0	9	clinical
Estimated Sleep Time	0	100	clinical
Total Sleep Time	0	850	data-driven & clinical
Sleep Latency	0	300	data-driven & clinical
REM Latency	0	610	data-driven & clinical
Sleep Efficiency	0	100	clinical
Awake Index	0	100	data-driven & clinical
Stage Transitions	0	500	data-driven & clinical
Sleep Cycles	0	15	data-driven & clinical
% Movement (30s epochs)	0	100	data-driven
% Wake	0	100	data-driven & clinical
% Stage 1	0	100	data-driven
% Stage 2	0	100	data-driven
% Stage 3	0	100	data-driven
% Stage 4	0	100	data-driven
% REM	0	100	data-driven
Apnea-Hypopnea Index	0	200	data-driven & clinical
Apnea Index	0	120	data-driven & clinical
Hypopnea Index	0	150	data-driven & clinical
Apnea-Hypopnea Duration	0	200	clinical
Mean Apnea Duration	0	100	data-driven
Maximum Apnea Duration	0	300	data-driven & clinical
Duration O2 < 90% (min)	0	600	data-driven & clinical
O2 < 80% (min)	0	600	data-driven
O2% Mean	0	100	data-driven
O2% Minimum	0	100	data-driven
Desaturation Index	0	200	data-driven & clinical
Heart Rate	0	200	data-driven & clinical
Movement Index	0	1200	data-driven & clinical
Periodic Movements	0	500	data-driven & clinical
PLMS Index	0	260	data-driven & clinical
Mean REM Latency	0	35	clinical
Cataplexies	0	1	clinical
Naps	0	1	clinical
Sleep Paralysis	0	1	clinical
Hallucination	0	1	clinical
Sleep Drunkenness	0	1	clinical
Headache	0	1	clinical
Psychiatric Disorders	0	1	data-driven & clinical
Adipositas	0	1	data-driven & clinical
Diabetes Mellitus	0	1	clinical
Cardial Comorbidities	0	1	data-driven & clinical
Pulmonary Comorbidities	0	1	clinical
Stroke/Intracerebral Hemorrhage	0	1	clinical
Neurodegenerative Disease	0	1	clinical

C.4 Clinical diagnosis of patients

C.4.1 Sub-diagnosis for the full cohort

Some of the diagnoses that were included in the full cohort of SWDs were a combination of multiple subcategories. For example, “insomnia” may include “chronic insomnia disorder”, “short-term insomnia disorder” and “other insomnia disorder”. Figure C.1 shows the breakdown of these sub-diagnoses (in shaded colors) and their prevalence within the main diagnosis (one circle per main diagnosis).

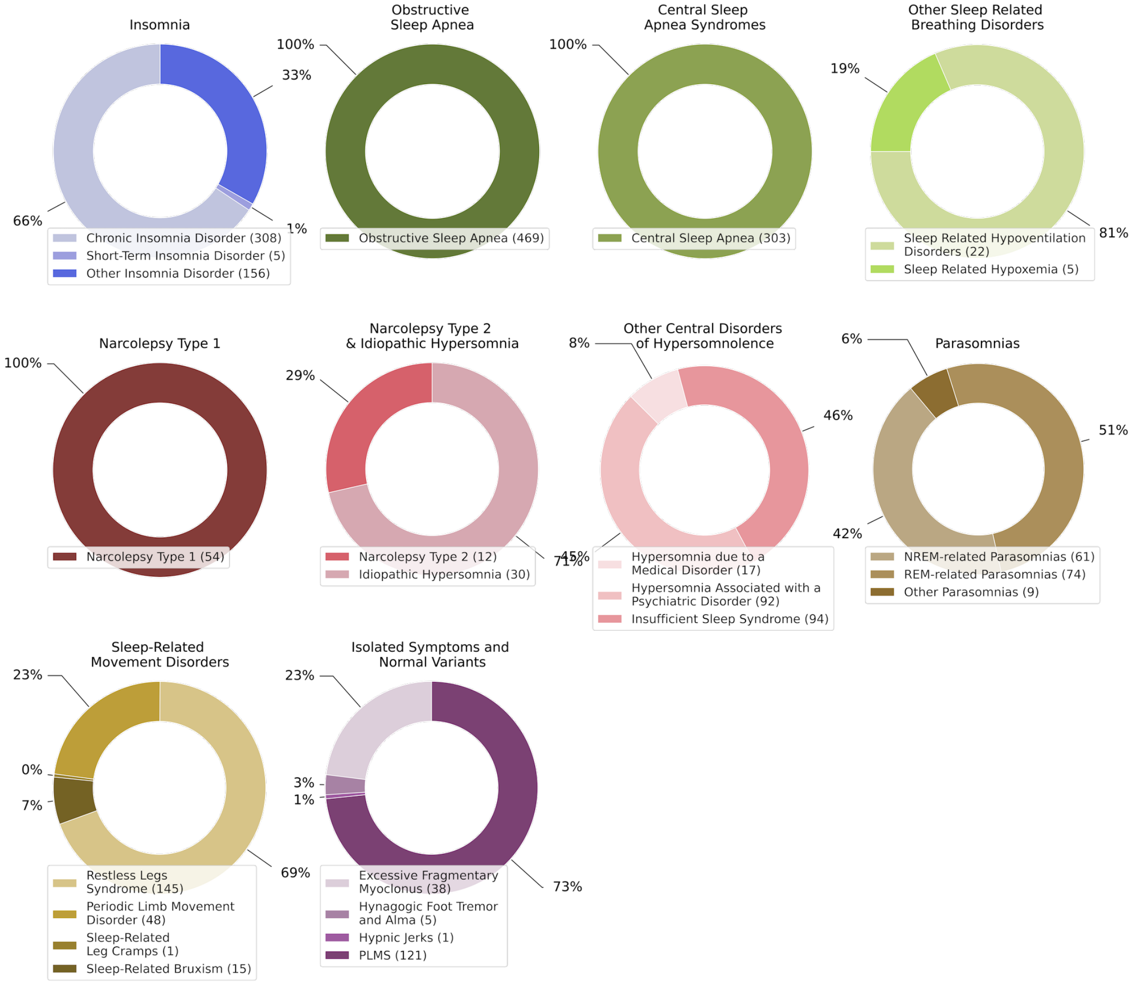
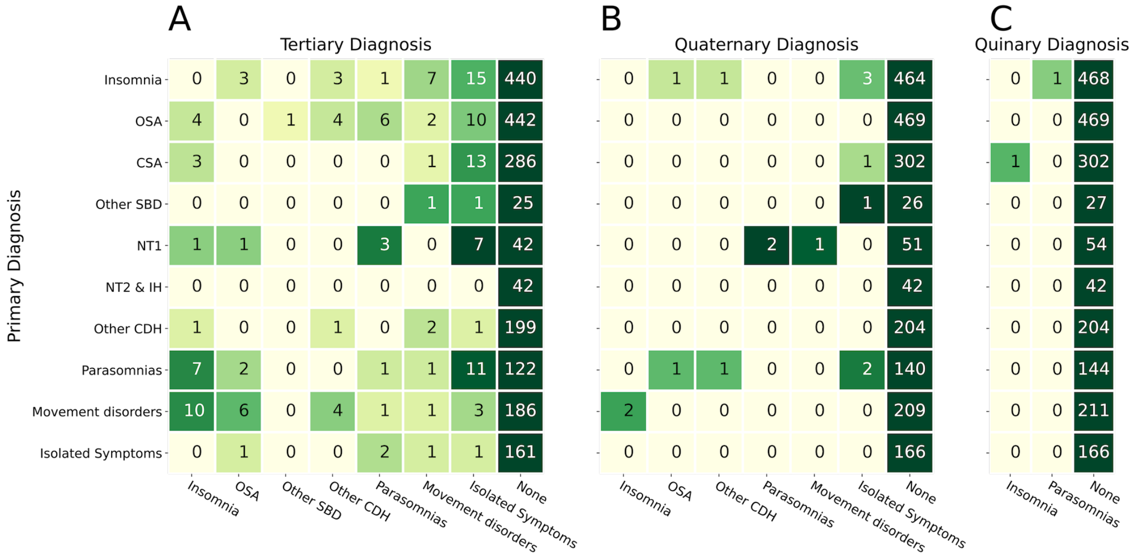


Figure C.1: Distribution of sub-diagnoses within the full cohort. Each main diagnosis is depicted in a separate circle, while the colored shades highlight the sub-diagnoses that were grouped together in our main analyses (Figure 4.2 and 4.4.)

