Universidad Zaragoza

Master's Thesis

Automated Sleep Stage Scoring using Bio-signals

Author

María Sierra Torralba

Supervisor

Luis Montesano del Campo

Master in Robotics, Graphics and Computer Vision

Escuela de Ingeniería y Arquitectura
2022

Abstract

Good quality sleep is vital for good health. It supports different physiological body functions, including immune, metabolic, and cardiovascular systems ([1], [2], [3]). Furthermore, adequate sleep facilitates optimal learning, memory, attention, mood, and decision-making processes ([4], [5]). Nevertheless, sleep disorders are prevalent worldwide.

Sleep monitoring and scoring is crucial in the study and diagnosis of these diseases. Today, the only widely accepted method in clinical practice is the polysomnography (PSG), which is both intrusive for patients and expensive to perform for health systems. Accurate monitoring requires at least one night in a sleep laboratory and a time-consuming setup by technicians. The classification of sleep stages across the night provides information on the overall architecture of sleep, as well as the duration and proportion of the sleep stages, all of which inform the diagnosis of sleep disorders. Currently, this task is performed visually by human experts, requiring each 30-second epoch of a full night recording to be assigned a sleep stage. As a result, wating times for diagnostics are often larger than six months, depriving many patients of effective treatment and, thus, representing a pragmatic bottleneck.

The main goals of this thesis are to automate such labor-intensive and routine process, leading to a great reduction in workload for clinicians, as well as addressing the increasing need for longitudinal monitoring in home environments. In order to accomplish them, an AI-powered technique is developed. This will constitute the main part of a wearable EEG monitoring device based on a new textile sensor technology, which can comfortably assess everyone's sleep at home.

For that purpose, the existing literature is explored. Machine learning algorithms and, in particular, emerging deep learning approaches have shown to be outstanding approaches. Accordingly, two different deep neural networks are proposed. After that, they are implemented and applied to sleep scoring. Their goodness is evaluated considering a wide range of datasets with very different characteristics, as well as applying diverse validation and testing methods.

The results presented in this project demonstrate the validness of the models to perform real-time sleep staging with a limited number of channels in realistic settings. Moreover, one of the designed approaches in particular leads to a performance very similar to human sleep experts.

Consequently, this work serves as a proof-of-concept for future sleep technology, and lays the foundation for a diverse scope of brain-computer interfaces for real-world applications.

Acknowledgements

I would like to express my sincere gratitude to all those people who have helped in one way or another to make this project a reality today.

First and foremost, I am deeply grateful to my supervisor Luis Montesano, for his assistance at every stage of this project, as well as for sharing his immense knowledge and valuable advice.

Besides, I would also like to offer my special thanks to Eduardo López-Larraz, for encouraging me from the beginning with his plentiful experience, providing technical support, insightful comments and suggestions.

Without them, this would have not been possible. It is a great luck to be able to continue learning from you.

Last but no least, I would like to extend my thanks to my family and friends, for making me the person I am nowadays. Specially, thanks to Nacho for the understanding in the past years, for being by my side in every decision and not doubting me for a second. Words cannot express my gratitude to my parents for giving me this opportunity, for their unwavering support and belief in me, you are the reason that I do what I do.

Forever grateful to all of them.

Contents

List of Figures

List of Tables

1. Introduction

Sleep is a natural behaviour that forms part of our daily routine, making up almost one third of our lives. During sleep, a significant number of brain and body functions remain active for restorative purposes. Consequently, good sleep is crucial in maintaining one's mental and physical health [9]. Without enough and proper sleep, the homeostasis of the sleep-wake cycle is seriously compromised, which can result in a number of disorders negatively impacting life quality and cognitive performance [10][11]. In fact, sleep disorders represent a significant and increasing public health problem. A considerable proportion of the world population is suffering from them and requires medical attention. Nonetheless, less than 20% of those individuals are diagnosed and treated [12]. This fact provides evidence on the challenges faced in current clinical practice.

The gold standard for assessing the structure and quality of sleep is the whole night polysomnography (PSG), originated in the late 1950s. The PSG monitors brain activity (EEG), eye movements (EOG), muscle activity or skeletal muscle activation (EMG derivations for chin and legs), body position (video camera) and heart rythm (EKG). Breathing functions (respiratory airflow, oxygen saturation, respiratory effort indicators) are also measured. A PSG typically requires that the patients sleep overnight at the hospital while those bio-signals are recorded. Figure 1.1 shows the standard channels that are universally present in PSG.

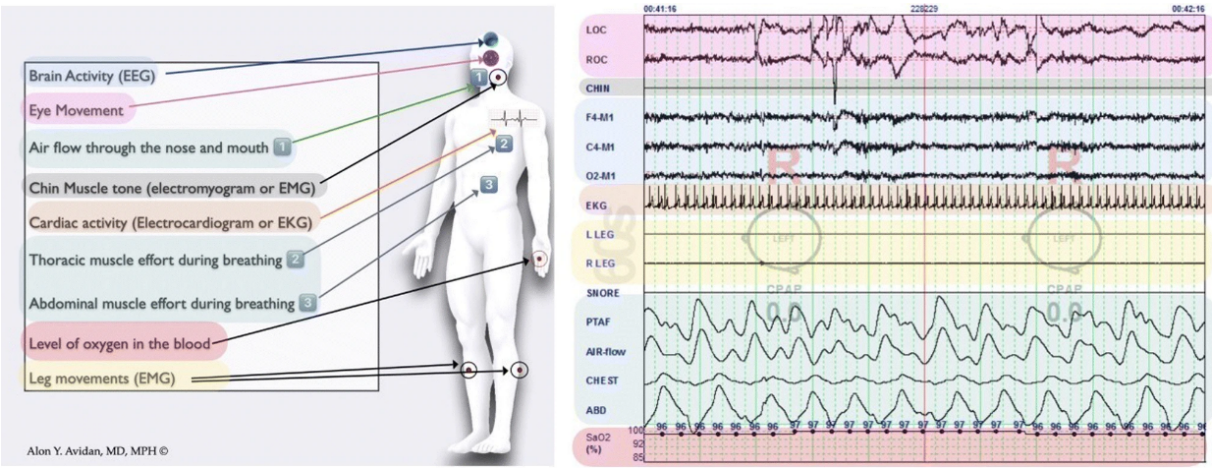


Figure 1.1: A standard PSG configuration consists of electroencephalography (EEG, measuring brain activity), electrooculography (EOG, measuring eye movements to assist in sleep staging), electromyography (EMG, measuring muscle tone in chin and limbs), electrocardiography (EKG, measuring cardiac activity), and respiratory channels (depicting airflow and effort) with pulse oximetry. These latter channels (respiratory and pulse oximetry) are most helpful in assessing for sleep-disordered breathing. From: [6].

Sleep scoring is the process of extracting sleep cycle information from the recorded electrophysiological signals. In current clinical practice it is performed manually. The aforementioned signals are noisy, have artifacts in addition to a non-stationary and subject-dependent nature, and they are not information that can be directly studied and analyzed (images, graphs, etc). To that extent, scoring an eight-hour PSG requires up to two hours of tedious, repetitive and time-consuming work that can only be performed by professionals highly trained to understand and extract meaningful information from this type of data. Even so, the task is very prone to subjective errors (i.e., the inter-scorer consistency of agreement is lower than desirable (about 70-80 % ([13]), ([14])), and the same scoring individual typically experiences low intra-scorer agreement of the same sleep recording (about 90 % ([15]))). Furthermore, it is estimated that the results do not come until 6 months, which prevents people from receiving early and effective treatment. Therefore, manual sleep scoring is limiting research on the field and is unsuited for handling large-scale data. It cannot be scaled to serve the needs of millions suffering from sleep disorders.

Last but not least, although PSG is a widely accepted method for recording ongoing brain and other physiological activity in all sleep stages, it is expensive, time-consuming to perform and disturbs participants' normal sleep.

Thus, all the above explains the critical requirement to simplify and speed up sleep scoring. These circumstances justify and motivate the realization of this thesis, which is carried out in collaboration with *Bitbrain Technologies*, a neurotechnology company that combines neuroscience, artificial intelligence (AI), and hardware to elaborate innovative products. Its objective within this field is to pilot an algorithm that allows new ergonomic, wearable, and mobile hardware based on smart textiles to be scaled to the market in order to monitor sleep. These novel, self-administered devices will allow individuals for longitudinal monitoring in home environments under uncontrolled conditions, without requiring trained professionals to apply the cumbersome and expensive PSG set up. Therefore, they can alleviate the most tedious aspects of performing a sleep study. The ultimate goal to be achieved with this technology is not only to improve the diagnosis of sleep disorders (see Appendix A for a description of the six major categories of sleep pathologies), but also to be able to perform interventions in closed-loop covering various applications from rehabilitating lost functions to increasing capabilities.

To achieve the above objectives and allow this technology to assess everyone's sleep at home, it is necessary to simplify the hardware. The technology must be simple to use and cheap to favor its scalability. In addition, it must have a reduced number of sensors and dispense with the application of conductive gels. However, these requirements lead to loss of information, as well as poorer signal quality, which is often accompanied by a large number of artifacts and noise. Therefore, this implies big challenges for the automatic sleep staging algorithm to be developed since data analysis is much more complex than with the standard system. Figure 1.2 shows both configurations.

Figure 1.2: Visualization of both types of configuration. Left: PSG, right: wearable 5 EEG channel headband from Bitbrain.

Furthermore, there is another challenge given that this algorithm will be used in closed-loop interventions. Thus, it must work online, scoring in real-time.

Those challenges will be tackled by developing an efficient and reliable approach capable of decoding and evaluating sleep in an automated way. As it is known from other disciplines, AI and machine learning (ML) tools are the perfect mechanism to automate, extend and improve data analysis. Moreover, the fact that sleep staging follows a predefined set of rules makes it a perfect task for automation with ML. For a few years now, AI is being used to optimize or reproduce visual sleep scoring.

Relevant state-of-the-art works devoted to this problem will be reviewed, studied, and finally summarized in a later chapter. After analyzing the existing work, several methods will be implemented and evaluated on different datasets with the aim of comparing them and understanding the challenges that still need to be addressed, as well as the future directions for automated sleep stage scoring to achieve clinical value. Anyway, before going into these topics, sections 1.1, 1.2 and 1.3 of the current chapter will be devoted to defining sleep, explaining what an EEG is and how it works, and introducing the EEG markers of each sleep stage, respectively.

1.1 Sleep

Sleep is a complex physiological process defined by several elements including reduced body movement and electromyographic activity, reduced responsiveness to external stimuli, closed eyes, reduced breathing rates, and altered body position and brain wave architecture. Accordingly, sleep onset is characterized by gradual changes in many behavioural and physiological characteristics. Based on physiological measurements, sleep is divided into two states with independent functions and controls: non rapid eye movement (NREM) and rapid eye movement

(REM) sleep alternating in a cyclic manner. In adult human, the first third of sleep is dominated by the slow-wave sleep and the last third is dominated by REM sleep. NREM sleep is subdivided into 4 stages (S1-S4) that become gradually deeper according to the traditional Rechtschaffen and Kales (RK) scoring manual. However, according to the recent American Academy of Sleep Medicine (AASM) scoring manual, on which this project will be based in order to be consistent with recent automatic scoring studies, this is subdivided into 3 stages (N1, N2, N3) mainly on the basis of the EEG criteria.

Currently, sleep scoring is performed manually by human experts. First of all, they divide the night of PSG recording into 30 second segments (known as epochs). Afterwards, they assign to each epoch a sleep stage, typically based on the standard rules defined by the AASM: wakefulness, N1, N2, N3 or REM. The successive representation of labeled epochs is recognized as hypnogram (see Figure 1.3).

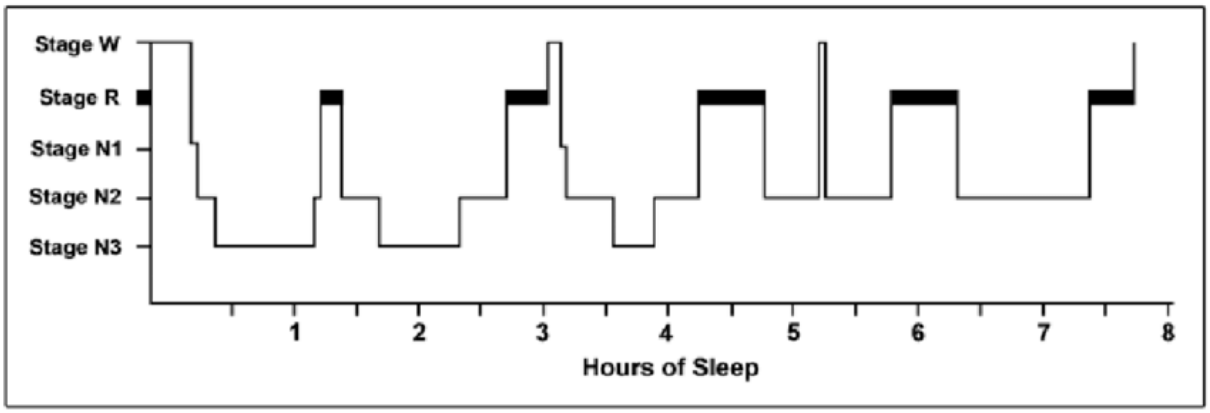


Figure 1.3: A hypnogram showing normal distribution of sleep stages. From: [7].

As far as sleep functions are concerned, all researchers agree that there is no single physiological role sleep serves [16]. Some of the most relevant charges of sleep are: development, energy conservation, consolidation of declarative as well as procedural and emotional memories, brain waste clearance, modulation of immune responses, cognition, performance, and vigilance.

1.2 Electroencephalogram (EEG)

The electroencephalogram (EEG) was first carried out in 1929 by the psychiatrist Hans Berger ([17]). Since then, it has been a widely used non-invasive method for monitoring the brain. Concisely, brain activity is characterized by the passing of electrical impulses along neurons and by postsynaptic responses as neurons communicate with one another. When many neurons "fire" (or become active) at the same time, small metal sensors placed on the scalp (called electrodes) can detect the cumulative electric fields associated with the previously mentioned impulses, and the potential differences produced. These differences are computed between two electrodes (any recording electrode and the electrode defined as reference) and are measured in μV. A typical EEG cap consists of many electrodes monitoring signals from a number of locations around the head. Further details about the types of electrodes and the EEG sensor positions will be detailed

in subsection 1.2.2. Besides, subsection 1.2.1 describes the EEG signal waves.

1.2.1 EEG waveforms

EEG waveforms can be classified according to their frequency, amplitude, and shape, as well as the sites on the scalp at which they are recorded. However, the most widely used method is by the frequency. As a consequence, brain or EEG waves are categorized into five different frequency bands: alpha, beta, theta, delta, and gamma. Delta waves (0.5-4 Hz) are the slowest EEG waves, normally detected during the deep sleep. Theta waves (4-8 Hz) are observed during some states of sleep and quiet focus. The alpha band (8-14 Hz) is originated during periods of relaxation with eyes closed but still awake. Beta band (14-30 Hz) emanates during consciousness and active concentration. Finally, gamma waves (30-100 Hz) are known to have stronger electrical signals in response to visual stimulation. Figure 1.4 depicts the different frequency bands of the EEG that have just been described.

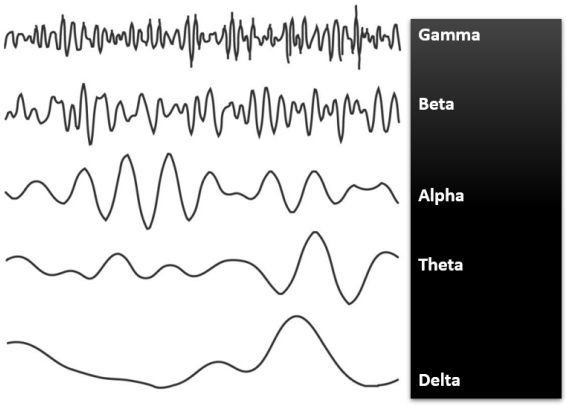


Figure 1.4: EEG signal frequency bands ([8]).

1.2.2 EEG technical features

The technical features of an EEG device are commonly divided into three principal areas: the sensor or headset area, the amplifier or acquisition layer, and the amplifier connectivity area (with other features such as data streaming, power or weight). This subsection focuses on the EEG sensor layer, explaining some of its main characteristics including the number of electrodes or channels, the placement of the electrodes, and the type of electrodes according to their contact with the skin.

Number of EEG electrodes or channels

As previously mentioned in this section, the EEG activity is measured as the voltage difference between two electrodes. An EEG headset is composed of three types of electrodes depending on their role: recording (placed in the positions of the scalp to be measured), reference (its signal is substracted from each of the recording electrodes), and ground (used to place both the amplifier and the body to the same potential, and to diminish common-mode interference).

The number of electrodes ranges between 8 and 128. This number determines the amount of information that can be measured from the brain, and always refers to the recording electrodes, since the reference and the ground are always required.

EEG electrode placement

The placement of the EEG sensors usually follows the International 10-20 System which is based on the relationship between the location of an electrode and the underlying area of the brain, specifically the cerebral cortex. The 10 and 20 refer to the fact that the actual distances between adjacent electrodes are either 10% or 20% of the total front-back or right-left distance of the skull. Based on this percentage, the 10-5, 10-10, and 10-20 systems are distinguished. Each electrode placement site has a letter to identify the lobe or area of the brain: pre-frontal (Fp), frontal (F), temporal (T), parietal (P), occipital (O), and central (C). There are also (Z) sites that refer to electrodes placed on the midline (zero line) of the skull, (Fpz, Fz, Cz, Oz), and specific anatomical locations of the ear namely the auricle (A) and the mastoid (M). Moreover, even-numbered electrodes (2, 4, 6, 8) refer to electrode placement on the right hemisphere, whereas odd numbers (1, 3, 5, 7) refer to those on the left hemisphere. Figure 1.5 shows the electrode locations of the 10-20 system for EEG recording and their relation with the different regions of the cerebral cortex. Commercial EEG systems can have fixed or interchangeable

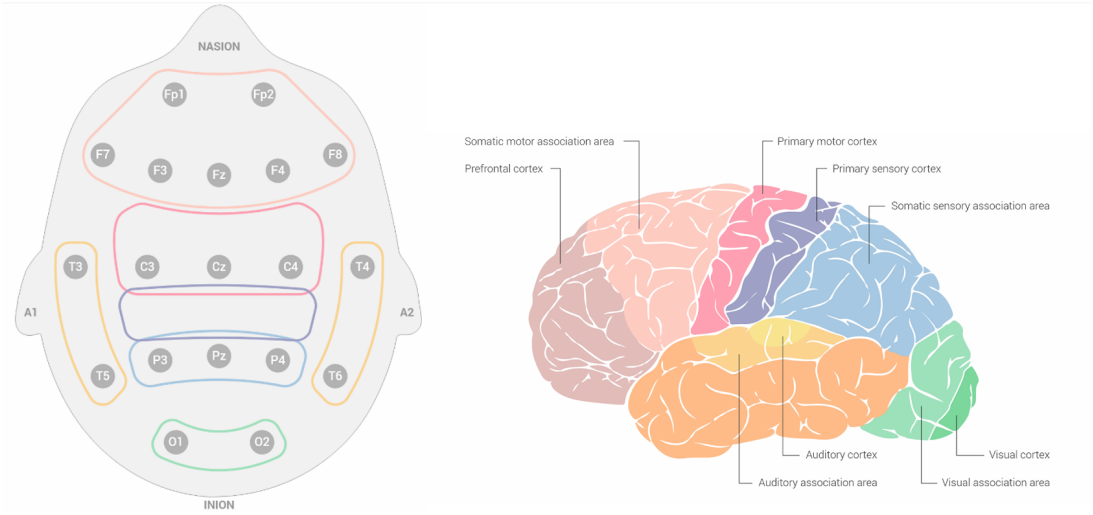


Figure 1.5: Left: placement of the standard electrodes of the 10-20 system. Right: regions of the cerebral cortex associated with brain functions.

sensor positions. The sensors of systems with fixed positions cannot be moved from one location to another, while systems with interchangeable positions can be moved to accommodate different experimental setups.

EEG electrode contact

Depending on the contact with the skin, the electrodes can be classified into two categories: dry and wet. On the one hand, dry EEG electrodes do not require the use of any electrolytic substance, making the contact directly with the scalp. On the other hand, wet EEG electrodes require the application of an electrolytic substance between the scalp and the electrode in order to improve the contact impedance. Different type of wet electrodes can work with electrolytic

gels, saline solutions, or just tap water.

1.3 Normal sleep EEG patterns

Normal healthy sleep lasts between 7-9 hours and is characterized by a certain regularity with the absence of sleep disruptions. During a normal sleep structure, stages progress cyclically from N1 through REM, taking from 90 to 110 minutes to complete a full cycle. Sleep scoring rules are based on the recognition of EEG frequencies in the different sleep stages, and on the presence of certain patterns. The characteristics of the sleep phases and some basics about the AASM scoring rules are summarized below ([18]). Finally, Figure 1.6 shows the brain wave patterns during sleep.

Stage Wake is characterized by the presence of alpha rhythm in the EEG signal, as well as by eye blinking, rapid eye movements (REMs) and normal or high chin muscle activity.

Stage N1 usually shows a shift from alpha frequency activity to theta activity. Alpha components should not exceed 50% of the total spectral band, and vertex sharp waves are often seen during transitions from other stages to N1. A mixed EEG pattern is present with low amplitude theta waves (3-7 Hz) along with slow-rolling eye movements. N1 lasts around 5% of total sleep and continues being scored until there is evidence of another stage.

Stage N2 exhibits very characteristic brain wave sleep patterns: sleep spindles and K-complex events. N2 should be scored if these patterns appear during the last half of the previous epoch or during the first half of the actual one. It should continue to be scored, also without spindles and K-complexes, until a new stage appears. N2 occupies approximately 50% of the entire night.

Stage N3 is also known as slow-wave sleep. It corresponds to the deepest NREM sleep stage. The EEG activity has recognizable patterns: slow waves (0.5-4 Hz) and slow oscillations (very large slow waves with their peak power at around 0.8 Hz). Spindles and K-complexes may remain during this stage, although they tend to decrease their presence. Stage N3 should be scored if more than 20% of the epoch consists of sleep waves and entails 20% of total sleep time.

Stage REM gets its name from the characteristic REMs seen in it (different from REMs in wake). The EEG brainwave features resemble those seen in in the awake state, but tend to be a bit slower and higher in amplitude. A particular type of theta activity termed sawtooth waves is likely to occur. A stage should continue to be scored as REM until one of the following occur: transition to Wake or N3, increase in chin EMG muscle tone, or appearance of a K-complex without arousal or a spindle in the first half of the epoch with no REMs. A person can spend about 20%-25% of their total sleep in this stage.

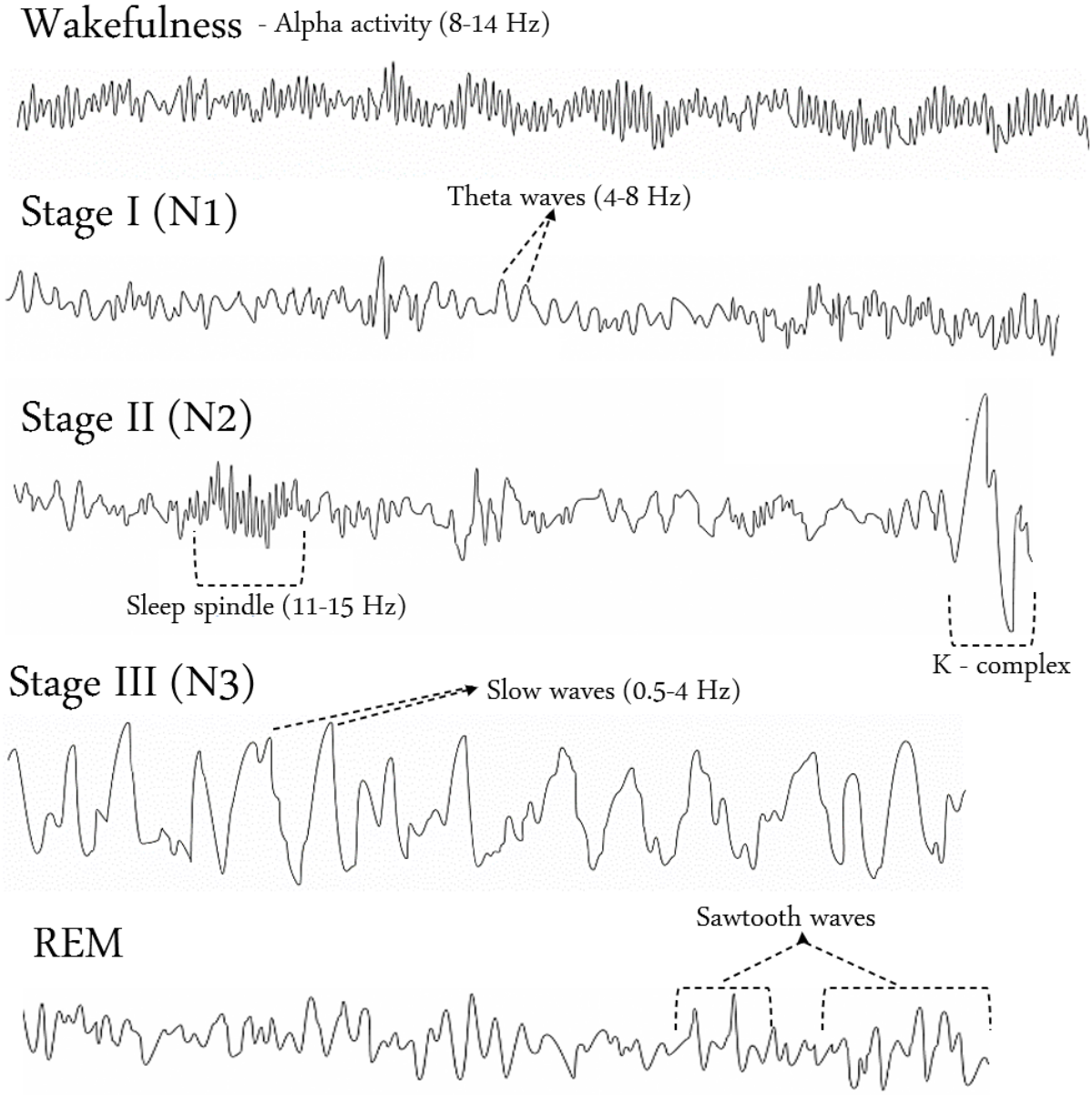


Figure 1.6: Typical EEG brain waves of sleep and wakefulness.

1.4 Objectives and scope of the project

The main objective of this master's thesis is the design and development of the algorithms and methods that are necessary for decoding the signals recorded, at home environments, by wearable devices during sleep. Consequently, this will result in the alleviation of the current burden on sleep experts, making sleep assessment and diagnostics more widely available. In order to accomplish it, the following specific objectives are established:

- Previous tasks of analysis of the state-of-the art methods for automatic sleep scoring, and study of directly related technical insights (chapter 2).
- Design and implementation of automatic sleep staging algorithms sufficiently accurate, robust and cost-effective (chapter 3, section 3.1).
- Application of signal processing and filtering techniques when necessary to remove noise, filter out artifacts while keeping as much EEG information as possible, or isolate an enhanced version of the signal of interest (chapter 3, section 3.2).
- Evaluation of the models' performance on a series of heterogeneous and independent datasets, as well as on a different number of channels and electrode positions in order to ensure their generalizability, and understand the role that electrode placement and combination of electrodes play in their functioning (chapters 4 and 5). Among these datasets, there is one recorded with a headband designed and manufactured by *Bitbrain* (which is still a prototype under evaluation. Such headband is shown in Figure 1.2).
- Validation and analysis of the results achieved, as well as comparisons to obtain the most efficient and reliable algorithm (chapter 5).

1.5 Planning and tools

The timeline followed to achieve the different objectives of the current thesis is shown in the Gantt chart of Figure 1.7.

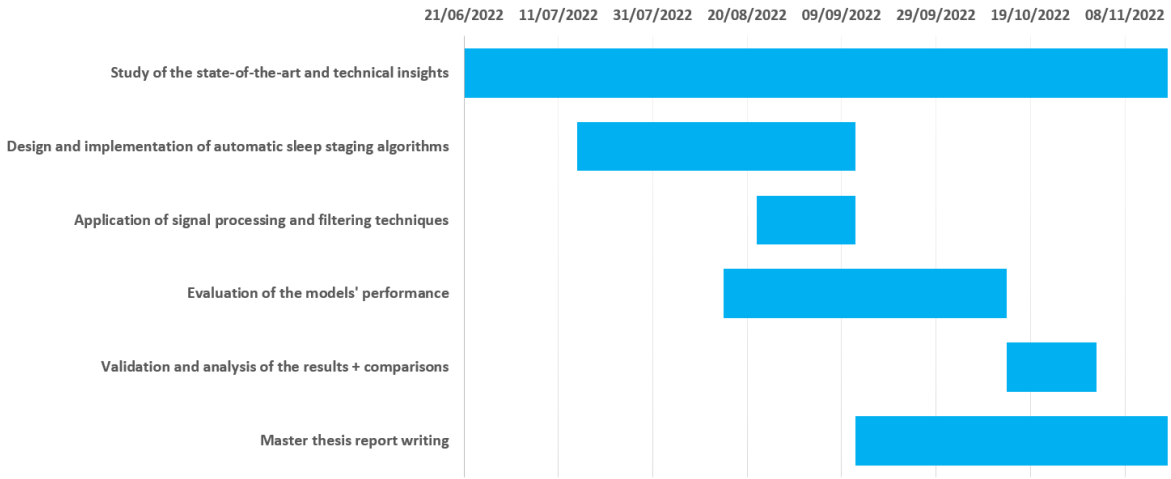


Figure 1.7: Gantt chart of the project schedule.

The implementation of the different deep learning models is carried out in Python programming language version 3.10 with Keras API version 2.8.0 and Tensorflow version 2.8.0 as the backend. Training, validation and testing of the algorithms are performed on a GPU NVIDIA GeForce RTX 3080 Ti. The datasets where they are trained on are either requested from the National Sleep Research Resource web page (NSRR, ([19])), found at Dreem Open Dataset ([20]) and ISRUC-Sleep ([21]) websites, or recorded by *Bitbrain*'s professionals. The MNE Python package is used to manage the data and to apply filtering techniques when required. The programming language and numeric computing environment MATLAB is used for data visualization purposes via the EEGLAB interactive toolbox.

2. Automatic sleep scoring: related work

AI refers to the simulation of human intelligence processes by machines. ML is an application of AI that enables systems to learn and improve from experience without being explicitly programmed. It focuses on developing algorithms that can access data and use it to learn for themselves. Thereby, these algorithms have a great potential to support and simplify the current manual sleep scoring procedure, and they have already been trained to mimic this task. However, despite progress in the field, this automatic approach has not been adopted widely in clinical environments. The ongoing chapter aims to give an overview of the state-of-the-art methods for automatic sleep staging, focusing mainly on the latest techniques.

2.1 Shallow learning

The main steps in a ML workflow are: data preprocessing, feature extraction, dimensionality reduction and classification. The preprocessing step allows to detect noise or artifacts present in the signals, meanwhile the feature extraction and dimensionality reduction phases allow to extract the most important information. Finally, the classification step uses all the previous knowledge to recognise sleep stages in this specific use case.

Initially, automatic sleep scoring was performed with traditional ML algorithms, also known as shallow learning processes. The first results close to that of a human scorer date back to 1972. On that date, in [22] a decision tree was applied using EEG and EOG data for scoring. Since then, several classification methods have been utilised in different studies including: K-means [23], Support Vector Machine (SVM) [24], Random Forest [25], Bootstrap Aggregating [26] and K-Nearest Neighbors (KNN) [27].

Some of the previous approaches reported results above 80% accuracy, showing thus potential. Nevertheless, they employ a small number of subjects from single datasets to validate their results, and, therefore, their performance may change when testing on larger datasets. Moreover, these learning processes are based on features extracted starting from the knowledge of the experts, so they can only be carried out by trained professionals. The obtained features can be affected by several factors, such as the dataset characteristics. In particular, sleep datasets are diverse and contain a large number of epochs. Therefore, this approach may not be worthy to satisfy a description of the heterogeneity of the subjects and the recorded signals ([18]). Due to this fact, the use of deep neural networks (section 2.2) for automatic sleep scoring started around 5 years ago, producing results that were never seen with conventional ML techniques for decades.

2.2 Deep learning

DL is a a subfield of ML based on artificial neural networks (ANNs) that learn through hidden layers. The more layers are stacked, the more complex features are produced. As opposed to task-specific algorithms, these methods are centered on learning data representations. Their ability to extract information from large amounts of data is the main reason why they began to be used for sleep stages classification.

The principal distinction between different ANNs is their architecture. That is, the way in which several neurons are arranged and connected to each other. Each of these neurons in the network performs a linear combination of the input followed by a non linear transformation.

Transitioning from shallow learning (section 2.1), the first attempts to use DL for automatic sleep scoring used standalone network architectures such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs, e.g., Long Short-Term Memory cells (LSTM), Bidirectional Long Short-Term Memory cells (Bi-LSTM) or Gated Recurrent Units (GRU)).

CNN models are frequently used to recognize images. However, their usage is not limited to 2D or 3D recognition tasks. 1D-CNNs share the same properties with other CNN models. The only difference is in the convolution operation, which is suitable for input data in 1D such as biomedical signals. A convolution takes a group of close samples from the input series and operates matematically with several matrices called kernels or weights. A group of kernels are called filters, and they loop through all the input neurons generating new output matrices. The output of a convolution process is called a feature map, and thus each of the filters can be thought of as a feature identifier. The feature maps can be sub-sampled with pooling layers, or they can be processed in other convolution layers. The mentioned sub-sampling layers reduce the dimensionality while keeping the relevant information. The final layer of the CNN model usually contains a fully connected layer that performs the classification task.
During training, the weights and bias of the network are continuously optimized to obtain the desired class for the input. Once the network is trained, it is ready to predict on new inputs. Regarding RNNs, their main difference with CNNs is that they do not consider only the current input but also the previously received inputs. Therefore, in this way, they can handle sequential data.

However, although the previous standalone architectures can learn useful features, they cannot capture long-term dependencies between sleep epochs. This is due to the fact that they use a few epochs around the target one to predict its sleep stage. In order to solve this issue, when any of the above networks have been trained and each epoch is encoded into an epoch-wise feature vector, an additional RNN is separately trained in a second stage to take into account a long sequence of feature vectors prior to a target epoch to classify it ([28]). These hybrid networks with two-stage training (epoch and sequence encoders) were initiated in [29], and can be generalized in a framework named sequence-to-sequence sleep staging. They surpassed the performance of other existing models, since their scheme is very similar to the way experts manually score. Specifically, they determine the label of a target epoch attending to a much larger context around it. Figure 2.1 depicts a schematic diagram of the sequence-to-sequence sleep staging framework.

