POLITECNICO DI TORINO

Master's Degree in Mechatronic Engineering

Politecnico
di Torino

Master's Degree Thesis

Machine Learning Strategies for Single-Channel EEG Automatic Detection of REM Sleep Behavior Disorder: a Model Based on REM and Slow Wave Sleep

Supervisors

Prof. Gabriella OLMO

Ing. Irene RECHICHI

Candidate

Gabriele S. GIARRUSSO

July 2023

Abstract

Human Sleep is the cyclic repetition of states characterized by different processes which play an important role in a wide range of activities, such as restoring the body's energy as well as supporting memory consolidation, and clearance of metabolic waste products generated by awake brain neural activity. It is mainly divided into two macro-stages, namely Rapid-Eye movement (REM) sleep and non-Rapid-Eye movement (NREM) sleep which, accordingly to the *American Academy of Sleep Medicine* (AASM) guidelines, is further characterized by three stages N1, N2, and N3 (or Slow Wave Sleep).

In recent years, Sleep Disorders got the researchers' attention and some studies demonstrated a strong correlation with some types of Neurodegenerative Disease pathogenesis. Remarkably, Idiopathic REM Sleep Behavior Disorder (iRBD) shows the strongest correlation with the family of α-synucleinopathies, e.g. Parkinson's Disease (PD), Dementia with Lewy Bodies (DLB), and Multiple System Atrophy (MSA), and it is supposed to represent an early symptom of these neurodegenerative conditions with particularly high rate of phenoconversion into PD (up to 80% after 14 years). Recent studies also highlighted the potential role of disturbed Slow Wave Sleep (SWS) as a predictive biomarker of neurodegenerative processes that involve both PD and Dementia.

This work aims to overcome the limits of the to-date available diagnostic tools proposing a fully-automatic EEG-based strategy able to detect RBD exploiting segments recorded during both REM and SWS sleep. Supervised Machine Learning models were trained and tested in a Leave-One-Out cross-validation framework, obtaining values of accuracy up to 91% and of sensitivity up to 94% (RBD class) and hence highlighting REM and SWS microstructures capabilities as RBD biomarkers. These results point to the potential of an EEG-based, low-cost, automatic RBD detection system which can be used in early diagnosis of neurodegeneration in order to spot prone individuals, and allowing them to join clinical trials of neuroprotective therapies to halt or at least delay the progression.

The last part of the work concerns the development of a three-stages semi-supervised-based method to qualify REM Sleep without Atonia (RSWA) as an intermediate pathological state supporting a finer characterization of the neurodegenerative progression from Healthy to RBD. The achieved values of Rand-Index (0.97) and Clustering Purity (0.99) confirm the existence of peculiar RSWA EEG-patterns as well as suggest the reliability of the process as a basis from which to carry out deeper analysis.

Table of Contents

List of Tables

List of Figures

Acronyms

A3P	Peak to Peak Period
AASM	American Academy of Sleep Medicine
AI	Artificial Intelligence
AL	Average Length
A.O.U	Azienda Ospedaliero-Universitaria
AP	Absolute Power
ARI	Arousal Index
ARi	Adjusted Rand Index
ASPS	Advanced Sleep-Phase Syndrome
AUC	Area Under the Curve
AZCP	Average Zero-up Crossing Period
BAG	Bootstrap Aggregation
CDF	Cumulative Distribution Function
CF	Crest Factor
CSAS	Central Sleep Apnea Syndrome
CV	Cross Validation
DBi	Davies-Boulding Index
DFA	Detrended Fluctuation Analysis
DLB	Dementia with Lewy Bodies
DPSS	Discrete Prolate Spheroidal Sequence
DSPS	Delayed Sleep-Phase Syndrome
DT	Decision Tree
DTFT	Discrete Time Fourier Transform
DWT	Discrete Wavelet Transform
EEG	Electroencephalography
EKG	Electrocardiography
EMG	Electromyography
EOG	Electrooculography

ESD	Energy Spectral Density
FD	Frequency Domain
FDS	Flexor Digitorum Superficialis
FF	Form Factor
FN	False Negative
FNR	False Negative Rate
FPR	False Positive Rate
FREM	Phasic REM Activity
FT	Fourier Transform
HFD	Higuchi Fractal Dimension
HPF	High Pass Frequency
H/C	Healthy-Control
ICSD-3	International Classification of Sleep Disorders 3
IF	Impact Factor
IQR	Interquartile Range
iRBD	Idiopathic REM Sleep Behavior Disorder
KNN	K-Nearest Neighbors
kSVM	Kernelized Support Vector Machines
LDA	Linear Discriminant Analysis
LOO-CV	Leave-One Out Cross Validation
LPF	Low Pass Frequency
MDF	Median Frequency
MI	Mutual Information
MID	Mutual Information Difference
MIFS	Mutual Information-based Feature Selection
ML	Machine Learning
MLST	Multiple Sleep Latency Testing
MMD	Maximum-Minimum Distance
MP	Mean Power
MREM	Minute of REM Sleep
mRMR	minimum-Redundancy Maximum Relevance
MSA	Multiple System Atrophy
MWT	Maintenance of Wakefulness