C.4.2 Tertiary, quaternary and quinary diagnosis

Each patient could be diagnosed with up to five different SWDs. The main text Section Figure 4.2 shows the distribution of the secondary diagnosis for patients of the full cohort, in the form of a confusion matrix. Figure C.2 additionally displays confusion matrices for the tertiary, quaternary and quinary diagnoses (panels A-C).



Primary Diagnosis	Insomnia	OSA	Other SBD	Other CDH	Parasomnias	Movement disorders	Isolated Symptoms	None
Insomnia	0	3	0	3	1	7	15	440
OSA	4	0	1	4	6	2	10	442
CSA	3	0	0	0	0	1	13	286
Other SBD	0	0	0	0	0	1	1	25
NT1	1	1	0	0	3	0	7	42
NT2 & IH	0	0	0	0	0	0	0	42
Other CDH	1	0	0	1	0	2	1	199
Parasomnias	7	2	0	0	1	1	11	122
Movement disorders	10	6	0	4	1	1	3	186
Isolated Symptoms	0	1	0	0	2	1	1	161

Primary Diagnosis	Insomnia	OSA	Other CDH	Parasomnias	Movement disorders	Isolated Symptoms	None
Insomnia	0	1	1	0	0	3	464
OSA	0	0	0	0	0	0	469
CSA	0	0	0	0	0	1	302
Other SBD	0	0	0	0	0	1	26
NT1	0	0	0	2	1	0	51
NT2 & IH	0	0	0	0	0	0	42
Other CDH	0	0	0	0	0	0	204
Parasomnias	0	1	1	0	0	2	140
Movement disorders	2	0	0	0	0	0	209
Isolated Symptoms	0	0	0	0	0	0	166

Primary Diagnosis	Insomnia	Parasomnias	None
Insomnia	0	1	468
OSA	0	0	469
CSA	1	0	302
Other SBD	0	0	27
NT1	0	0	54
NT2 & IH	0	0	42
Other CDH	0	0	204
Parasomnias	0	0	144
Movement disorders	0	0	211
Isolated Symptoms	0	0	166

Figure C.2: Primary versus Tertiary (A), Quaternary (B) and Quinary (C) diagnosis for the full cohort of SWDs used in the main text.

C.5 Technical control analysis for the full cohort of sleep disorders

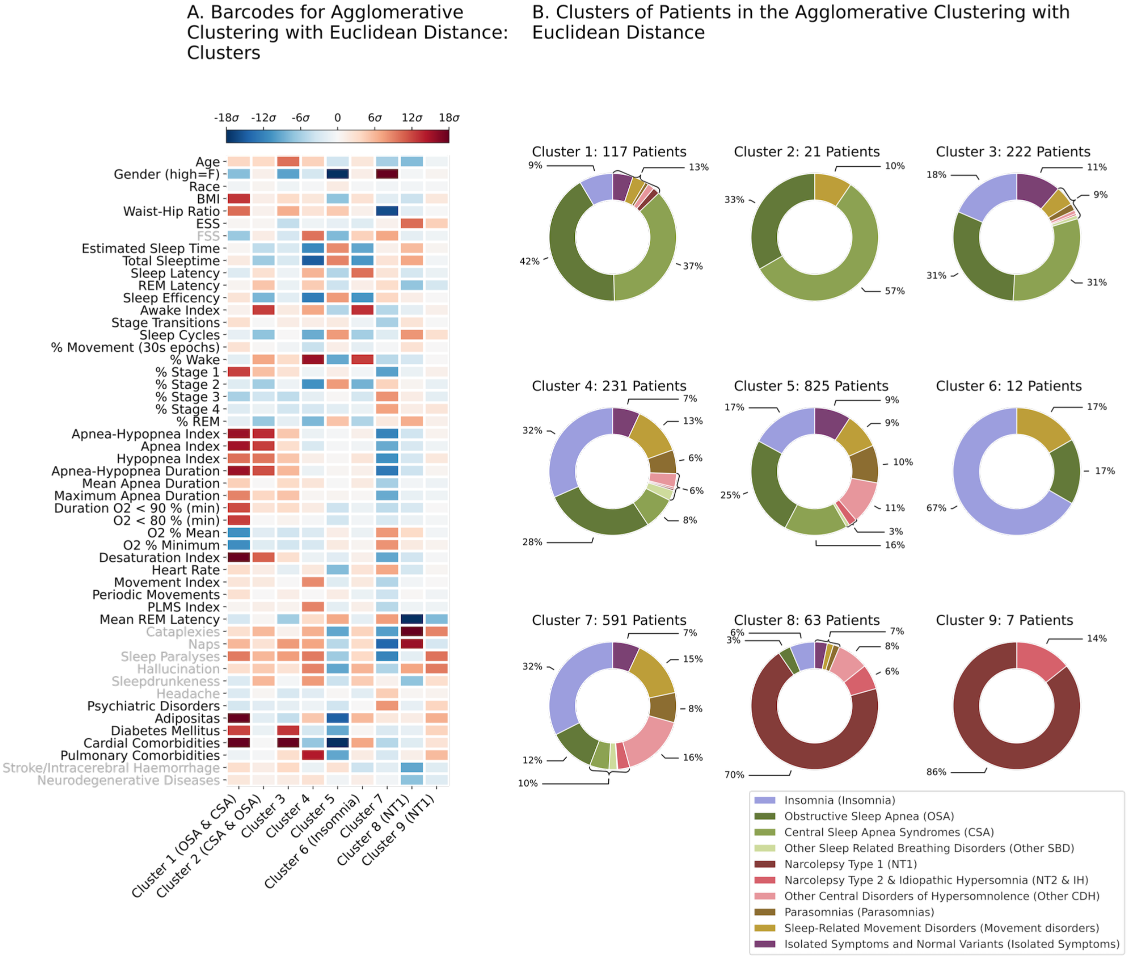
As our main clustering analysis includes multiple steps and choices (i.e. clustering algorithm, imputation technique, distance metric, marker selection), we provide a multitude of technical control analyses. With these analyses we examine whether alternative choices of analysis metrics or steps would better disentangle the SWDs compared to the pipeline that we present in the main text. Table C.3 shows a summary of the technical controls that were performed, with the first column indicating the subsection where a given control analysis can be found.

Table C.3: Summary of technical controls that were performed for the full spectrum of SWDs, and information about the section of the main manuscript where these controls relate to.

Control Analysis	Imputation	Clustering Algorithm	Metric	Marker selection
Section C.5.1	Random Forest	Agglomerative	Euclidean	combined
Section C.5.2	Random Forest	Agglomerative	Gower's	combined
Section C.5.3	None	Agglomerative	Gower's	combined
Section C.5.4	Mean	KMeans	Euclidean	combined
Section C.5.5	Median	KMeans	Euclidean	combined
Section C.5.6	Random Forest	KMeans	Euclidean	data-driven
Section C.5.7	Random Forest	KMeans	Euclidean	clinical

C.5.1 Agglomerative clustering with euclidean distance

This subsection covers the clustering of the same cohort used for the main manuscript, but done with an agglomerative clustering (implemented by scipy (Virtanen et al., 2020)) and with euclidean distance, same as for the clustering described in the main text (Section 4.3.2). Similar to the main text, which used k-means, with agglomerative clustering we obtained three OSA and CSA clusters (1, 2 & 3), where however the second cluster only contained 21 patients. Cluster 6 was dominated by insomnia patients, but only contained 12 patients. Two NT1 clusters were identified (clusters 8 & 9), where the second only contained 7 patients total. The rest of the clusters were rather mixed, with cluster 5 and 7 containing the vast majority of patients, with 825 and 591 patients respectively.


A. Barcodes for Agglomerative Clustering with Euclidean Distance: Clusters

B. Clusters of Patients in the Agglomerative Clustering with Euclidean Distance

Figure C.3: (A) Barcode and (B) clusters of the agglomerative clustering with euclidean distance on the same cohort selected for the main text.

C.5.2 Agglomerative clustering with Gower's distance

Here also an agglomerative clustering was performed on the original cohort, but with a Gower's distance (Gower, 1971) instead of euclidean. Because all values were imputed, the Gowers distance was equivalent to an L_1 metric. This control analysis showed similar clusters as the one in Figure C.3 which also relied on an agglomerative algorithm. Also in this analysis, there were some clusters with very few patients (2, 3, 5, 8 & 9), and one cluster with surprisingly large amounts of patients (4). There were two OSA and CSA (1 & 2), two insomnia (3 & 6) and two NT1 (8 & 9) clusters. The remaining clusters were mixed.