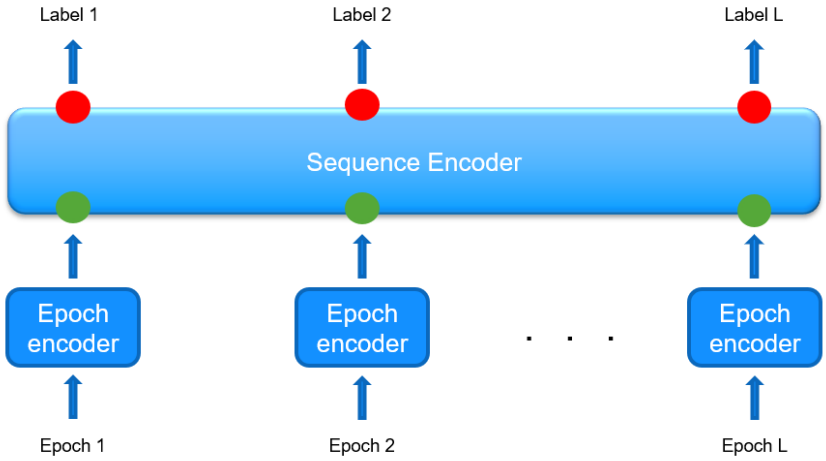


Figure 2.1: Sequence-to-sequence sleep staging framework. The epoch encoder takes an epoch and converts it into a feature vector representation (green circle). The sequence encoder improves this representation by incorporating in its output (red circle) the interaction of each epoch with other epochs in its context.

Inspired by the previous results, since late 2017, studies have focused on these architectures. Nonetheless, although RNNs are the models commonly used for sequence encoding and CNNs for epoch enconding, recent approaches have shown that emerging network architectures such as Transformers and Self-attention ([30]) are also useful as this type of encoders.

Table 2.1 provides a collection of automatic sleep scoring systems existing in the recent literature, all of them with long-term modeling capabilities. The mentioned systems are presented in chronological order along with the type of input they receive (raw data or time-frequency images), the architectures used as epoch and sequence encoders, and the performance they have demonstrated on various public datasets (SleepEDF ([31]), MASS ([32]), SHHS ([33]), and DREEM (DOD-H, DOD-O) ([20])), almost all of them containing a majority of healthy subjects. These results are obtained either from the original works or from other works where the systems are evaluated.

Network	Year	Input	System	SleepEDF	MASS	SHHS	DOD-H	DOD-O
DeepSleepNet ([29])	2017	Raw	CNN+RNN	82.0	86.2	-	-	-
SeqSleepNet ([34])	2018	Time-freq	RNN	82.6	82.8	86.5	80.4	77.2
SleepEEGNet ([35])	2019	Raw	CNN+RNN	82.83	-	-	-	-
SimpleSleepNet ([20])	2019	Time-freq	RNN	-	-	-	85.7	79.3
FCNN+RNN ([36])	2020	Raw	Fully CNN+RNN	82.8	-	86.7	-	-
XSleepNet ([36])	2020	Raw, Time-freq	CNN, RNN+RNN	84	85.2	87.5	-	-
RecSleepNet ([37])	2021	Raw	CNN+RNN	86.1	-	-	-	-
AttnSleep ([38])	2021	Raw	CNN+Self attention	81.3	-	86.6	-	-
SleepTransformer ([39])	2022	Time-freq	Transformer	81.4	-	87.7	-	-

Table 2.1: State-of-the art in automatic sleep scoring. Overview of the latest approaches.

Note that the performances presented in the table above should not be used to justify the effectiveness of a network or to compare one with another. Discrepancies in the testing settings, as well as in the modeling mean that they cannot be analyzed as equals and, therefore, that this is not a direct comparison, but an informative collection of results.

Finally, despite the performance in healthy individuals is at the level of experts' scoring, progress is still needed when subjects under inspection are patients. Moreover, most of the previous models are evaluated using several channels (EEG, EOG, and EMG). However, those results should be achieved with just a few EEG electrodes in order to apply the technique in home environments. Lastly, all of them use large temporal sequences to obtain their respective accuracies. This is possible since they are tested in an offline way, but the temporal resolution must be reduced and future information cannot be provided to allow for real-time applications.

3. Methods

Having discussed previous studies on automatic sleep stage classification, it is obvious that recent advances made in DL have shown remarkable results. Therefore, this chapter describes the deep learning-based methods that have been implemented from scratch or applied for the development of this project. Furthermore, several techniques used with the objective of improving the performance of the designed frameworks are explained.

3.1 Model architectures

The current section aims to introduce the model architectures developed and employed for automatic sleep staging in this thesis. They are intended to use raw data as input.

As mentioned in the previous chapter, the main drawback of performing real-time sleep scoring is that, when batching the data during training, future information cannot be provided. Thereby, more computational resources are needed for convergence.
If future information is provided during training but not during testing, accuracy will be decreased. With the aim of solving this issue, the TimeDistributed layer of the Keras API is used in all the proposed architectures. This layer allows to apply one or several layers to independent time frames (30 seconds of data in this case). In this way, neither future nor past context is given during training, but the data can still be batched in large groups in order to accelerate the loss function optimization. Particularly, the optimization algorithm used in all the cases is Adam, and the loss function is the sparse categorical cross-entropy due to the presence of more than 2 labels.

No major modifications will be applied to the described models. The only changes will be specific to the test being carried out consisting of adjustments to the input layers and data batching procedure.

3.1.1 CNN

The first architecture was employed in a previous work ([40]) and, since it showed promising results in different tests, is used here in order to complete its validation.

The network's structure can be seen in Figure 3.1. It consists of an expanded version of the CNN model proposed in [41] in order to achieve real-time sleep staging of 30 second epochs of

data. These epochs are used as input for the initial sub-model, referred to as 'Time distributed model' (see Figure 3.2), which is responsible for feature extraction of each individual epoch, independently of when it may occur during night and disregarding any information in its vicinity. Therefore, the weights are applied in a temporally independent way. Within this model, data points of each epoch initially pass through a structure of two 1D convolutions of 16 filters with kernel size of 5. Then, a max-pooling of size 2 and a dropout with 0.01 rate are performed. This structure is repeated 4 times with an increasing number of filters (from 16, to 32 (twice) and finally 256) while reducing the kernel size to 3 in all cases. Lastly, a fully-connected dense layer of 64 units is used as output of the first sub-model for every epoch. From this point on, the weights are no longer applied independently to each 30 seconds epoch, but to the entirety of the training, validation or testing batch. This is performed by the second sub-model, which is referred to as 'Global model' (see Figure 3.2), with the aim of rearranging all epochs simultaneously into 5-unit vectors, corresponding to the probability of selecting each class. In order to accomplish it, the following operations are performed: 2 sets of 1D convolutions (128 filters and kernel size of 3) and dropout with rate 0.01, followed by a final 1D convolution (5 filters and kernel size of 3). These operations produce as output the aforementioned 5-unit vector. Eventually, the maximum value of this vector is computed to obtain the sleep label.

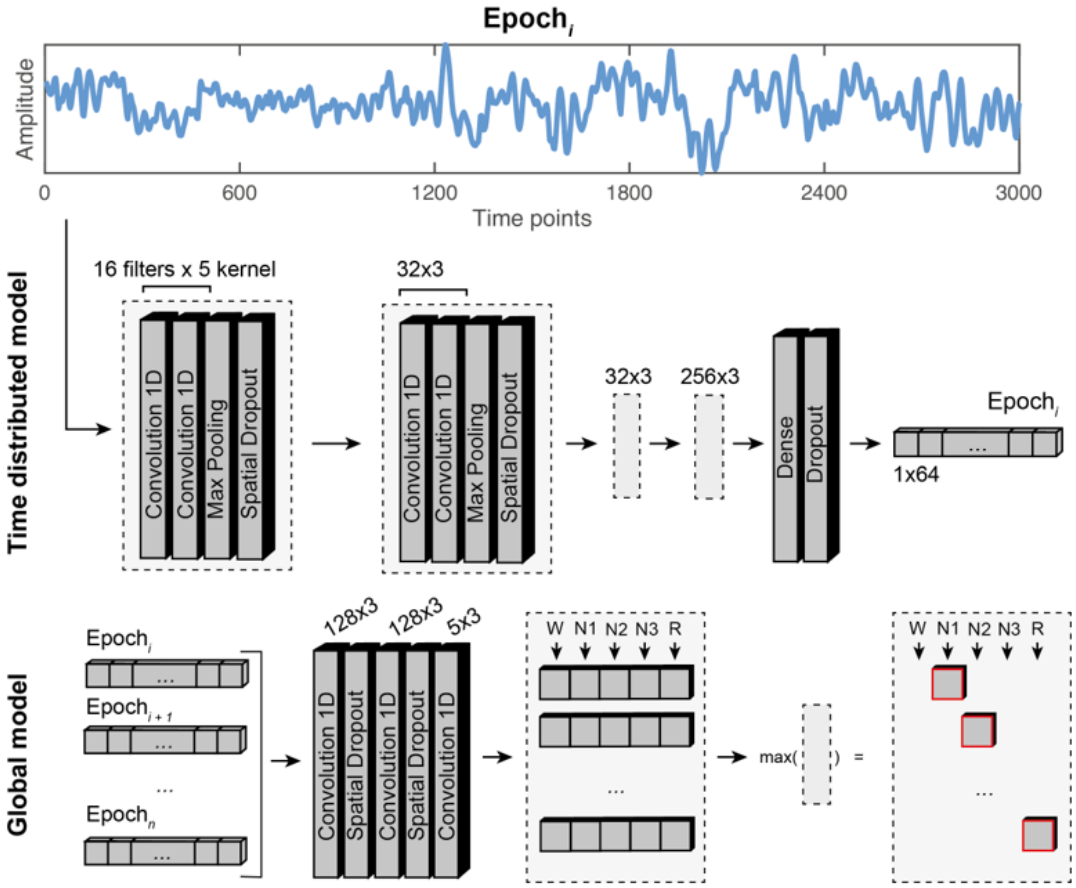


Figure 3.1: Illustration of the proposed CNN architecture.

The activation function of all the above convolutional layers is a rectified linear unit (ReLU), except for the last layer of 5 filters in which a softmax function is used.

3.1.2 CNN + RNN

The second architecture is inspired in [29] and it is made up of two different parts. An overview of this model can be observed in Figure 3.2.

The first part (Figure 3.2 A) is used to learn filters to extract time-invariant features from the raw signals. It is a CNN consisting of 4 convolutional layers, interleaved with 2 max-pooling layers and two dropout layers. (the specifications of the number of filters, filter sizes, pooling sizes, dropout probabilities and stride sizes can be seen in the figure). Each of the mentioned convolutional layer performs 3 operations sequentially: 1D convolution, batch normalization, and activation. On the other hand, the second part of the network (Figure 3.2 B) encodes temporal information in the extracted features. This part is located close to the output sleep stages, and it is a unidirectional RNN consisting of a single LSTM layer followed by a dropout layer. Bidirectional-LSTMs are not considered since they double the computational resources required for the sequence learning, and they exploit information from both the past and the future. In the end, a densely connected layer with 5 nodes outputs the desired 5-unit vector. As it was done with the previous architecture, the maximum value of this vector is selected and corresponds to the final label.

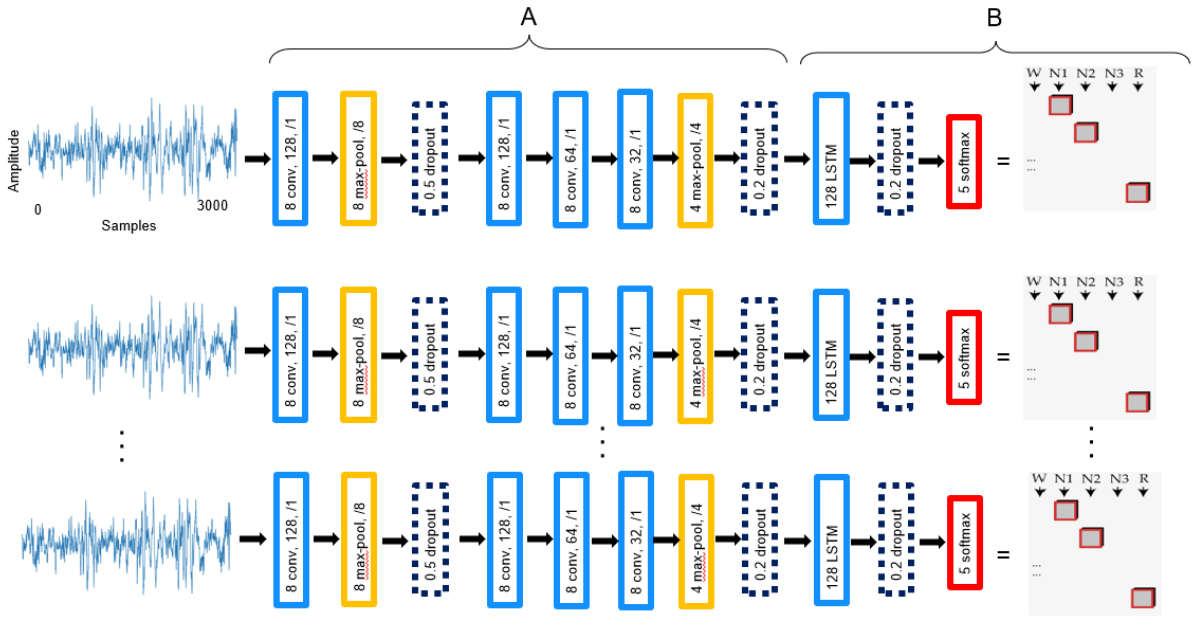


Figure 3.2: Illustration of the proposed CNN (A) + RNN (B) architecture. The specifications of the convolutional and max-pooling layers of the CNN are as follows: [filter size (conv), number of filters, /stride size] and [pooling size (max-pool), /stride size], respectively.

In order to avoid overfitting to noises or artifacts in the data, L2 weight decay is used in addition to dropout. This technique adds a penalty term into the loss function to prevent large

values of the parameters in the model (exploding gradients). Nevertheless, it is only applied on the first layer of the CNN, since in [42] it is stated that it can limit the model capabilities of learning long-term dependencies.

Finally, a slight variation of this model will also be tested. Since all layers are changed during an update, the update procedure is forever chasing a moving target. Batch normalization is used to coordinate the update of multiple layers in the model. It does this by scaling the output of the layer, standardizing the activation of each input variable per mini-batch (this standardization refers to re-scaling the data to have a mean of 0 and a standard deviation of 1). Batch normalization is thus making the network more stable during training. Related to this, the first modification that can be performed to the described architecture is to establish a larger than the previous learning rate, speeding up the learning process. Furthermore, a second modification is carried out because it is not a good idea to use batch normalization with dropout in the same network. Therefore, dropout layers are removed from the scheme. This is due to the fact that batch normalization is offering some regularization effect, and the statistics used to normalize the activation's of the prior layer may become noisy given the random dropping out of nodes during the dropout procedure.

3.2 Filtering

All the algorithms described above will be fed with raw EEG data. One of the main problems when dealing with this kind of signals is that their amplitude is in the order of micro volts and can then be easily contaminated with noise, i.e., artifacts. These artifacts are not of interest to the analysis and contribute to a lack of precision of the DL techniques. Therefore, they must be filtered from the neural processes to keep the valuable information needed for the current task.

An artifact can be denoted as any component recorded by the EEG that is not directly produced by human brain activity. They are classified depending on their source, which can be physiological or external to the human body (non-physiological/technical). For more details on artifact types, see Appendix B. As a very brief summary, low frequency noise comes from sources such as movement of the head and electrode wires, and perspiration originated by sweat glands of the skin. Low frequency noise appears as slow drifts in the EEG signal over many seconds. On the other hand, high frequency noise comes from sources including electromagnetic interference, and muscle contractions. High frequency noise looks like rapid up-down changes in the EEG.

With the aim of attenuating the power of the collected signals at the frequencies below and above the range of experimental relevance, a filtering process will be applied before giving them as input to the network. With this preprocessing step, the low SNR presented by EEG systems will be compensated. In particular, a notch filter will be used to remove the 50Hz or 60Hz AC electrical interference, and then a bandpass filter with low and high frequency cutoffs of 0.1-0.5 and 30-45 Hz, respectively, according to the vast majority of researchers. If necessary, EEG artifact rejection will also be applied. This consists of selecting and rejecting EEG epochs with artifacts. It will be accomplished by computing automated statistics in the time domain (e.g., defining a threshold to remove epochs that have a significantly higher or lower amplitude).

3.3 Data augmentation

Data augmentation (DA) is a technique that is used to artificially expand the size of a training set by creating modified data from the existing one. It started to be popular because of DL algorithms, which need a huge amount of data to properly train the models. Exposing a model to varied representations of its training samples makes it more invariant and robust to transformations of the type that is likely to encounter when attempting to generalize to unseen samples. Moreover, increasing the size of the training set allows using more complex models and prevents overfitting. This method has been demonstrated to achieve considerable performance gains for DL in many fields, such as computer vision ([43]) or speech ([44]), increasing the accuracy and stability of the classification. In recent years, DA techniques have received widespread attention and substantially increased accuracy when using DL for EEG analysis ([45]).

The reasons why artificial data is generated can be varied. Examples of them are: economical cost (acquiring data can be very expensive), technical challenge (acquiring data can be difficult), and uncommon data (some data, such as minority diseases, is very rare). These reasons are applicable to EEG data since acquiring it requires subjects to undergo long calibration sessions, it can be difficult and expensive to collect the desired number of samples, and large openly available EEG datasets are uncommon. However, in this case, the motivation behind using DA is to mitigate the imbalance that occurs very frequently in time series. Specifically, the datasets that will be used for this thesis are highly unbalanced toward the N2 class. This is something to be expected because the majority of sleep is usually spent in that stage. On the contrary, N1 accounts for the least proportion of the night. Being more specific, of the total sleep time, N2 constitutes 50%, N3 and REM, 20% , and N1 5%, leaving a variable degree of Wake percentage depending primarily on the quality of sleep. Therefore, DA will be used to get a balanced dataset and thus check if this has a positive impact on the network's performance. For that purpose, different methods will be applied to generate artificial samples of minority classes: geometric transformations (drift), noise addition, overlapping windows, and sampling. These techniques will be employed only on training data to avoid corrupting the prediction results.

3.4 Adding information sources: IMU

As discussed previously in this document, the standard PSG equipment is cumbersome, expensive, and requires a trained professional to apply the set up. Novel, self-administered devices can alleviate the most tedious aspects of performing a sleep study. However, current DL algorithms suffer from low accuracy, especially for sleep stage transitions, minority sleep stages (N1), and when only EEG is available, such as in the mentioned setups.

One of the datasets that will be used in this thesis, which will be detailed in the next chapter, is composed of nights recorded with one of these types of devices. Specifically, it is the comfortable, self-managed headband that has been previously introduced. Therefore, adding further sources of information to the input of the previous models can alleviate the issues that such setups entail. Accordingly, with the goal of improving the performance of the sleep scoring algorithms on the headband recordings, head movement data measured by an inertial measurement unit (IMU) will be provided.

The reason for this idea is that this kind of data contains information about sleep stages. Figure 3.3 shows evidence that movement is clearly able to differentiate between them. The plots display the number of acceleration peaks per minute (A) and the average acceleration magnitude (B) for each sleep stage. As it can be observed, both measurements are larger during Wake and light NREM.

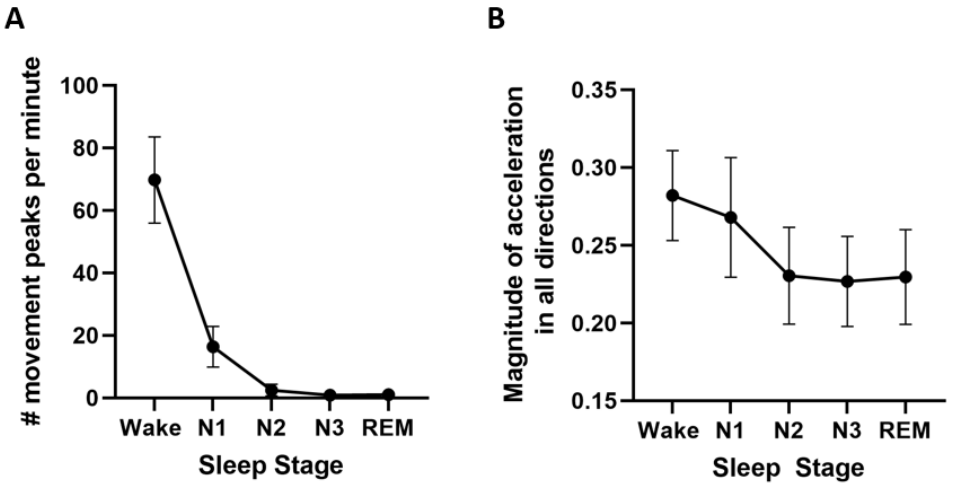


Figure 3.3: **A. Number of peaks per minute for each sleep stage.** Peaks were defined as acceleration magnitude scores greater than 1.5 g (after preprocessing). Peaks were 4 to 6 times greater than average magnitude scores. The figure shows the number of peaks per minute (± SEM) for each sleep stage. **B. Average acceleration magnitude for each sleep stage.** Acceleration magnitude was calculated by taking the root-sum-square of the acceleration values for each axis of the triaxial accelerometer. The figure shows the average (± SEM) for each sleep stage.

Consequently, the objectives are comparable to those sought with DA, although the technique is different. In this case, it will be checked if acquiring complementary information from different sources for the same event can overcome some of the limitations of individual systems. In short, the aim is to improve the performance of the DL architectures by adding distinct information about underrepresented stages and balancing skewed performance towards more abundant classes.

3.5 Uncertainty quantification

DL has demonstrated its benefits to healthcare: superior performance, capability of handling large and complex data, data-driven learning ability, etc. However, communications of sleep experts state that the complex nature of these algorithms and the skepticism of DL models being a black box have been hampering medical professionals' trust ([46]). Both are the main obstacles impeding DL-based automatic sleep scoring from obtaining a widespread clinical value and adoption.

In order to make the AI system more reliable, a simple yet efficient method to quantify uncertainty in the model's decisions is proposed.
In the current classification problem, the network outputs a vector ($\hat{y}$) whose elements are 5

probability values, corresponding to the 5 sleep stages. As described previously (section 3.1), the sleep stage matching the maximum probability constitutes the final output of the model. However, the predicted discrete label does not inform about the confidence of the network in the decision. Therefore, the multi-class probability distribution over the sleep stages encoded in $\hat{y}$ can provide a value for the confidence measure. On the one hand, if $\hat{y}$ assigns a probability of 1 to a sleep stage and a probability of 0 to the rest, the network is very confident in its decision. On the other hand, when the output distribution is flat (all the elements in $\hat{y}$ are equal), the network has no confidence in its decision. All other distributions represent varying levels of confidence between these two extremes.
Accordingly, the likelihood of the most likely class will be used to obtain a measure of confidence. In this manner, it will be tested if the number quantifying the confidence aligns well with the model's successes and mistakes. If that is the case, epochs with lower confidence will be provided to human experts for manual verification and correction. This also allows the system to work alongside practitioners in an interactive and collaborative way.

4. Evaluation

The performance of the previous neural networks will be studied on 5 different datasets, and characterized on a different number of channels and electrode position. Moreover, different tests will be carried out with the goal of assessing the capability of the architectures to perform transfer learning between datasets.

This chapter explains the characteristics of the datasets being used, the information sources (number of channels and electrode placement) employed, how the transfer effectiveness is checked, and the metrics computed to evaluate the performance of the models. Finally, a summary of all the evaluation tests to conduct is given.

4.1 Dataset characteristics

The data analyzed in this project can be distinguished into two types: data recorded in *Bitbrain*'s sleep laboratories using our own equipment as well as publicly available datasets. The latter are 4, including both healthy volunteers and patients suffering from different types of sleep pathologies. Furthermore the last one is a massive dataset .

4.1.1 *Bitbrain*'s data recordings

Data is acquired in a fully equipped sleep laboratory. This lab features a full, medical-grade polysomnographic , which acts as a gold-standard comparation device. Furthermore, data is collected concurrently with a *Bitbrain* wearable and highly comfortable headband together with an inertial measurment unit (IMU). The recording setup can be observed in Figure 4.1.
Specifically, the dataset is formed by 30 recordings, each of them containing information from:

- 6 EEG channels from the PSG device (F3, F4, C3, C4, O1, and O2), along with 2 EOG channels (bipolar derivations left and right of the eyes (EOGV, vertical eye movements), as well as above and below the eyes (EOGH, horizontal eye movements)), 1 EMG channel (face muscle activity, bipolar derivations under the mouth), and blood oxygen saturation information measured by a pulse oximeter on finger.
- 5 EEG channels from the headband: AF8, Fp2, Fp1, AF7, and T8.
- Head movement data, measured by an IMU.

The recordings are sampled at 256 Hz, and resampled to 128 Hz. Furthermore, they are band-pass filtered applying two different boundaries (0.5-45 Hz, 2-45 Hz), leading to 2 different versions of the same data.

All data were scored as result of an epoch-wise majority vote of three experienced sleep scoring experts according to the AASM scoring rules. In cases in which no majority vote could be found (i.e., when an epoch was scored differently by all three scorers), a fourth scorer made the final decision.

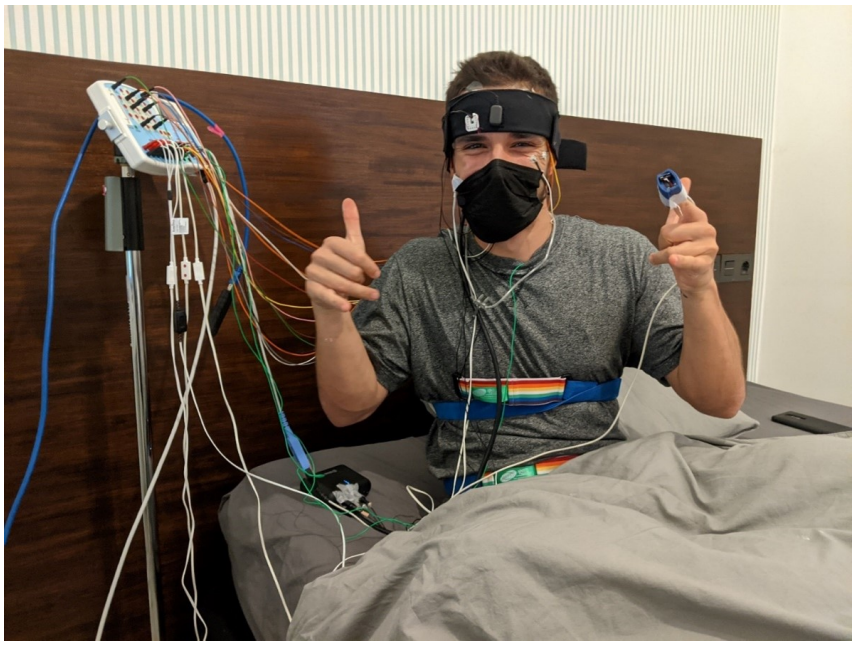


Figure 4.1: Recording setup. Volunteer in one of *Bitbrain*'s sleep laboratories with a full medical PSG setup and a *Bitbrain* headband.

4.1.2 Publicly available datasets

ISRUC-Sleep ([21]) is a publicly available dataset composed of 3 subgroups (I-III). Subgroup I and II are formed by 100 and 8 PSG recordings, respectively, from adults having evidence of sleep disorders. Subgroup III is formed by 10 healthy subjects. PSG recordings contain information from the EOG, EEG (F3-A2, F4-A1, C3-A2, C4-A1, O1-A2 and O2-A1), EMG, EKG, SaO2, position, airflow and abdominal effort, sampled at 200 Hz. The 3 subgroups of participants are scored by two well-trained technicians according to the AASM guidelines. However, given the large imbalance in the number of subjects, data from subgroup I is used exclusively.

Dreem Open Dataset ([20]) is an open-access sleep archive with 2 subsets. The first one is referred to as Dreem Open Dataset - Healthy (DOD-H), comprised of 25 healthy subjects and the second as Dreem Open Dataset - Obstructive (DOD-O) comprised of 50 subjects suspected of having sleep related breathing disorders. Both datasets are used. They contain EOG, EMG, EKG and EEG (C3-M2, F4-M1, F3-F4, F3-M2, F4-O2, F3-O1, Fp1-F3, Fp1-M2, Fp1-O1, Fp2-F4, Fp2-M1, Fp2-O2) sampled at 250 Hz and scored by 5 expert sleep technologists following AASM criteria in order to minimize inter-rater variability.

Stanford Technology Analytics and Genomics in Sleep (STAGES) ([19]) is a dataset publicly available for use by any interested researcher, provided a request is submitted to the National Sleep Research Resource and approved. It explores sleep through genetics and technology in a large-scale patient-oriented study, collecting objective and subjective sleep data, and biological samples. This cross-sectional, multi-site study involves data collected from 14 centres including Stanford University, Bogan Sleep Consulting, Geisinger Health, Mayo Clinic, MedSleep, and St. Luke's Hospital. Specifically, the project has collected data on 1500 adult/adolescent patients evaluated for sleep disorders, including: objective nocturnal sleep polysomnography (PSG) recordings (EEGs, chin and leg EMGs, nasal and oral breathing, chest movements, leg movements, position, EKG) , comprehensive subjective sleep symptoms assessment through an online sleep questionnaire, continuous actigraphy over several weeks, 3D facial scans to extract craniofacial features predictive of sleep apnea, online neuropsychological assessments and psychovigilance tests, and medical record data.

All the previous data were resampled to 100 Hz, band-pass filtered between 0.5 and 30 Hz, and notch filtered at 50 Hz (ISRUC, DOD-H, DOD-O) or at 60 Hz (STAGES).

In the particular case of the massive dataset, data preprocessing was more complicated. After eliminating subjects in which most of the signal was of poor quality, 13 folders are available, corresponding to 13 different medical centers. The total number of subjects among all of them is 1223. For detailed information on the process carried out, go to Appendix C.

4.2 Electrode placement, channels & presence/absence of future information

In order to understand the role that electrode positions play in the automatic sleep scoring process, three tests will be performed. These tests consist of training and testing on single channel EEG information coming from a frontal (F3), central (C3) and occipital (O1) electrode, respectively. After that, two tests will be carried out in order to find out if combining multiple channels has a positive impact in the global accuracy: one training and testing using only EEG data (frontal, central and occipital) and another one mixing EEG, EMG, EKG and EOG. In all the aforementioned evaluations, data will fed into the network in a real-time manner (online) during testing. This means that these tests are performed in the absence of future information during classification. Therefore, for this online emulation approaches, epochs will be fed through an overlapping sliding window of 2.5 minutes, where every new epoch is appended to the window, inputted to the network, and labelled.

Furthermore, an additional test is designed to comprehend if future information (i.e., data collected after the epoch being scored) affects classification performance. This test will employ the single-channel frontal derivation and it will be performed in a fully offline manner. For this purpose, data will be batched in non-overlapping windows of 100 epochs, where all epochs are labelled at once rather than one by one.

All information above refers to the DOD-H, DOD-O, and ISRUC datasets. In the case of STAGES dataset, it was necessary to find common electrode positions among all the recordings

from each of the medical centers, in order to combine all of them, generating a massive set. Thinking of favoring a future transfer to the data of the *Bitbrain*'s headband, the selected ones were F3 and F4. All the tests will be carried out using the information from these two EEG channels.

Finally, the *Bitbrain*'s data recordings will be employed. Different tests will be performed for a further comprehension of the effect of the number of channels on the network's performance: training and testing on all the headband channels, on two channels with frontal electrode position (Fp2, Fp1) or on a single frontal EEG channel (Fp1). All these evaluations will be repeated adding the IMU channel. In this way, it will be evaluated if providing head movement data (chapter 3, section 3.4) actually contributes in a positive way to the sleep scoring algorithm. In the last place, two tests will be done using: all the EEG channels from the PSG device and only two EEG channels (frontal electrode location, F3 and F4). The purpose of these experiments is to quantify the drop in accuracy that can be experienced when using wearable, self-managed devices such as the band.

4.3 Transfer learning

As it was commented before, an evaluation of the capability of the networks to perform transfer learning between datasets is carried out. The different tests that have this intention are listed below:

- Within the STAGES dataset itself: train using the subjects belonging to all the medical centers except one, and test with the remaining medical center. In this way, it can be observed how much the accuracy changes with respect to training and testing with the same center, checking if an overfitting to the smaller datasets was taking place. Moreover, this test will serve to be compared with another one in which the centers are divided proportionally into training/validation and test so that there is the same percentage of each of them in the three sets. Thereby, this will allow to check if a higher heterogeneity (data from independent centers involved different recording devices, montage, sampling rate, and participants from a wide age range, sex, health status, and sleep disorders) is paramount to ensure a high reliability and generalizability of the algorithm.

- Train with DOD-H and test on DOD-O, as well as the same test but taking the weights from DOD-O and testing on DOD-H. Performing this transfer learning test in both directions will give an insight on whether one of the datasets generalizes better than the other. This will also make it possible to find out if the knowledge acquired by the model is transferable between healthy subjects and subjects suffering from sleep-related pathologies.

- Transfer learning from ISRUC to DOD-O. Since both datasets include patients suffering from sleep disorders, it can be assessed if, in this case, the results are better than when transferring the weights from DOD-H to DOD-O.

- Train with the PSG data acquired during *Bitbrain*'s recordings and test with the data collected by the headband (using 2 frontal electrodes from the PSG: F3, F4, and 2 from the band: Fp1, Fp2). This could be useful to evaluate if learning from higher quality

recordings helps to increase accuracy results when using the signals collected during the same night by a device with lower SNR.

- Transfer learning from STAGES to *Bitbrain*'s dataset. Specifically, this will be accomplished by training with two frontal electrodes from STAGES (F3 and F4, as mentioned in section 4.2), and testing on two frontal electrodes of the band (Fp1 and Fp2). This is done to study if the limited number of subjects was restricting the accuracy achieved. Several variations of this same test will be performed: in addition to transfer learning, fine-tuning will be carried out unfreezing and retraining different numbers of layers.

4.4 Performance

In any test that is carried out, the same DL pipeline will always be implemented. It can be seen in Figure 4.2. Firstly, every recording is preprocessed by applying a certain filtering technique when necessary, resampling to the desired frequency (typically 100 Hz), and reshaping into a matrix where the first dimension is the number of epochs, the second is the number of time points (3000 if the signal has been resampled to 100 Hz), and the third is the employed number of channels (which, as discussed in previous subsections, will vary depending on the dataset and the particular test). Subsequently, each dataset is divided into training and testing with 90% and 10% ratio, respectively, to carry out a 10-fold cross-validation. Each training set is further divided into a train (85%) and validation set (15%). After this splitting, the recordings are z-scored according exclusively to the training data. In this way, the information from the test set is not seen ahead of time in any way. As the last step, the data is passed onto the models.