NB Naive Bayes
ND Neurodegenerative Disease
NFI NREM Fragmentation Index
NND Nonlinear Domain
NP Number of Peaks
NREM non-Rapid-Eye Movement

OSAS Obstructive Sleep Apnea Syndrome

PD Parkinson's Disease
PGO Ponto Genicular Occipital
PKF Peak Frequency
PLMD Periodic Limb Movement Disorder
PP Peak Prominence
PSD Power Spectral Density
PSG Polysomnography
PW Peak Width

QoL Quality of Life

RAI REM Atonia Index
RASb Ratio between Successive sub Bands
RBAs Relief-based algorithms
RBD REM Sleep Behavior Disorder
RBDQ REM Sleep Behavior Disorder Questionnaire
REM Rapid-Eye Movement
REN Renyi's Entropy
RF Random Forest
RFI REM Fragmentation Index
RK Rechtscheffen & Kales
RLS Restless Legs Syndrome
RMSe Root Square Mean error
ROC Receiving Characteristic Curve
RP Relative Power
RSWA REM Sleep without Atonia

SampEn Sample Entropy
SCe Spectral Centroid

SCr Spectral Crest
SE Sleep Efficiency
SEN Spectral Entropy
SK Spectral Kurtosis
SLD Sublaterodorsal Tagmental Nucleus
SM Spectral Moment
SOL Sleep Onset Latency
SOs Slow Oscillations
SSk Spectral Skewness
SSP Sleep Stage Proportion
SSp Spectral Spread
STFT Short Time Fourier Transform
STI Sleep Transition INdex
SubC Subcoeruleus Nucleus
SVM Support Vector Machines
SWS Slow Wave Sleep
SWA Slow Wave Activity

TD Time Domain
TFD Time-Frequency Domain
TIB Time In Bed
TKEO Teiger-Kaiser Energy Operator
TN True Negative
TNR True Negative Rate
TP True Positive
TPR True Positive Rate
TREM Tonic REM Activity
TST Total Sleep Time

UoM Unit of Measure

VMM Ventromedial Medulla
vPSG video-Polysomnography

WASO Wake After Sleep Onset
WP Proportion of Wake
WT Wavelet Transform

ZCR Zero Crossing Rate

Chapter 1

Sleep and REM Sleep Behavior Disorder

From a physiological perspective, sleep is the cyclic succession of states characterized by different physiological, biochemical, and psychological processes resulting in a complex neurological condition. Although its exact function is not totally explored yet, it is possible to state that the main role played by sleep concerns providing rest and restoring the body from an energy level and homeostatic point of view as well as supporting memory consolidation, and clearance of metabolic waste products generated by awake brain neural activity. Spending a sufficient amount of time in sleep is essential and sleep deprivation, resulting both from lifestyle or sleep disorders, may negatively affect different processes such as attention, blood pressure regulation, immune control, hormone production, and memory, worsening physical and mental health and bringing to a wide spectrum of causes ranging from short-term (impaired attention) to long-term consequences (neurological disorders, strokes, and depression)[1], [2]. In general it is estimated that about one-third of life is spent sleeping even if there is a wide variation in quality and amount among individuals, based on various factors like age, sex, and habits.

1.1 Sleep Macrostructures

Based on physiological Polysomnographic measurements involving Electroencephalography (EEG), Electrooculography (EOG), and Electrocardiography (EKG) normal human sleep is divided into two macro-stages (sleep macrostructures) with independent function and control mechanisms, namely non-Rapid-Eye movement (NREM) sleep and Rapid-Eye movement (REM) sleep which alternate defining sleep cycles that in adult humans can last an average of 90-110 minutes for a total of 4-6 cycles per night.

Figure 1.1: Sleep cycle division.
https://www.sleepfoundation.org/stages-of-sleep/nrem-sleep

In physiological conditions, NREM sleep accounts for about 75% of the total sleep time and according to *American Academy of Sleep Medicine* (AASM) is subdivided into three stages N1, N2, and N3 or Slow Wave Sleep (SWS) while REM sleep (R) accounts for the remaining 25% (according to the old RK the stages are more finely divided into N1-N4 and R). Normally each cycle begins in NREM sleep starting with N1 and switching to deeper stages (N2, and then N3), before finally reaching REM sleep. In healthy adult humans the first one-third of sleep is mainly dominated by SWS while the last one-third by REM sleep [3] as it is possible to notice by analyzing a Hypnogram (Figure 1.2) i.e. the Medical representation of the cyclic progression of the various sleep stages over time.