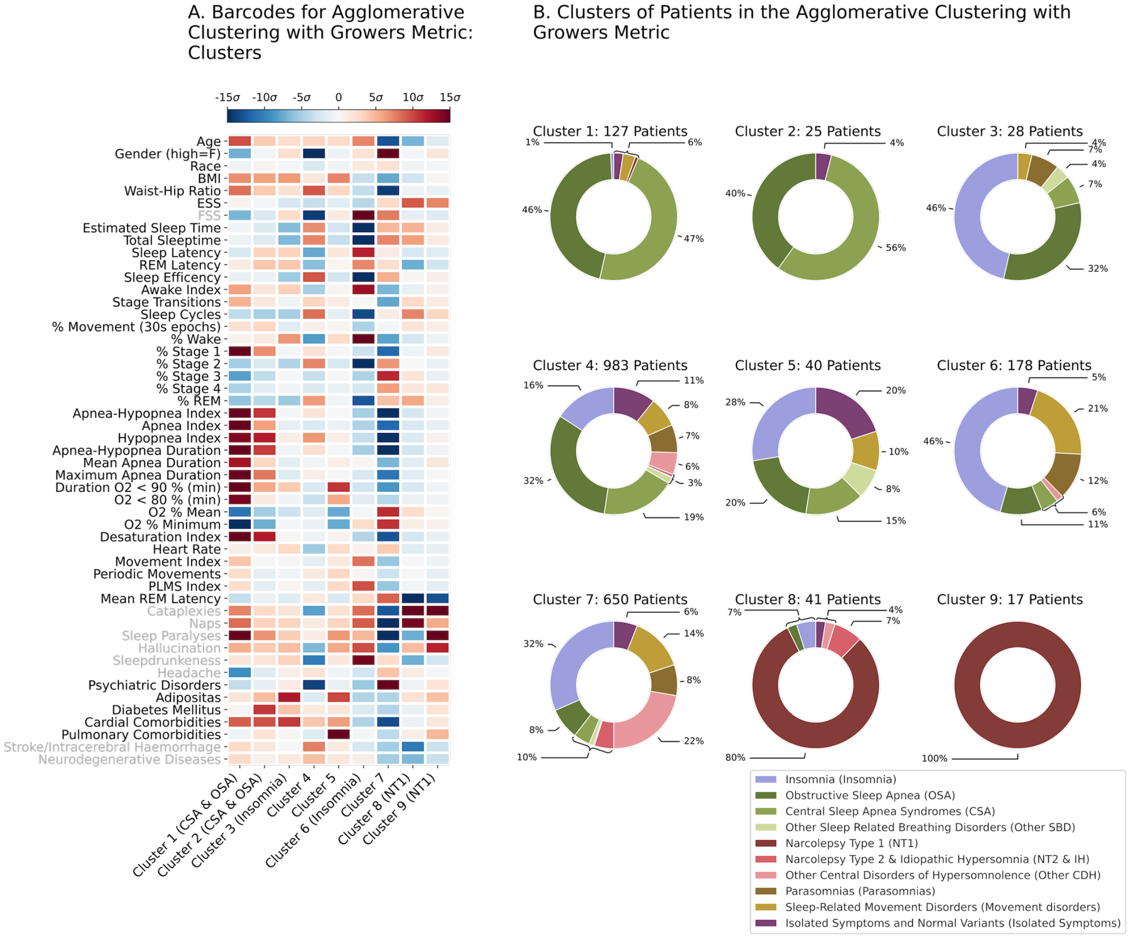
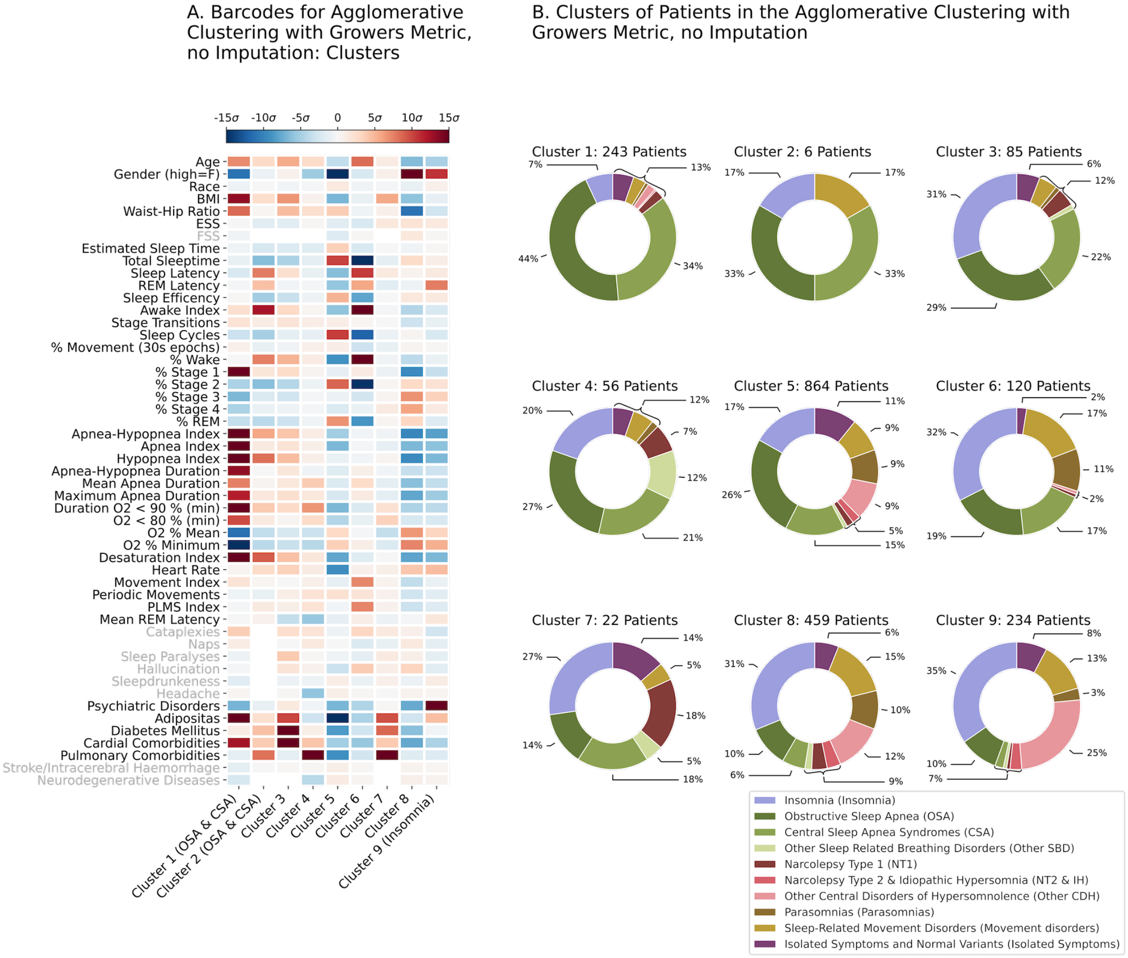


Figure C.4: (A) Barcode and (B) clusters of the agglomerative clustering with Gowers distance on the same cohort selected for the main text.

C.5.3 Agglomerative clustering with Gower's distance and no data imputation

This subsection covers the results of the control analysis where the 5'th step of the pipeline (Figure 4.1) was skipped and no data imputation was performed. An agglomerative clustering with Gower's distance was used for this control analysis. When calculating the distance between two patients any dimension with missing values was ignored and only markers where both patients had a value given were considered. This clustering identified two OSA & CSA (1 & 2) and one insomnia (9) cluster. One cluster (2) contained only a few patients and one cluster contained more than 800 patients (5). No cluster of NT1 patients was detected.


A. Barcodes for Agglomerative Clustering with Growers Metric, no Imputation: Clusters

B. Clusters of Patients in the Agglomerative Clustering with Growers Metric, no Imputation

Figure C.5: (A) Barcode and (B) clusters of the agglomerative clustering with Gowers distance on a cohort, selected the same as for the main text, but no data was imputed.

C.5.4 KMeans clustering with euclidean distance and data imputation with mean

Here the cohort was selected in the same way as for the main text, but the Random Forest data imputation in the fifth step of our pipeline (Figure 4.1) was replaced, by imputing missing values with the mean value per marker. This clustering found one OSA cluster (1), and four clusters with many OSA and CSA patients (2,3,4 & 5), but no single diagnosis reached one third in any of these clusters. Clusters 8 and 9 contained a big amount of patients suffering from insomnia. No cluster with NT1 patients was detected.