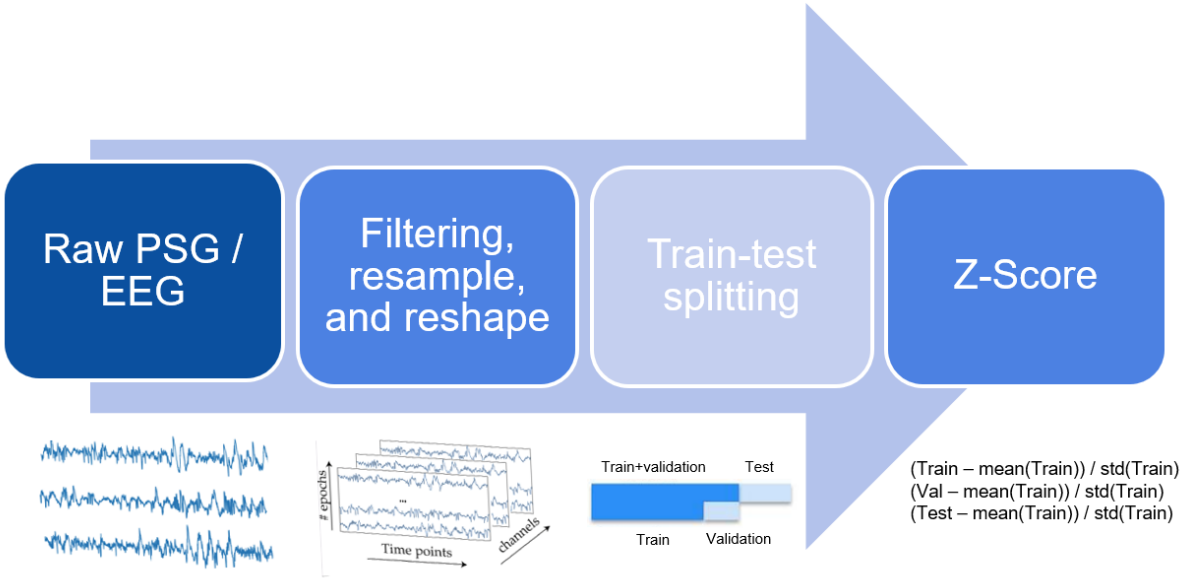


Figure 4.2: Deep learning pipeline. Each recording containing several channels is first filtered, resampled and reshaped into a 3D matrix (epochs, time points and channels). Then the data is split and z-scored using data from the training set.

The performance of the models is evaluated with the confusion matrix and a classification

report in order to assess the agreement between the expert's labels and the algorithms. The metrics employed in the classification report are:

- Accuracy: measures the percentage of cases that the model has got right. It is computed as the number of correct predictions divided by the total number of predictions. More specifically, it is the sum of the true positives (TP) and true negatives (TN) divided by TP, TN, false positives (FP) and false negatives (FN). It can be seen in Equation 4.1.

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \tag{4.1}$$

- Precision: answers the question of what proportion of positive identifications was correct. It is defined as follows:

$$Precision = \frac{TP}{TP + FP} \tag{4.2}$$

- Recall: also known as sensitivity or true positive rate. It allows to see how many positive instances the model was able to correctly identify. From a mathematical point of view, it is described as shown in Equation 4.3.

$$Recall = \frac{TP}{TP + FN} \tag{4.3}$$

- F1: the F1 value is used to combine the precision and recall measurements into a single value. This is handy because it makes it easier to compare the combined precision and recall performance between various solutions. It is calculated by taking the harmonic mean between precision and recall:

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{4.4}$$

4.5 Evaluation tests overview

The tests that will be conducted on each dataset (section 4.1) in order to characterize the performance of the networks, and that have been explained in the previous subsections (chapter 4:section 4.2, section 4.3, chapter 3: section 3.3, section 3.4), are summarized below. Table 4.1 contains a list of the tests that will be performed on the publicly available datasets, while Table 4.2 summarizes the tests that will be carried out on the dataset collected by *Bitbrain*. Each test will be identified with a number, which is shown in the first column. For each of them, the motivation for performing it is commented, as well as a brief description of it.

Test number	Motivation	Description
	DOD-H	
1	Test on single-channel information with electrode placement in frontal location (understand the role that electrode positions play)	Training and testing on F3-M2
2	Test on single-channel information with electrode placement in central location (understand the role that electrode positions play)	Training and testing on C3-M2
3	Test on single-channel information with electrode placement in occipital location (understand the role that electrode positions play)	Training and testing on O1-M2
4	Understand if combining multiple channels has a positive impact	Training and testing on F3-M2 + C3-M2 + O1-M2
5	Understand if combining multiple channels (not only EEG) has a positive impact	Training and testing on F3-M2 + C3-M2 + O1-M2 + + EOG + EMG + EKG
6	Offline testing with past and future context	Batch data in windows of 100 non-overlapping epochs (training and testing on F3-M2)
7	Test if weights from one dataset are transferable to other datasets	Transfer learning on DOD-O dataset
8	Get a balanced dataset and check if it yields improvement	Generate artificial samples of minority classes using different data augmentation methods (training and testing on F3-M2)
	DOD-O	
1	Test on single-channel information with electrode placement in frontal location (understand the role that electrode positions play)	Training and testing on F3-M2
2	Test on single-channel information with electrode placement in central location (understand the role that electrode positions play)	Training and testing on C3-M2
3	Test on single-channel information with electrode placement in occipital location (understand the role that electrode positions play)	Training and testing on O1-M2
4	Understand if combining multiple channels has a positive impact	Training and testing on F3-M2 + C3-M2 + O1-M2
5	Understand if combining multiple channels (not only EEG) has a positive impact	Training and testing on F3-M2 + C3-M2 + O1-M2 + + EOG + EMG + EKG
6	Offline testing with past and future context	Batch data in windows of 100 non-overlapping epochs (training and testing on F3-M2)
7	Test if weights from one dataset are transferable to other datasets	Transfer learning on DOD-H dataset
	ISRUC	
1	Test on single-channel information with electrode placement in frontal location (understand the role that electrode positions play)	Training and testing on F3-A2
2	Test on single-channel information with electrode placement in central location (understand the role that electrode positions play)	Training and testing on C3-A2
3	Test on single-channel information with electrode placement in occipital location (understand the role that electrode positions play)	Training and testing on O1-A2
4	Understand if combining multiple channels has a positive impact	Training and testing on F3-A2 + C3-A2 + O1-A2
5	Understand if combining multiple channels (not only EEG) has a positive impact	Training and testing on F3-A2 + C3-A2 + O1-A2 + + EOG + EMG + EKG
6	Offline testing with past and future context	Batch data in windows of 100 non-overlapping epochs (training and testing on F3-M2)
7	Test if weights from one dataset are transferable to other datasets	Transfer learning on DOD-O dataset
	STAGES	
1	Test on two channels information with electrode placement in frontal location. Data belonging to a single clinical center, folder name: BOGN	Training and testing on F3-M2 + F4-M1
2	Test on two channels information with electrode placement in frontal location. Data belonging to a single clinical center, folder name: GSBB	Training and testing on F3-A2 + F4-A1
3	Test on two channels information with electrode placement in frontal location. Data belonging to a single clinical center, folder name: GSDV	Training and testing on F3-A2 + F4-A1
4	Test on two channels information with electrode placement in frontal location. Data belonging to a single clinical center, folder name: GSLH	Training and testing on F3-A2 + F4-A1
5	Test on two channels information with electrode placement in frontal location. Data belonging to a single clinical center, folder name: GSSA	Training and testing on F3-A2 + F4-A1
6	Test on two channels information with electrode placement in frontal location. Data belonging to a single clinical center, folder name: GSSW	Training and testing on F3-A2 + F4-A1
7	Test on two channels information with electrode placement in frontal location. Data belonging to a single clinical center, folder name: MSMI	Training and testing on F3-A2 + F4-A1
8	Test on two channels information with electrode placement in frontal location. Data belonging to a single clinical center, folder name: MSNF	Training and testing on F3-A2 + F4-A1
9	Test on two channels information with electrode placement in frontal location. Data belonging to a single clinical center, folder name: MSQW	Training and testing on F3-A2 + F4-A1
10	Test on two channels information with electrode placement in frontal location. Data belonging to a single clinical center, folder name: MSTH	Training and testing on F3-A2 + F4-A1
11	Test on two channels information with electrode placement in frontal location. Data belonging to a single clinical center, folder name: MSTR	Training and testing on F3-A2 + F4-A1
12	Test on two channels information with electrode placement in frontal location. Data belonging to a single clinical center, folder name: STLK	Training and testing on F3-M2 + F4-M1
13	Test on two channels information with electrode placement in frontal location. Data belonging to a single clinical center, folder name: STNF	Training and testing on F3-M2 + F4-M1
14	Comprehend if there was overfitting to the smaller datasets (tests 1-13: training and testing with the same center)	Train using the subjects belonging to all the medical centers except one, and test with the remaining medical center
15	Understand if greater heterogeneity is paramount to ensure a high reliability and generalizability	Divide the centers proportionally into training/validation and test so that there is the same percentage of each of them in the three sets
16	Test if weights from one dataset are transferable to other datasets & check if a limited number of subjects restricts accuracy	Transfer learning / Fine-tuning on Bitbrain dataset

Table 4.1: Overview of evaluation tests performed on publicly available datasets.

Test number	Motivation	Description
	BITBRAIN	
1	Further comprehesion of the effect of the number of channels on the performance & undertsand the effects of applying different filter boundaries to the input data	Training and testing on Fp1 + Fp2 + AF7 + AF8 + T8 (data band-pass filtered between 0.5 and 45 Hz)
2	Further comprehesion of the effect of the number of channels on the performance & undertsand the effects of applying different filter boundaries to the input data	Training and testing on Fp1 + Fp2 + AF7 + AF8 + T8 (data band-pass filtered between 2 and 45 Hz)
3	Further comprehension of the effect of the number of channels on the performance	Training and testing on Fp1 + Fp2
4	Further comprehension of the effect of the number of channels on the performance	Training and testing on Fp1
5	Quantify the drop in accuracy that can be experienced when using wearable, self-managed devices & undertsand the effects of applying different filter boundaries to the input data	Training and testing on F3 + F4 + C3 + C4 + O1 + O2 (PSG, data band-pass filtered between 2 and 45 Hz)
6	Quantify the drop in accuracy that can be experienced when using wearable, self-managed devices & undertsand the effects of applying different filter boundaries to the input data	Training and testing on F3 + F4 (PSG, data band-pass filtered between 2 and 45 Hz)
7	Understand the effects of applying different filter boundaries to the input data	Training and testing on F3 + F4 + C3 + C4 + O1 + O2 (PSG, data band-pass filtered between 0.5 and 45 Hz)
8	Understand the effects of applying different filter boundaries to the input data	Training and testing on F3 + F4 (PSG, data band-pass filtered between 0.5 and 45 Hz)
9	Understand if providing head movement data improves yields improvement of sleep scoring on headband recordings	Training and testing on Fp1 + Fp2 + AF7 + AF8 + + T8 + IMU
10	Understand if providing head movement data improves yields improvement of sleep scoring on headband recordings	Training and testing on Fp1 + Fp2 + IMU
11	Understand if providing head movement data improves yields improvement of sleep scoring on headband recordings	Training and testing on Fp1 + IMU
12	Evaluate if learning from higher quality recordings helps to increase accuracy results when using the signals collected during the same night by a device with lower SNR	Train with the PSG data and test with the headband data (PSG: F3 + F4, headband: Fp1 + Fp2)

Table 4.2: Overview of evaluation tests performed on Bitbrain's data recordings.

5. Results and Discussion

This chapter shows and discusses the results obtained by using the methodology explained in chapter 3 to perform the tests described in chapter 4. In this way, the performance of the proposed architectures will be characterized in different scenarios (section 5.1). After that, the code will be integrated in the *Bitbrain*'s software platform in order to make it work in closed-loop, carrying out real-time decoding of EEG data by using a model previously trained (section 5.2).

5.1 Automatic sleep staging

The performance of the networks employed in this thesis are studied on DOD-H, DOD-O, ISRUC, STAGES and BITBRAIN datasets and visualised in the different Tables and Figures that are shown throughout the current section. Most of them display classification reports as well as confusion matrices containing more delailed information regarding each sleep stage. Until the first findings are made, the first part of each table and the first image of each pair correspond to the outcomes of the CNN architecture, while the latter characterize the performance of the CNN+RNN model.

5.1.1 Results on publicly available datasets: DOD-H, DOD-O & ISRUC

Tests 1-6 - Electrode placement, channels & presence/absence of future information

To begin, the results obtained for DOD-H dataset will be commented. The first tests that are presented (1-3) consist of assessing the models with single-channel information. First of all, the test on F3-M2 shows a global accuracy of 77.12% (see Table 5.1) with the CNN architecture, and of 75.08% employing the CNN+RNN architecture. On the other hand, the accuracy achieved using information from the C3-M2 channel are 77.93% and 75.55%, whereas that the results with the occipital location (O1-M2) are 67.54% and 64.49%, respectively for the CNN and CNN+RNN. Precision, recall and F1-score values are within a close range of accuracy in all the cases. Regarding the confusion matrices, it is clearly visible that the main confusion during the 3 tests is in N1 for both models. The sleep stage that shows the second worst accuracy is Wake in the case of the CNN (see Figure 5.1), and REM in the CNN+RNN (see Figure 5.2). Both N2 and N3 show very good values in terms of specificity, where true labels that are not N2 or N3 are rarely predicted as so. In N3, there is confusion with N2 but with practically no other phases. In the case of the O1-M2 channel, N2 is confused more with REM than with N3.

Results from combining multiple channels is assessed in tests 4-5. The combination of the three EEG channels (F3-M2 + C3-M2 + O1-M2) causes a slight improvement with both architectures (CNN: 79.86%, CNN+RNN: 77.99%). However, using EOG, EMG and EKG channels does not provide an increase in the algorithms performance, leading to worse results (CNN: 72.74%, CNN+RNN: 55.94%).
Regarding test 6, in which testing is done in an offline way, the highest accuracy is obtained in the case of the CNN, 81.43%. This result is remarkable since it is above inter-rater reliability. Nevertheless, for the CNN+RNN architecture the result is 76.82%, which is lower than the one obtained in test 4.

CNN				
Test	**Accuracy**	**Precision**	**Recall**	**F1-score**
DOD-H 1	0.7712	0.7923	0.7712	0.7668
DOD-H 2	0.7793	0.7965	0.7793	0.7798
DOD-H 3	0.6754	0.7304	0.6754	0.6733
DOD-H 4	0.7986	0.8102	0.7986	0.7952
DOD-H 5	0.7274	0.7573	0.7274	0.7297
DOD-H 6	0.8143	0.8193	0.8143	0.8111
CNN + RNN				
Test	**Accuracy**	**Precision**	**Recall**	**F1-score**
DOD-H 1	0.7508	0.7805	0.7508	0.7538
DOD-H 2	0.7558	0.7832	0.7558	0.7556
DOD-H 3	0.6449	0.6874	0.6449	0.6427
DOD-H 4	0.7799	0.7972	0.7799	0.7793
DOD-H 5	0.5594	0.6767	0.5594	0.5764
DOD-H 6	0.7682	0.7908	0.7682	0.7662

Table 5.1: Results of the classification report in DOD-H, for both the CNN and the CNN+RNN architectures.

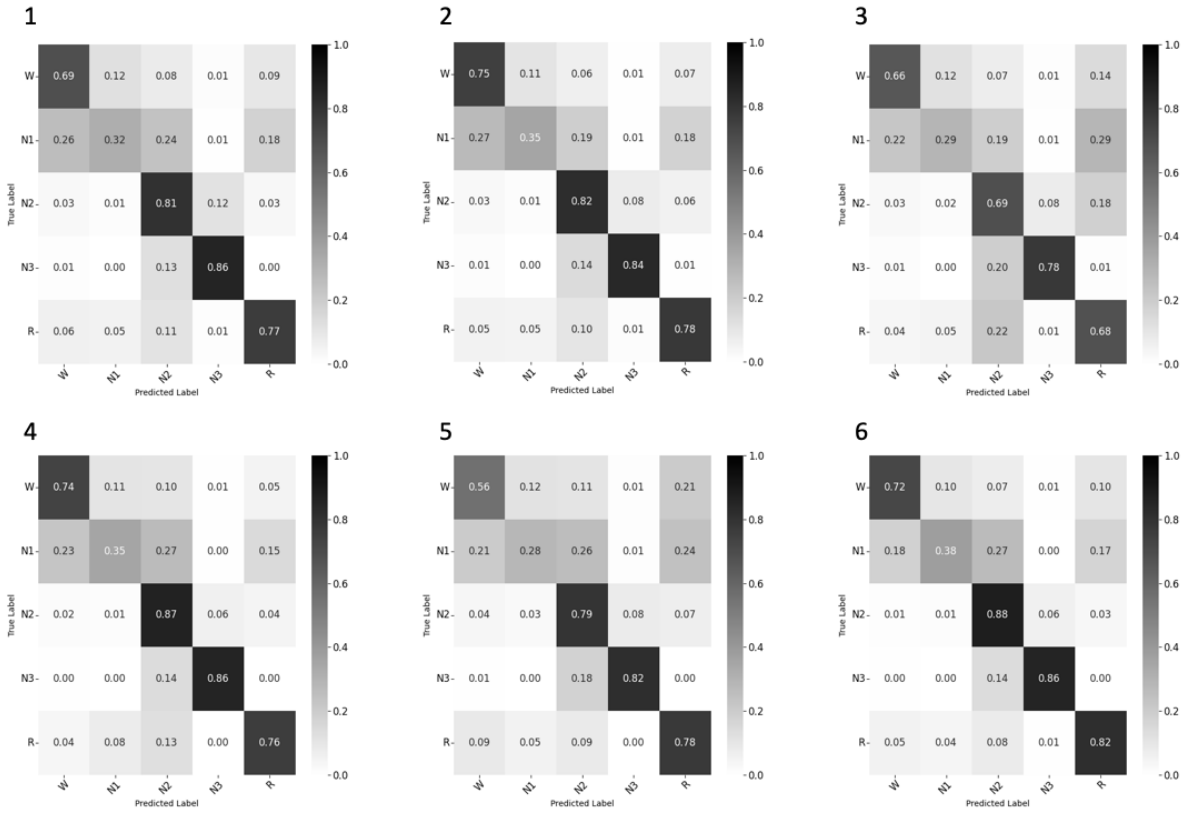


Figure 5.1: CNN results: confusion matrices for DOD-H tests 1-6 (the test number is indicated in the upper left corner of each matrix).

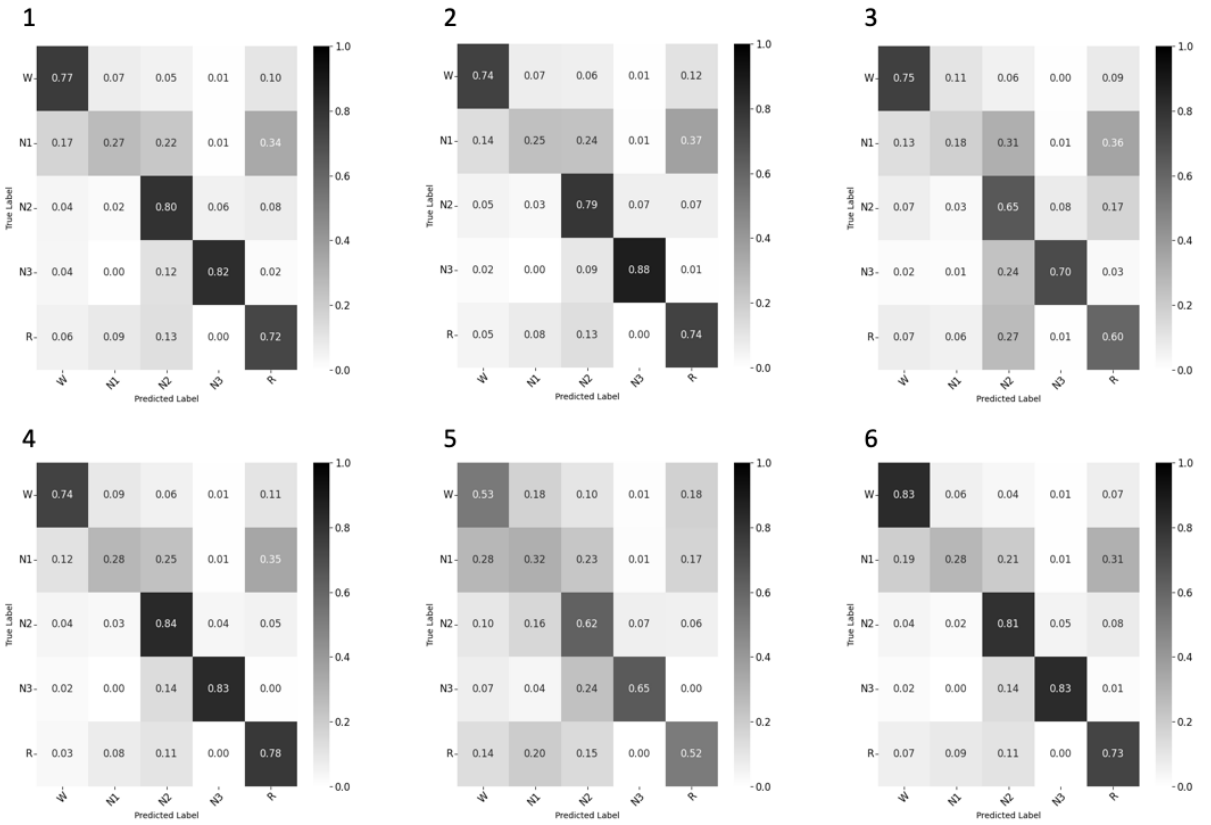


Figure 5.2: CNN+RNN results: confusion matrices for DOD-H tests 1-6 (the test number is indicated in the upper left corner of each matrix).

Next, the results obtained for the DOD-O dataset are described. It can be seen that the performance of both architectures is lower in all cases, compared to the previous dataset. As it happened before, precision, recall and F1-score values are within close range of accuracy.
First of all, the presented results will focus on the classification report (see Table 5.2), and more specifically on the CNN architecture. The three single-channel tests on F3-M2, C3-M2, O1-M2 show an accuracy of 75.63%, 74.97%, and 65.57%, respectively. Results from combining multiple channels are assessed in tests 4 (75.6%) and 5 (75.61%). They are practically the same and also very similar to those achieved with just one channel (frontal and central electrode position). The last test does not show any improvement over the first (75.13%).
In respect of the CNN+RNN architecture, the accuracies obtained for the three first tests are 70.91%, 71.84%, and 61.43%. In this case, combining multiple EEG channels improves the global accuracy, showing a value of 72.22%. However combining EEG, EOG, EMG, and EKG leads to impaired performance (64.01%) compared to the test employing the F3-M2 channel and to the one that uses the C3-M2 channel. Finally, testing offline (test 7) shows no increase resulting in a 70.65% of accuracy.
Regarding the confusion matrices (see Figure 5.3 and Figure 5.4), N1 has the lowest agreement between scorer and network. In particular, the worst outcome is found for the CNN+RNN model (test 3), where predictions only make up 11% of the total number of true labels. However, high agreement is reached at Wake, being even higher for the CNN+RNN model, where N1, N2, and REM show lower values compared to the CNN. Errors can be appreaciated between REM and

N2, and between N2 and N3.

CNN				
Test	**Accuracy**	**Precision**	**Recall**	**F1-score**
DOD-O 1	0.7563	0.769	0.7563	0.7539
DOD-O 2	0.7497	0.7688	0.7497	0.7467
DOD-O 3	0.6557	0.7075	0.6557	0.6573
DOD-O 4	0.756	0.7809	0.756	0.7544
DOD-O 5	0.7561	0.7746	0.7561	0.7554
DOD-O 6	0.7513	0.7807	0.7513	0.7441
CNN + RNN				
Test	**Accuracy**	**Precision**	**Recall**	**F1-score**
DOD-O 1	0.7091	0.7354	0.7091	0.7043
DOD-O 2	0.7184	0.7527	0.7184	0.7213
DOD-O 3	0.6143	0.6809	0.6143	0.6176
DOD-O 4	0.7222	0.7516	0.7222	0.7212
DOD-O 5	0.6401	0.7105	0.6401	0.6459
DOD-O 6	0.7065	0.7405	0.7065	0.7038

Table 5.2: Results of the classification report in DOD-O, for both the CNN and the CNN+RNN architectures.

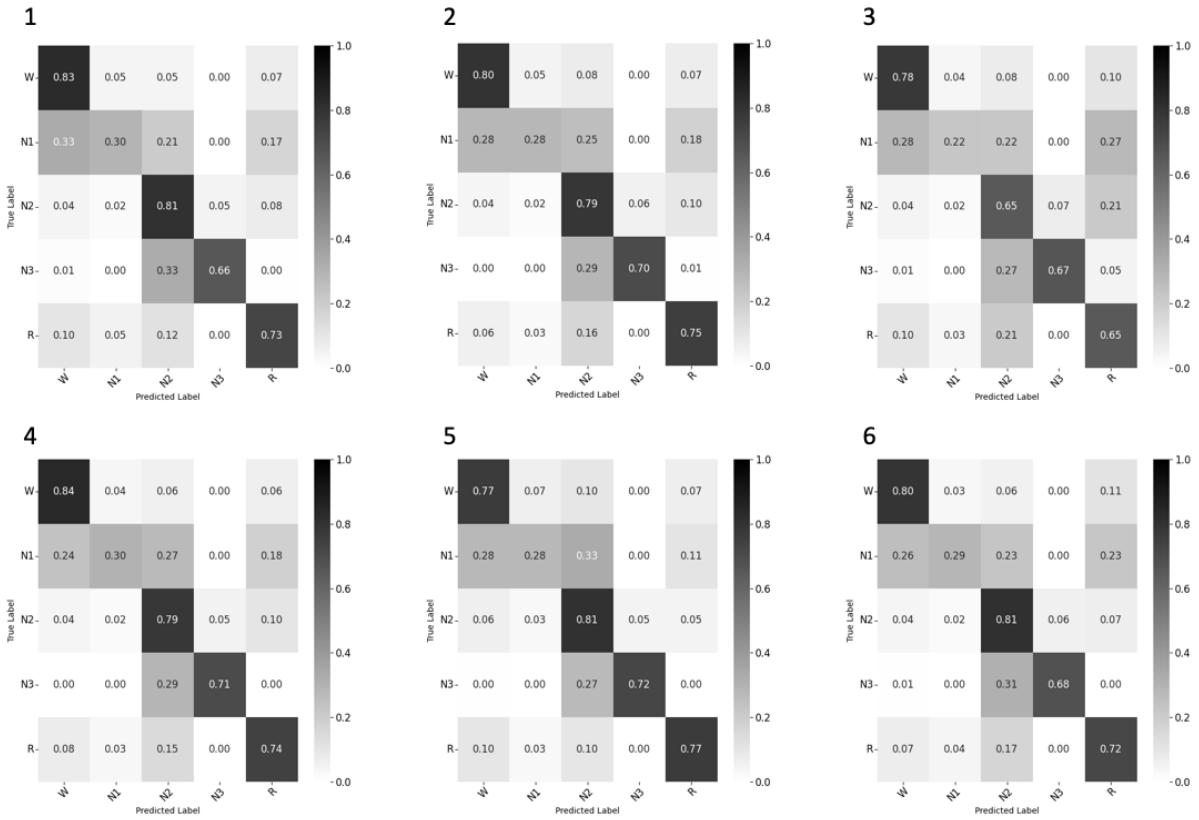


Figure 5.3: CNN results: confusion matrices for DOD-O tests 1-6 (the test number is indicated in the upper left corner of each matrix).

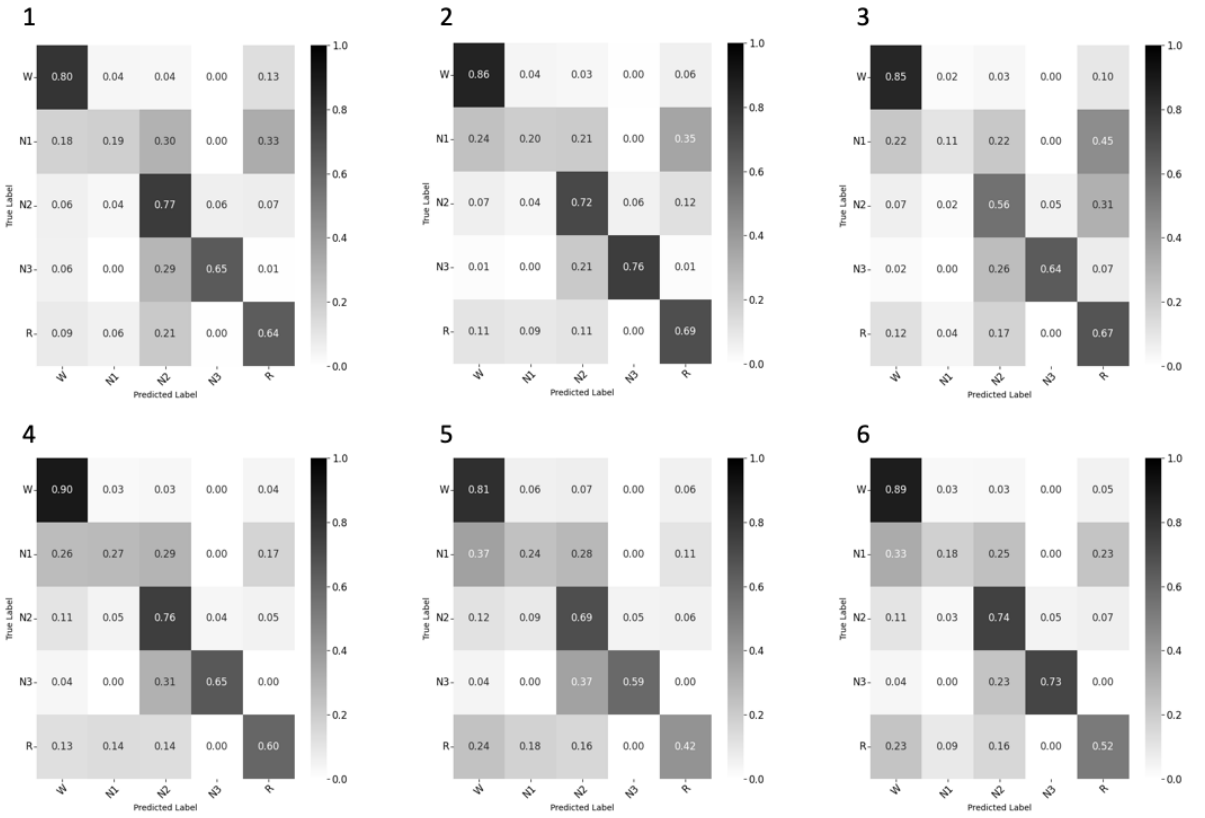


Figure 5.4: CNN+RNN results: confusion matrices for DOD-O tests 1-6 (the test number is indicated in the upper left corner of each matrix).

The same six tests that have been evaluated until now are performed for the last time on the ISRUC dataset. Using this data, both models show their worst performance.
Regarding the CNN, testing single-channel decoding yields an accuracy of 70.72% on the F3-A2 electrode, of 67.34% on the C3-A2 electrode, and of 64.73% on the O1-A2 electrode. In this case, using a higher number of channels improves moderately the accuracy value. The results are 71.84% and 75.73% for tests 4 and 5, respectively. The final test leads to a very similar result to the first.
When it comes to the CNN+RNN architecture, it can be seen that it achieves an accuracy of 62.27% on test 2, 4, and 5. The tests performed on F3-A2 show an accuracy of 59.9% and 61.63% for the online and offline testing, respectively. Lastly, the test using occipital electrode location results in 56.84%.
With respect to the confusion matrices, the same fact is repeated again: the main confusion is in N1. Nevertheless, the agreement is not so low in the case of the CNN+RNN, the erroneous predictions are more spread out among the different classes.

CNN				
Test	**Accuracy**	**Precision**	**Recall**	**F1-score**
ISRUC 1	0.7072	0.707	0.7072	0.695
ISRUC 2	0.6734	0.6888	0.6734	0.6618
ISRUC 3	0.6473	0.6451	0.6473	0.6275
ISRUC 4	0.7184	0.7225	0.7184	0.7083
ISRUC 5	0.7573	0.7605	0.7573	0.749
ISRUC 6	0.7025	0.7218	0.7025	0.6932
CNN + RNN				
Test	**Accuracy**	**Precision**	**Recall**	**F1-score**
ISRUC 1	0.599	0.695	0.599	0.6134
ISRUC 2	0.6227	0.6762	0.6227	0.6261
ISRUC 3	0.5684	0.6201	0.5684	0.5715
ISRUC 4	0.6227	0.694	0.6227	0.6253
ISRUC 5	0.6227	0.7064	0.6227	0.6359
ISRUC 6	0.6163	0.6874	0.6163	0.6317

Table 5.3: Results of the classification report in ISRUC, for both the CNN and the CNN+RNN architectures.

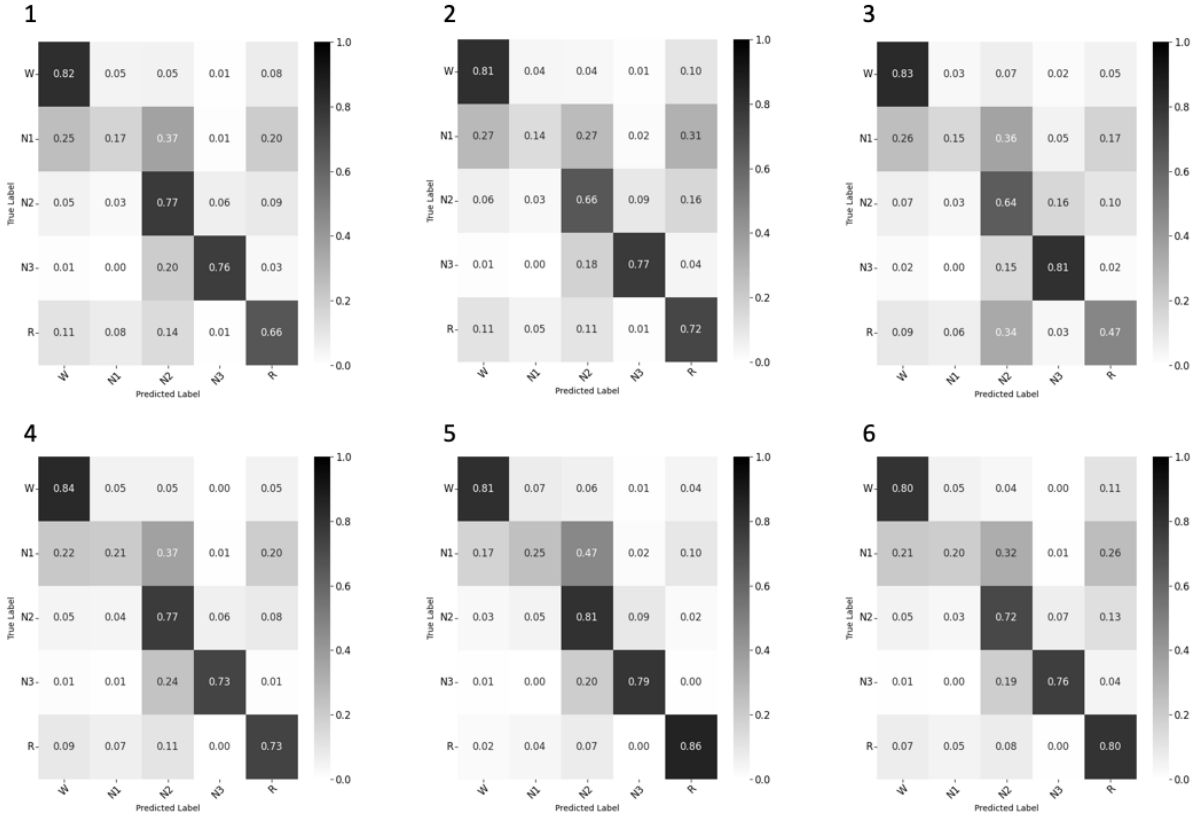


Figure 5.5: CNN results: confusion matrices for ISRUC tests 1-6 (the test number is indicated in the upper left corner of each matrix).