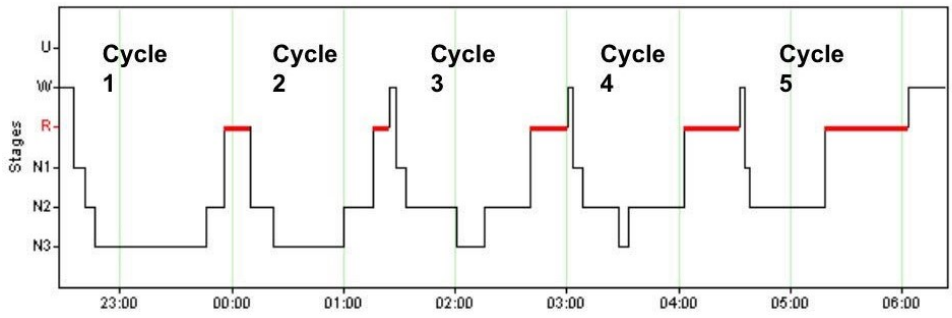


Figure 1.2: Hypnogram of physiological sleep.

From the EEG analysis, each sleep stage can be associated with a specific spectral power content, characterized by frequencies and amplitudes of the brain waves which occupy the following bands: δ(0.5-4 Hz, 20-100 uV), θ(4-8 Hz, 10 uV), α(8-13 Hz, 2-100 uV), β(13-30 Hz, 5-10 uV), and γ(>30 Hz, <2 uV).

1.1.1 NREM

During NREM sleep, EEG waves show increased voltage and lower frequency components as the sleep goes from N1 to N3, representing progressively deeper sleep stages characterized by a reduced responsiveness to external stimuli. Throughout the deeper phases the muscles are relaxed, the heart rate and blood pressure decline, while the gastrointestinal mobility and the parasympathetic activity become higher.

N1 (5% of sleep) is the lightest stage of sleep and is mainly characterized by low-amplitude mixed-frequency activity (4-7Hz). During this stage muscle tone is present and breathing rate is regular.

In N2 (45% of the sleep) the heart rate and the temperature drop, and characteristic structures such as K-complexes (long δ waves) and Sleep Spindles (powerful bursts in β band) are present, coupled or alone. Different studies suggest that these structures play an important role in memory consolidation while only the former has been shown to perform a sleep stabilization function [4].

N3 (25% of sleep) is the deepest stage of sleep with activity in δ band, where the body repairs and regrows tissues as well as builds bones and muscles and strengthens the immune system. This stage is the most difficult to interrupt with subjects awoken during this phase experiencing transient mental fogginess (sleep inertia) and moderately impaired mental performances for up to one hour.