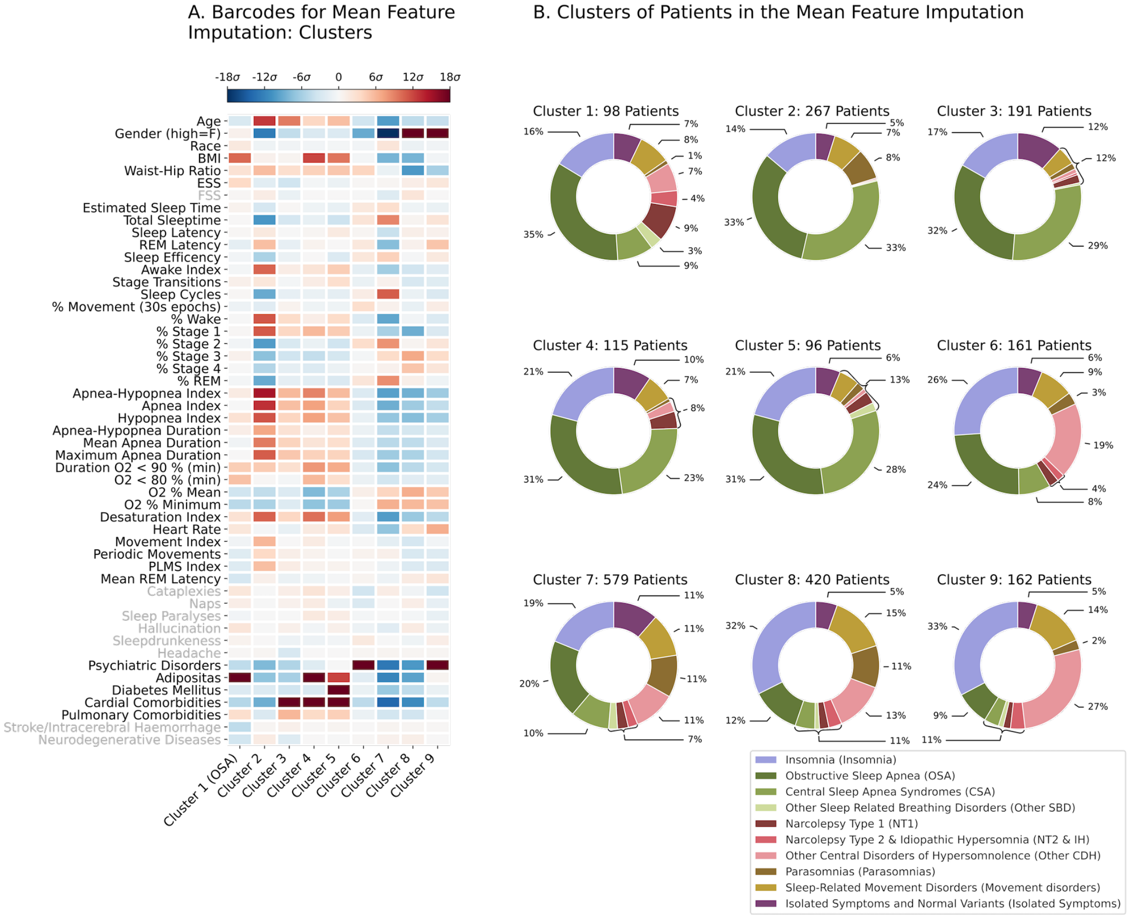


Figure C.6: (A) Barcode and (B) clusters of a clustering on a cohort, selected the same as for the main text, but any missing data was imputed with the mean value per marker.

C.5.5 KMeans clustering with euclidean distance and data imputation with median

For this subsection the cohort selection was based on the same parameters as the one for the main text. However in the 5'th step of the pipeline (Figure 4.1) the imputation was replaced by a median imputation per marker. This clustering found three OSA clusters (1, 2 & 3), where the last also contained many CSA patients. Cluster 4 also contained many CSA and OSA patients. 8 was identified as an insomnia cluster and 9 as a NT1 cluster.

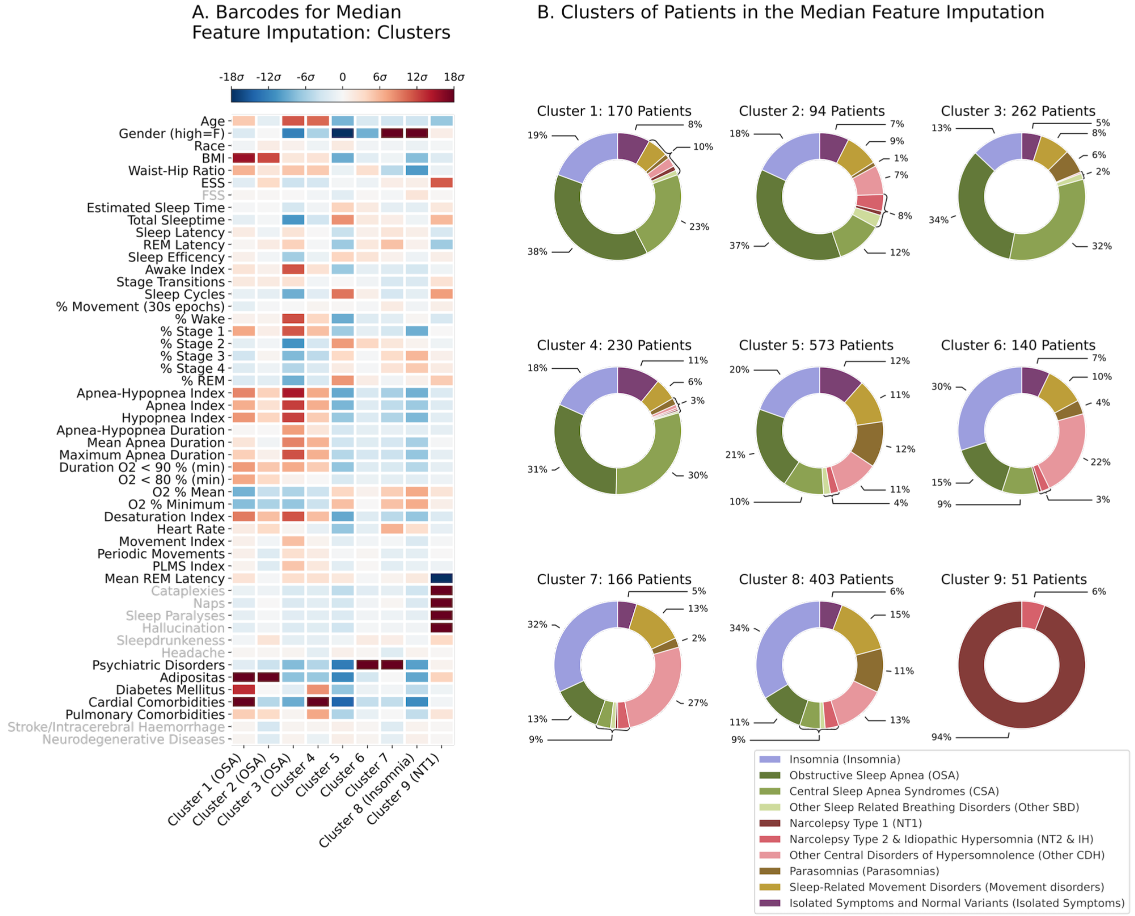


Figure C.7: (A) Barcode and (B) clusters of a clustering on a cohort, selected the same as for the main text, but any missing data was imputed with the median value per marker.

C.5.6 KMeans clustering with euclidean distance and only data-driven marker selection

Here we selected a slightly different sub-cohort of patients, using data-driven marker selection. This was based on markers that had no more than 50% of values missing across all patients and no more than 90% of the values were the same. The cohort contained 36 markers and 2'090 patients The clustering identified two OSA (1 & 2), and one CSA cluster (3), also containing many OSA patients and one OSA and CSA cluster (4), but no single diagnosis reached one third of the patients. Cluster 9 contained more than a third of patients diagnosed with insomnia. No cluster of NT1 patients was found.