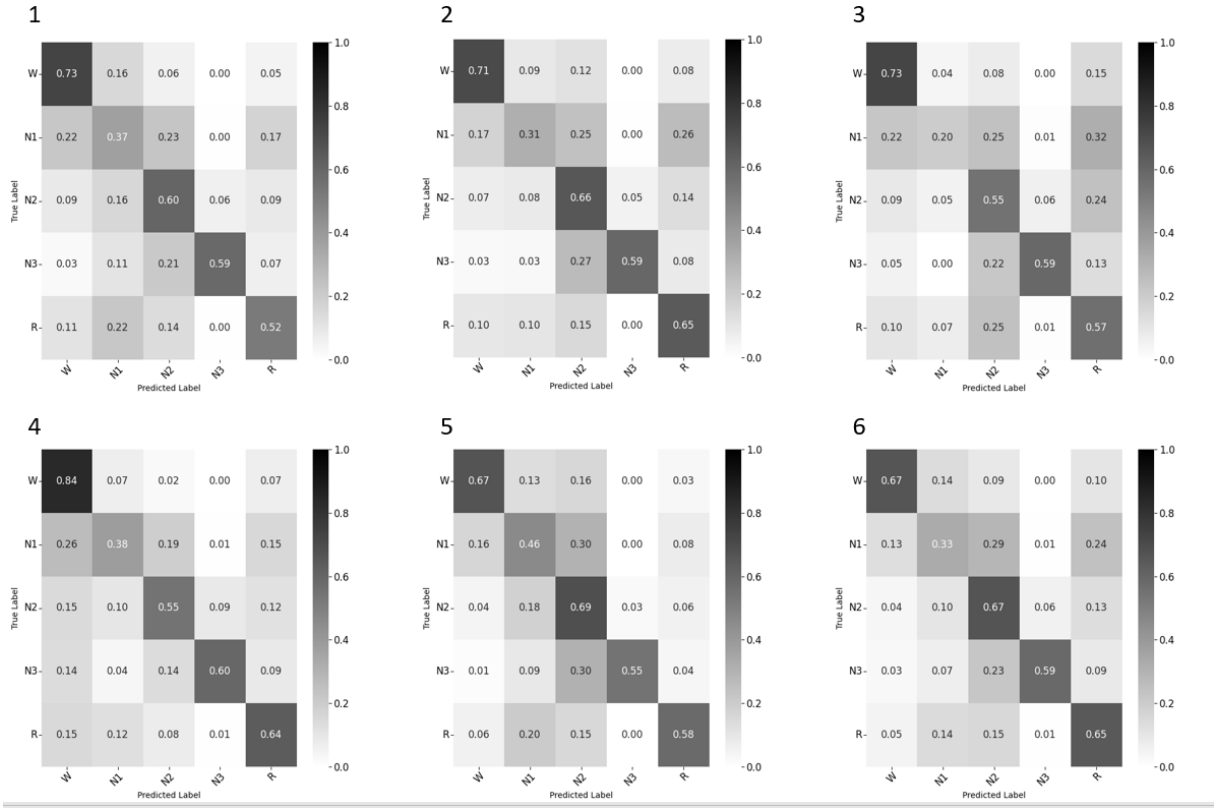


Figure 5.6: CNN+RNN results: confusion matrices for ISRUC tests 1-6 (the test number is indicated in the upper left corner of each matrix).

As it can be observed from the previous tables and figures, there are commonalities in the results obtained in each of the datasets. As it has been mentioned before, precision, recall and F1-score values are within a range very close to the accuracy. On the other hand, regarding the confusion matrices, the agreement between the labels predicted by the algorithm and those established by the experts decreases significantly for the N1 phase in all the analyzed cases. This is mainly due to the low number of events of this type that are available to train the model, since N1 only occupies 5% of the night. Furthermore, it is a transition phase between W and N2, which is normally labeled by experts taking into account both future epoch and EOG information. Since in most of the tests, with the exception of 5 and 6, the classification is carried out in real time without future context or EOG channels, the pronounced drop in accuracy that has just been commented on is observed. In most of the cases, specially for test 5 in which EOG sensors are used, an increase in the coincidence between true and predicted labels can be observed for the stage under consideration.

Another interesting aspect is the high specificity of N3 labels, which is practically only confused with N2. This is because the N3 phase is characterized by the slow oscillations (SOs) that also occur, albeit to a lesser extent, during N2. However, these slow waves with their peak power at around 0.8 Hz are only present in N2 and N3. Thereby, this fact explains why the confusion with other stages is so low in N3.

From the different tests carried out, different conclusions can be drawn to the questions that

were wanted to be resolved. First, the location of the EEG electrodes is clearly impacting scoring. The performance of both models is very similar when using frontal and central derivations, and always superior compared to that obtained with an occipital electrode. Accordingly, it can be stated that both frontal and central electrode locations are the optimal recording sites for staging, meaning that valuable hallmarks for sleep classification are better captured from the EEG derived from them.
To continue, the combination of EEG channels from the three derivations does not lead to a significant improvement in the results in any case. Therefore, the use of multiple EEG channels from different locations does not provide additional information useful to improve scoring outstandingly. This proves that even a single-channel approach is enough to differentiate between the different sleep stages. The same finding occurs when using additional sensors (EOG, EMG and EKG), which are routinely employed in clinical practice. This test, in all cases except for ISRUC dataset, even leads to worse results compared to single-channel setups. This may indicate that these channels do not add enough specific information related to sleep stages to increase the accuracy of the models, but rather introduce greater confusion into it.
On the other hand, comparing the results obtained in tests 1 and 6, it can be seen that the greatest difference in their results is when using the DOD-H dataset. The increase in accuracy when using past and future context is about 4% and 2% for the CNN and for the CNN+RNN, respectively. For the two other datasets, the change in accuracy is completely negligible. This is an indicator of the versatility of the architectures. As a consequence, it can be concluded that the lack of future information is not limiting decoding accuracy. Therefore, the models are suitable for achieving real-time scoring, matching the final objectives of the project. Consequently, the remaining tests will be performed with the online approach.

Therefore, from all the previous insights it is clear what information is crucial for staging and which is redundant. In fact, a single-channel EEG-based real-time setup, which constitutes the most limiting scenario in terms of available information, could be sufficient for decoding. This modality of staging can offer an increase in the subjects comfort as well as potential in clinical and research settings. Besides, future algorithms could also benefit from this insight by optimizing their design towards relevant futures.

Additionaly, comparing the results achieved for the different datasets, it can be observed that DOD-O and ISRUC, which are formed by patients with sleep disorders, lead to greater difficulties effectuating the classification. All the values shown for DOD-H (see Table 5.1), which is formed in its entirety by healthy subjects, are superior to those obtained in DOD-O (Table 5.2) and ISRUC (Table 5.3). This is because abnormal patterns of brain activity derived from sleep pathology are hindering training, making it less effective and resulting in the mentioned drop in classification performance.

Finally, a comparison can be made between the two deep learning techniques that have been proposed. As it can be observed, the CNN architecture is undoubtedly more effective performing sleep staging than the CNN+RNN. Its performance is higher for any test and recording modality. That being the case, some slight modifications are made to the CNN+RNN model, with the aim of improving its performance. With the motivations discussed in section 3.1 (subsection 3.1.2) a higher learning rate is employed and dropout is removed in the intermediate layers. In order to check the impact of these variations, test 1 is performed again using both DOD-H and DOD-O datasets. The results obtained can be seen in Table 5.4 (classification report) and Figure 5.7 (confusion matrix). There is an increase in accuracy of approximately 1% in both

cases. Specifically, the classes that improve their sensitivity are Wake, N3 and REM for DOD-O, and Wake, N2 and N3 for DOD-H. However, despite this small enhancement, the CNN model continues to be more distinguished and skillful at decoding.

Test	Accuracy	Precision	Recall	F1-score
DOD-H 1	0.7773	0.7972	0.7773	0.7707
DOD-O 1	0.7199	0.735	0.7199	0.7168

Table 5.4: Results of the classification report in DOD-H and DOD-O for the modified version of the CNN+RNN.

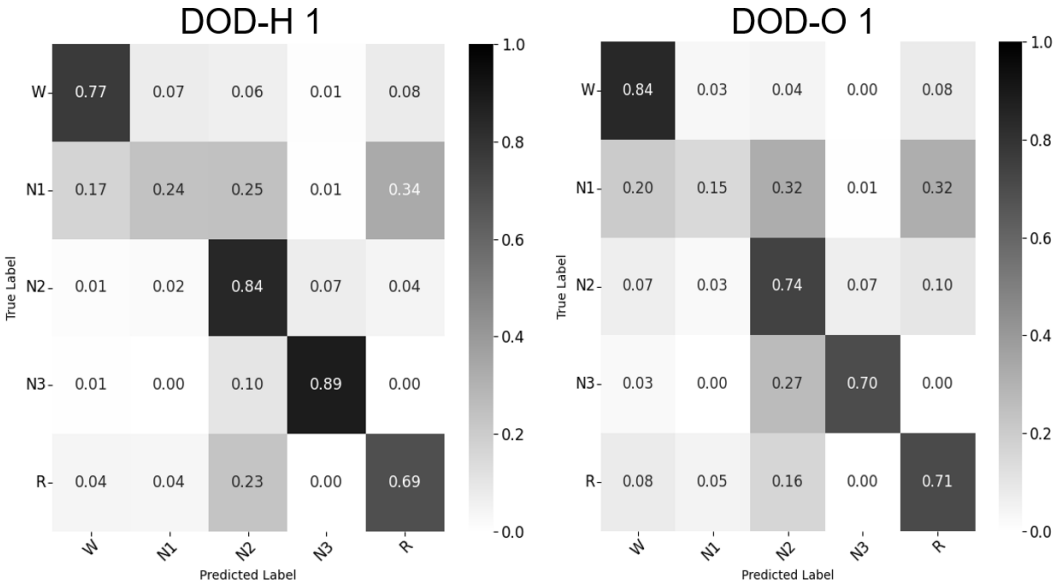


Figure 5.7: Modified version of CNN+RNN results: Confusion matrix for DOD-H and DOD-O test 1.

Therefore, and contrary to what was thought, the strong temporal component of the RNNs could be penalizing the classification. Besides, it is possible that the combination of a CNN and a RNN constitutes a model that is too complex for the data that is being handled, causing an over fit to them. Hence, the tests with the remaining datasets will be performed only with the purely convolutional model.

Finally, results regarding uncertainty quantification, data augmentation and transfer learning tests can be found in Appendix E, Appendix G, and Appendix H (section H.1), respectively.

5.1.2 Results on publicly available datasets: STAGES

Tests 1-13 - Single-sleep center

The current subsection starts showing the results obtained when training and testing employing data belonging to a single clinical center. These can be found in the classification report shown

in Table 5.5 and also in Figure 5.8. The latter, simply for a better visualization, displays a bar plot of the accuracy acquired for the 10-fold cross validation depending on the medical center to which the data used to train and test pertains. The confusion matrices showing specific values for each sleep stage can be found in Appendix D (Figure D.1 to D.13). As it was mentioned before, all of the following evaluations are only performed for the CNN architecture.

Looking at the extracted metrics it can be observed that all the accuracy values are between 67% and 77%, except for the first and last medical center in which the accuracy is of 80.08% and 43.64%, respectively. The first number surpasses inter-rater reliability, while the second one is close to the chance level, which was established at 30% in [40]. The latter is obtained for STNF center, comprising the smallest number of subjects. The results achieved for precision, recall, and f1-score are proximate to accuracy.

Taking into account the unusual succession of sleep phases and the high number of obstructive apneas and hypopneas that occur throughout the night, practically all the results are more than acceptable. This is due to the fact that data is recorded on patients suffering from sleep disorders. Nevertheless, inspecting the confusion matrices, the low specificity of the N3 class is surprising. In some cases such as GSDV and GSSA there is no agreement between scorer and network for this class, always being confused with N2. As a consequence, the sleep architecture is computed for each directory/medical center (see Appendix D, Figure D.14) and it is noted that the percentage of N3 stage is much lower than usual (it is not higher than 10% in any case and lower than 1% in 4 of them). Therefore, the previous result is something to be expected given the low number of training examples that the model has for that class.

Test	Accuracy	Precision	Recall	F1-score
STAGES 1 - BOGN	0.8008	0.8097	0.8008	0.799
STAGES 2 - GSBB	0.7615	0.7725	0.7615	0.7505
STAGES 3 - GSDV	0.759	0.7574	0.759	0.7431
STAGES 4 - GSLH	0.7504	0.7551	0.7504	0.7392
STAGES 5 - GSSA	0.6896	0.7408	0.6896	0.6654
STAGES 6 - GSSW	0.7203	0.711	0.7203	0.7
STAGES 7 - MSMI	0.7316	0.7222	0.7316	0.7133
STAGES 8 - MSNF	0.6774	0.6993	0.6774	0.6612
STAGES 9 - MSQW	0.6921	0.6955	0.6921	0.6791
STAGES 10 - MSTH	0.7389	0.7436	0.7389	0.7175
STAGES 11 - MSTR	0.6984	0.7115	0.6984	0.6893
STAGES 12 - STLK	0.7238	0.7338	0.7238	0.7166
STAGES 13 - STNF	0.4364	0.4975	0.4364	0.418

Table 5.5: Results of the classification report in STAGES for each clinical center.

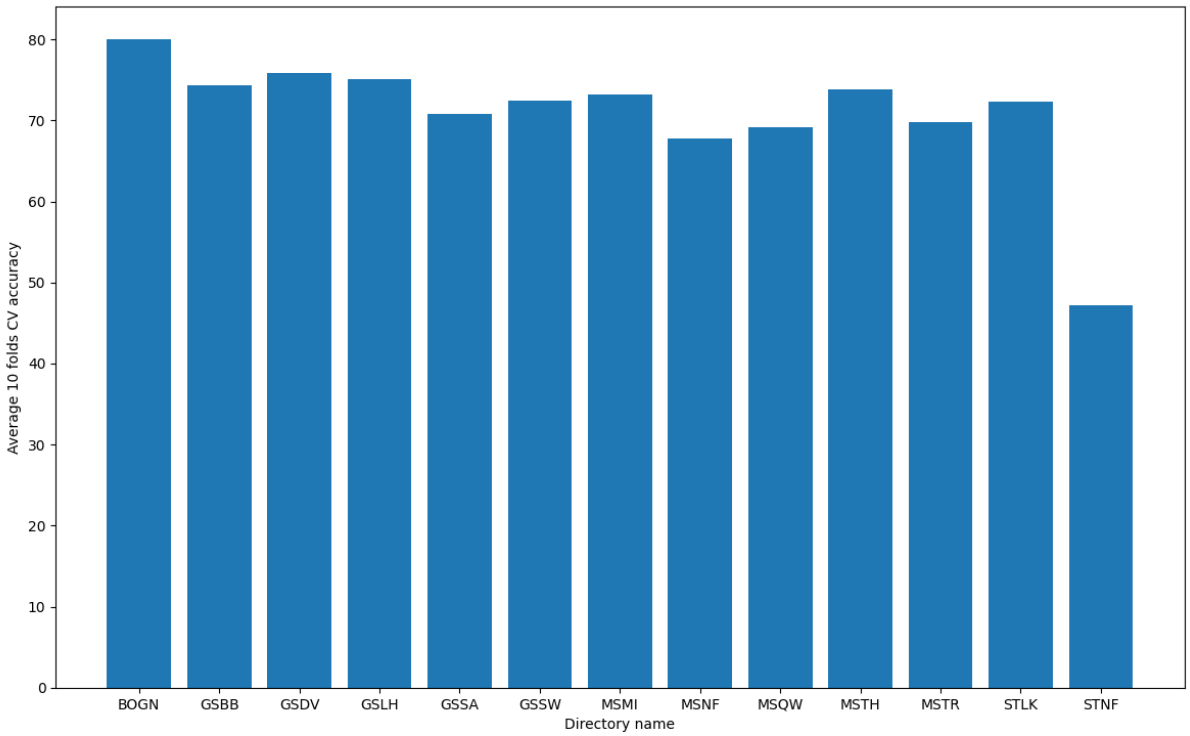


Figure 5.8: Accuracy per STAGES clinical center.

Tests 14-15 - Massive dataset

Once the data from each center has been used to train and test in an individual manner, it is time to build the massive dataset by combining all of them.
The first approach to accomplish this consists of joining the subjects from every directory except one, which will make up the test set. Therefore, a 13-fold cross validation is performed in this test. The accuracy achieved in each of those can be seen in Figure 5.10, in which it is indicated the specific clinical center that was employed as test set in each fold. In most of the cases, the accuracy is under the result achieved when using that center for both training and testing. This is specially noticeable for the MSTR clinical center, which experiences a drop in accuracy of approximately 15%. Accordingly, the power of the algorithm is compromised by the characteristics of the underlying dataset, meaning that the high performance could be indicating overfitting to the homogeneous and smaller sample sizes. The global accuracy acquired is of 66.96% (see Table 5.6), value which is negatively influenced by the fold in which STNF constitutes the test set. Each of the folds performed with this technique can be seen as a transfer learning to the specific center being tested.
The second approach consists of dividing all the centers proportionally into training and testing so that there is the same percentage of each of them in the three sets. A 10-fold cross validation is performed in this case. This test shows an accuracy of 69.5%, result higher than the previous one by 2.54% (see Table 5.6). Subsequently, making the sets more independent and heterogeneous increases the reliability and generalizability of the algorithm.
These results are interesting since training on data from a single-sleep center or population can

reduce the ability of the algorithms to generalize to other recording systems and/or populations, resulting on high accuracies just for the small training datasets.
Regarding the confusion matrices displayed in Figure 5.9, they show common patterns for both tests. As it was mentioned before, N3 is rarely predicted by the algorithm since there are very few examples from such sleep stage. Thereby, most of the errors come from labelling as N2 epochs that actually correspond to N3. There are also inaccuracies concerning N1, which is the second minor class in this dataset. However, the confusions are more spread in this case, being distributed among Wake, N2 and N3. On the other hand, the agreement between scorers and network is high for the remaining classes, and specially for N2.

Test	Accuracy	Precision	Recall	F1-score
STAGES 14	0.6696	0.6868	0.6696	0.6524
STAGES 15	0.695	0.6845	0.695	0.667

Table 5.6: Results of the classification report in STAGES for the entire dataset.

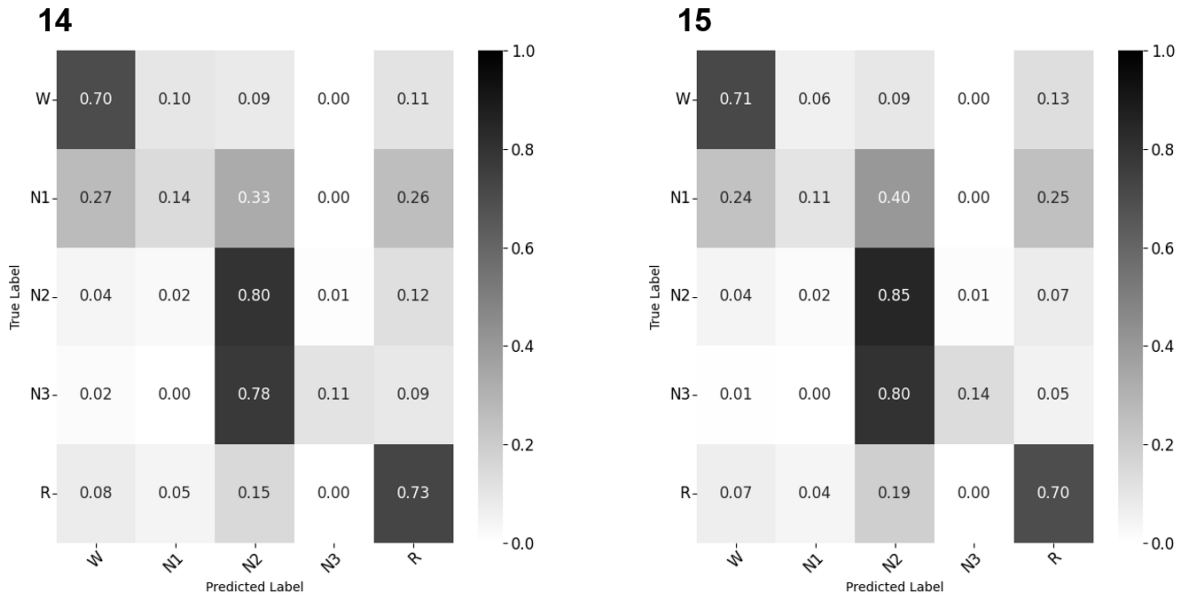


Figure 5.9: Confusion matrices for STAGES tests 14-15 (the test number is indicated in the upper left corner of both matrices).

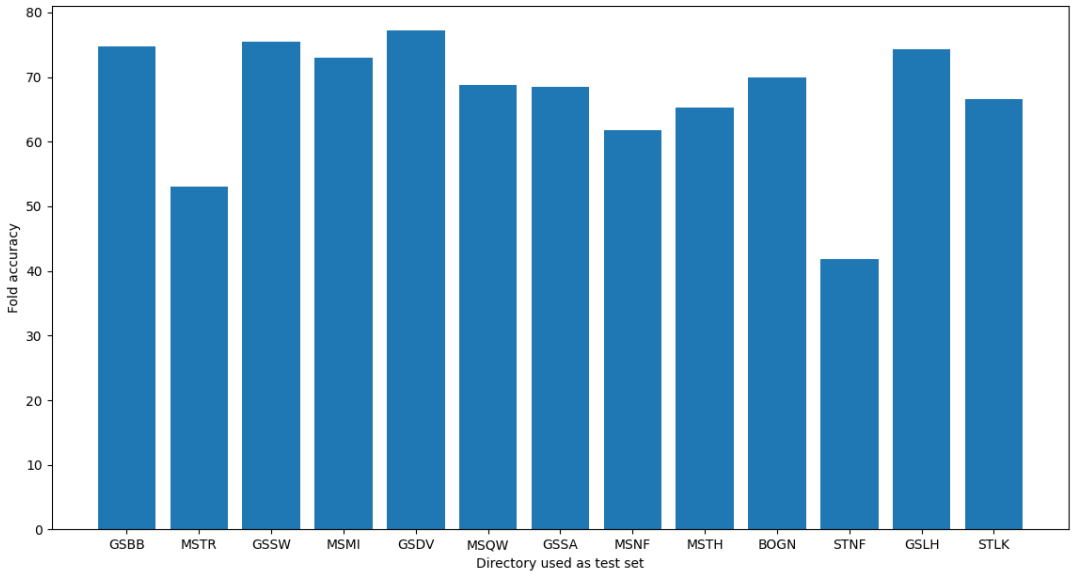


Figure 5.10: Accuracy per fold throughout the 13-fold cross validation performed in test 14. Each medical center constitutes the test set of each fold.

5.1.3 Results on Bitbrain's data recordings

Tests 1-4 - Headband: channels & filter boundaries

The results obtained for BITBRAIN dataset are commented and displayed from here on. First of all, different experiments are performed with the EEG channels from the headband. The results concerning the first 4 tests are shown in Table 5.7 and Figure 5.11. Regarding test 1 and 2, it can be seen that the performance of the network is better when the recordings are filtered between 2-45 Hz rather than 0.5-45 Hz. This is due to the fact that the signals are highly contaminated with sweat artifacts, which manifest as low frequency noise. With the high-pass at 2 Hz, the slow oscillations that characterize NREM sleep (more specificially N2 and N3) are filtered out. A human expert would not be able to score correctly without such information. However, the algorithm manages to do so even with greater accuracy at any stage. Consequently, for tests 3 and 4 (and the next ones employing the headband), the data is band-pass filtered between 2 and 45 Hz.
In this case, comparing tests 2, 3, and 4, the performance achieved with 5 (test 2: Fp1, Fp2, AF7, AF8 and T8) or 2 electrodes (test 3: Fp1 and Fp2) is very similar. Improvements occur in Wake and N3 when employing double-channel information. Nevertheless, and although the difference is small, the results are poorer with the single-channel approach (test 4: Fp1). This is specially noticeable in N3, which is predicted as N2 42% of the time.
The commented results are quite good considering the bad quality of the signal and the high content of artifacts in practically all the subjects. This is due to the fact that the band with which the recordings were made was an initial version of a prototype under evaluation. These rehearsals serve for later modifications and improvements of it.

Test	Accuracy	Precision	Recall	F1-score
BITBRAIN 1	0.7083	0.7221	0.7083	0.7
BITBRAIN 2	0.7344	0.7463	0.7344	0.7271
BITBRAIN 3	0.7392	0.7499	0.7392	0.7341
BITBRAIN 4	0.6996	0.7094	0.6996	0.6865

Table 5.7: Results of the classification report for the Bitbrain's headband data.

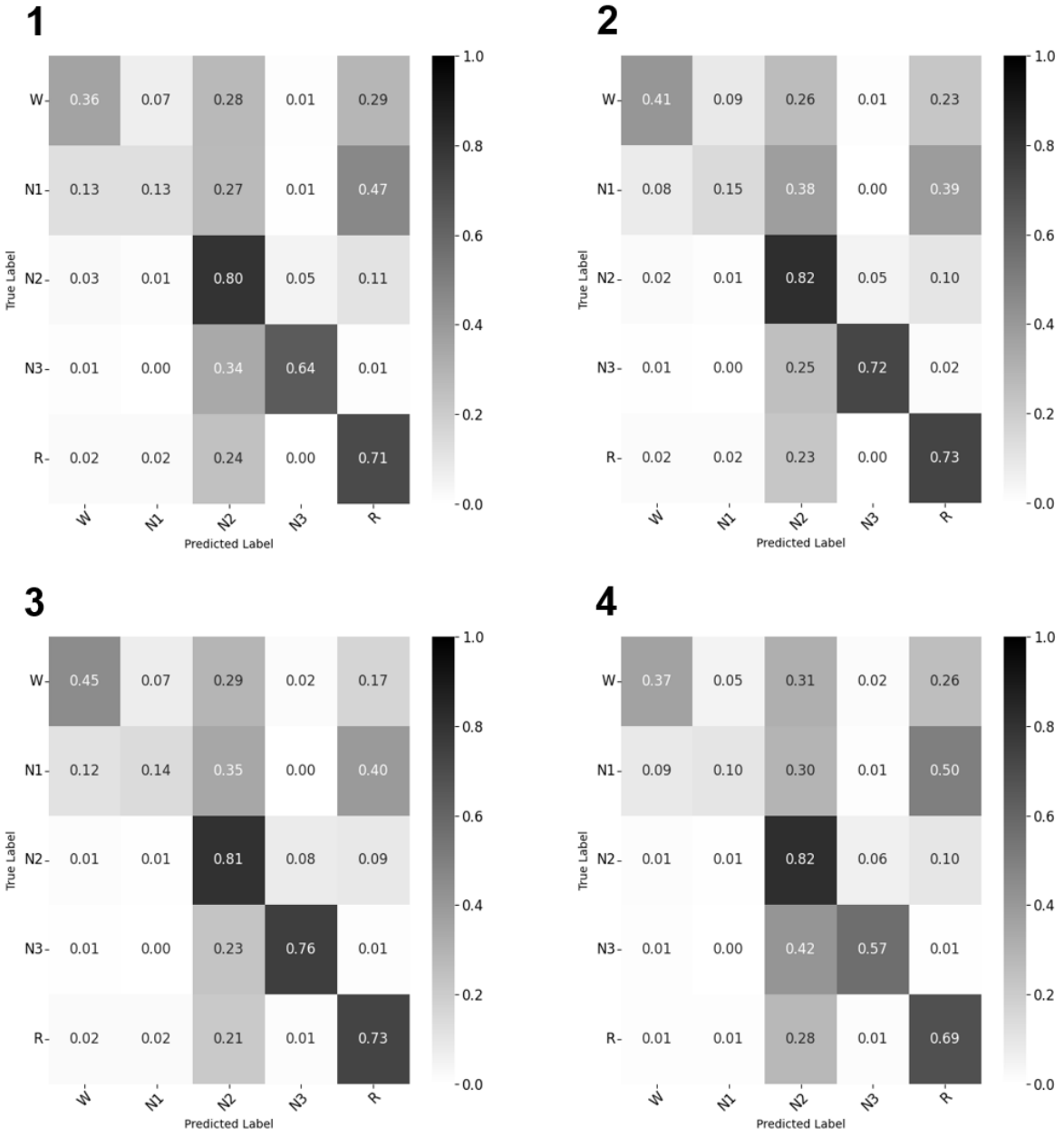


Figure 5.11: Confusion matrices for BITBRAIN dataset tests 1-4 (the test number is indicated in the upper left corner of the matrices).

Tests 9-11 - Headband: head movement data

Table 5.8 and Figure 5.12 show the performance of the sleep scoring algorithm on the headband recordings when head movement data is also provided. Comparing tests 2-9 (five EEG channels, five EEG channels + IMU), 3-10 (two EEG channels, two EEG channels + IMU), and 4-11 (single-channel, single-channel + IMU) it can be seen that, despite containing information about the sleep stages (e.g., more acceleration peaks during Wake and light NREM (section 3.4)), the movement data does not improve the accuracy of the classifier (the drop in accuracy is of 0.6%, 4.26% and 1.49%, respectively). The reason behind this could be that movement information is already reflected in the EEG in the form of movement artifacts. The additional movement data would thus be redundant.

Test	Accuracy	Precision	Recall	F1-score
BITBRAIN 9	0.7284	0.7357	0.7284	0.7225
BITBRAIN 10	0.6966	0.7093	0.6966	0.6944
BITBRAIN 11	0.6847	0.7132	0.6847	0.6818

Table 5.8: Results of the classification report for the Bitbrain's headband recordings providing head movement data.

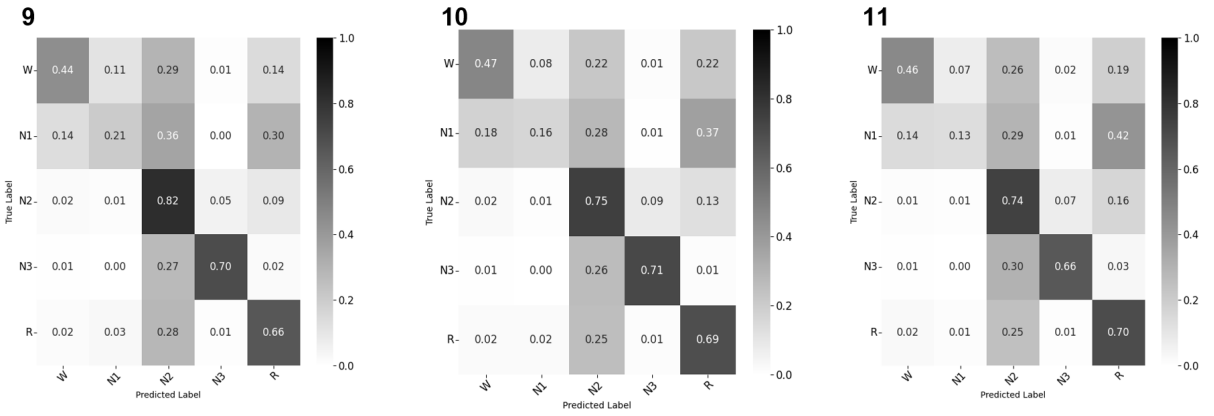


Figure 5.12: Confusion matrices for BITBRAIN dataset tests 9-11 (the test number is indicated in the upper left corner of the matrices).

Tests 5-8 - PSG

The results achieved using the PSG channels are shown in Table 5.9 and in Figure 5.13. For both filter settings, employing information from two frontal electrodes (F3, F4: tests 6 (2-45 Hz) and 8 (0.5-45 Hz)) leads to finer outcomes than using the 6 available EEG channels (F3, F4, C3, C4, O1, O2: tests 5 (2-45 Hz) and 7 (0.5-45 Hz)). These results with just two channels in frontal derivations are above inter-rater reliability. Furthermore, the performance in this case is better when the data is filtered between 0.5-45 Hz (tests 7 and 8, 6 and 2 EEG channels, respectively). This is because the PSG recordings are not contaminated with sweat artifacts that the 2-45 Hz filter would alleviate, and the SOs can be detected with the 0.5 high-pass cutoff. In fact, looking at the confusion matrices, the biggest difference is in the increased accuracy of the N3 labels,

which are less confused with N2.
Comparing the results of the PSG with those of the headband there are appreciable improvements concerning Wake and N1 sleep stages. The difference in accuracy between tests 2 and 5 (2-45 Hz, all the channels) is of 35% and 12% for Wake and N1, respectively. Regarding tests 3 and 6 (2-45 Hz, two channels) the drop in accuracy that occurs using the headband is of 24% for Wake and of 16% for N1. This may be important, because it implies that there is considerable scope for improvement in Wake. Therefore, it may make sense to use features such as IMU, which correlate well with Wake-Sleep. However, the IMU signals that were recorded with the band, and which lead to the only features available at the moment, are not adequate.

Test	Accuracy	Precision	Recall	F1-score
BITBRAIN 5	0.7844	0.8019	0.7844	0.7817
BITBRAIN 6	0.8129	0.8226	0.8129	0.808
BITBRAIN 7	0.7874	0.8076	0.7874	0.7907
BITBRAIN 8	0.8268	0.8363	0.8268	0.8237

Table 5.9: Results of the classification report for the Bitbrain's PSG recordings.

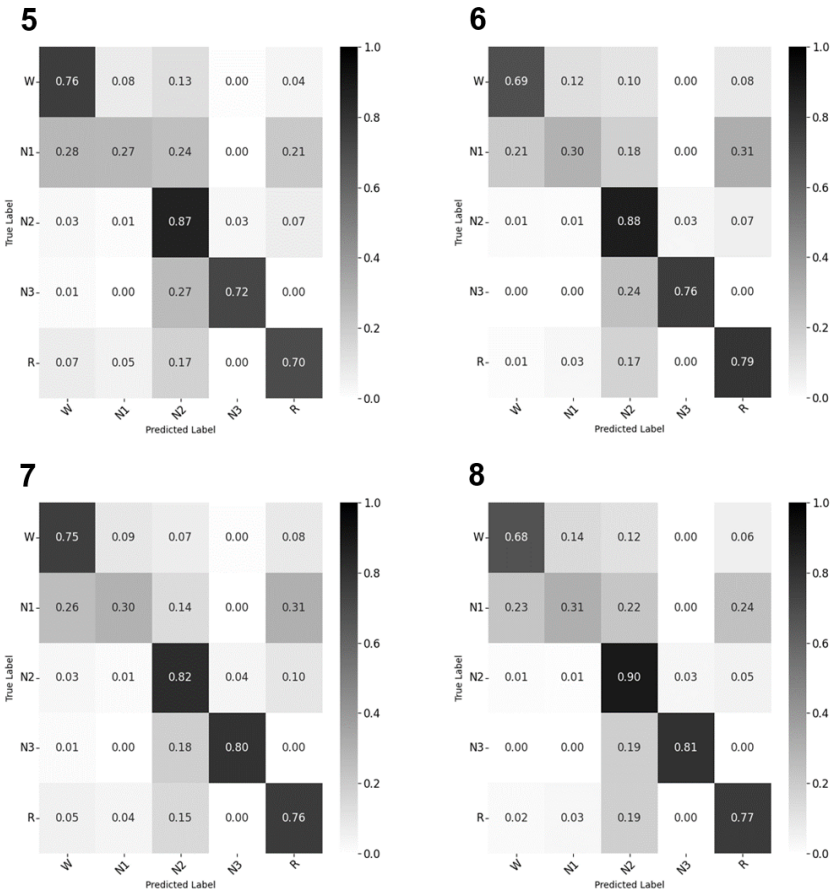


Figure 5.13: Confusion matrices for BITBRAIN dataset tests 5-8 (the test number is indicated in the upper left corner of the matrices).