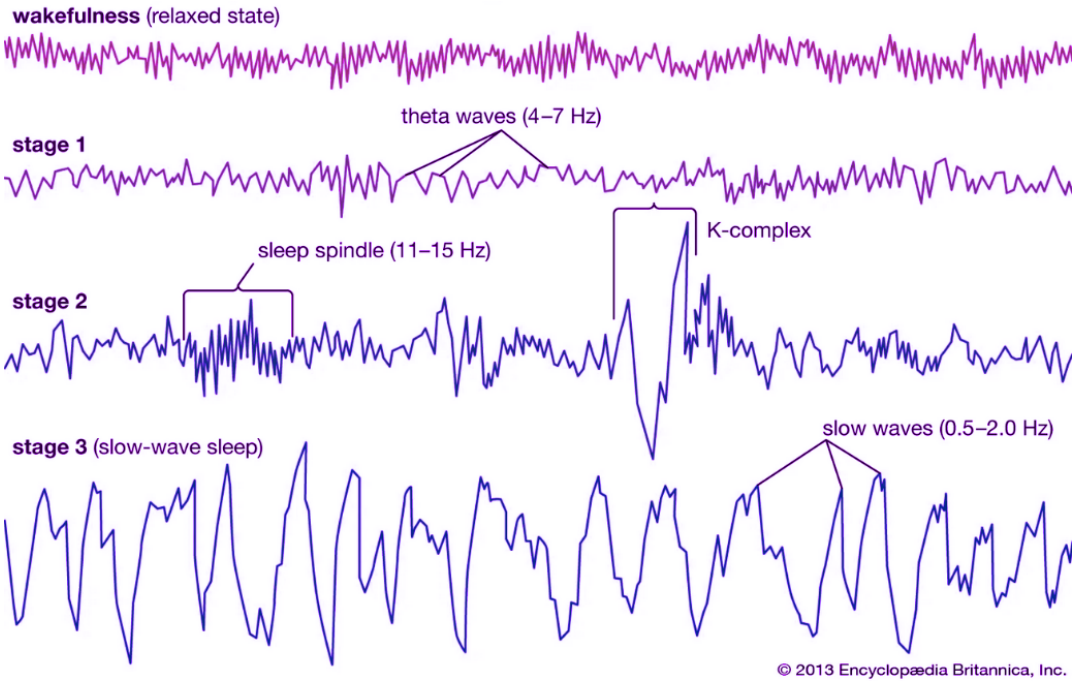


Figure 1.3: Typical brain waves of NREM [5].

1.1.2 REM

REM (25% of sleep) contains low-amplitude mixed-activity in almost all the fundamental frequency bands. It is associated with the most intense dreaming activity and it is not considered a restful stage. While REM EEG activity presents similarities with wake (for this reason is also said paradoxical sleep) (Figure 1.4), muscle activity is almost absent (atonia) due to brainstem-mediated inhibition of motor neurons, except for peripheral muscle twitches, burst-like saccadic eye movements, and diaphragmatic breathing muscles involved in irregular breathing rate. Moreover, phasic swings in blood pressure and heart rate are present as well as potentially a few periods of apnoea and hypopnoea [2].

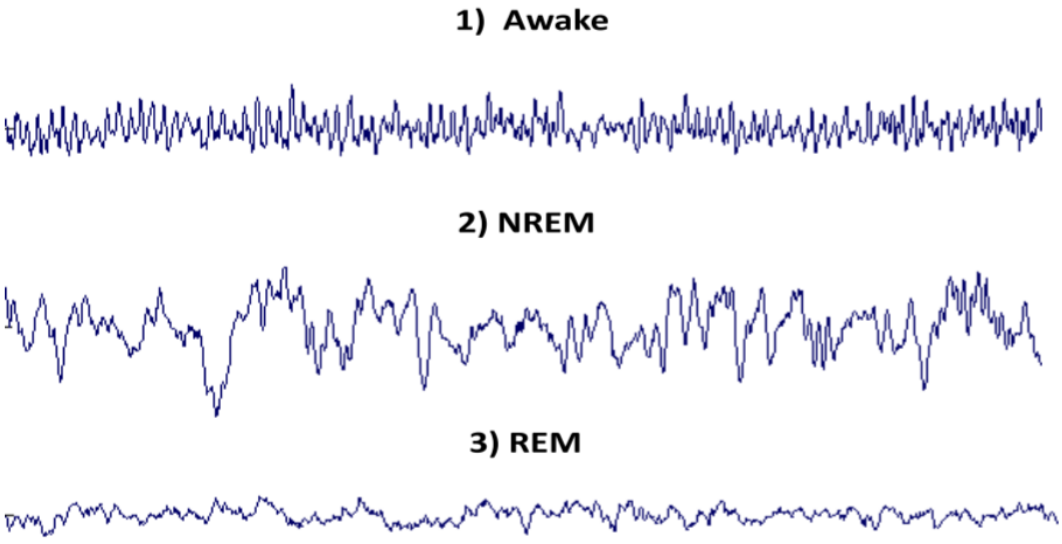


Figure 1.4: Comparison of Awake, NREM and REM wave shapes [6].

This stage usually starts at the end of the first sleep cycle and gets longer throughout the night, lasting about ten minutes in the first cycle up to one hour in the final one after which people tend to awaken spontaneously [4]. The large amount of REM sleep in mammals led to different hypotheses about the importance of its role in brain maturation, concerning the development of the sensorimotor system and ranging from "basic mechanisms such as regulation of brain temperature, modulation of receptor sensitivity, and synaptic plasticity to more complex processes such as procedural and declarative learning, emotional memory processing, or the development of consciousness" [7].