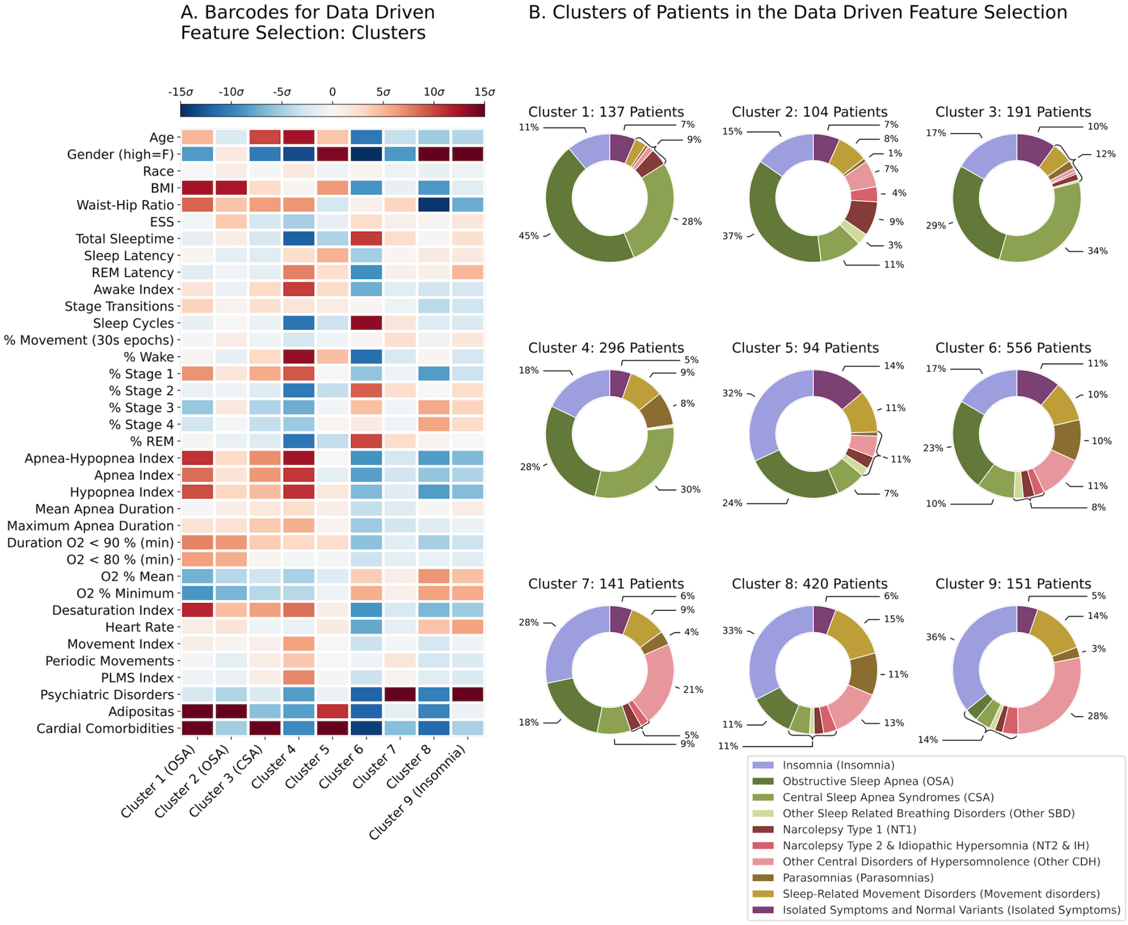


Figure C.8: (A) Barcode and (B) clusters of a clustering on a cohort, where marker selection was only based on a data-driven approach.

C.5.7 KMeans clustering with euclidean distance and data marker selection only clinical

For this subsection the clustering was performed on a cohort where marker selection was based solely on clinical expertise. For this control analysis we worked with 2'136 patients and 41 markers. The clustering found two OSA (1 & 2) and one CSA cluster (3) also containing many OSA patients. There were quite a lot of clusters (4, 6, 7 & 8) with many insomnia patients, but not reaching the threshold of one third of patients. Cluster 9 was found to be a NT1 cluster.

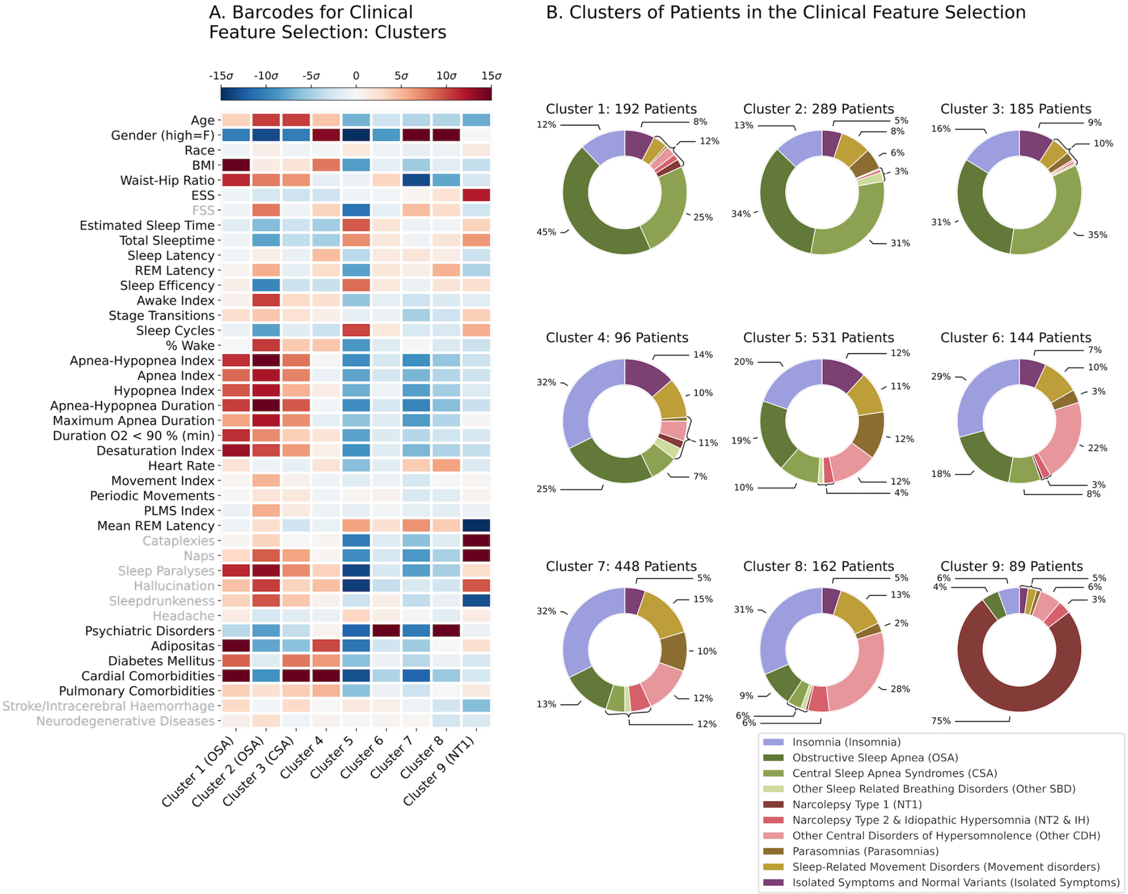


Figure C.9: (A) Barcode and (B) clusters of a clustering on a cohort, where marker selection was only based on clinical expertise.

C.6 Clinical control analysis for full cohort of sleep disorders

Next to the technical controls, focussing on hyperparameters related to the selected algorithms, we also performed four clinical controls, exploring questions arising from a clinical focus. Table C.4 shows all clinical controls, indicating the subsection covering the results found for a clustering with nine clusters, the same number as selected in the main manuscript.

Table C.4: All clinical controls and information in which section the results can be found

Control Analysis	Description
Section C.6.1	Not downsampling obstructive sleep apnea patients
Section C.6.2	Control analysis for narcolepsy type 1 and CDH diagnosis (removing clinical measures cataplexy and REM latency)
Section C.6.3	Control analysis for obstructive sleep apnea diagnosis (removing clinical measure apnea-hypopnea index)
Section C.6.4	Control analysis for PLMS diagnosis (removing clinical measure PLMS Index)

C.6.1 Not downsampling OSA patients

Because patients with OSA make up more than 60% of patients of the full cohort (2'834 out of 4'454 patients), for our main analysis, we downsampled them, because they would otherwise over-dominate every cluster. Here, in a control analysis, the OSA patients were not downsampled and the cohort contained 4'454 patients, out of which 2'834 were diagnosed with OSA. As expected, the clustering found that all nine clusters consisted of more than one third of OSA patients, showing the need to downsample this cohort.