To conclude, the results obtained for the transfer learning tests can be seen in Appendix H (section H.2).

5.2 Real-time staging - Bitbrain interface coupling

Finally, in order to make the algorithm work in real-time with *Bitbrain*'s EEG devices, the code is integrated in the company's software platform, which is developed in C++ and Javascript. This step is necessary to allow the decoder to work in closed-loop applications. For that purpose, data is sampled at 256 Hz and filtered using a causal forward-in-time technique, which cannot be zero phase (i.e., shifts the signal as it filters). This data is handled as it comes in blocks of 8 in 8 samples. Each time a new block arrives, it is introduced into a circular buffer of 30 seconds size (960 epochs, 7680 samples). This leads to a "warm-up" period until 960 epochs of EEG information have been acquired. After this, all epochs are inputted into the previously trained CNN model, which outputs its first predictions. From this point on, each new epoch of 8 samples is added and the oldest one is discarded. In this way, a constant length of 30 seconds is maintained. Every time a new second of data is available (32 epochs, 256 samples), the model predicts a new label.

Note that, in the just explained manner, the predictions are made based on the previous 30 seconds, instead of looking 15 seconds ahead and 15 seconds back as the experts do. The reason behind all this is disclosed below.

The decoder will be incorporated, together with a slow wave detector, as part of a real application. We are currently developing this system, which will be used to apply auditory stimulation during sleep. The main goal of this technique is to reduce the age-associated cognitive deterioration and memory loss in elderly people. This can be achieved since that type of stimulation evokes physiological responses in the brain, including slow waves, that are vital for many important cognitive processes ([47]). For a detailed description, see Appendix J.

A very important detail of this approach is that auditory stimulation can only have the desired benefits if applied during deeper NonREM sleep, more specifically the so-called N2 and N3 sleep phases. This is because stimulation in other sleep stages can more easily wake the person (N1) or will not have the desired neurophysiological effect (REM sleep). Accordingly, the automatic sleep scoring algorithm will be in charge of providing an input signal to the slow wave decoder when it detects the desired sleep stages. This input signal will allow the slow wave decoder to start working. If it finds a slow wave, the stimulation is triggered in order to generate new slow waves that would not occur naturally. The sleep stage decoder cannot score every 30 seconds, since many stimulation opportunities could be lost and the interesting thing is to maximize them. This is why it does it every second.

As a consequence of all of the above, the technique must be applied in a real-time closed-loop manner, in which the recorded brain activity is continuously analyzed and used to synchronize the stimulation (see Figure 5.14).

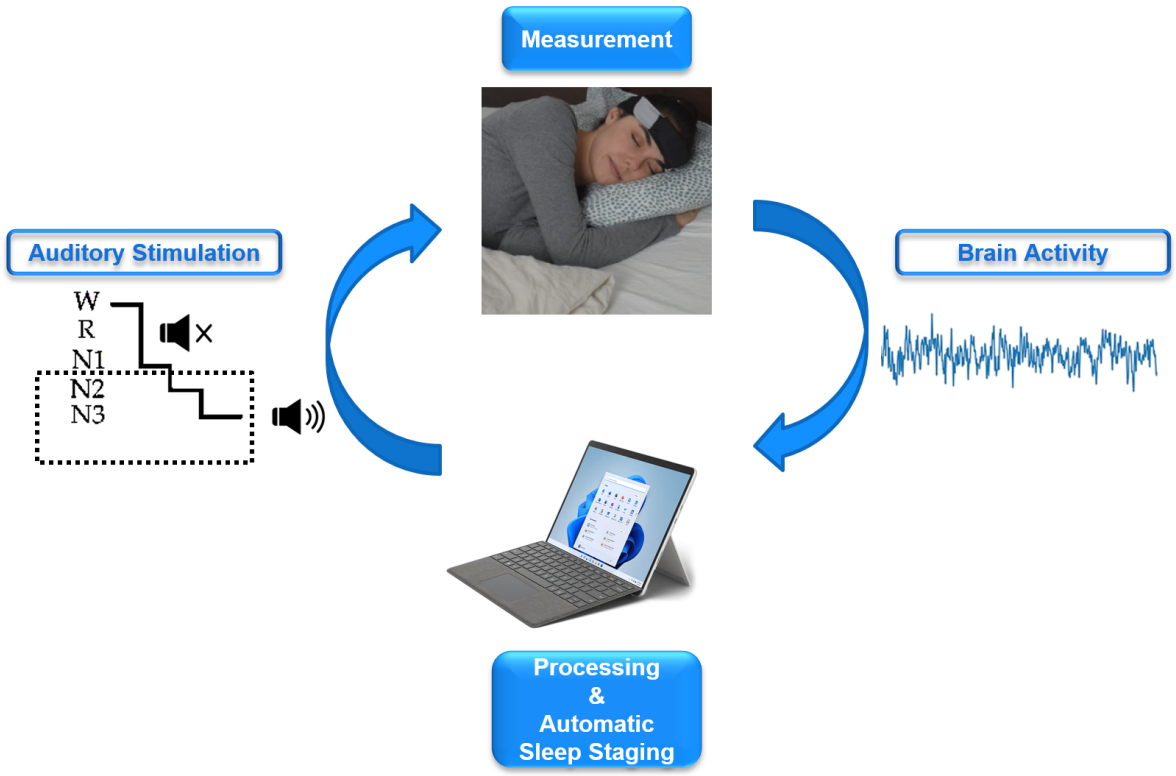


Figure 5.14: Closed-loop schematic. EEG activity is recorded with the headband and passed onto a surface contaning pre-processing scripts. Then, the data is scored in real-time by the automatic sleep staging algorithm. Depending on the outputted sleep stage, the slow wave detector is enabled. Tones are delivered when those are found by a second decoder, in order to hit the positive half of the oscillation.

A nap test has already been performed with the basic hard- and software setup available so far. However, before using this version of the decoder adapted to the platform, it is desired to train the CNN model using data recorded with an improved design of the headband, which will be the one employed for these studies (this headband has already been manufactured and the first recordings are beginning. Remember that the one used during the analysis of this thesis was a prototype under evaluation). Consequently, in this test, the slow wave decoder is started if enabled manually by a experimenter.

Figure 5.15 displays a screenshot of the platform during the mentioned test. The blue shaded area indicates the information that would be stored in the circular buffer at the time the model would predict N2 or N3: 21 seconds that are still shown in the screen (green rectangle) and 9 seconds that have already passed (black rectangle). This prediction would activate the slow wave decoder, which founds a slow wave and plays a sound (vertical white line on the right). Therefore, this figure tries to clarify the procedure that will be followed after developing the complete application, but also demonstrates that the method actually works. It can be seen how the brain of the sleeping participant responded to the the delivered tone with a slow oscillation (red mark).

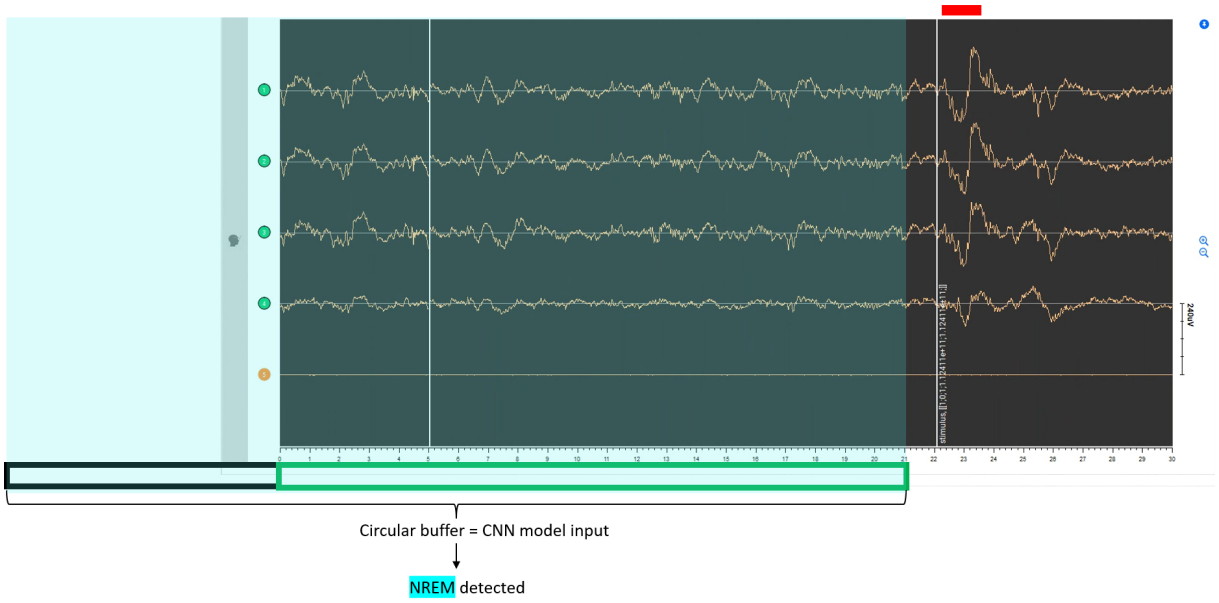


Figure 5.15: Screenshot of *Bitbrain*'s software platform: the release of a tone results in the creation of a slow oscillation.

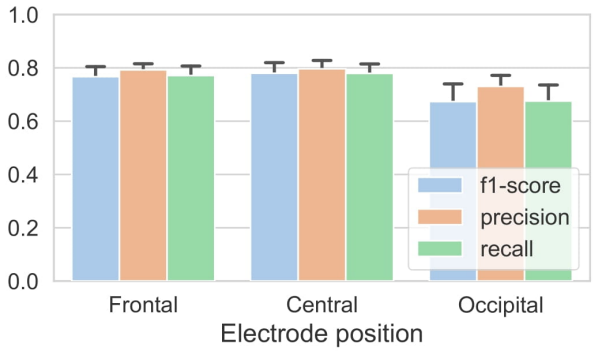
6. Conclusions

This master's thesis has successfully designed and developed the algorithm that exploits data from a novel, wearable EEG system that aims at revolutionizing sleep monitoring, allowing easy recordings as well as reducing costs and individuals' discomfort. This algorithm is already functional and capable of scoring in real-time.

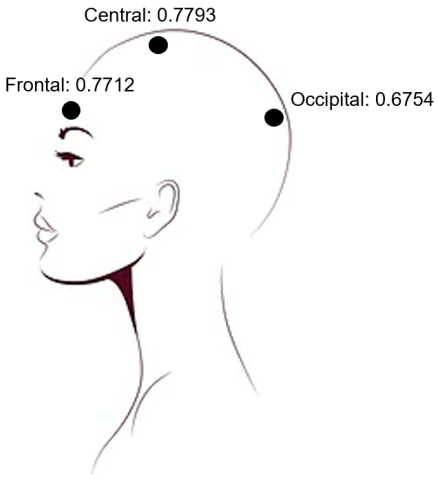
For that purpose, the existing literature around current ML approaches, which use a high number of channels and score in a purely offline way, has been explored and studied. In particular, it has been seen that DL has produced meaningful and promising results, overcoming those achieved with conventional ML algorithms.

Given that DL is concentrating all the attention in the processing and analysis of EEG signal in automatic sleep scoring tasks, two DL-based approaches are implemented and presented in the project. Specifically, these architectures consist of a CNN and a CNN+RNN. Moreover, signal processing and filtering techniques are applied to their input data to reduce noise and artifacts, keeping the significant information.

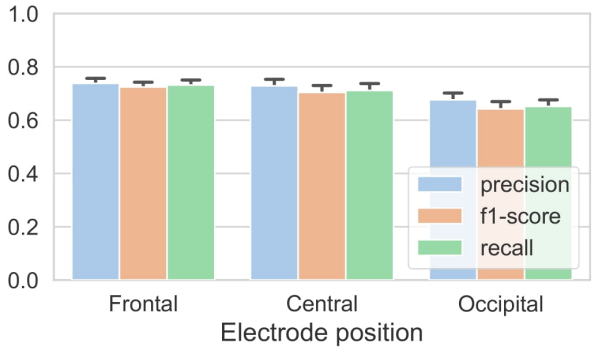
The performance of the developed models is studied on 4 public datasets. This has allowed them to be evaluated and characterized on various sample sizes, including both healthy individuals and patients, as well as on different electrode positions and number of channels. Furthermore, using these recordings, it is assessed the impact of not using future context for scoring. By way of conclusion, a graphical summary has been made based on the results obtained for these data. Concretely on those considered relevant to develop the desired functional neurotechnology, and achieved with the approach that has been shown to provide better results in a considerable number of tests.
The overview is shown in Figures 6.1, 6.2, and 6.3. All of them distinguish between healthy subjects (DOD-H): a), b), and patients (DOD-O + ISRUC): c), d). Electrode placement (see Figure 6.1) and different combinations (see Figure 6.2) clearly impacts performance. Averaging both cohorts, frontal locations are working best. Furthermore, additional channels such as EMG, EOG and ECG are introducing more confusion, resulting in a decreased performance. As a consequence, it is concluded that the developed algorithm is capable of practically reaching the level of experts using only frontal EEG channels. Those are the ones desired to maximize comfort with wearable devices. This is achieved even with single-channel information.
Additionally, it is important to notice that the system has reached those results scoring in real-time. In fact, the performance drop experienced in absence of future context is very small (see Figure 6.3). This demonstrates the validity of the network to work in closed-loop.



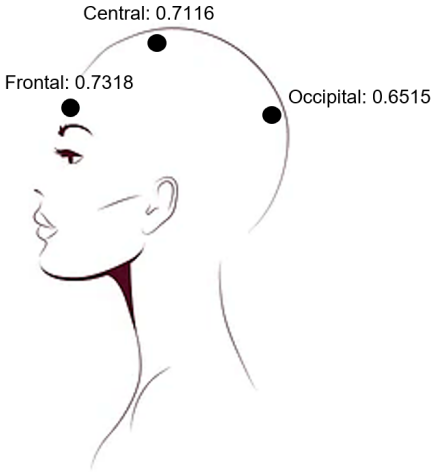
(a) Healthy subjects (DOD-H): Barplot of precision, recall, and F1 score.



(b) Healthy subjects (DOD-H): Accuracy value.

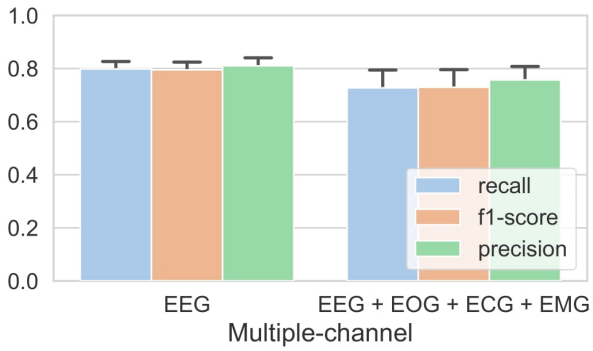


(c) Patients (DOD-O and ISRUC): Barplot of precision, recall, and F1 score.



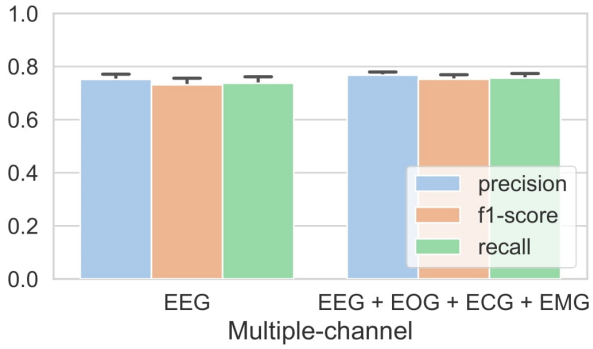
(d) Patients (DOD-O and ISRUC): Accuracy value.

Figure 6.1: Results overview in single-channel tests (frontal, central and occipital electrode position).



(a) Healthy subjects (DOD-H): Barplot of precision, recall, and F1 score.

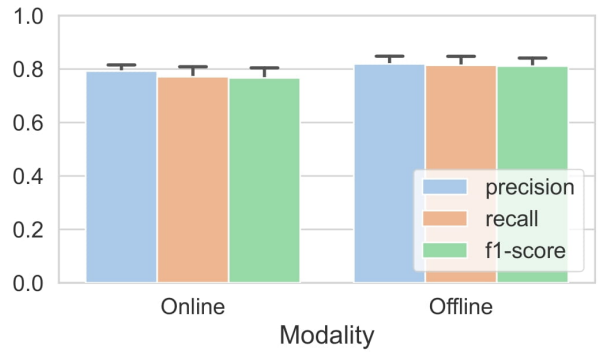
(b) Healthy subjects (DOD-H): Accuracy value.



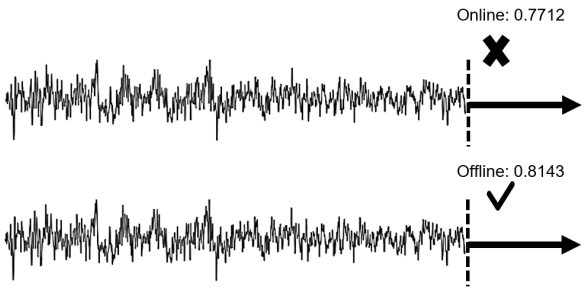
(c) Patients (DOD-O and ISRUC): Barplot of precision, recall, and F1 score.

(d) Patients (DOD-O and ISRUC): Accuracy value.

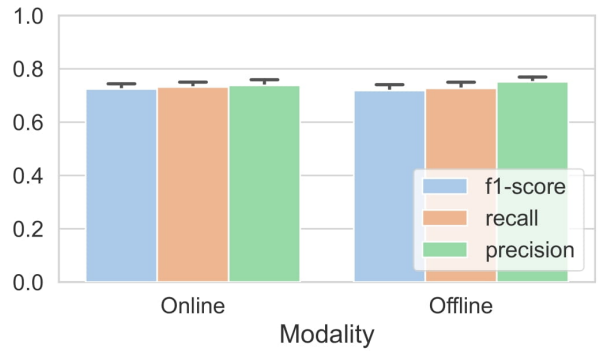
Figure 6.2: Results overview in multiple-channel tests (EEG (frontal, central and occipital) and EEG (frontal, central and occipital) + EOG + ECG + EMG.



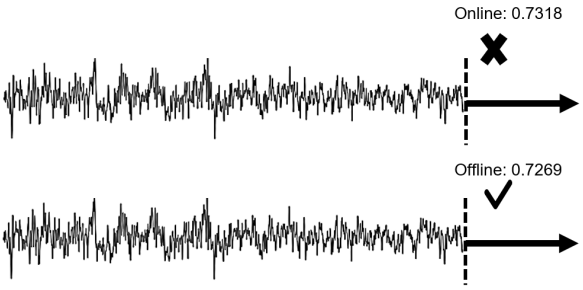
(a) Healthy subjects (DOD-H): Barplot of precision, recall, and F1 score.



(b) Healthy subjects (DOD-H): Accuracy value.



(c) Patients (DOD-O and ISRUC): Barplot of precision, recall, and F1 score.

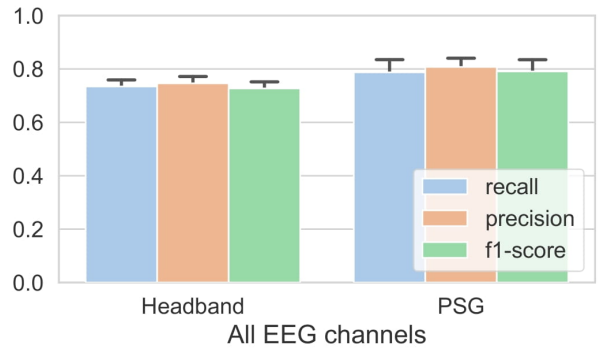


(d) Patients (DOD-O and ISRUC): Accuracy value.

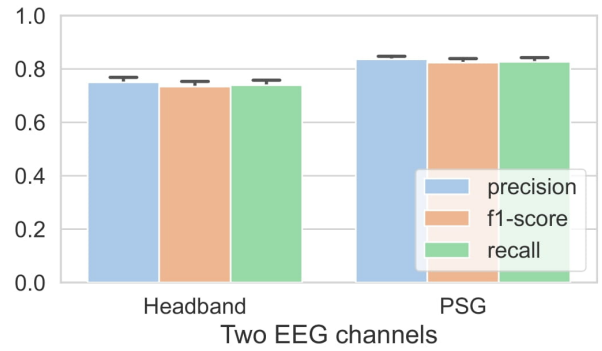
Figure 6.3: Results overview for the online and offline tests with frontal electrode position.

The approach showing better results is further evaluated on data recorded in *Bitbrain*'s sleep laboratories. This is specially relevant since it permits testing the network on data collected by both a full medical-grade polysomnographic and a wearable, highly comfortable headband. An abbreviated version of the results is performed again and shown in Figure 6.4. The fall in performance seen when using the mobile sleep monitoring device is between 5-9%. The loss is not excessive considering that these devices have the limitation of poorer signal quality, given that no electrolytic substances are used to improve the conductivity between the surface electrodes attached and the scalp. Besides, the headband employed in this case has more limitations caused by a bad decision in the reference. These consist of the generation of a huge amount of artifacts.

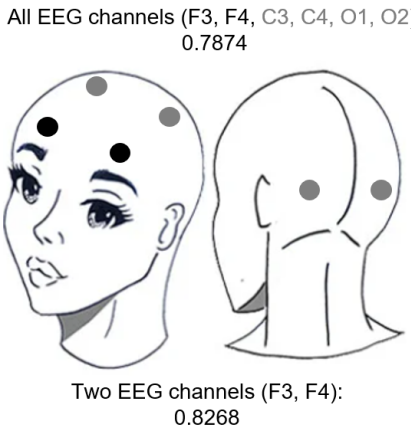
Eventually, in addition to the initially proposed objectives, different methods consisting of data augmentation were applied with the goal of improving the performance of the designed techniques. Along the same lines, the information measured by an IMU was incorporated to the headband data. Besides, an approach is proposed to quantify uncertainty in the model's decisions, with the aim of ending the black-box skepticism that hampers professionals' trust. Finally, different tests were carried out in order to assess their capability to perform transfer learning between datasets.



(a) PSG & Headband (all available EEG channels): Barplot of precision, recall, and F1 score.



(b) PSG & Headband (2 EEG channels): Barplot of precision, recall, and F1 score.



(c) PSG: Accuracy value.

(d) Headband: Accuracy value.

Figure 6.4: Results overview in tests using all available EEG channels (PSG: F3, F4, C3, C4, O1, O2. Headband: AF8, Fp1, Fp2, AF7, T8) and two EEG channels (PSG: F3, F4. Headband: Fp1, Fp2).

Establishing a comparison with state-of-the-art scoring systems, most of them have an accuracy above ours (accuracies range between 77.2% and 87.7% in the overview of the latest approaches performed in Table 2.1). However, none achieve it by labelling epochs in real-time and using a reduced number of channels. Moreover, despite their well performing, they have not been given a practical use. Therefore, although the presented results are not significantly better than those of the literature, the novel feature of complete real-time staging in data recorded with few EEG channels is introduced.

6.1 Future work

Given the research nature of the thesis and the promising results achieved by the proposed methodology, showing a comparable performance to sleep experts on the sleep scoring tasks, the lines open for future work are broad. There are still challenges to overcome in order to bring sleep monitoring outside sleep laboratories to daily living environments. Future steps to address them would be:

- Improve scoring results in stages such as N1 and reduce confusion as much as possible. This could be achieved by using additional sensors (e.g., pulse oximetry).
- Train and test the algorithm on data recorded by a new *Bitbrain*'s EEG system with higher reliability.
- Estimate the sufficient number of subjects that needs to be recorded to guarantee or surpass expert level classification.
- Achieve the same performance in subjects suffering from sleep disorders as in healthy individuals.

Additionally, although the RNN model works worse in all shown cases, it cannot be affirmed that this is the case. In future work, it would be interesting to continue exploring these architectures and exploit their strong temporal component, something that could not be done in this project given the period of time that has been arranged for it.

Something similar happens with data augmentation and uncertainty quantification techniques. Although the results obtained have not been as desired, it is worth continuing to investigate them in future projects. Besides, in addition to this quantification of uncertainty, it could be useful to introduce interpretability in order to prevent the black-box behaviour of the models, which could limit the end-user acceptance.

Finally, once the decoder has been trained with new *Bitbrain*'s data, it will make up one of the main components of the auditory closed-loop stimulation system previously explained. This step will be fast since the code has already been integrated into the company's platform (section 5.2). Nevertheless, it will be necessary to make many recordings with several participants, to be able to carry out an exhaustive analysis of the data, as well as of the neurophysiological effects generated, and thus be able to perform such interventions in the future.

Bibliography

[1] Haack M Besedovsky L, Lange T. The sleep-immune crosstalk in health and disease. *Physiological Reviews 99: 1325–1380*, 2019.

[2] Miller MA Cappuccio FP. Sleep and cardio-metabolic disease. *Current Cardiology Reports 19: 110*, 2017.

[3] Wisden W Harding EC, Franks NP. Sleep and thermoregulation. *Current Opinion in Physiology 15: 7–13*, 2020.

[4] Barnes CM Walker MP Ben Simon E, Vallat R. Sleep loss and the socio-emotional brain. *Trends in Cognitive Sciences 24: 435–450*, 2020.

[5] Walker MP. The role of sleep in cognition and emotion. *Annals of the New York Academy of Sciences 1156: 168–197*, 2009.

[6] Leslie C. Markun Ajay Sampat. Clinician-focused overview and developments in polysomnography. *Current Sleep Medicine Reports 6, 309-321*, 2020.

[7] Smitha George Karen Lorraine Acosta Seithikurippu Ratnas Pandi-Perumal Ahmed S. Ba-Hammam, Divinagracia E. Gacuan and Ravi Gupta. Polysomnography i: Procedure and technology. *Synopsis of Sleep Medicine (pp.443-456)*, 2016.

[8] Siavash Sakhavi. Application of deep learning methods in brain-computer interface systems. *https://www.researchgate.net/publication/323142453_APPLICATION_OF_DEEP_LEARNING_METHODS_IN_BRAIN-COMPUTER_INTERFACE_SYSTEMS*, 2017.

[9] P. Maquet. The role of sleep in learning and memory. *Science, vol. 294, no. 5544, pp. 1048-1052*, 2001.

[10] Sudhansu Chokrovery et al. Overview of sleep & sleep disorders. *Indian J Med Res 131.2, pp. 126-140*, 2010.

[11] Lori A Panossian and Alon Y Avidan. Review of sleep disorders. *Medical Clinics of North America 93.2, pp. 407-425*, 2009.

[12] Maurice M Ohayon. Prevalence and comorbidity of sleep disorders in general population. *La Revue du Praticien 57.14, pp. 15121-1528*, 2007.

[13] Steven Van Hout Richard S Rosenberg. The american academy of sleep medicine inter-scorer reliability program: sleep stage scoring. *J Clin Sleep Med 9(1):81-7*, 2013.

[14] Hanly P Younes M, Raneri J. Staging sleep in polysomnograms: analysis of inter-scorer variability. *J Clin Sleep Med 12(06):885-94*, 2016.

[15] Schmidt C Vandewalle G Jaspar M Devillers J et al. Muto V, Berthomier C. 0315 inter- and intra-expert variability in sleep scoring: Comparison between visual and automatic analysis. *Sleep, Volume 41, Issue suppl*$_1$, *Page A*121, 2018.

[16] Robert W. McCarley Mark R. Zielinski, James T. McKenna. Functions and mechanisms of sleep. *AIMS Neuroscience; 3(1): 67-104*, 2016.

[17] H. Berger. Über das elektrenkephalogramm des menschen. *European Archives of Psychiatry and Clinical Neuroscience, 87(1):527–570*, 1929.

[18] Fiorillo et al. Automated sleep scoring: A review of the latest approaches. *Sleep Medicine Reviews, Volume 48, 101204*, 2019.

[19] R. Mueller S. Tao M. Kim M. Rueschman S. Mariani D. Mobley G. Q. Zhang, L. Cui and S. Redline. The national sleep research resource: towards a sleep data commons. *J Am Med Inform Assoc., vol. 25, no. 10, pp. 1351–1358*, 2018.

[20] Emmanuel H During Antoine Guillot, Fabien Sauvet and Valentin Thorey. Dreem open datasets: Multi-scored sleep datasets to compare human and automated sleep staging. *IEEE Engineering in Medicine and Biology Society. PP(99):1-1*, 2020.

[21] José Moutinho Santos Urbano Nunes. Sirvan Khalighi, Teresa Sousa. Isruc-sleep: A comprehensive public dataset for sleep researchers. *Computer methods and programs in biomedicine 12.4, pp. 180-192.*, 2006.

[22] Johnson L. Viglione S. Naitoh P. Joseph R. Martin, W. and J Moses. Pattern recognition of eeg-eog as a technique for all-night sleep stage scoring. *Electroencephalogr. Clin. Neurophysiol. 32, 417–427*, 1972.

[23] P. Wen M. Diykh, Y. Li. Eeg sleep stages classification based on time domain features and structural graph similarity. *IEEE Transactions on Neural Systems and Rehabilitation Engineering 24, 1159–1168*, 2016.

[24] M. Bonkovic M. Cic, J. Soda. Automatic classification of infant sleep based on instantaneous frequencies in a single-channel eeg signal. *Computers in biology and medicine 43, 2110–2117*, 2013.

[25] F. Faradji P. Memar. A novel multi-class eeg-based sleep stage classification system. *IEEE Transactions on Neural Systems and Rehabilitation Engineering 26, 84-95*, 2018.

[26] A. Subasi A. R. Hassan. A decision support system for automated identification of sleep stages from single-channel eeg signals. *Knowledge-Based Systems 128, 115-124*, 2017.

[27] Panteleimon Chriskos et al. Automatic sleep stage classification applying machine learning algorithms on eeg recordings. *IEEE 30th International Symposium on Computer-Based Medical System (CBMS). IEEE. 2017, pp. 435–439*, 2017.

[28] Kaare Mikkelsen Huy Phan. Automatic sleep staging of eeg signals: Recent development, challenges, and future directions. *Physiol Meas 43(4)*, 2022.

[29] C. Wu A. Supratak, H. Dong and Y. Guo. Deepsleepnet: A model for automatic sleep stage scoring based on raw single-channel eeg. *IEEE Trans. on Neural Systems and Rehabilitation Engineering, vol. 25, no. 11, pp. 1998–2008*, 2017.

[30] N. Parmar J. Uszkoreit L. Jones A. N. Gomez L. Kaiser A. Vaswani, N. Shazeer and I. Polosukhin. Attention is all you need. *NIPS 2017, pp. 5998-6008*, 2017.

[31] Amaral L. Glass L. Hausdorff J. Ivanov P. C. Mark R. ... Stanley H. E Goldberger, A. Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation [Online]. 101 (23), pp. e215–e220*, 2013.

[32] Mahault Garnerin Éric Le Ferrand Laurent Besacier Marcely Zanon Boito, William Havard. Mass: A large and clean multilingual corpus of sentence-aligned spoken utterances extracted from the bible. *Proceedings of the 12th Language Resources and Evaluation Conference, pages 6486–6493*, 2020.

[33] Iber C Kiley JP Nieto FJ O'Connor GT Rapoport DM Redline S Robbins J Samet JM Wahl PW Quan SF, Howard BV. The sleep heart health study: design, rationale, and methods. *Sleep 20(12):1077-85. PMID: 9493915*, 1997.

[34] Navin Cooray Oliver Y Chen Maarten De Vos Huy Phan, Fernando Andreotti. Seqsleepnet: End-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging. *IEEE Trans Neural Syst Rehabil Eng . 27(3):400-410*, 2018.

[35] U. Rajendra Acharya Sajad Mousavi, Fatemeh Afghah. Sleepeegnet: Automated sleep stage scoring with sequence to sequence deep learning approach. *PLoS ONE 14(5): e0216456*, 2019.

[36] Minh C. Tran Philipp Koch Alfred Mertins Huy Phan, Oliver Y. Ch´en and Maarten De Vos. Xsleepnet: Multi-view sequential model for automatic sleep staging. *arXiv preprint arXiv:2007.05492, 2020*, 2020.

[37] S. Tu H. Nie and L. Xu. Recsleepnet: An automatic sleep staging model based on feature reconstruction. *IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 1458-1461*, 2021.

[38] E. Eldele et al. An attention-based deep learning approach for sleep stage classification with single-channel eeg. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 29, pp. 809-818*, 2021.

[39] Oliver Y. Chén Huy Phan, Kaare Mikkelsen. Sleeptransformer: Automatic sleep staging with interpretability and uncertainty quantification. *IEEE transactions on bio-medical engineering PP(8)*, 2022.

[40] Jens G Klinzing Luis Montesano Javier Minguez Martin Esparza-Iaizzo, Ion Álvarez-Guerrico and Eduardo López-Larraz. Sleepbci: a platform for memory enhancement during sleep based on automatic scoring. *Conference: XXXIX Annual Congress of the Spanish Society of Biomedical Engineering*, 2021.

[41] Amores J. Maes-P Koushik, A. Real-time smartphone-based sleep staging using 1-channel eeg. *IEEE 16th International Conference on Wearable and Implantable Body Sensor Networks (BSN) (pp. 1-4)*, 2019.

[42] Bengio Y Pascanu R, Mikolov T. On the difficulty of training recurrent neural networks. *30th Int Conf Mach Learn ICML 2013, vol. PART 3; p. 2347–55*, 2012.

[43] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data, 6(1):60*, 2019.

[44] V Goel X Cui and B Kingsbury. Data augmentation for deep neural network acoustic modeling. *IEEE/ACM TASLP, 23(9):1469–1477*, 2015.

[45] D.; Maoz U Lashgari, E.; Liang. Data augmentation for deep-learning-based electroencephalography. *J. Neurosci. Methods 346, 108885*, 2020.

[46] J. Amann et al. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Medical Informatics and Decision Making, vol. 20, pp. 310*, 2020.

[47] Niels Niethard Klinzing, Jens G and Jan Born. Mechanisms of systems memory consolidation during sleep. *Nature Neuroscience 22(10): 1598–1610*, 2019.