Although commonly treated as a homogeneous state mainly characterized by desynchronized EEG activity, REM is based on the alternation of two different

microstructures namely phasic and tonic REM periods which in recent years have attracted the attention of researchers thanks to their relevance in the study of sleep disorders. Phasic REM (FREM) periods are mainly characterized by bursts of eye movements (REMs) related to so-called ponto-genicular-occipital (PGO) waves, contraction of the ear muscles, myoclonic (brief, involuntary, and irregular) twitches of skeletal muscles, sawtooth waves (triangular, 2-4Hz), and irregularities in both respiratory and cardiac activity. On the other hand, tonic REM (TREM) periods concern longer and quiescent segments mainly present in between periods of phasic activity and are characterized by feature muscle atonia and no significant ocular movements. Although the latter seems a less activated state, studies demonstrated its higher behavioral alertness and stimuli responsiveness, lower arousal threshold, as well as external information processing, in contrast to the phasic state that appears pretty incompatible with sensory stimulation (high arousal threshold) as this may inhibit REMs and promote the transition to the tonic state suggesting phasic REM as the one linked to intense dream experiences [7]. These findings seem coherent with the studies on EEG Power Spectral Density (PSD) in REM microstates characterized by a decrease in α-β activity and a predominance of activity in δ and θ bands (2-8 Hz) for FREM, with the last presumably involved in memory consolidation processes and neuronal plasticity, as well as a prevalence in α and β bands (7-16 Hz) for TREM [8], [9]. Another interesting finding concerns the FREM power relative increase in higher frequencies included in the γ band, assumed "to reflect intense sensorimotor, emotional and cognitive processes that individuals experience in the form of vivid dreams" [7]. The coexistence of both slow frequency oscillations (sensory disconnection) and high-frequency power (intense cortical activity) in FREM activity as well as the contrast between the quiescent character of TREM and its high stimuli responsiveness, suggests the paradoxical denomination attributed to REM can be extended also to its microstates.

The focus on the heterogeneity of REM through the partition into its two sub-states may be relevant in some diagnostic contests since they show very recognizable characteristics helping the classification of different sleep disorders, like REM sleep behavior disorder (RBD) [8]. For example, RBD patients show a loss of distinction between the brain's tonic and phasic REM sleep activity as well as a reduced decrease in β activity during phasic compared to tonic REM sleep suggesting "a reduction of the normal arousal suppression that protects active dreaming states" [9].

1.2 Sleep Disorders

Sleep Disorder is a term used to describe each anomaly that results in changes in the physiologic way to sleep and in its macrostructure, reducing the proportion

of restorative sleep and thus worsening the Quality of Life (QoL). They can be caused by a wide spectrum of physical (e.g. respiratory problems, chronic pain), psychic (e.g. depression, stress, and anxiety disorders), environmental, circadian (e.g. working the night shift), genetic, and aging factors, as well as the use of medications like antidepressants. Subjects with sleep disorders often experience struggle to stay awake when inactive, difficulties in paying attention, reduced work/school performances, and memory issues.

Although the current work will focus on a specific type of sleep disorder i.e. RBD, it can be useful to know other families of sleep anomalies to contextualize the general framework and to highlight elements that may be useful for the discussion. Following the *International Classification of Sleep Disorders 3 ed.* (ICSD)-3 more than 80 different Sleep Disorders exist, and they are mainly classified into seven sections [10], [11]:

1. **Insomnia**: it includes difficulty starting and/or maintaining sleep despite adequate time and opportunity to do so, which results in an increased proportion of nocturnal wakefulness and a worse sleep quality, bringing to daytime impairment. In ICSD-2 insomnia is divided into primary (isolated from any possible other overt disorder) and secondary or comorbid (due to drugs, medical conditions, or neurological disorders) although in ICSD-3 these diagnoses are grouped under the name of "chronic insomnia", next to "short-term insomnia" and "other insomnia". Insomnia is mainly associated with daily activities which are incompatible with good quality sleep like irregular sleep onset and wake times, stimulation of activities before bedtime, or the usage of substances such as alcohol and caffeine near sleep time

2. **Sleep-related breathing disorders**: this family comprises a large variety of disorders like Central Sleep Apnea Syndrome (CSAS) which includes all problems in which respiratory effort diminishes or is absent, cyclically or intermittently, as a result of dysfunction in the central nervous system due (in its secondary form) to pathological and environmental causes (Cheyene-Stokes breathing pattern, high-altitude periodic breathing), medical disorders, or medication/substance use. When CSAS is primary, its causes are unknown and recurrent episodes of cessation of sleep-breathing without associated ventilatory effort occur with patients complaining of excessive daytime sleepiness, insomnia, or difficulty breathing during sleep.

   Another category belonging to the current family concerns Obstructive Sleep Apnea Syndromes (OSAS) characterized by obstruction in the airway and resulting in increased breathing effort and non-adequate ventilation. Due to different diagnosis and treatment requirements, OSAS are classified in adult and pediatric forms. OSAS in adults are marked by frequent episodes of apnea (cessation of breathing) or hypopneas (partial upper airway obstruction) and

are thus associated with reduced blood saturation which causes snoring and sleep disruption (due to increased breathing effort), resulting in excessive daytime sleepiness or insomnia. Pediatric OSAS features are similar to the adult ones, but cortical arousals do not occur, possibly due to a higher arousal threshold.