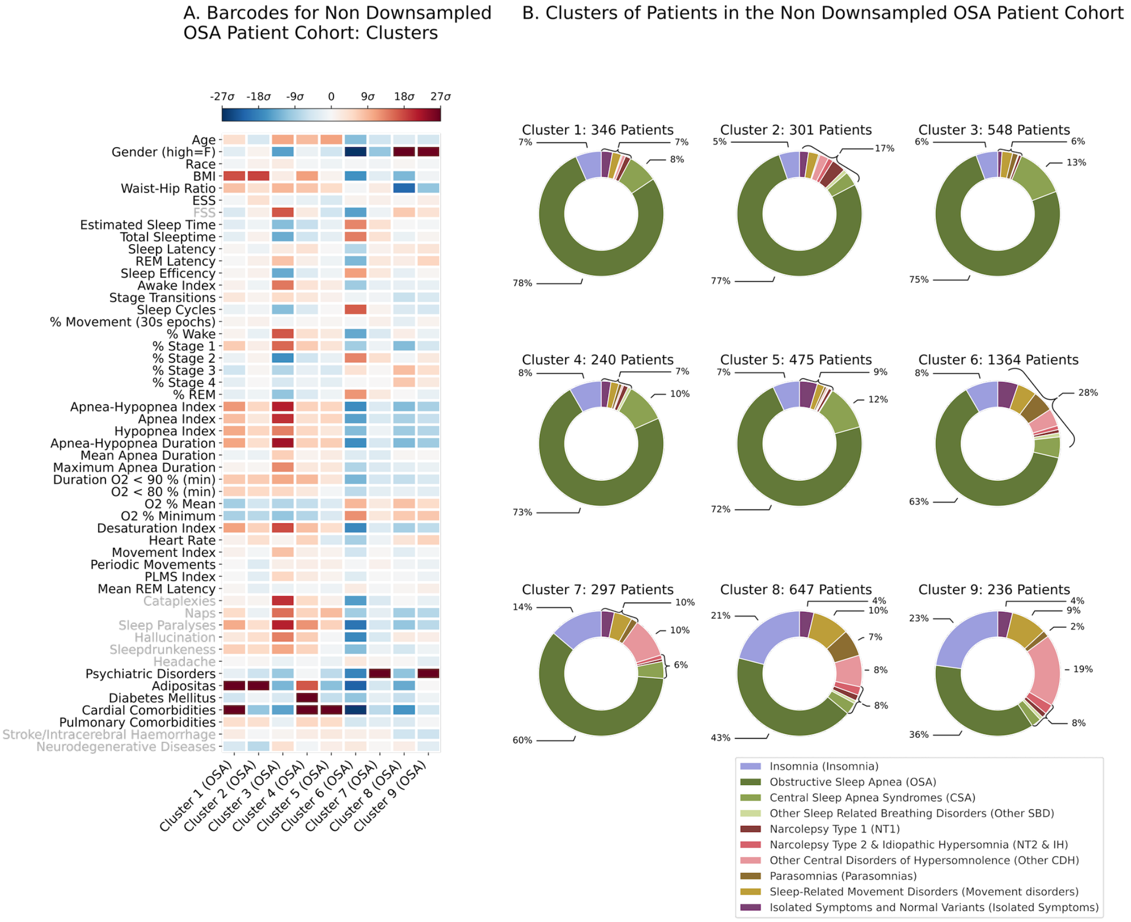


Figure C.10: (A) Barcode and (B) clusters of a clustering on a cohort, where all OSA patients were included.

C.6.2 Control analysis for NT1 and CDH diagnosis

As an additional clinical control, we excluded all markers that were formally used by clinicians to diagnose NT1, NT2&IH and Other CDH. In particular, we excluded the markers cataplexy and REM latency. Similar to our main findings, this clustering found one OSA and CSA cluster (1) and one OSA cluster (2). Clusters 7 & 8 were identified as clusters of insomnia. As expected, when excluding the markers cataplexy and REM latency, no cluster of NT1 was found, highlighting the importance of these clinical criteria in diagnosing NT1.

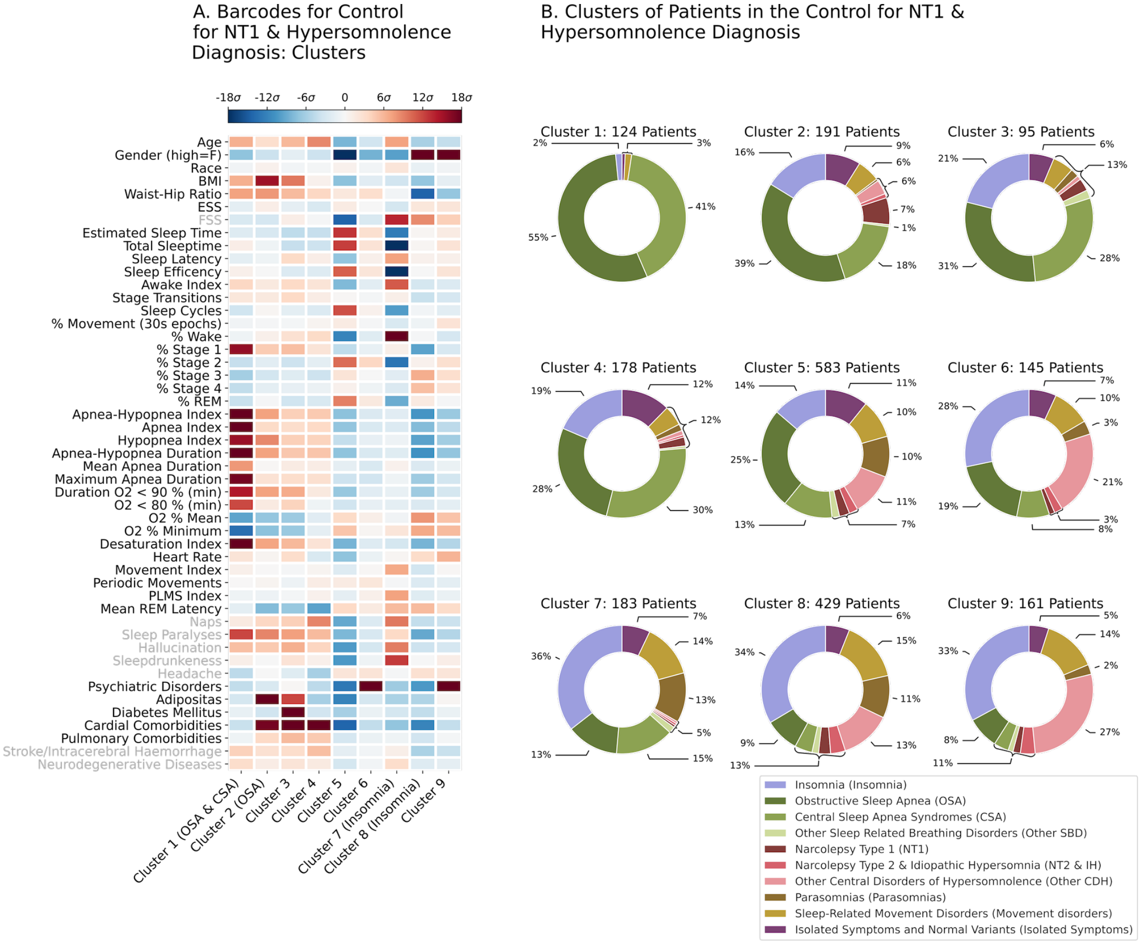


Figure C.11: (A) Barcode and (B) clusters of a clustering on a cohort, where markers used for the diagnosis of NT1, NT2&IH and Other CDH were excluded.

C.6.3 Control analysis for OSA diagnosis

Similar to C.6.2, as an additional clinical control, we excluded the apnea-hypopnea index, which is used to diagnose OSA. Despite this exclusion, our clustering was still able to identify clear clusters of OSA (1, 2 & 3), where the first one also contained CSA. These were predominantly identified via markers related to respiratory functions in sleep, such as apnea index, desaturation index, or O2 levels. Cluster 9 was identified as a cluster of insomnia.