[48] Nava Zisapel Alan Wade and Patrick Lemoine. Prolonged-release melatonin for the treatment of insomnia: targeting quality of sleep and morning alertness. 2008.

[49] Naresh M Punjabi. The epidemiology of adult obstructive sleep apnea. *Proceedings of the American Thoratic Society 5.2, pp. 136-143*, 2008.

[50] Adam V Benjafield et al. Estimation of the global prevalence and burden of obstructive sleep apnea: a literature-based analysis. *The Lancet Respiratory Medicine 7.8, pp. 687-698*, 2019.

[51] Christer Hublin et al. The prevalence of narcolepsy: an epidemiological study of the finnish twin cohort. *Annals of neurology 35.6, pp.709-716.*, 1994.

[52] Daniele Manfredini et al. Epidemiology of bruxism in adults: a systematic review of the literature. *J Orofac Pain 27.2, pp. 99-110*, 2013.

[53] José Haba-Rubio et al. Prevalence and determinants of rapid eye movement sleep behavior disorder in the general population. *Sleep 41.2, zsx197*, 2018.

[54] Magdolna Hornyak et al. Periodic leg movements in sleep and periodic limb movement disorder: prevalence, clinical significance and treatment. *Sleep medicine reviews 10.3, pp.169-177.*, 2006.

[55] ABEN QEEG-D D. Corydon Hammond, Ph.D. and QEEG-D Jay Gunkelman. The art of artifacting. *An ISNR Research Foundation Publication*, 2001.

[56] Hugo Tito-Chura Anibal Flores and Honorio Apaza-Alanoca. Data augmentation for short-term time series prediction with deep learning. *Intelligent Computing. Springer, 492–506*, 2021.

[57] Khandakar M Rashid and Joseph Louis. Window-warping: a time series data augmentation of imu data for construction equipment activity identification. *ISARC. Proceedings of the International Symposium on Automation and Robotics in Construction, Vol. 36. IAARC Publications, 651–657*, 2019.

[58] Lawrence O Hall Nitesh V Chawla, Kevin W Bowyer and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research, 16:321–357*, 2002.

[59] Thomas Martinetz Jan Born Ngo, Hong-Viet V and Matthias Mölle. Auditory closed-loop stimulation of the sleep slow oscillation enhances memory. *Neuron 78(3): 545–53*, 2013.

[60] Hong-Viet V et al. Ngo. Driving sleep slow oscillations by auditory closed-loop stimulation-a self-limiting process. *Journal of Neuroscience 35(17): 6630–38.*, 2015.

[61] Saure E Paajanen T Zee PC Santostasi G Hublin C Müller K Porkka-Heiskanen T Huotilainen M Paunio T Leminen MM, Virkkala J. Enhanced memory consolidation via automatic sound stimulation during non-rem sleep. *Sleep; 40 (3)*, 2017.

[62] Manousakis JE Nicholas CL Drummond SPA Anderson C. Diep C, Ftouni S. Acoustic slow wave sleep enhancement via a novel, automated device improves executive function in middle-aged men. *Sleep; 43 (1)*, 2020.

[63] Malkani RG Braun R Weintraub S Paller KA Zee PC. Papalambros NA, Santostasi G. Acoustic enhancement of sleep slow oscillations and concomitant memory improvement in older adults. *Front Hum Neurosci. 11:109*, 2017.

[64] Chen T Grimaldi D Santostasi G Paller KA Zee PC Malkani RG. Papalambros NA, Weintraub S. Acoustic enhancement of sleep slow oscillations in mild cognitive impairment. *Ann Clin Transl Neurol. 6(7):1191-1201*, 2019.

[65] Susanne Ruf Markus Wolff Jan Born Hong-Viet V. Ngo Jens G. Klinzing, Lilian Tashiro. Auditory stimulation during sleep suppresses spike activity in benign epilepsy with centrotemporal spikes. *Cell Reports Medicine 2, 100432*, 2021.

Appendices

Appendix A. Sleep related disorders

As previously mentioned, normal sleep effectuates a restorative function in the brain, helping the integration of many cognitive processes and the modulation of emotional stress. Examination of EEG brainwave activity allows researchers and clinicians to obtain a more comprehensive assessment of brain activity patterns associated with normal sleep (which are shown in section 1.3). These patterns are essential in the detection of sleep disorders, helping health care professionals to create a treatment plan. This section briefly describes the six major categories of sleep pathologies:

- Insomnia: difficulty initiating or maintaining sleep. It is characterized by frequent body movement, enhanced levels of autonomic functioning, reduced levels of REM sleep, and, in some cases, the intrusion of waking rhythms (alpha waves) throughout the different sleep stages. This disorder affects 30-35% of the population ([48]).

- Sleep-related breathing disorders: one of the more-common sleep problems is obstructive sleep apnea (OSA), affecting close to 1 billion people worldwide ([49], [50]). In this disorder, the upper airway repeatedly impedes the flow of air due to a mechanical obstruction. This can happen dozens of times per hour during sleep. Consequently, there is impaired gas exchange in the lungs, leading to reductions in blood oxygen levels and unwanted elevations in blood levels of carbon dioxide. In addition, there are frequent disruptions of sleep that can lead to chronic sleep deprivation unless treated.

- Hypersomnia of central origin: narcolepsy is thought to involve specific abnormal functioning of subcortical sleep-regulatory centres, in particular a specialized area of the hypothalamus that releases a molecule called hypocretin. People who experience attacks of narcolepsy (0.05%, ([51])) commonly have the following symptoms: cataplexy (sudden loss of muscle tone), hypnagogic (sleep onset) and hypnopompic (awakening) visual hallucinations of dreamlike sort, and hypnagogic or hypnopompic sleep paralysis (the person is unable to move voluntary muscles for several seconds or minutes). Sleep attacks lead to the precocious triggering of REM sleep.

- Parasomnias: among the episodes that are considered problematic in sleep are somniloquy (sleep talking), somnambulism (sleep walking), enuresis (bed-wetting), bruxism (teeth grinding, affects approximately to 8-10% people ([52])), snoring, and nightmares. REM sleep behaviour disorder (RBD) is also in this same category (0.5-1% ([53])). In it, the sleeper acts out the dream content. The main characteristic of this disorder is a lack of the typical muscle paralysis seen during REM sleep. The condition is thought to be a degenerative brain disorder, increasing the risk for later developing parkinson disease.

- Sleep-related movement disorders: restless legs syndrome (RLS) and a related disorder known as periodic limb movement disorder (PLMB) are examples affecting 4-11% of the population ([54]). People with RLS experience an uncomfortable sensation in the legs that makes movement irresistible. During sleep, subtle periodic movements result in its disruption.
- Circadian rhythm disorders: phase-advanced sleep and phase-delayed sleep are the two prominent types of sleep-schedule disorders. In the former the sleep onset and offset occur earlier than social norms, while in the latter sleep onset is delayed and waking is also later in the day than is desirable.

All the previous disorders negatively affect human mental abilities and emotional state, finding more difficult to concentrate and increasing the risk for accidents. A lack of sleep also has an unfavorable impact in both short- and long-term memory. In fact, sleep disruptions can accelerate the aging process and symptoms of dementia. Other psychological risks also include anxiety, depression, paranoia, and suicidal thoughts. Besides, long-term sleep deprivation enlarges the risk for chronic conditions, such as diabetes mellitus and heart disease.

Therefore, both the high proportion of people affected by these diseases and the drastic consequences they have on their lives, emphasize the relevance of an adequate identification and treatment of sleep pathology.

Appendix B. Types of EEG artifacts

An artifact refers to a modification of the EEG tracing that is due to an extra-cerebral source ([55]). Artifacts are inevitably present in every EEG, but they may obscure EEG activity and lead to misinterpretations. Therefore, it is of great important to distinguish artifact from brain waves. This appendix defines the most usual artifacts that are encountered in clinical practice.

EEG artifacts can be classified depending on their origin. Physiologic artifacts originate from the patient and non-physiologic artifacts originate from the environment of the patient. For each of both types, The most common are summarized in Figure B.1 and Figure B.2, respectively.

	Origin	Why it affects the EEG	Types of effects	Effect on time domain	Effect on frequency domain
Ocular activity	Muscle can be electrically modeled as a magnetic dipole and it distorts the electric field in the region when it moves.	Has an amplitude usually one order of magnitude larger than the EEG signal.	Blinking, lateral movement, eye movements.	Blinking produces a quick change with high amplitude on the EEG signals in the electrodes on the frontal area, more pronounced in those closer to the eyes. Lateral movements of the eye affect also the frontal areas.	Effect in low frequencies that can be confused with delta and theta bands.
Muscle activity	Muscle produce electrical activity when they are contracted	Interferes with the actual EEG activity leading to high frequency artifacts.	Clenching the jaw, neck and shoulder tensión, frowning, swallowing, chewing, talking, sniffing, hiccuping.	High frequency signal that overlaps with the EEG signal. The Amplitude depends on the strength of the contraction.	Overlapping in beta and gamma EEG bands.
Cardiac activity	Electrical activity from the heart.	Creates a rhythmic distortion on the EEG signals.	Cardiac activity, pulse.	A rhythmic pattern, corresponding with the hearbeats that overlaps the EEG signal.	The frequency components ovarla EEG band frequencies so they are difficult to visualize.
Perspiration	Sweat glands of the skin.	Produces changes in the electrical baseline of the electrodes. It could even create shorts between electrodes.	Sweat glands, skin potentials.	Slow waves overlapping the EEG signal.	Low frequency artifact that overlaps deta and theta bands.
Respiration	Movement of chest and head when breathing.	It is more common in sleep recordings as these movementsmodify the contact bewteen electrodes and scalp.	Inhale, exhale.	Slow waves synchronized with breathing rhythm.	Low frequency artifact that overlaps theta and theta bands.

Figure B.1: Overview of physiological artifacts.

	Origin	Why it affects the EEG	Types of effects	Effect on time domain	Effect on frequency domain
Electrode pop	Temporary failures in the contact between the EEG sensor and the scalp.	Due to changes in contact potential between the scalp and the electrode.	Electrode pop.	Abrupt and high amplitude interference on the EEG signal usually localized on a single channel.	Difficult to see due to a wide range of possible distortions.
Cable movement	Movement of the cables connecting the electrodes and the amplifier.	Distortion in the recorded signal and in the scalp-sensor contact.	Cable movement, cable touch.	Distortions overlapping EEG signals with the same rhythm that the cable movement.	Non-EEG related frequency peaks.
Incorrect reference placement	Reference channel not placed or bad contact.	The recorded signal is not EEG.	Reference sensor not placed.	Abrupt changes in all channels with high amplitude. All channels will converge slowly to actual EEG signals when the reference is placed correctly.	Very high power in all channels.
AC electrical and electromagnetic interferences	The signal can be affected by surrounding fields like AC power sources and wires due to insufficient or lack of wire shielding.	50 or 60 Hz.	High frequency noise continuously overlapping the EEG signal.	Big spike around 50 or 60 Hz depending on the AC frequency standard for the country you are in.	Low frequency artifact that overlaps deta and theta bands.
Body movements	Body movements, principally affected by head movements.	When moving, although unintentionally, the contact between electrode and skin is affected and the EEG signal is corrupted.	Head and arm movements, walking, running.	Temporary slow waves corresponding with the rhythm of the movement.	Effect localized in lower frequencies overlapping delta and theta bands.

Figure B.2: Overview of non-physiological artifacts.

Appendix C. STAGES dataset preprocessing

Given the limited information available, an exhaustive analysis of the STAGES dataset was necessary to be able to use it as input for the neural network and perform automatic sleep staging. Afterwards, it was found that it was necessary to apply the steps described below:

- Electrodes F3 and F4 were selected from each of the subjects from the different centers. They were also rearranged so that the order was the same in all of them.

- Some of the files containing the labels were found to be empty and others incorrectly delimited, therefore the corresponding subjects were eliminated. Furthermore, in other cases, the start time of the EEG signal recording did not coincide with the start time of the labels. They were also deleted.
  The percentage of subjects with respect to the total that were deleted from each of the directories/clinical centers are shown in Figure C.1.

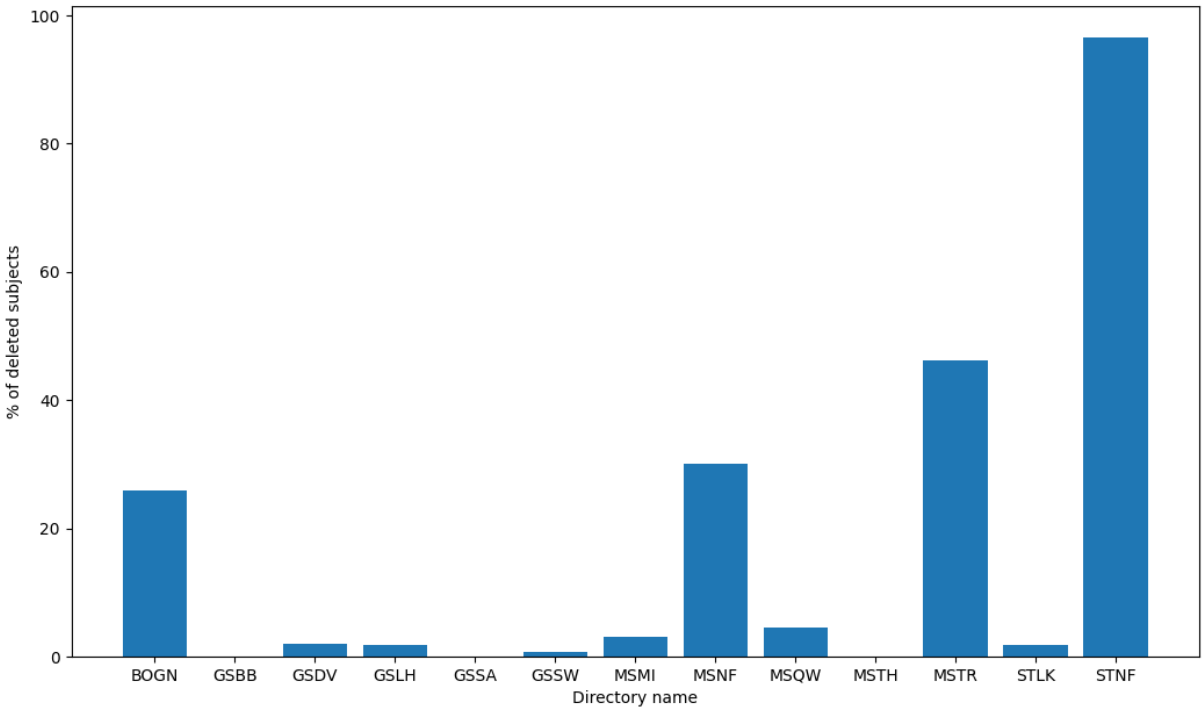


Figure C.1: Bar plot showing the percentage of subjects (Y axis) that were deleted from each directory (X axis).

- After visualizing the data (both the EEG signal in the time domain and the power spectra density (PSD)), it was concluded that it was necessary to apply two filters: notch at 60 Hz, and band-pass between 0.5 and 30 Hz. Then, the data was resampled to 100 Hz.
- It was necessary to manage the annotations and align them with the signal. An example of a fragment from an original annotation file is shown in Figure C.2. As it can be seen, Stage2 is labelled for 660 seconds. Taking into account its start time, the next label should take place at 23:57:33. However, there are many markers in between related with events that are not sleep stages, and thus uninteresting for the purposes of the project. Accordingly, the annotations relative to sleep stages were saved as 0 (Wake), 1 (N1), 2 (N2), 3 (N3) or 4 (REM) while markers were assigned a 5, all of them according to its duration.

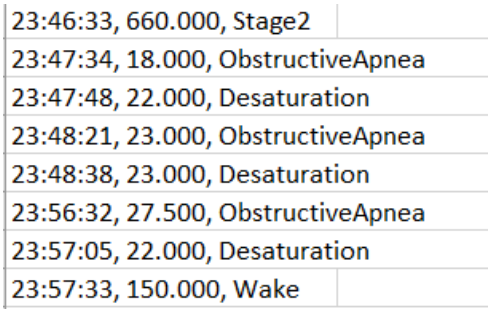


Figure C.2: Example 1 (start time, duration (seconds), event): annotation fragment extracted from an original .csv file.

Moreover, it is noted that some tests were performed to ensure the setting was correct. Therefore, the beginning of the annotation and signal files contained half an our or more of irrelevant data (look up and down 5 times, look left and right 5 times, blink 5 times, grit teeth for 5 seconds, simulate a snore for 5 seconds, hold breath for 10 seconds, breathe normally...). This is shown in Figure C.3 with another example.

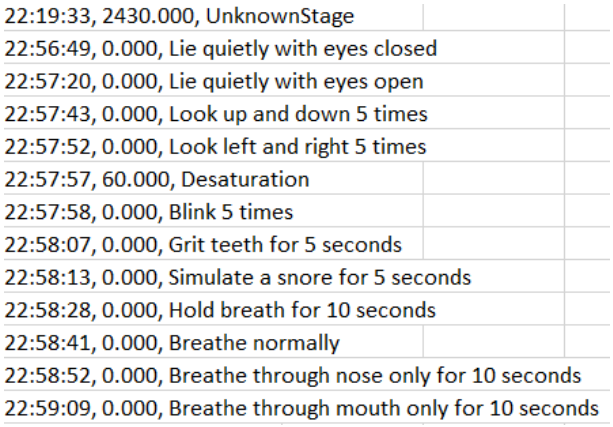


Figure C.3: Example 2 (start time, duration (seconds), event): annotation fragment extracted from an original .csv file.

All this time was scored as 5 and it was removed from both the signal and the labels, until the first epoch corresponding to a sleep stage appears. The average percentage of deleted

epochs until the first sleep stage is scored is shown in Figure C.4. This average (and the following ones) is computed across all the subjects from each center. Later, the remaining 5's were deleted exclusively from the vector of labels. If after these steps the dimensions of both vectors (i.e., signal and labels) are not coincident, the necessary number of epochs is removed from the longest one (from the end since the beginning is now perfectly aligned). This happens due to the fact that the annotations often did not run until the recording end time, though the annotations ended in a series of wake epochs, which would be a natural time to stop annotating a recording. However, the signal continues to be recorded.
The average percentage of epochs that were deleted to make data and labels be of the same length is displayed in Figure C.5.

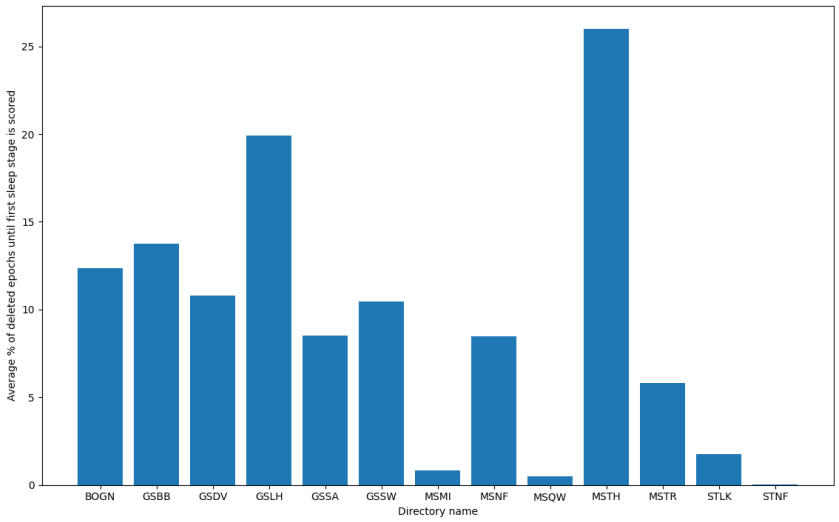


Figure C.4: Bar plot showing the average percentage of epochs across subjects (Y axis) that were deleted until the first sleep stage is scored. This is done separately for each center (X axis).

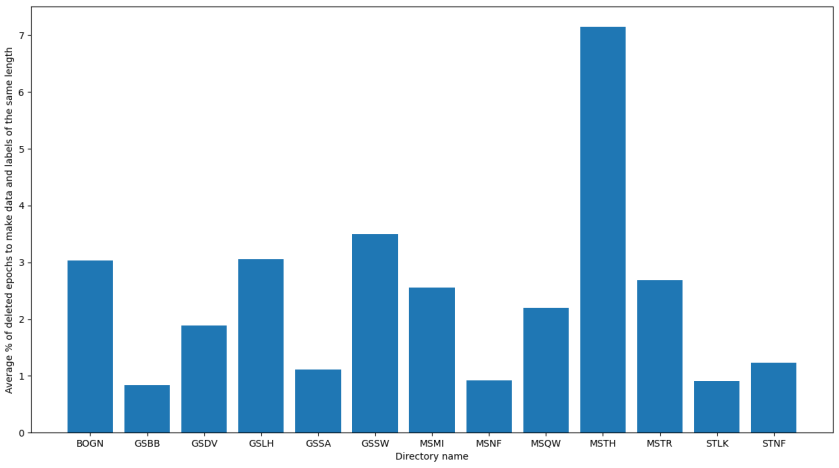


Figure C.5: Bar plot showing the average percentage of epochs across subjects (Y axis) that were deleted with the aim of making the signal and the labels of the same length. This is done separately for each center (X axis).

- All subjects in all folders/centers were made to be on the same scale (micro volts), as it

was not the case initially.

- Looking in more detail at the EEG, it was observed that it contained epochs of very bad signal quality. Consequently, after trying different parameters, the epochs with more than 0.2% of the signal above 100 micro volts or below -100 micro volts were removed, as well as those in which the signal was completely flat (they were removed from both F3 and F4 channels, regardless of where it took place).
  The average percentage of epochs that were deleted due to their bad EEG signal quality is depicted in Figure C.6.

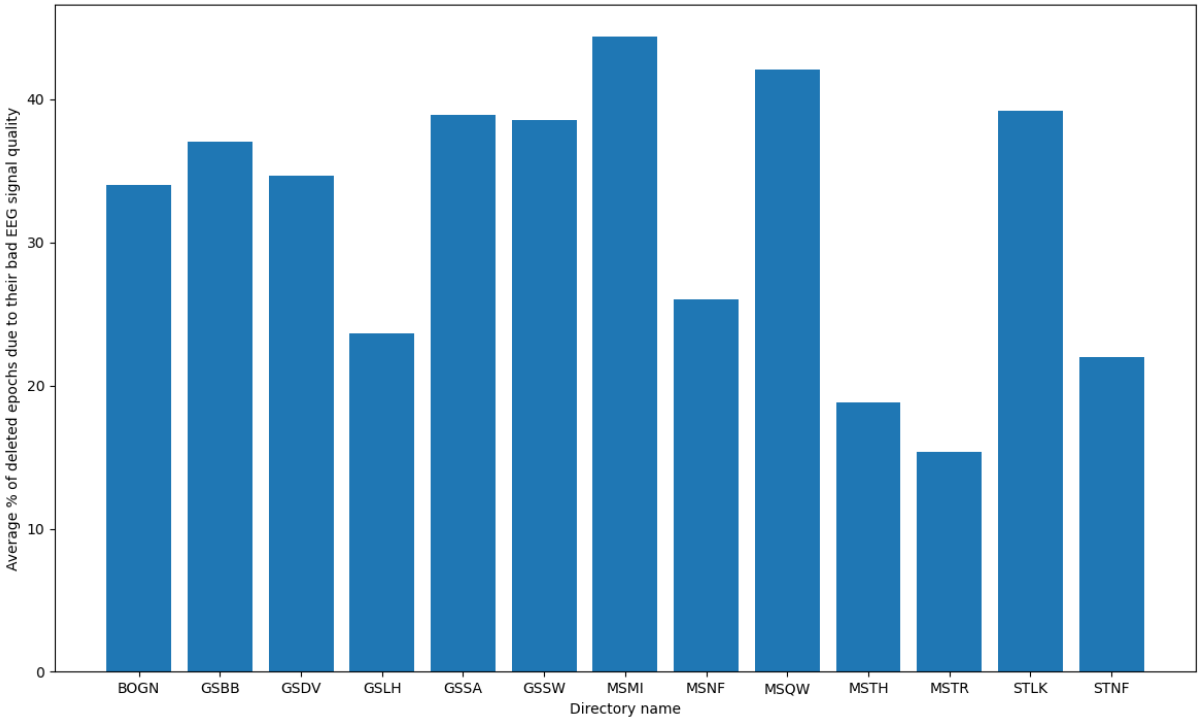


Figure C.6: Bar plot showing the average percentage of epochs across subjects (Y axis) that were deleted due to their bad EEG signal quality. This is done separately for each center (X axis).

Appendix D. STAGES results

In the following figures you can see the confusion matrices obtained for the 10-fold cross validation tests performed on each one of the clinical centers that compose the STAGES dataset. Next to them you will find bar charts showing the precision, recall and f1-score values for each sleep stage (see Figure D.1 to D.13). Finally, Figure D.14 displays the average percentage of each sleep across subjects in each medical center.

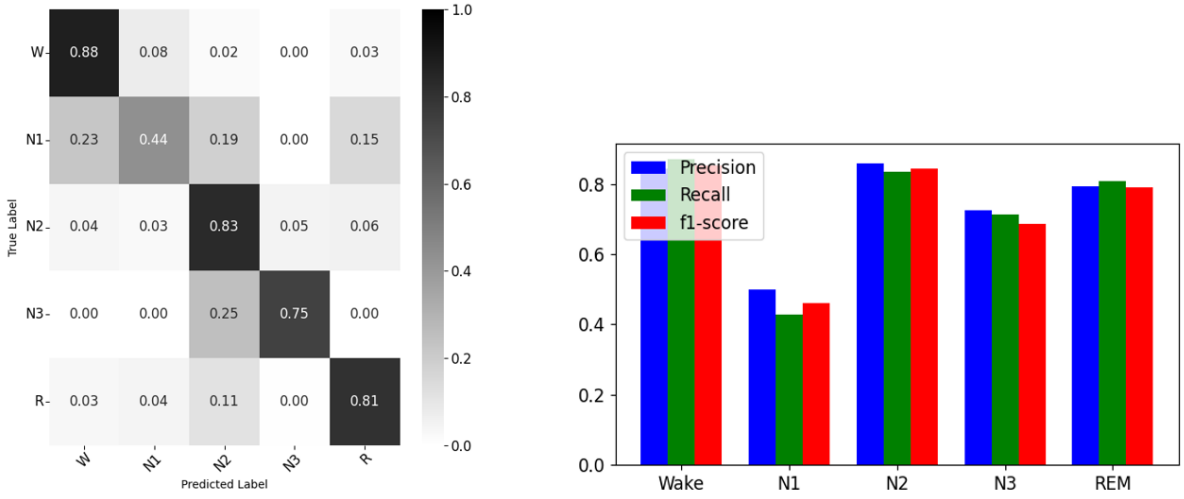


Figure D.1: Confusion matrix (left) and bar chart (right) obtained for STAGES test 1: BOGN directory/clinical center.

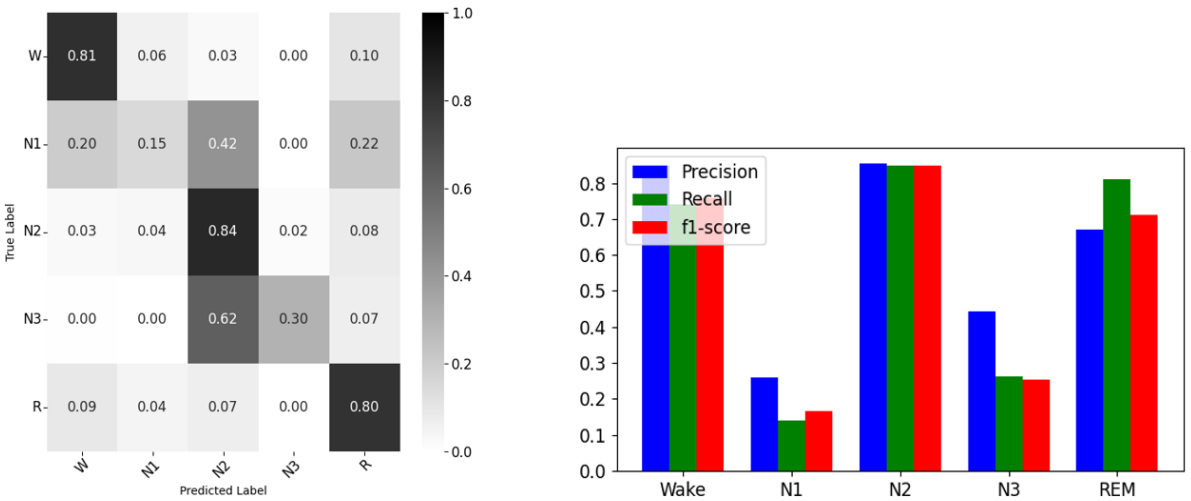


Figure D.2: Confusion matrix (left) and bar chart (right) obtained for STAGES test 2: GSBB directory/clinical center.

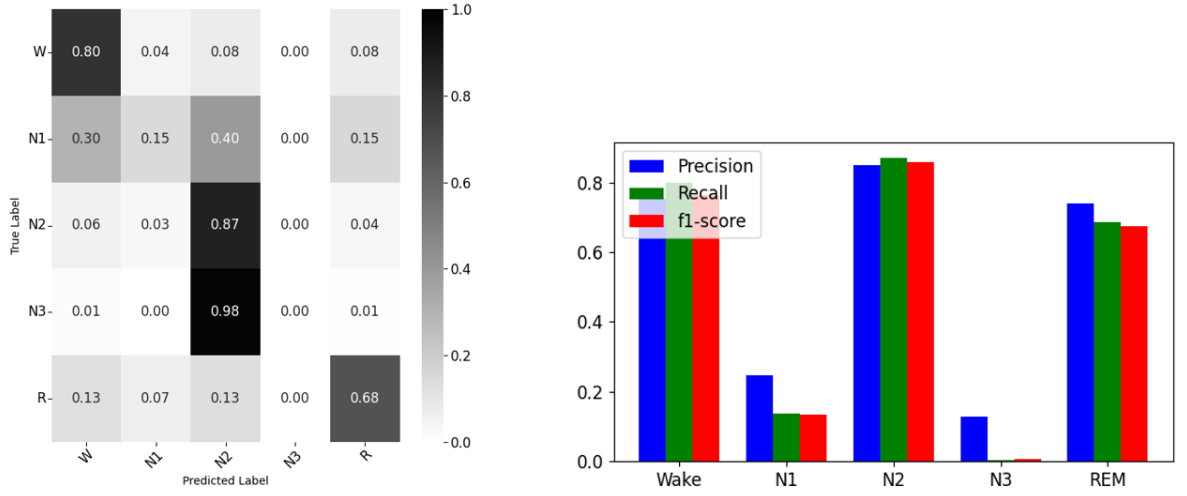


Figure D.3: Confusion matrix (left) and bar chart (right) obtained for STAGES test 3: GSDV directory/clinical center.

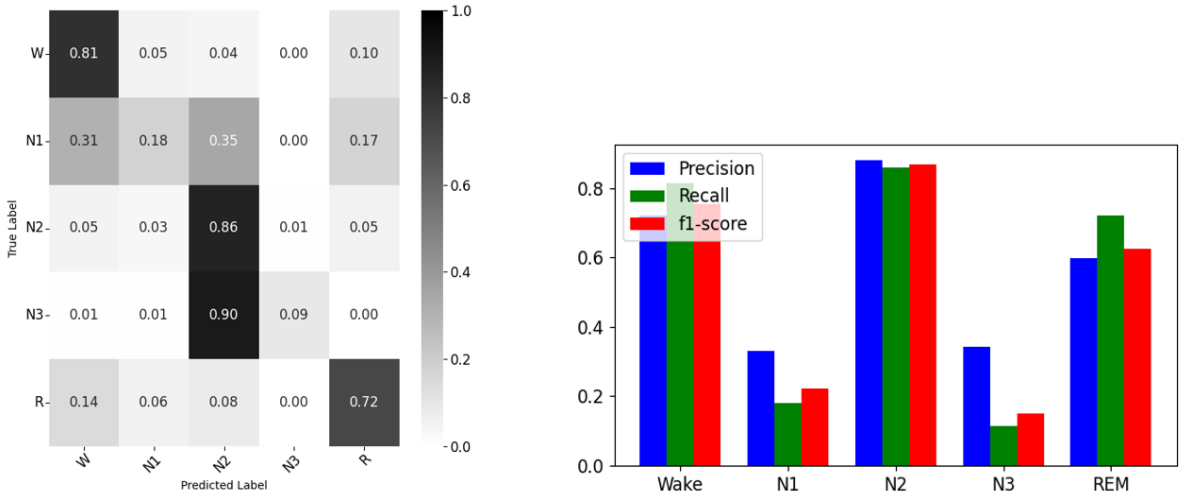


Figure D.4: Confusion matrix (left) and bar chart (right) obtained for STAGES test 4: GSLH directory/clinical center.

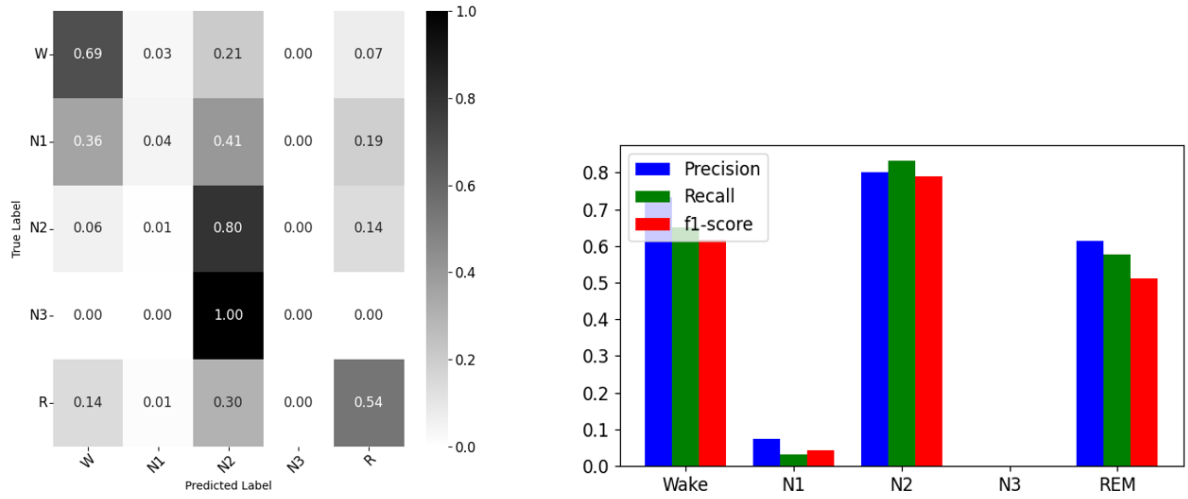


Figure D.5: Confusion matrix (left) and bar chart (right) obtained for STAGES test 5: GSSA directory/clinical center.