Sleep-related hypoventilation/hypoxemia syndromes comprise six disorders whose hallmark is hypoventilation or hypoxemia during sleep (elevated arterial carbon dioxide pressure PaCO2 and reduced oxygen saturation). The idiopathic form concerns alveolar hypoventilation "resulting in sleep-related arterial oxygen desaturation in patients with normal mechanical properties of the lungs" while congenital central alveolar hypoventilation syndrome is related to a failure in automatic central control of breathing in infants. The other four subtypes of this syndrome are related to medical conditions such as obesity, hypothalamic dysfunction, medical disorders (pulmonary parenchymal or vascular pathologies), or medication or substances.

3. **Central disorder of hypersomnolence**: it is a group of disorders characterized by excessive daytime sleepiness (inability to stay awake and alert during the primary waking periods of the day) not attributable to other sleep disorders, specifically those resulting in nocturnal sleep disturbance or a misaligned circadian rhythm. The main exponent of this subgroup is Narcolepsy divided into "Narcolepsy type 1" and "Narcolepsy type 2" (ICSD-3) whose main cause is a deficiency in the Hypocretin (or Orexin) levels which in physiological conditions regulate circadian rhythm. Other forms of Hypersomnia exist, like idiopathic Hypersomnia, caused by medical disorders, due to medication or substances, and associated with psychiatric disorders.

4. **Circadian rhythm sleep-wake disorders**: all the disturbs belonging to this category share a common underlying chronophysiologic basis i.e. persistent or recurrent mismatch between the sleep pattern experienced by the patient and the one desired or suggested as the societal norm, making it impossible for the subject to sleep when desired, needed or expected and resulting in Insomnia and excessive sleepiness. The wake-sleep schedule disorders can be primary (malfunction of the biological clock per se) or they could be secondary when associated with environmental causes such as lack of synchronized factors (light, physical and social activities), as well as jet lag, shift work, personal behavior, or medical conditions (neurological disorders) [1]. The most common types of circadian disorders are Delayed Sleep-Phase Syndrome (DSPS) and Advanced Sleep-Phase Syndrome (ASPS) which cause patients to fall asleep late (early respectively) and rise late (early) and more frequently brought on by the person's decision to stay up late or by setting an earlier bedtime.

5. **Parasomnias**: these Sleep Disorders are undesired occurrences that predominately happen during sleep rather than being anomalies of the systems responsible for sleep and awake states per se. They consist of abnormal sleep-related movements, emotions, behaviors, dreaming, perceptions, and autonomic nervous system functioning which result from disorders of arousal and transitions of sleep stages. Autonomic nervous system changes and skeletal muscle activity are the main features of this category of sleep disorders. Parasomnias are divided into three clusters i.e. NREM related, REM related, and "other".

   NREM-related are linked to disorders of arousal commonly from SWS (Confusional arousal, Sleepwalking, Sleep terrors with the last two often coexisting together, blending in a unique phenomenon) or to activity like recurrent eating and drinking episodes occurring during arousals from nocturnal sleep.

   REM-related comprise REM Sleep Behavior Disorder which involves the absence of normal muscle atonia and abnormal and injurious behavior during REM sleep, Recurrent Isolated Sleep Paralysis characterized by the inability to perform voluntary movements and by hallucinatory experiences (occurs in REM during sleep onset or at awakening), and Nightmare Disorder which brings patients to wake up with intense anxiety, fear, and other negative and involving feelings (occurs in REM and N2). In contrast to arousal disorders which in the majority of cases involve healthy individuals, many cases of REM-related disorders arise from serious neuropathology (RBD) or psychopathology.

   The last category concerns all parasomnias not fitting the previous classification and comprises disorders in which the integrative components of consciousness, memory, identity, and perception of the environment are disrupted, sleep enuresis, Exploding Head Syndrome, sleep-related hallucinations, parasomnias due to drugs or substances, or due to underlying psychiatric disorders (unspecified parasomnia).

6. **Sleep-related movement disorders**: are characterized by sleep-related relatively simple, elementary, and stereotyped movements, which may result in arousal and hence sleep disturbance and decreased sleep quality. The most common are Periodic Limb Movement Disorder (PLMD) concerning repetitive, highly stereotyped limb movement, and Restless Legs Syndrome (RLS) which consists of a strong urge to move the legs accompanied often by painful or uncomfortable symptoms which can be relieved through walking or leg movements. Sleep-related cramps, sleep-related bruxism, and sleep-related rhythmic movement disorder (during drowsiness or light sleep) belong to this subgroup as well as unspecified, due to medical disorders and to a drug or substance sleep-related movement disorders. Head and limb injuries can be the result of violent movements.