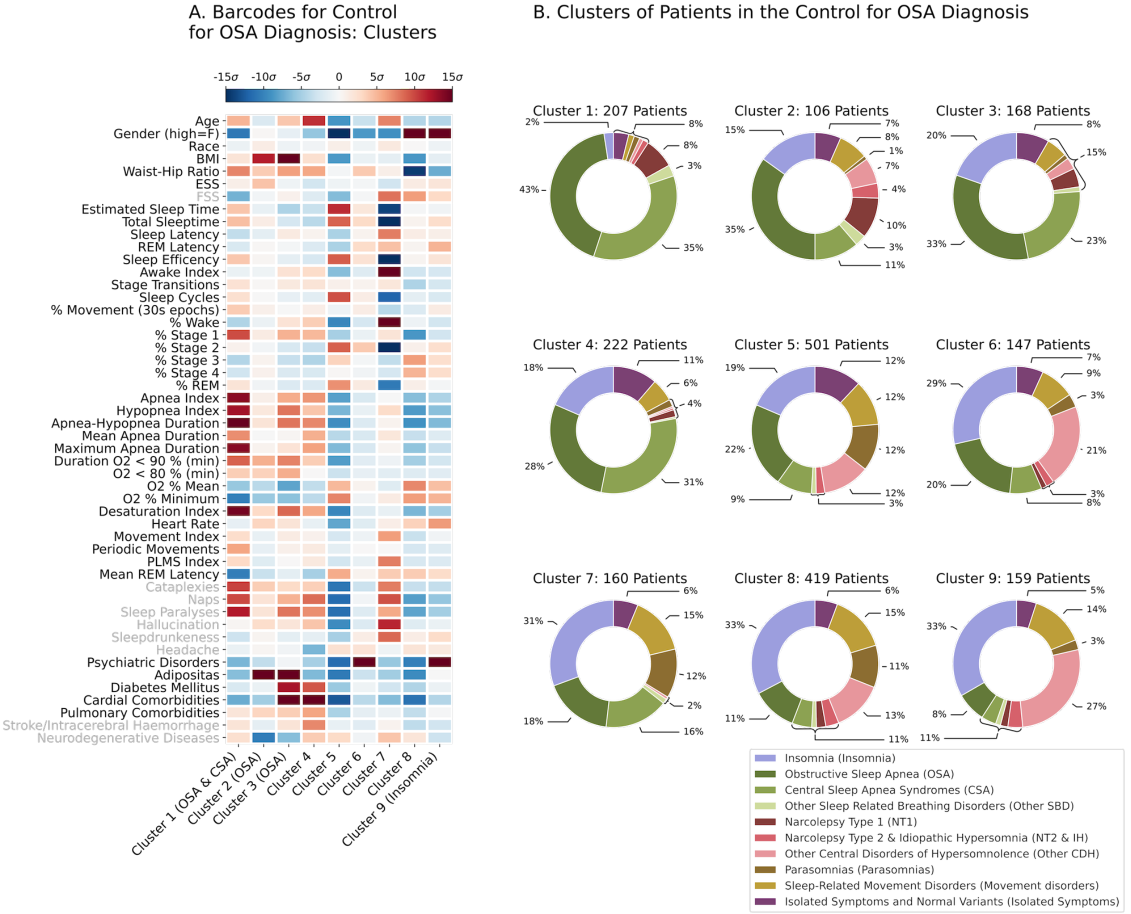


Figure C.12: (A) Barcode and (B) clusters of a clustering on a cohort, where markers used for the diagnosis of OSA were excluded.

C.6.4 Control analysis for PLMS diagnosis

As a final clinical control, we excluded the marker PLMS index, which is typically used to diagnose PLMS. This clustering found one OSA and CSA cluster (1) and two OSA clusters (2 & 3). Cluster 8 was identified as a cluster of insomnia. As in our main analysis, no clear movement disorder cluster was identified.

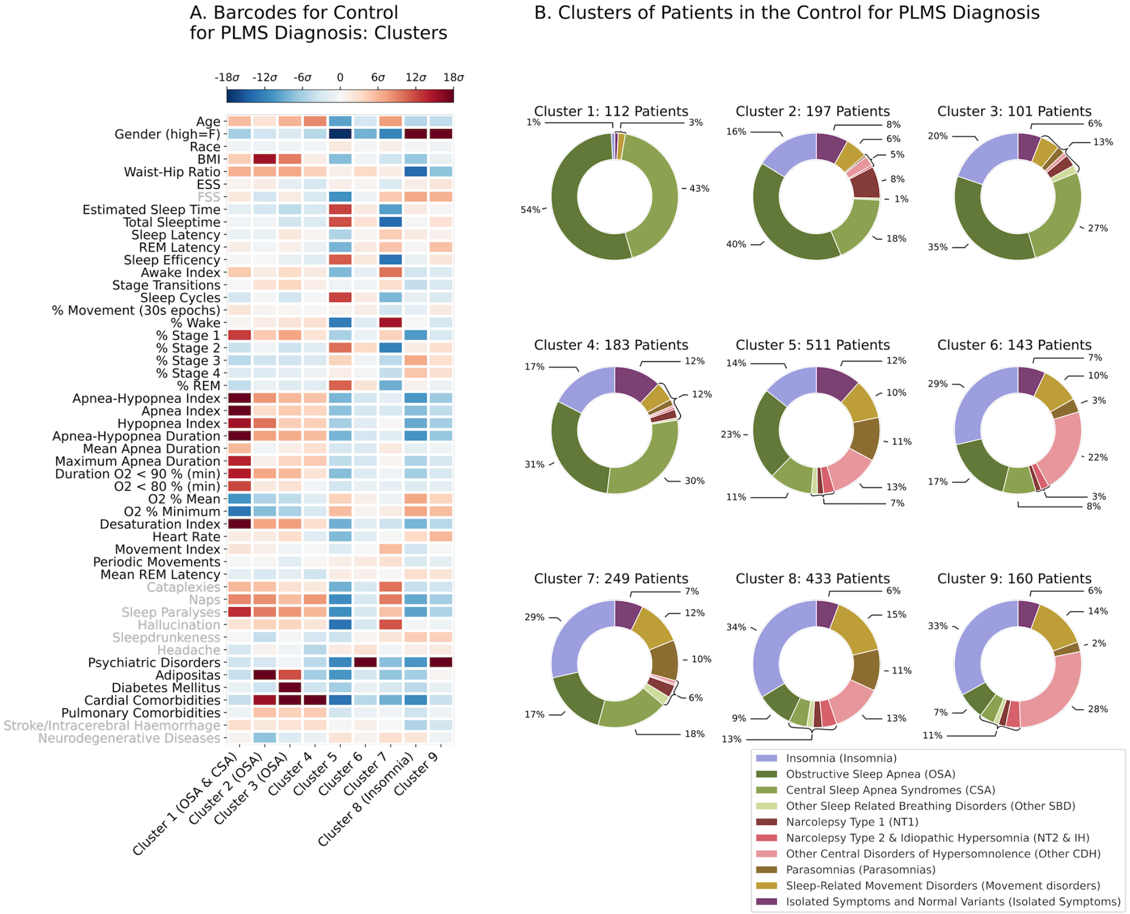


Figure C.13: (A) Barcode and (B) clusters of a clustering on a cohort, where markers used for the diagnosis of PLMS were excluded.

C.7 Selection of number of clusters

This section shows the explained variance and difference in explained variance for the hypersomnolence and full cohort mentioned in the main text of the manuscript. The final number of clusters was selected where the explained variance showed a plateau, i.e. the in including another cluster would improve the clustering only slightly. We confirmed that the remaining metrics were in accordance. For all further analysis in the main text (the well defined cohort) as well as the control analysis in this appendix the same number of clusters as for the full cohort were chosen, as a direct comparison.