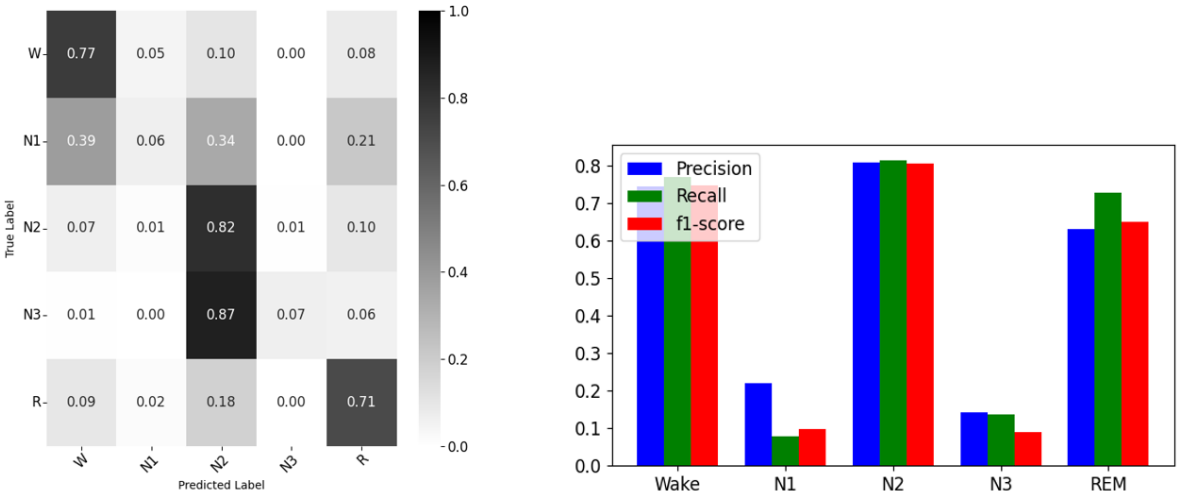


Figure D.6: Confusion matrix (left) and bar chart (right) obtained for STAGES test 6: GSSW directory/clinical center.

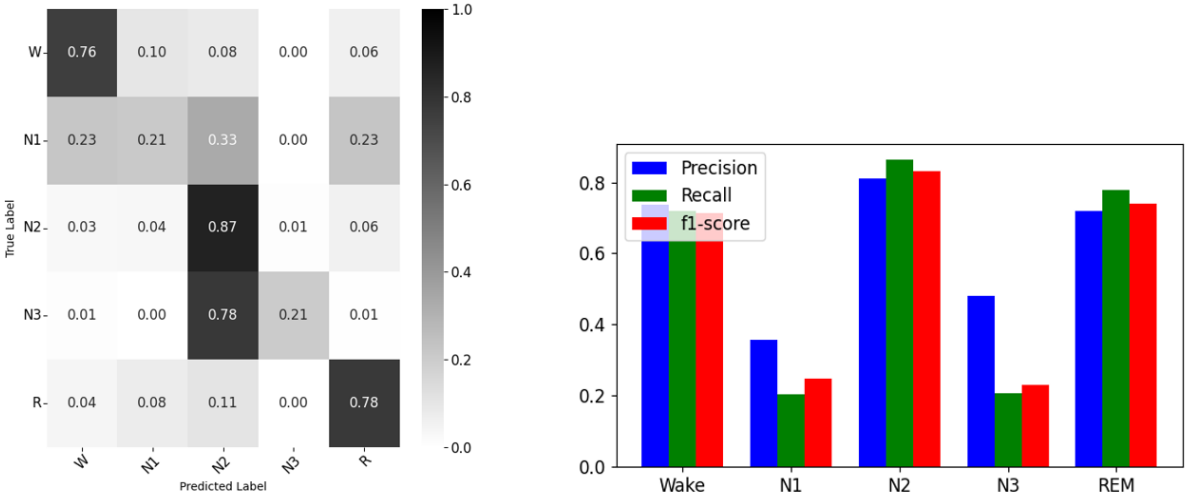


Figure D.7: Confusion matrix (left) and bar chart (right) obtained for STAGES test 7: MSMI directory/clinical center.

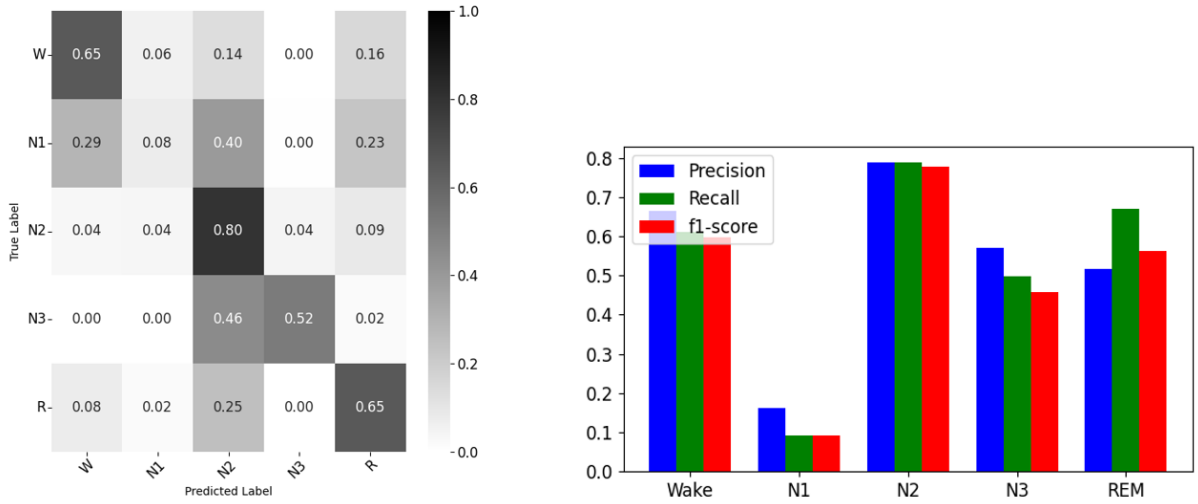


Figure D.8: Confusion matrix (left) and bar chart (right) obtained for STAGES test 8: MSNF directory/clinical center.

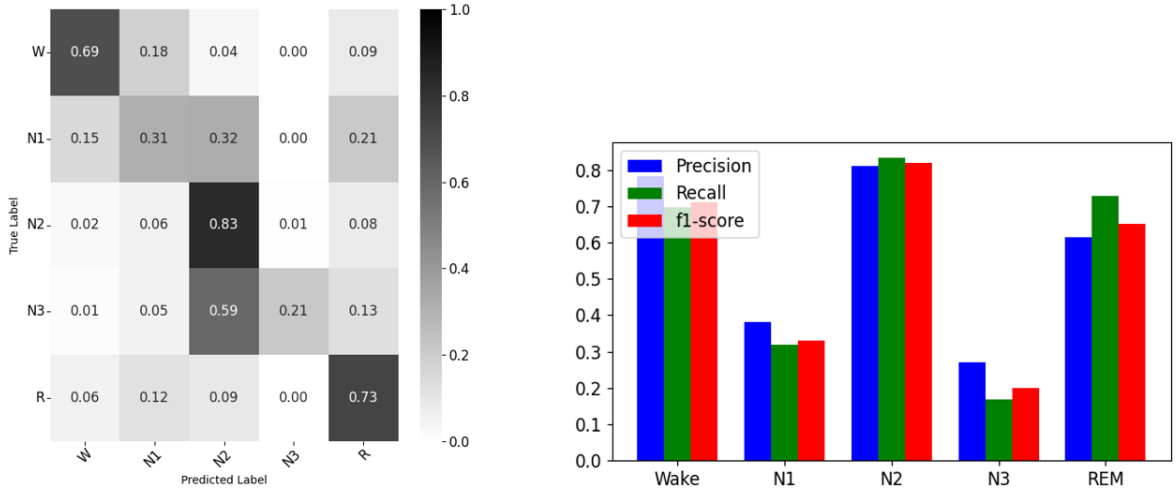


Figure D.9: Confusion matrix (left) and bar chart (right) obtained for STAGES test 9: MSQW directory/clinical center.

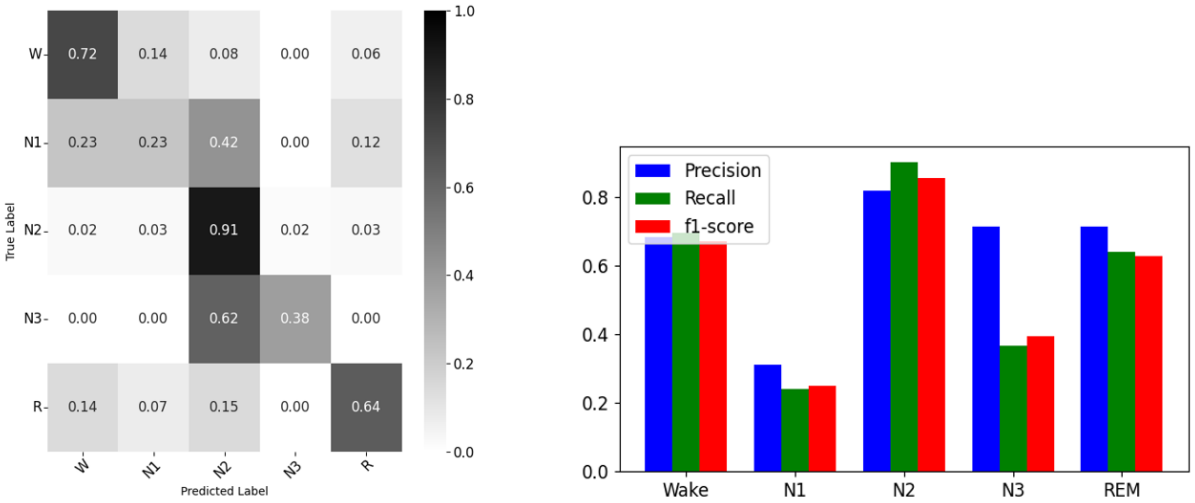


Figure D.10: Confusion matrix (left) and bar chart (right) obtained for STAGES test 10: MSTH directory/clinical center.

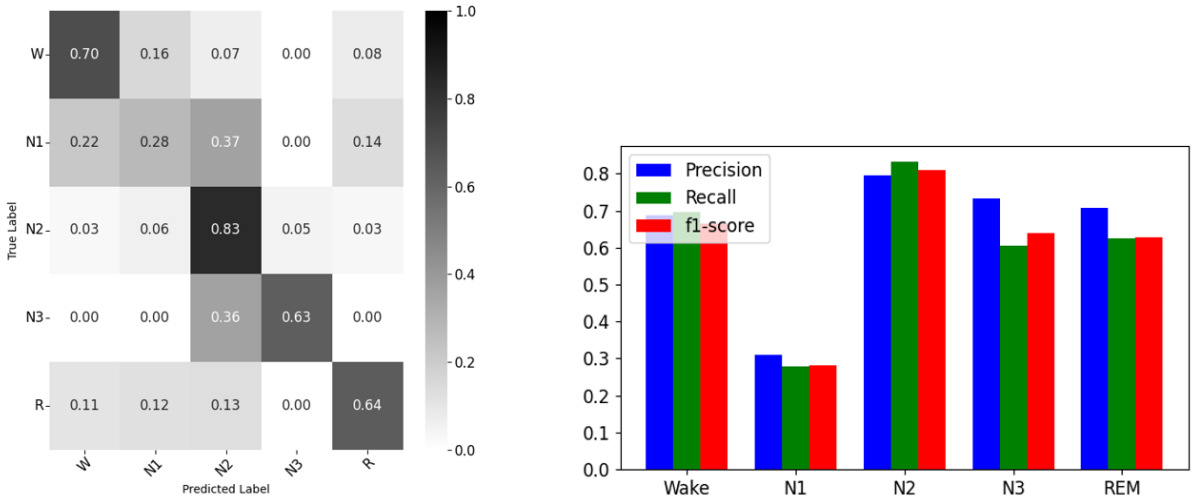


Figure D.11: Confusion matrix (left) and bar chart (right) obtained for STAGES test 11: MSTR directory/clinical center.

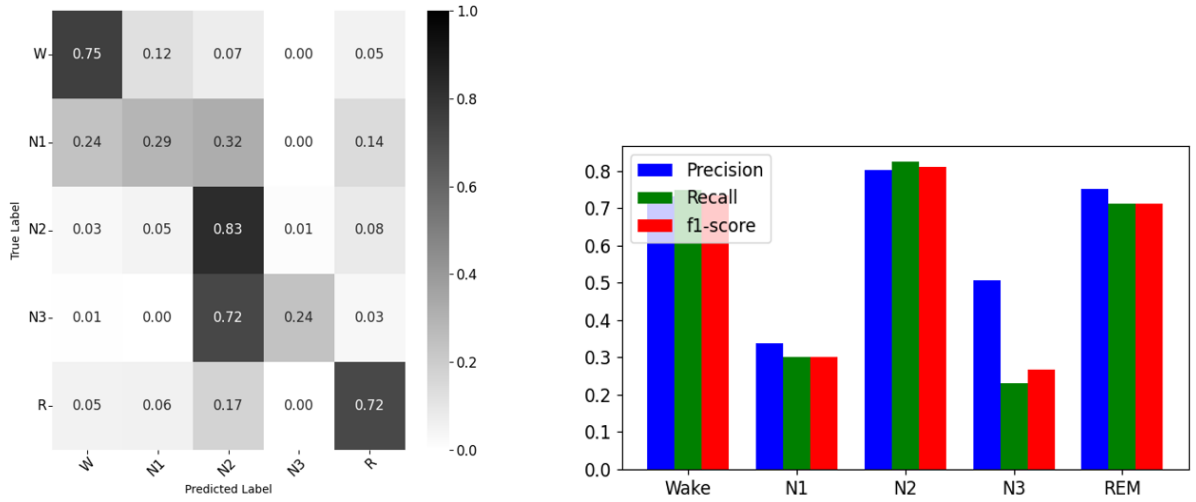


Figure D.12: Confusion matrix (left) and bar chart (right) obtained for STAGES test 12: STLK directory/clinical center.

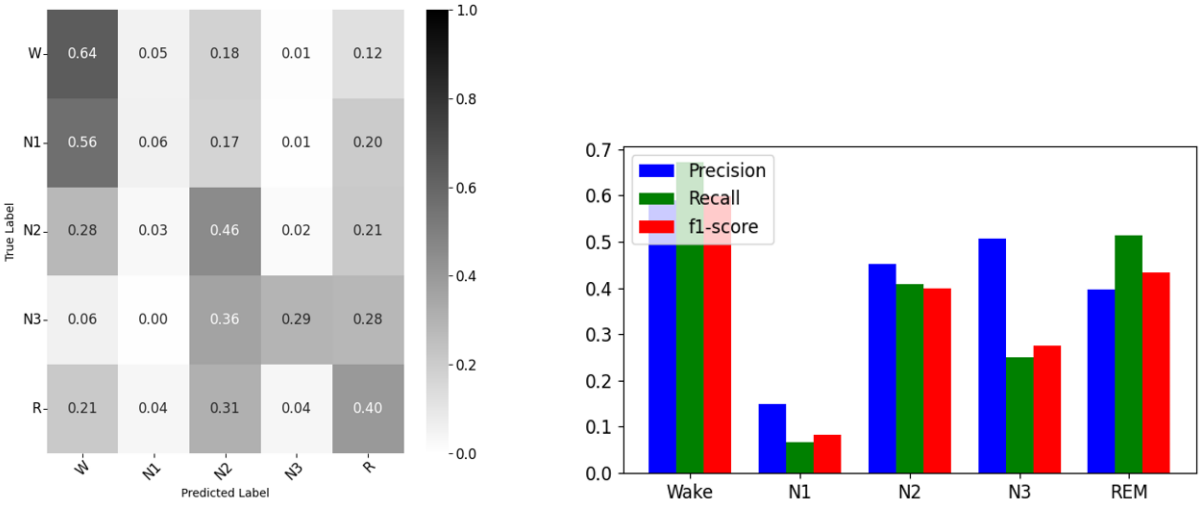


Figure D.13: Confusion matrix (left) and bar chart (right) obtained for STAGES test 13: STNF directory/clinical center.

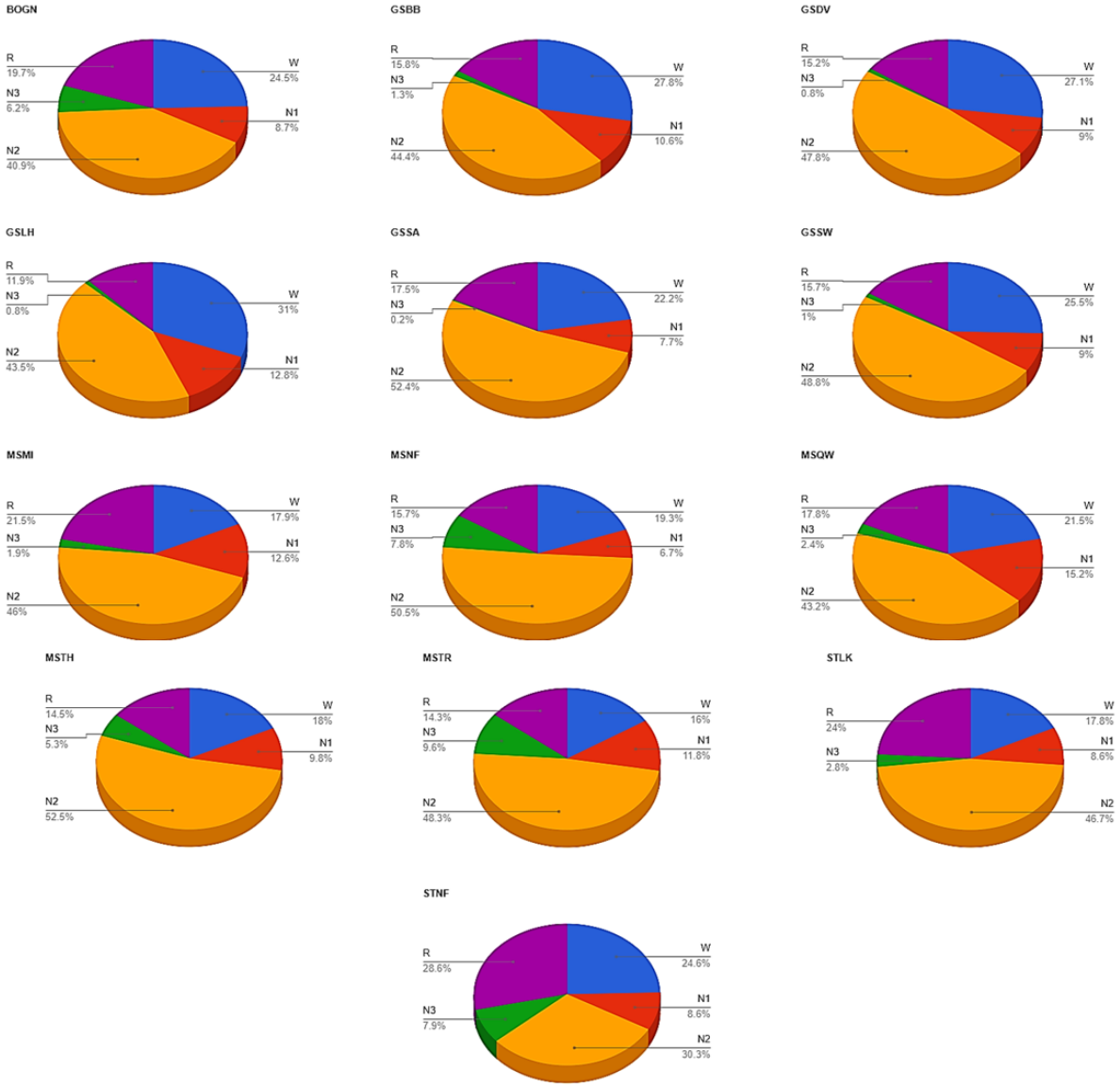


Figure D.14: Sleep architecture for each directory/clinical center.

Appendix E. Uncertainty quantification results

The uncertainty quantification procedure described in section 3.5 is performed on the DOD-H dataset. Specifically, it is applied for the test that uses single-channel information coming from a frontal electrode. The confidence of the network in its decisions is established as the likelihood of the most likely class. Then, the epochs with confidence below a threshold of 0.5 are chosen as low-confidence epochs. Figure E.1 and Figure E.2 showcase the findings for two different subjects. Both of them portray the ground-truth hypnogram and the predicted hypnogram alongside the quantified confidence. Besides, both the output hypnogram and the confidence through the epochs contain "x" markers pointing out the misclassified epochs.

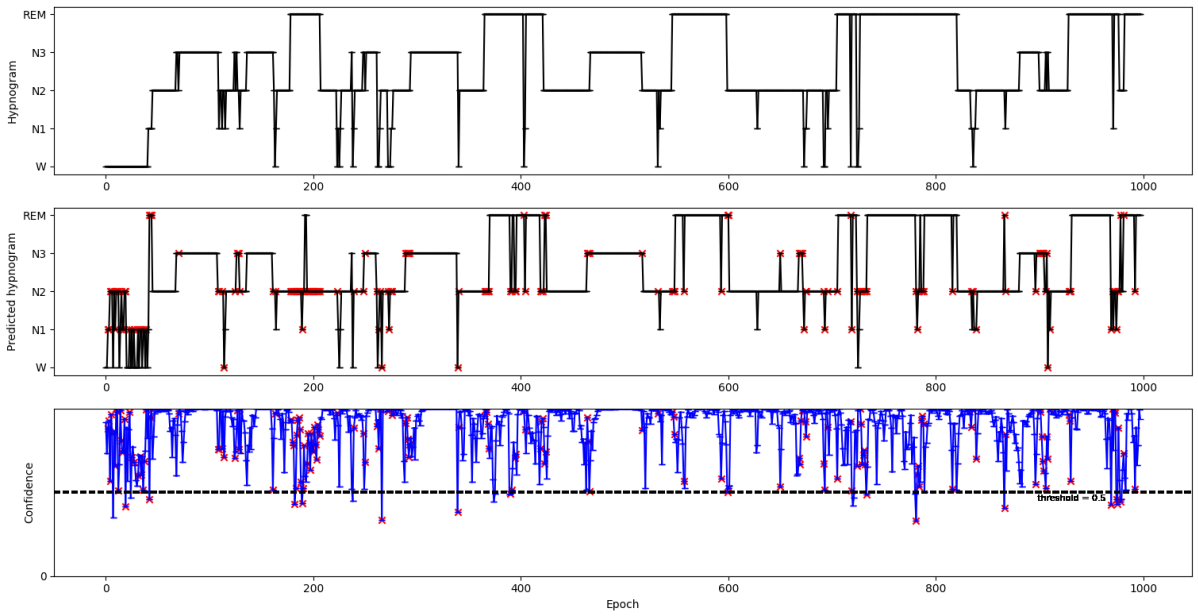


Figure E.2: Visualization of the estimated confidence for subject 19 of DOD-H.

It can be seen that most of the errors come from transitioning epochs as well as from confusing N1-N2, N1-Wake, and REM-N2, something which was also observed in the confusion matrix (Figure 5.1: 1). Moreover, the epochs considered as low-confidence ones are always mislabeled and their accuracy remains lower. This implies that the misclassified epochs are often associated with low confidences. Therefore, the confidence metric could be useful and meaningful in helping

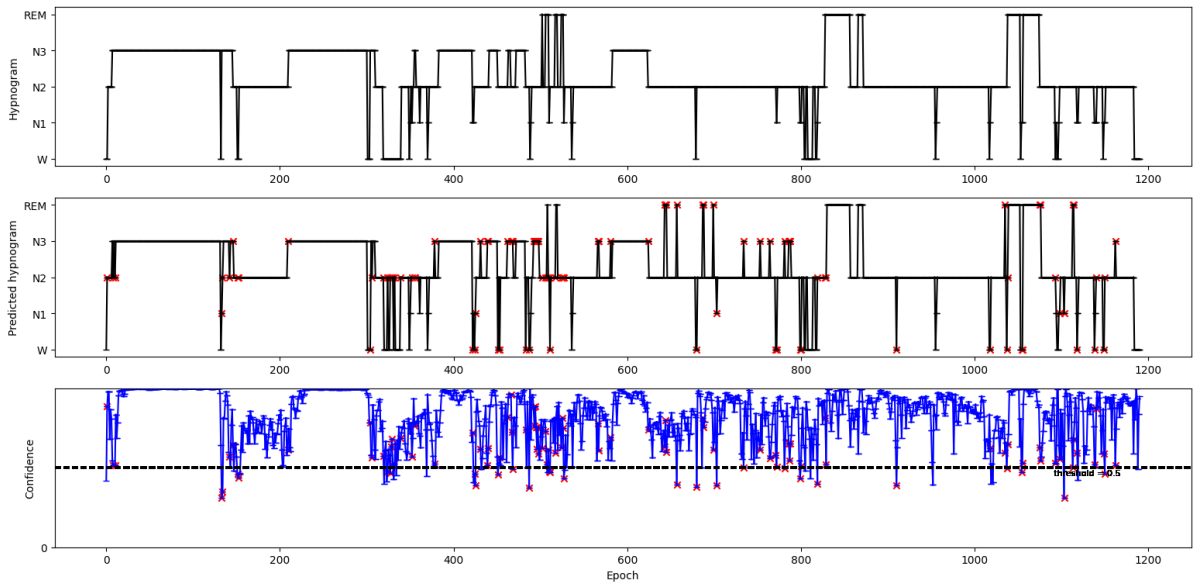


Figure E.1: Visualization of the estimated confidence for subject 1 of DOD-H.

to filter out those epochs for additional manual verification and correction. However, further inspection is needed to find out the optimal threshold in such a way that this relationship between error and low confidence is maximized. In these two cases, the mean confidence of correctly predicted epochs is 0.85 for subject 1 and 0.92 for subject 19. On the other hand, the mean confidence of misclassified epochs is 0.59 and 0.77, respectively. Accordingly, the selected threshold is lower than it should be in order to defer as many mistaken epochs as possible to a human expert.

Appendix F. Data augmentation methods

Data augmentation (DA) comprises the generation of new samples to augment an existing dataset by transforming existing samples. Exposing the classifiers to varied representations of its training examples makes the model less biased and more robust to such transformations when attempting to generalize the model to new datasets. The current appendix details the methods that have been used in this thesis to augment EEG signals. In particular, the following approaches were implemented with the aim of addressing imbalanced classification problems by artificially inflating the size of the minority classes.

F.1 Jittering (noise addition)

One of the simplest, yet effective, transformation-based data augmentation methods is jittering, or the act of adding noise to time series ([56], [57]). The jittering process can be defined as follows:

$$x' = \{x_1 + \epsilon_1, ..., x_t + \epsilon_t, ..., x_T + \epsilon_T\} \tag{F.1}$$

where ϵ refers to the noise addition at each step of the signal.
Adding noise to the inputs is a well-known method of increasing the generalization of neural networks. It is able to do this by effectively creating new patterns with the assumption that the unseen test patterns are only different from the training patterns by a factor of noise. Specifically, this technique assumes that it is normal for the time series patterns of the particular dataset to be noisy, which in fact is often true when dealing with EEG data.
Figures F.1 and F.2 show an example of an original epoch and its artificially generated version (achieved in this case by adding random noise up to 5% - 15% with 50% probability), respectively.

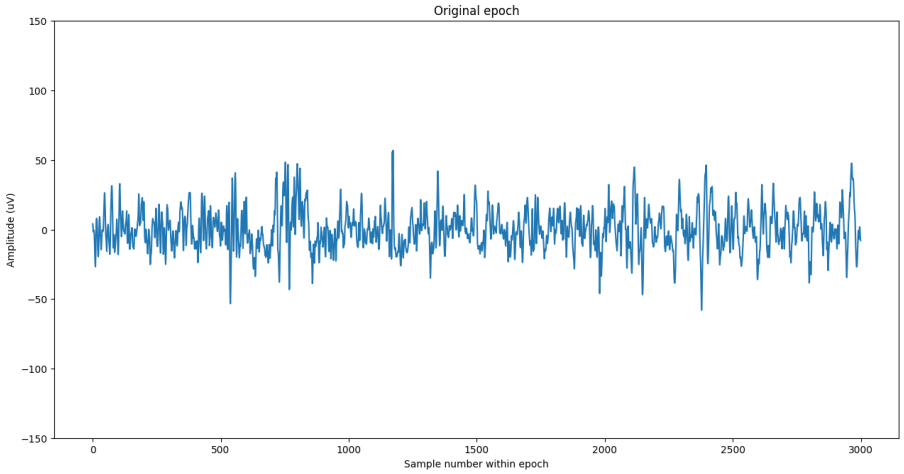


Figure F.1: Original N1 epoch from subject 21 in DOD-H dataset, channel F3-M2.

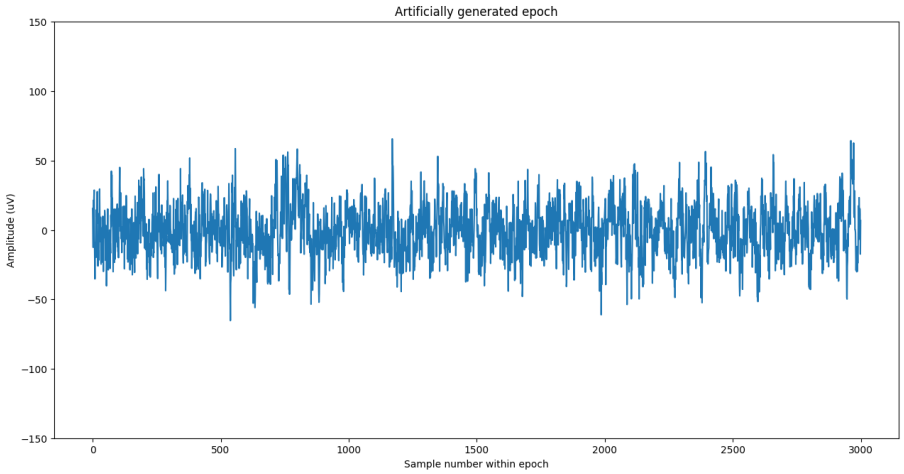


Figure F.2: Artificially generated N1 epoch from subject 21 in DOD-H dataset, channel F3-M2. Such epoch is obtained by adding random noise.

Another technique that was used together with the addition of noise is drift. The results achieved for the same epoch that has been shown below is displayed in Figure F.3. Specifically, random drift is applied to the signal up to 15% in addition to random noise up to 5% - 15% with 50% probability.

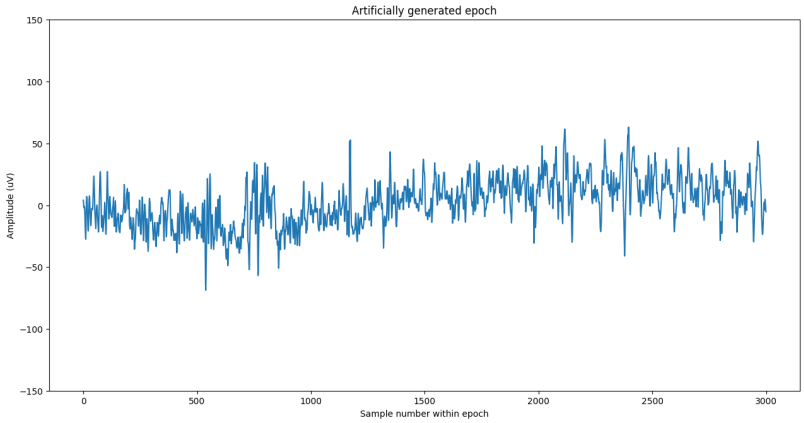


Figure F.3: Artificially generated N1 epoch from subject 21 in DOD-H dataset, channel F3-M2. Such epoch is obtained by adding random noise and drift.

F.2 Oversampling

One approach to addressing imbalanced datasets is to oversample the minority class. The simplest approach involves duplicating examples in the minority class. This can balance the class distribution but does not provide any additional information to the model.
An improvement on duplicating examples is to synthesize them from the minority class. The most widely used approach to synthesizing new examples is called the Synthetic Minority Oversampling TEchnique, or SMOTE for short ([58]). This technique is the one being used in this project. Using this method, the minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbors (i.e., by interpolation, selecting examples that are close in the feature space). Specifically, a random example from the minority class is first chosen. Then k of the nearest neighbors for that example are found (in this case, k=3). A randomly selected neighbor is chosen and a synthetic example is created at a randomly selected point between the two examples in feature space.
The approach is effective since the generated examples are plausible. This is due to the fact that they are relatively close in feature space to existing examples from the minority class. However, a downside of the technique is that the synthetic samples are created without considering the majority class, possibly resulting in ambiguous examples if there is a strong overlap for the classes.

F.3 Overlapping windows

The last approach that was implemented is called overlapping windows. In this method, all contiguous 30-s epochs of the same sleep stage are concatenated in the time domain. The concatenated blocks are then redivided into new 30-s epochs, all overlapping by a certain percentage. In this case, different overlap percentages were used, concretely: 10%, 20%, 30%, and 50%. The

process is illustrated in Figure F.4 using two concatenated 30-s epochs of REM sleep stage as example. The three newly generated epochs obtained as a result are shown in Figure F.5.

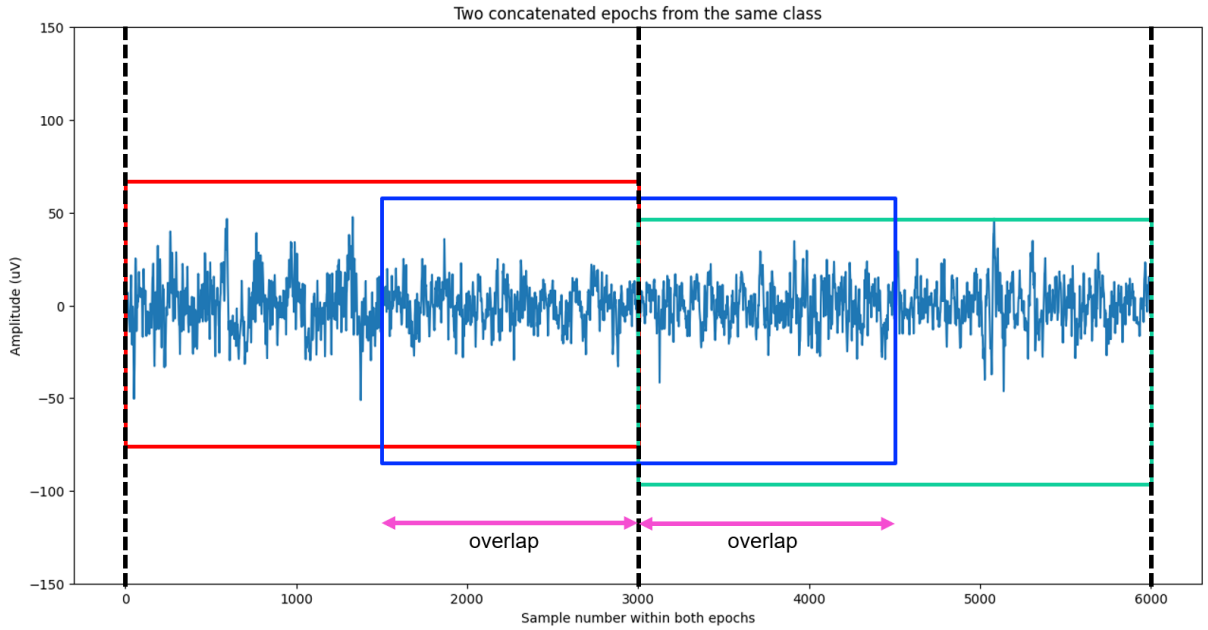


Figure F.4: Concatenation of two contiguous 30-s epochs of REM sleep stage: subject 19 in DOD-H dataset, channel F3-M2. The dotted vertical lines delimit the two epochs. The red, blue and green rectangles correspond to the newly generated epochs after applying the specified overlap (pink).