7. **Other Sleep Disorders**: all the Sleep-related Disorders which not fit inside the above classification.

1.2.1 REM Sleep Behavior Disorder

RBD is a type of Parasomnia of the REM stage mainly characterized by the loss of the normal REM sleep muscle atonia which manifests itself through vocalization and dream enactment [12]. The dream content reported by patients is often vivid, violent, and emotionally involving, concerning recurrent themes of being chased or protecting themselves or a loved one from an attack, which makes the motor activity range from elementary and simple limb twitches to complex and violent movements (punching or kicking) that may represent a danger for patients and also for their sleeping partners [13]. The dream enactment occurrences, if present, can vary in frequency (nightly to annually) and appear at least 90 minutes after sleep onset and more often during the last half of the sleep.

Population-based, although not numerous epidemiologic studies on RBD suggest a prevalence of 0.5-1.25% in the general population which remarkably increases in the elderly population where RBD turns out to be present between 5% and 13% in adults aged 60-99 years old. Among the older adult population, approximately 60% of cases are idiopathic, or isolated RBD (iRBD) while 40% are more likely to be suggestive of an underlying neurological disorder [14]. The incidence of RBD is higher in the male population after 50 years old with a male-to-female predominance which can rise to 8:1 in some sleep-clinic ascertained cohorts, although these results can be affected by biases related to the often more aggressive and violent RBD episodes in male with respect to female population [15], and to the fact older adult females RBD is less likely to be witnessed considering that statistically they outlive their male bed partners more frequently. The distribution tends to be equal between males and females under 50 years with RBD female adults more commonly associated with autoimmune disorders, antidepressant medication exposure, or narcolepsy [16]. Some works reported links between RBD in childhood and narcolepsy with 30% of young subjects with narcolepsy type I having RBD or idiopathic hypersomnia even if RBD in young children can be considered rare. Patients with REM sleep without Atonia (RSWA) but without any overt clinical RBD symptom are common, about 2% of the general population, 12% among antidepressant exposed pones, and up to 25% in older men [14].

As mentioned before a prime feature of RBD is the intermittent loss of atonia during REM sleep present instead in normal and healthy subjects both as a protective measure to avoid acting of what one dreams and as a way to facilitate the sleep-memory consolidation mechanisms. This physiologic motor activity suppression during REM sleep is the result of multiple neuronal circuits located in the brainstem which mainly originate in the pons and terminate in the spinal

cord motor neurons, namely Sublaterodorsal Tegmental Nucleus (SLD) also called Subcoeruleus Nucleus (SubC), and Ventromedial Medulla (VMM) forming what is called REM sleep atonia circuitry (fig. 1.5) [17].

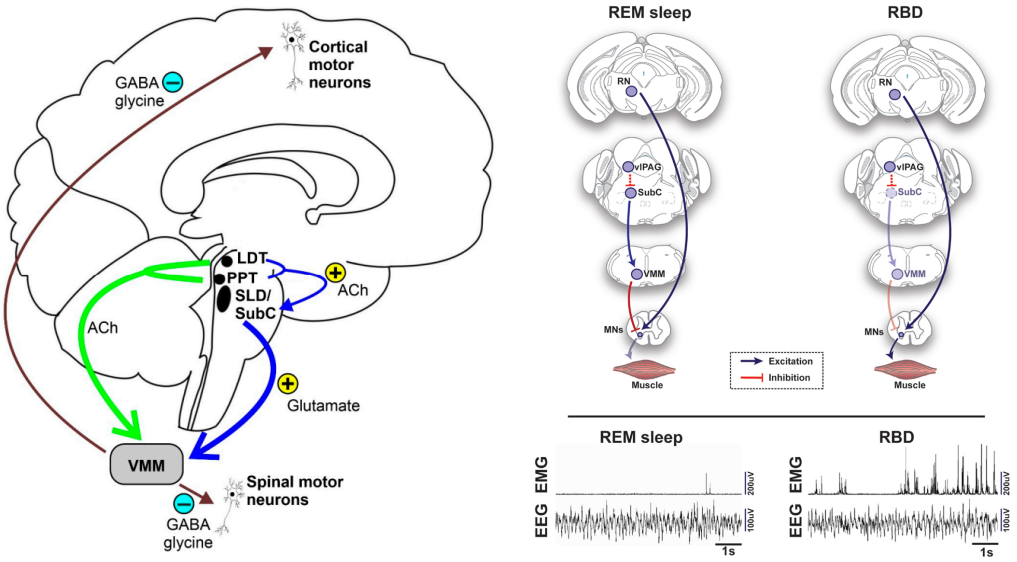


Figure 1.5: REM atonia circuitry schemes [17], [18].
REM atonia circuitry basic scheme [17] (left), comparison of pathways and circuits that regulate atonia during Healthy/Control (H/C) REM sleep and RBD in the rodent brain [18] (right).