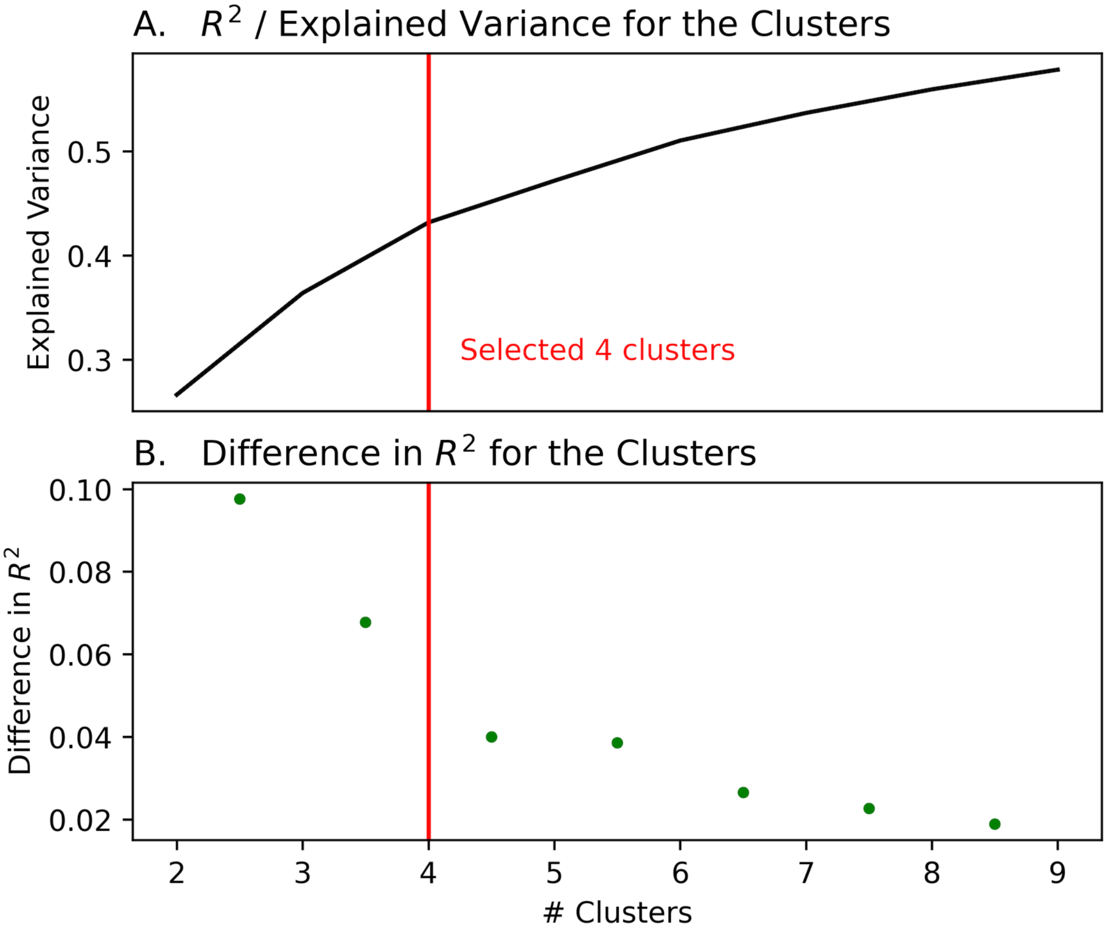


Figure C.14: R^2 distance or explained variance for the clusters of the hypersomnolence cohort (A) explained variance per cluster (B) difference in explained variance per cluster. The clustering showed a slight plateau going from 4 to 5 clusters.

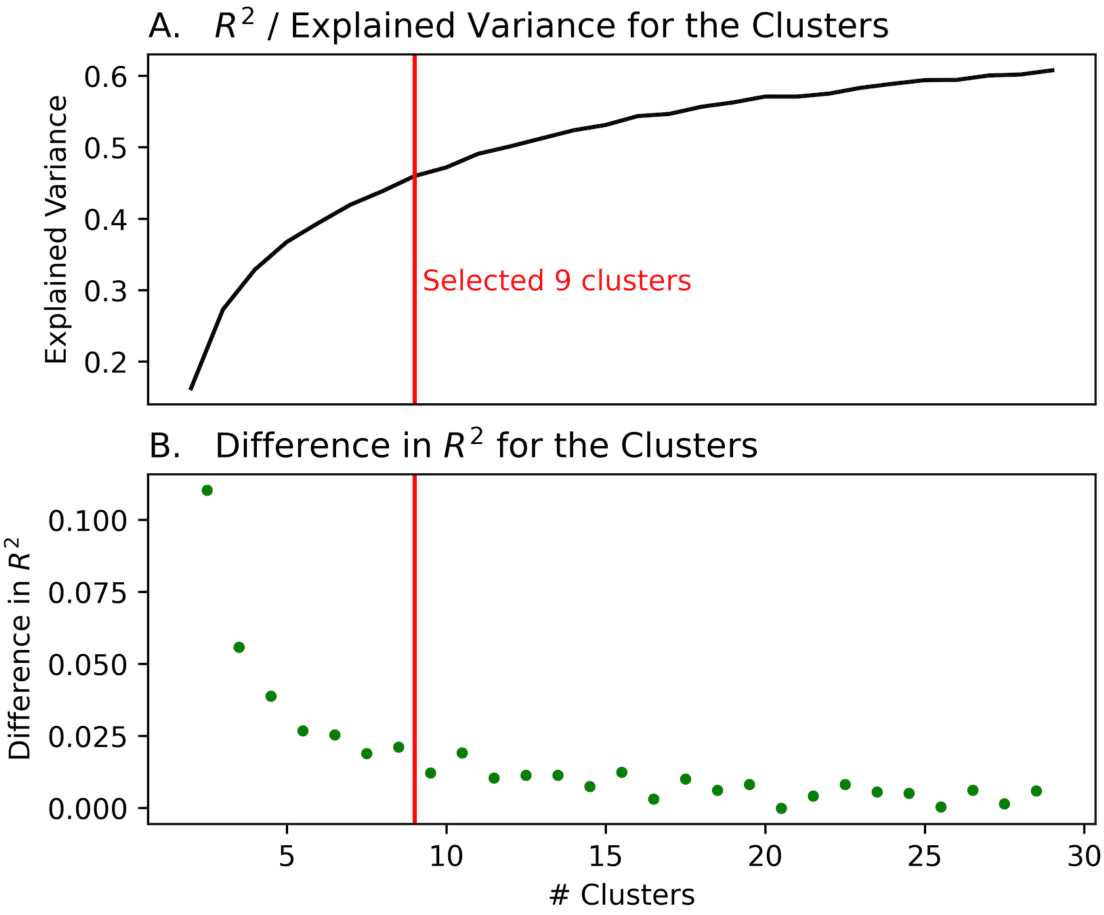


Figure C.15: R^2 distance or explained variance for the clusters of the full cohort of sleep disorders (A) explained variance per cluster (B) difference in explained variance per cluster. The clustering showed a plateau going from 9 to 10 clusters.

C.8 Restless leg syndrome versus parasomnia

The analyses in the main text were not able to distinguish clear clusters for neither restless leg syndrome (RLS) nor parasomnia. These two disorders have clear and distinct symptoms, and their distinction is based on self reported questionnaires. For the present study these questionnaires were only available in a digitized format for five parasomnia and four RLS patients. Because of the extremely low sample size, we did not include them as variables in the main analysis. Here, we plot the responses to these questionnaires by sleep disorder, to highlight that they do indeed differentiate the two diagnoses (Figure C.16). The two questionnaires are:

b_43: *"How often do you have a burning, biting or smarting or tingling sensation in the legs , which forces you to move or rub your legs?"*

b_85d: *"Are there people in your family that have/had these sleep problems: restless, biting or itchy legs?"*

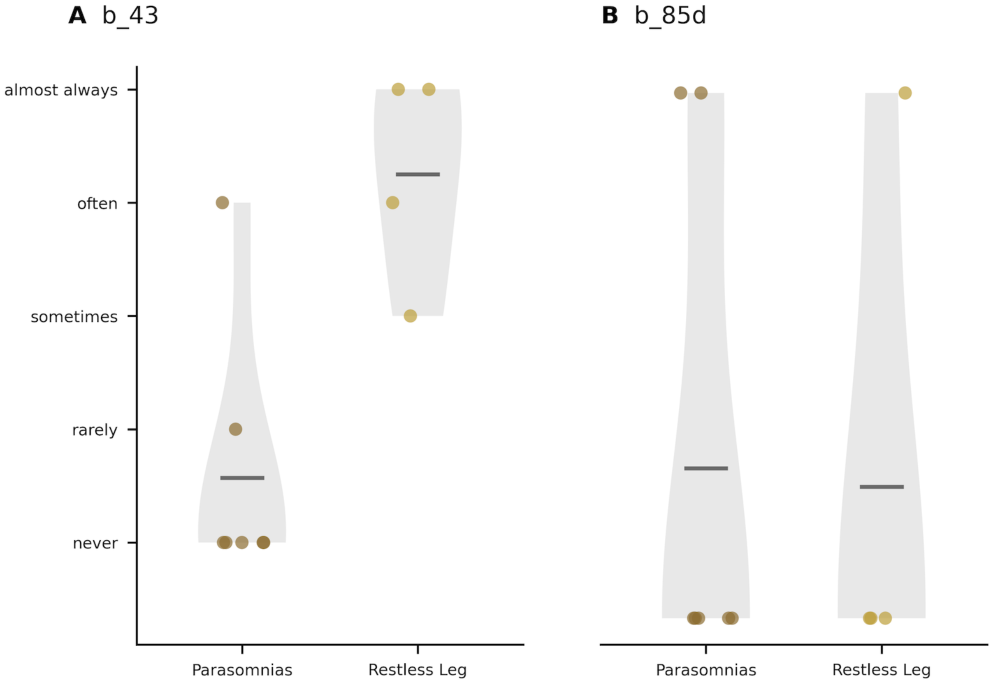


Figure C.16: Answers of patients with parasomnia and RLS for the questions b_43 and b_85d of the Bern Sleep Database.