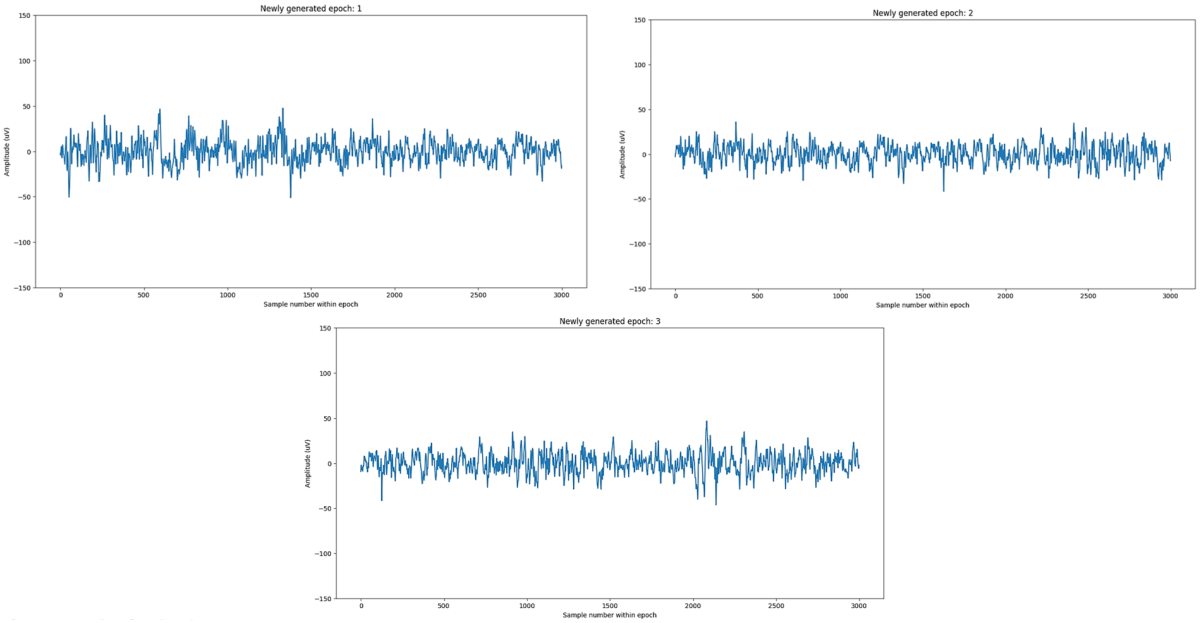


Figure F.5: Newly generated REM epochs after applying an overlap of 50%: subject 19 in DOD-H dataset, channel F3-M2.

Appendix G. Data augmentation results

As mentioned throughout the project, DA is used in order to balance the data. Concretely, this method is applied for DOD-H dataset, employing single-channel information from the frontal electrode location. Artificial samples of the minority classes are generated using different DA techniques: drift, noise addition, overlapping windows (using different overlap percentages), and oversampling. For a brief explanation of these methods, see Appendix F. Figure G.1 reports the distribution of the assigned sleep stage labels in such dataset before (A) and after (B) applying data augmentation. Specifically, the oversampling approach was used in this case.

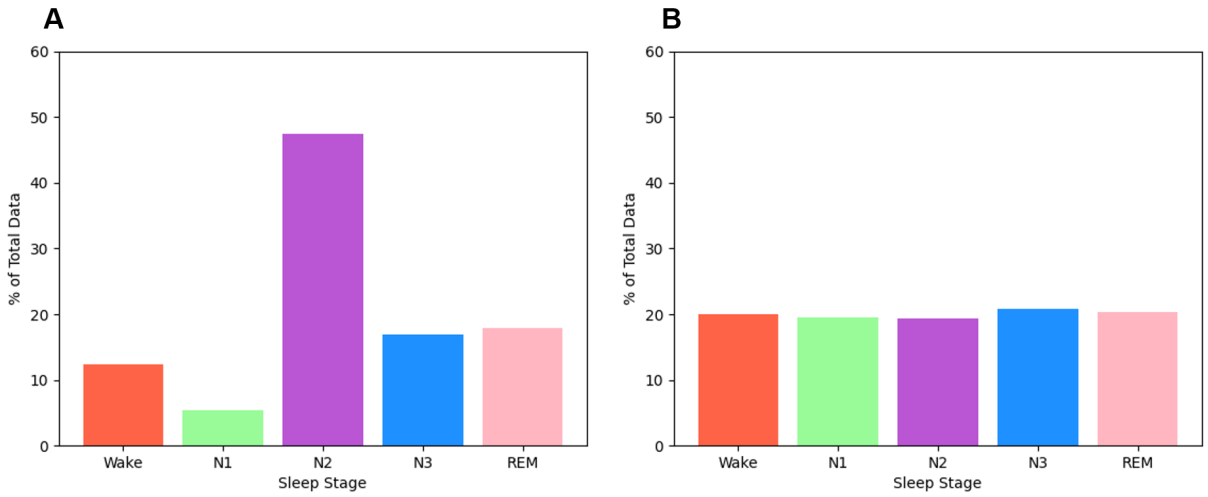


Figure G.1: Distribution of sleep stages in the DOD-H dataset before (A) and after data augmentation (B). This example is obtained applying the oversampling technique.

As it can be observed, the dataset is no more highly unbalanced toward the N2 class, which was roughly 3 times N3 and REM classes, 4 times the Wake class, and 10 times more frequent than the N1 class.

The classification reports achieved after applying the mentioned DA approaches are shown in Tables G.1 and G.2, while the resulting confusion matrices can be seen in Figures G.2 and G.3 for the CNN and CNN+RNN, respectively.
In the case of the CNN, none of the methods used exceeds the performance obtained without generating artificial samples. On the other hand, oversampling and overlapping (20% and 10% overlap) provide small performance gains for the CNN+RNN (approximately 2%, 0.5%, and 0.7%, respectively). Looking at the confusion matrices, it can be seen that both Wake and N1

show better values in terms of specificity (in the case of the CNN+RNN, this is not fulfilled for Wake class when applying oversampling, noise addition and noise addition along with drift). This is specially noticeable in the accuracy achieved for N1 sleep stage with the CNN+RNN model when using the overlapping approach (20% overlap). The agreement between scorer and network is almost double than without data augmentation. In contrast, the remaining sleep stages (N2, N3 and REM) do not always experience this increase and are even harmed. A clear example of the latter are the poor results of the CNN+RNN architecture for REM at any overlapping test. Its confusion with N1 is plainly larger.

Accordingly, it can be concluded that these techniques are not impacting the performance of the networks in the desired way. These strategies may be changing the distribution of raw data and causing overfitting. Besides, it is raising the computational complexity required during training. Data augmentation is commonly used in image classification tasks, achieving considerable performance gains. However, unlike images, EEG is a collection of noisy, non-stationary time-series. Therefore, it is possible that the transformations applied are not directly suitable for EEG data because those may impact and even destroy time-domain features. Something that could be done to overcome these issues in a future project is to play around with the loss function. The idea would be to weight the loss computed for different samples differently based on whether they belong to the majority or the minority classes. Higher weights would be assigned to the loss encountered by samples associated with minor classes.

Test	Accuracy	Precision	Recall	F1-score
DOD-H 8 - Without DA	0.7712	0.7923	0.7712	0.7668
DOD-H 8 - Oversampling	0.7705	0.8024	0.7705	0.7725
DOD-H 8 - Noise addition	0.7645	0.8063	0.7645	0.767
DOD-H 8 - Noise addition + drift	0.7536	0.7789	0.7536	0.7532
DOD-H 8 - Overlapping 50%	0.7412	0.7818	0.7412	0.7397
DOD-H 8 - Overlapping 30%	0.762	0.814	0.762	0.7648
DOD-H 8 - Overlapping 20%	0.7186	0.7826	0.7186	0.7151
DOD-H 8 - Overlapping 10%	0.7709	0.8065	0.7709	0.7753

Table G.1: CNN results: classification report for DOD-H data augmentation tests employing different approaches.

Test	Accuracy	Precision	Recall	F1-score
DOD-H 8 - Without DA	0.7508	0.7805	0.7508	07538
DOD-H 8 - Oversampling	0.7691	0.7819	0.7691	0.7653
DOD-H 8 - Noise addition	0.7456	0.786	0.7456	0.7526
DOD-H 8 - Noise addition + drift	0.7446	0.7788	0.7446	0.7515
DOD-H 8 - Overlapping 50%	0.7459	0.7898	0.7459	0.7487
DOD-H 8 - Overlapping 30%	0.7425	0.7979	0.7425	0.7545
DOD-H 8 - Overlapping 20%	0.7551	0.8079	0.7551	0.769
DOD-H 8 - Overlapping 10%	0.7576	0.787	0.7576	0.7609

Table G.2: CNN+RNN results: classification report for DOD-H data augmentation tests employing different approaches.

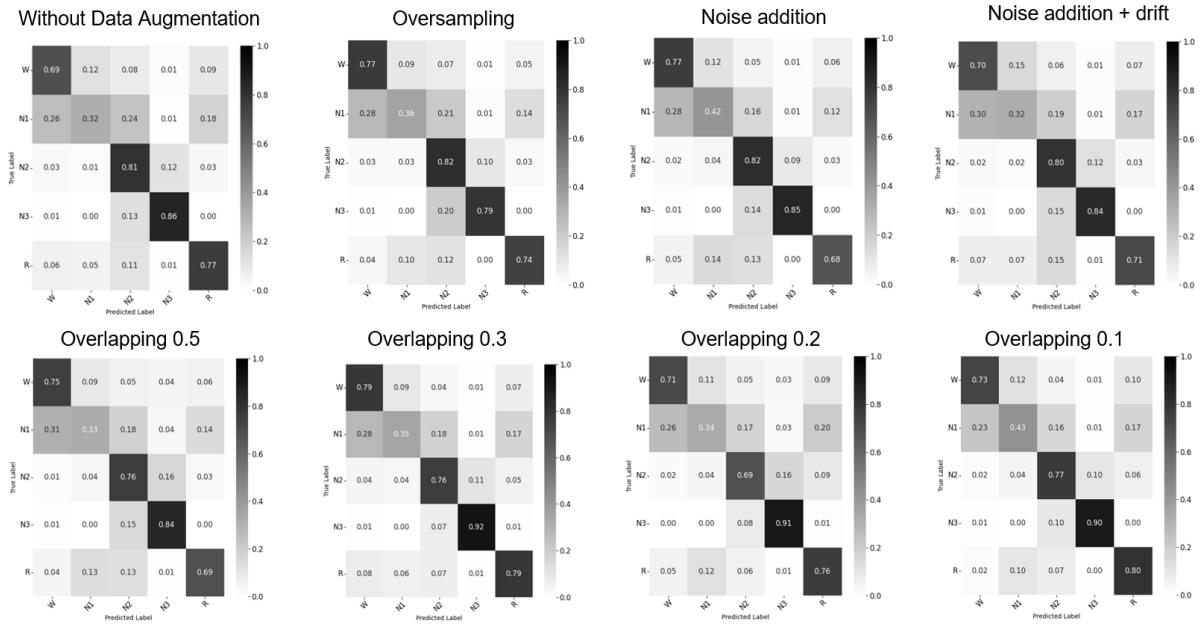


Figure G.2: CNN results: Confusion matrix for DOD-H data augmentation tests employing different approaches.

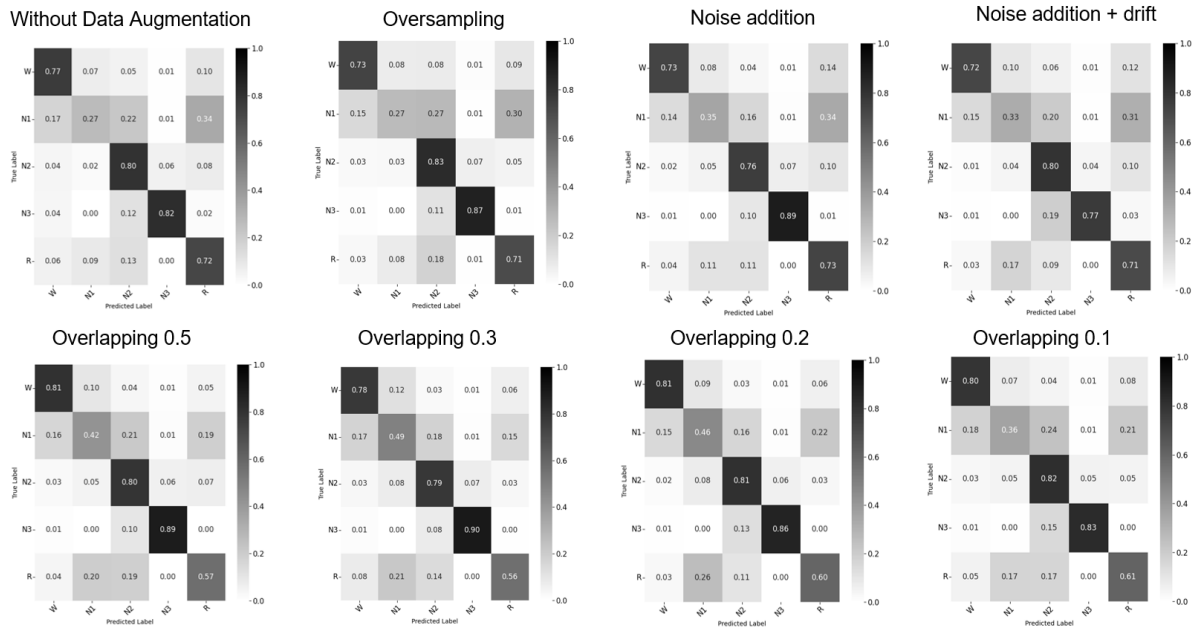


Figure G.3: CNN+RNN results: Confusion matrix for DOD-H data augmentation tests employing different approaches.

Appendix H. Transfer learning results

H.1 Results on publicly available datasets: DOD-H, DOD-O & ISRUC

The capability of the CNN model to perform transfer learning between datasets is assessed. In the first evaluation, weights from DOD-H 1 are directly tested on DOD-O. The accuracy shows a decrease from 75.63% to 46.78 % (DOD-O 1). Then, the same test is performed again but in the opposite direction, obtaining an accuracy of 61.8%. This implies a fall of 15.32% under the results obtained for DOD-H 1. Finally the weights from ISRUC 1 are tested on DOD-O, yielding an accuracy of 61.58% (14.05% less than in DOD-O 1).
Afterwards, a modification is performed in the workflow of the three previous tests. It consists of removing the output layer of the pre-trained model and creating a new model of two layers on top of it. The latter are trained on the new dataset, while the layers of the base model are frozen. This variation leads to the performance shown in Table H.1 (classification report) and in Figure H.1 (confusion matrices). As it can be observed, the results are closer to those achieved training and testing with the same dataset (DOD-H 1: 77.12%, DOD-O 1: 75.63%, ISRUC 1: 70.72%). Furthermore, both modus operandi lead to the conclusion that datasets comprised of patients generalize better than those formed by healthy subjects. This is logical, since training entails larger difficulties and then it is easier to score examples that present healthy sleep with a normal structure, certain regularity and absence of sleep disruptions. On the contrary, it is very difficult to predict on nights containing abnormal patterns coming from sleep related pathology if they have never been seen before.

Test	Accuracy	Precision	Recall	F1-score
DOD-H 7 - Transfer to DOD-O	0.7049	0.7158	0.7049	0.6849
DOD-O 7 - Transfer to DOD-H	0.7804	0.8047	0.7804	0.7785
ISRUC 7 - Transfer to DOD-O	0.7187	0.7375	0.7187	0.7097

Table H.1: Classification report results for the transfer learning tests with DOD-H, DOD-O and ISRUC.

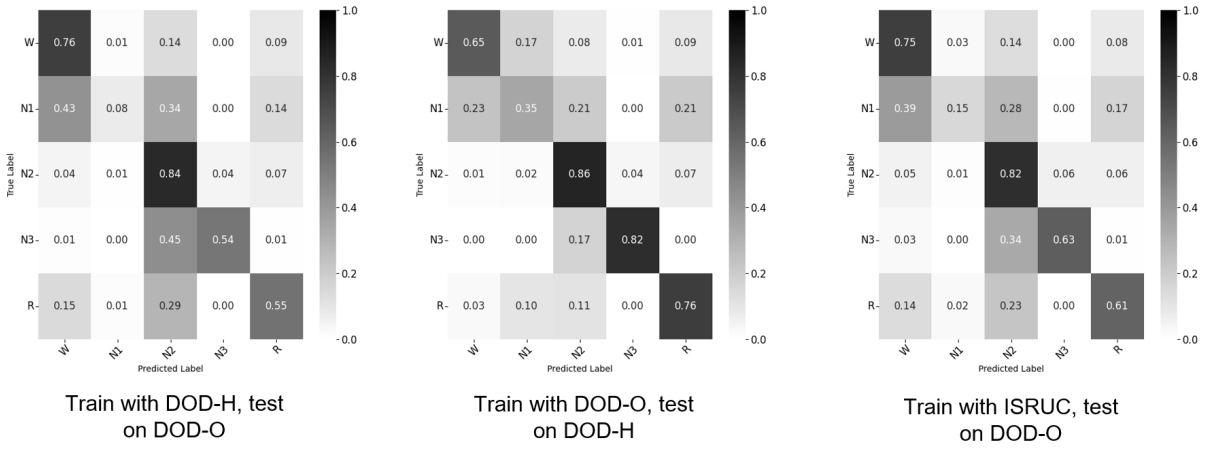


Figure H.1: Confusion matrices for the transfer learning tests with DOD-H, DOD-O and ISRUC.

H.2 Results on Bitbrain's data recordings

Four different transfer learning tests were performed. The workflow is the same as before, the output layer of the pre-trained model is removed and a new model of two layers is created on top of it. The two new layers are trained on the new dataset, while the layers of the base model are frozen.

In the first test, weights learned from PSG data are tested on the headband. The remaining three employed the weights learned from the massive STAGES dataset. Of these three, in the last two, a part of the base model is unfreezed and retrained with a lower learning rate on the new data (this fine tuning is performed from layer 4 and 3 in the two different tests, respectively). The results obtained are shown in Table H.2 and in Figure H.2. Accuracy and recall are low in all the cases (around 50%). Besides, precision and f1-score are even worse (less than 40%). The poor performance is corroborated looking at the confusion matrices, where it can be seen that the network is labelling practically all epochs as N2 in a random way.

This transfer learning approach is not suitable for the real-time purposes of the project. Nevertheless, the results of a direct transfer would be less acceptable. This implies that the algorithm is sensitive to the format of the data and the system employed to record each night. This is a strong limitation considering that in clinical settings, training the model on one dataset and using it to stage another which is unlabelled, may be of interest. However, the network has the intended use of being coupled, trained and tested on *Bitbrain*'s EEG hardware. Accordingly, this kind of transfer scenarios will not take place.

Test	Accuracy	Precision	Recall	F1-score
BITBRAIN 12 - Transfer PSG to HB	0.5327	0.3984	0.5327	0.3908
STAGES 16 - Transfer STAGES to BITBRAIN (HB)	0.5093	0.3378	0.5093	0.369
STAGES 16 (2) - Fine tuning from layer 4	0.5209	0.3363	0.5209	0.3794
STAGES 16 (3) - Fine tuning from layer 3	0.5235	0.3371	0.5235	0.3744

Table H.2: Classification report results for the transfer learning tests on Bitbrain's headband data .

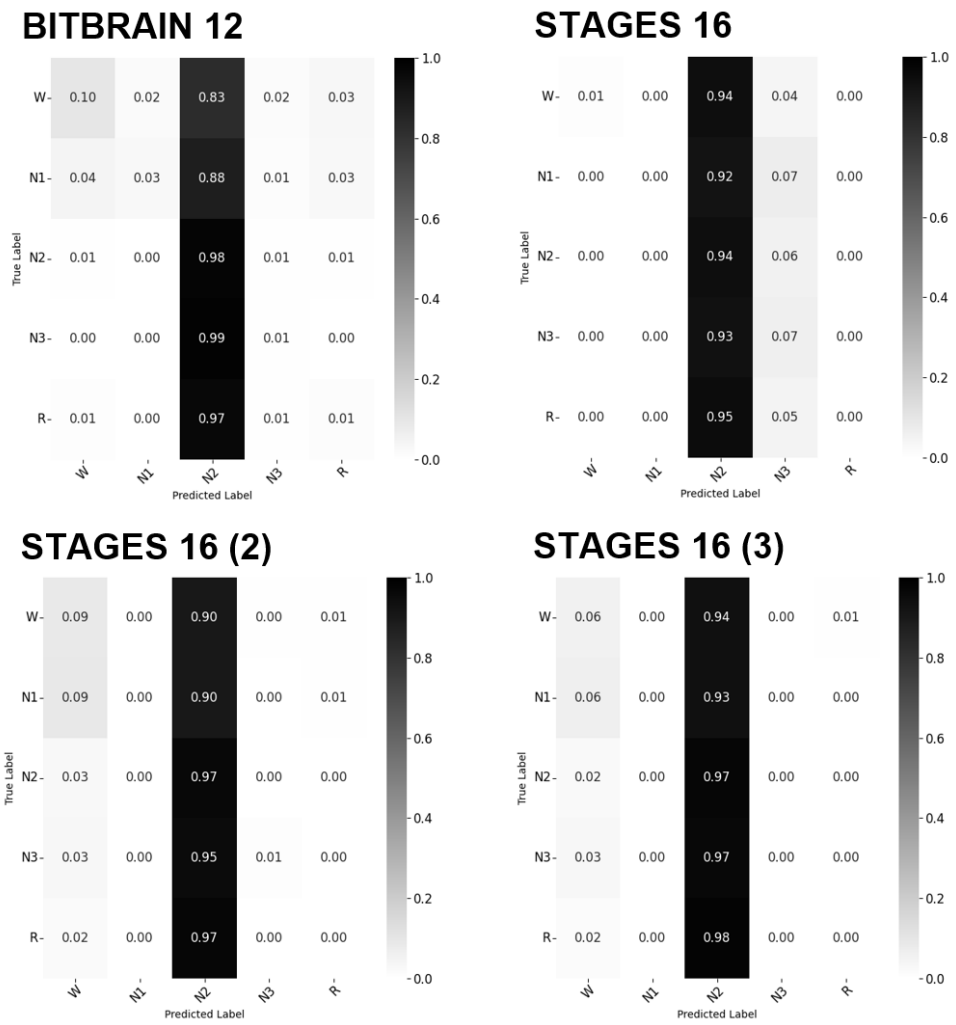


Figure H.2: Confusion matrices for the transfer learning tests on Bitbrain's headband recordings (the test number is indicated in the upper left corner of both matrices).

Appendix I. Bitbrain interface coupling

After integrating the code in *Bitbrain*'s software platform, two screenshots (Figures I.1 and I.2) were made in order to show examples of two different recordings being scored in real-time (it can be seen how highly contaminated with artifacts and noise the signals are). These recordings are performed with the headband and sampled at 256 Hz.

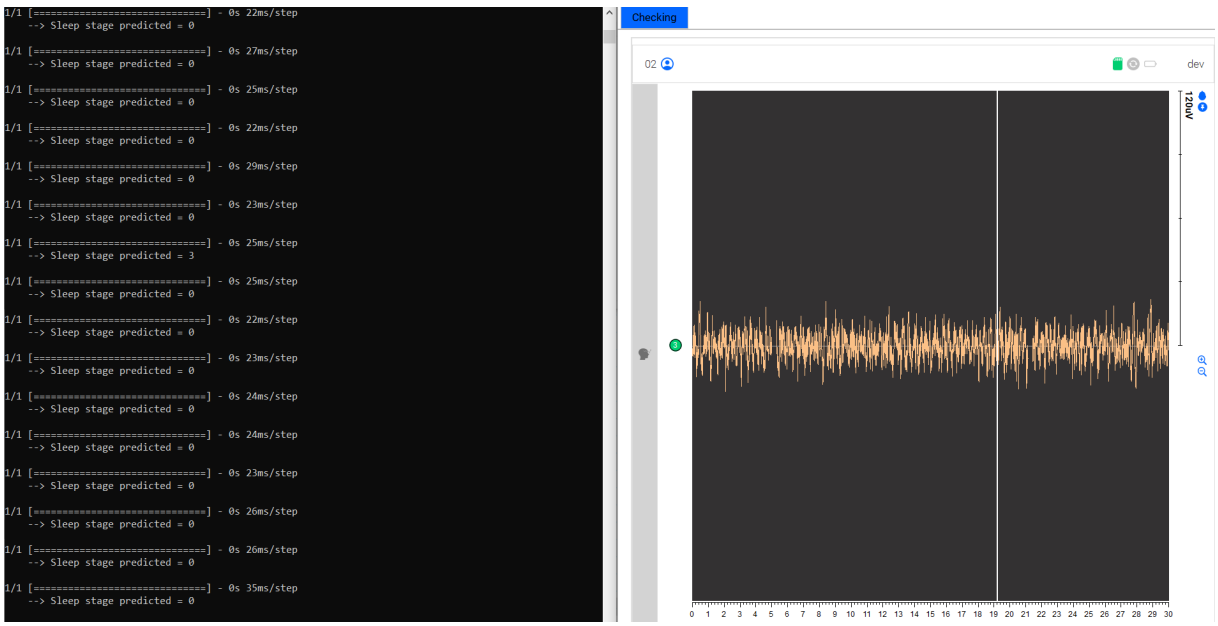


Figure I.2: Example 2: Real-time decoding of EEG performed in *Bibtrain*'s software platform. Left: labels predicted by the pre-trained model on DOD-H (F3 electrode). Right: headband signal being acquired in real-time (showing: Fp1 electrode). Zooming is recommended.

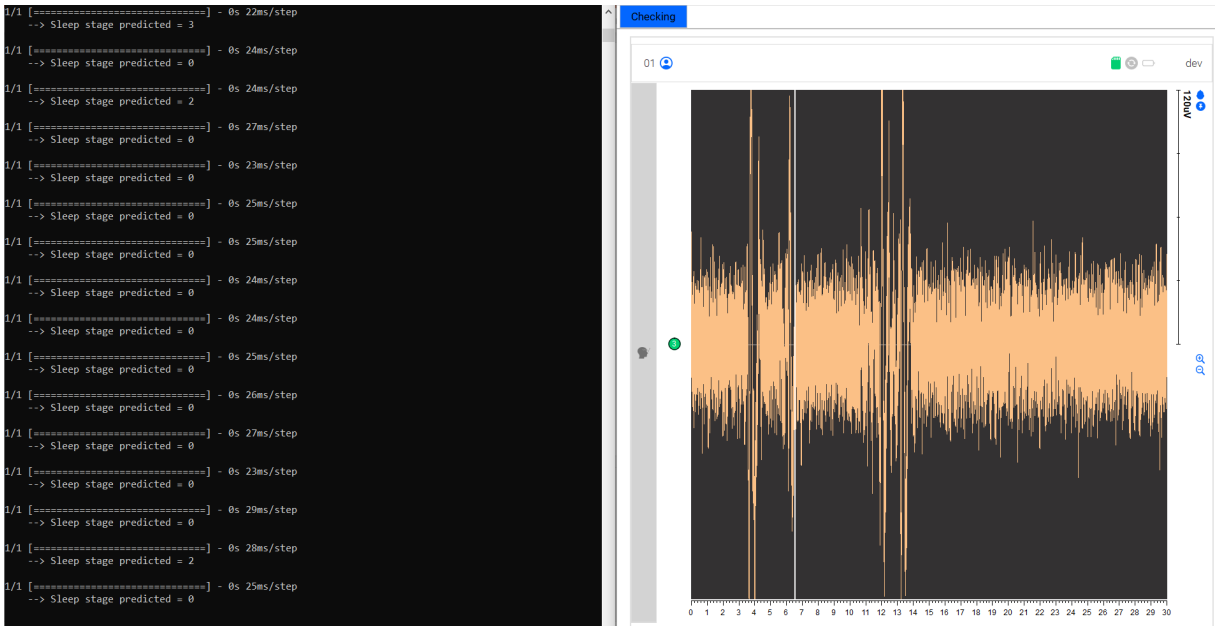


Figure I.1: Example 1: Real-time decoding of EEG performed in *Bibtrain*'s software platform. Left: labels predicted by the pre-trained model on DOD-H (F3 electrode). Right: headband signal being acquired in real-time (showing: Fp1 electrode). Zooming is recommended.

Appendix J. Auditory closed-loop stimulation

Traditional transcranial brain stimulation techniques, such as transcranial magnetic stimulation (TMS), transcranial direct current stimulation (tDCS), and invasive deep brain stimulation (DBS) have been widely investigated in neuroscience for decades. A new, less invasive technique, auditory stimulation during sleep, is based on presenting short tones at particular times during the Non-Rapid Eye Movement (NonREM) sleep stage. This type of stimulation evokes physiological responses in the brain, including slow waves and sleep spindles, that are vital for many important cognitive processes ([47]), but are impaired in the elderly and particularly in people with dementia. Auditory stimulation during sleep has been shown to have a series of benefits, including improving memory consolidation ([59], [60], [61]). Other studies showed similar benefits in the elderly ([62], [63]) and even in mild cognitive impairment (MCI) patients ([64]). Auditory stimulation during sleep has also been applied in other clinical populations, e.g., to suppress epileptic activity in children with Rolandic epilepsy ([65]).

The idea of our project consists of an advanced implementation of the auditory stimulation technique that can be used flexibly to improve memory consolidation and other cognitive abilities in healthy young or elderly individuals, as well as patients with memory impairments (e.g., due to mild cognitive impairment).

The stimulation will be performed as trains of 3 tones, since even without stimulations (i.e., spontaneously), slow oscillations tend to come in trains of 2-4 events. During the study, people will come to *Bitbrain*'s laboratories for 3 nights: a calibration night (detect slow waves and use fixed, predefined stimulation parameters (taken from literature) to stimulate with single tones. This will help to estimate optimal values for the stimulation parameters), and two experimental nights (using the values estimated in the calibration night, the stimulation will be performed in following experimental nights: stimulation and placebo nights).

Importantly, auditory stimulation can only have the desired benefits if applied during deeper NonREM sleep, more specifically the so-called N2 and N3 sleep phases. This is because stimulation in other sleep stages can more easily wake the person (N1) or will not have the desired neurophysiological effect (REM sleep). Accordingly, when the correct phase is detected by the automatic sleep stage decoder, auditory stimulation is started. To have its strongest effects, it must be applied during the excitatory up states (see Figure J.1) of ongoing slow waves ($\sim$0.5-4 Hz).

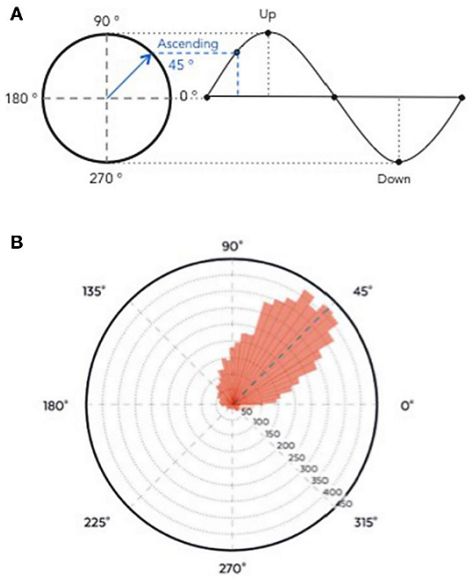


Figure J.1: Stimulation is best applied during early slow wave up states. A) Phase convention for the polar plot. B) The goal is to hit the signal at around 45^{o} (between the down-up zero crossing and the up peak at 90^{o}).

The recordings will be done with electrodes over frontal areas of the brain (i.e., using the new *Bibtrain*'s headband). To hit the up phase of the slow waves, a second decoder is needed in order to detect a spontaneously occurring slow wave down peak. This detection will trigger a tone with the delay after which down peaks are typically followed by a subsequent up state ("down-to-up peak delay" or "delay I"). After another, longer delay, a tone is delivered again to hit the next up state, and once more using the same delay to hit the one after that ("up-to-up peak" or "delay II"). Regarding the stimulation volume, it is desired to stimulate as loud as possible to elicit a neural response without waking up the participants. This volume is individualized to the person and typically ranges at around 60 dB.

The most important aspects about the implementation of the slow wave detection and stimulation algorithm are briefly described in the following sections. The exact details and parameters might be subject to change throughout the project, optimal values might be found during development.

J.1 Online slow wave detection

The main steps performed by the slow wave decoder must be:

- Filtering the incoming signal in the slow wave range.
- Detecting down peaks using an amplitude threshold, and rejecting them if they are under a specific artifact rejection threshold.
- Detect subsequent up peaks from a continuously updated average wave. This step is useful to compute delay I and delay II and re-calculate them since they may change during a

single night.

Important parameters for the detection are the filter boundaries and type as well as the detection and artifact thresholds. Currently, this algorithm starts only if it is enabled manually by the experimenter and once the participant has entered stable NREM sleep. Once the automatic sleep scoring algorithm has been trained with the new data, it will be in charge of providing an input signal to the slow wave decoder.

J.2 Stimulation

The stimulation routine is triggered when slow wave down peaks are found by the online detection algorithm. The routine will produce 3 tones timed to the slow wave up states that follow the detected down peaks (see Figure J.2). Important parameters for the stimulation are the stimulation volume, delay I (usually between 480 and 580 ms), delay II (usually between 1 and 1.2 s), and pause duration (after a stimulation train has been completed, further stimulations are paused for 2.5 s).

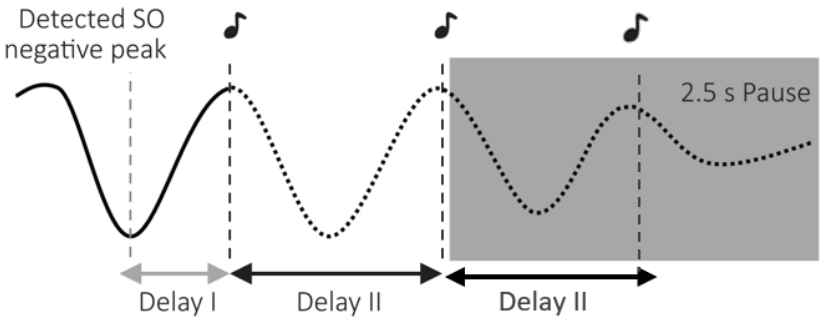


Figure J.2: Down peak detection (light grey vertical line) and a first tone that is triggered during the subsequent up state after delay I (first dark grey vertical line). Another tone is triggered during the second up state after delay II. A third tone is applied after another delay II. Stimulation is followed by a pause, before detection and stimulation are continued.

J.3 Data analysis and parameter estimation

In addition to the previous algorithm, an offline detector is designed. It consists of an improved slow wave detection that can use the entire EEG signal. This algorithm will help to assess the sleep and stimulation success for each night as well as to estimate the optimal stimulation parameters for the subsequent ones.