Different studies on animal models, post-mortem analysis, and diagnostic imaging of case reports suggest the interruption or the disinhibition of these brainstem areas preclude normal REM sleep paralysis thus highlighting them as the main pathophysiological causes of RBD onset [14]. However this malfunction alone does not account for the wide number of other non-motor symptoms seen in RBD which seem instead somehow related to synuclein deposition in other brainstem nuclei (Braak's ascending model), suggesting that many aspects of neurodegeneration in RBD are not clear yet [9]. The failure of the above-mentioned systems in secondary RBD may be due to different factors like α-synuclein-based (formation of insoluble protein aggregates within neurons or glial cells) or other neurological disorders (lower prevalence), structural lesions of pontine regions (vascular, demyelinating, and traumatic etiologies), narcolepsy (deficit of Orexin), medications (especially antidepressant like Serotonin reuptake inhibitors and tricyclic antidepressants), and autoimmune disorders which hence represent RBD risk factors.

RBD shows the strongest correlation with the so-called α-synucleinopathies, a family of degenerative disorders related to intracellular accumulation due to the non-correct disposal of α-synuclein. Secondary RBD is identified in 25-58% of patients already diagnosed with Parkinson's Disease (PD), up to 80% of those with

Dementia with Lewy Bodies (DLB), and in 80-100% of those already diagnosed with Multiple System Atrophy (MSA) [15]. Recent studies demonstrate iRBD is a prodromal syndrome of some α-synuclein degeneration subtypes, representing hence an early symptom of these neurodegenerative conditions, with a particularly high rate of phenoconversion into PD and DLB [16] (more than 50% and 40% respectively, after 10 years fig. 1.6). Longitudinal studies on patients diagnosed with iRBD showed in fact that the risk of developing a neurodegenerative disease was about 32% with a follow-up duration of approximately 5 years, and the most frequent condition represented by Parkinson's disease (44% of converters), followed by Dementia with Lewy Bodies (25%), no better-specified Dementia (7%), Multiple System Atrophy (5%), Mild Cognitive Impairment (3%), Alzheimer's disease (3%) and by other types of degenerative diseases (0.5%). The overall estimated conversion risk for RBD is higher than 90% throughout a long-term follow-up (97% after 14.2 years) suggesting also a strong positive correlation between follow-up duration and conversion rate [12].

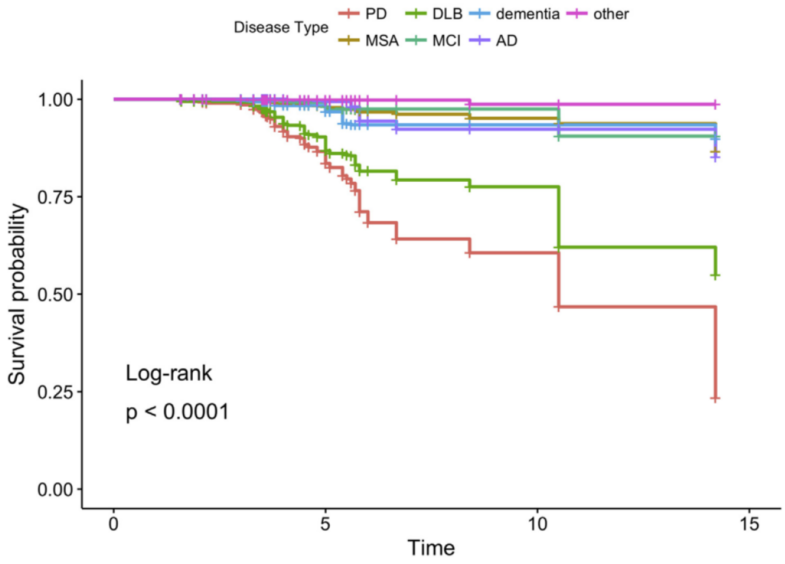


Figure 1.6: Kaplan-Meier curve for RBD phenoconversion rate [12].
The Kaplan-Meier Curve represents the probability for a RBD patient to remain disease-free after a certain follow-up time, stratified for different neurodegenerative diseases (can be interpreted also as the risk of conversion from RBD to a neurodegeneration). For Parkinson's Disease this probability is halved after 10 years from RBD diagnosis.

These findings, confirmed by different studies involving video-Polysomnography (vPSG), suggest iRBD features as early biomarkers for the onset of the above-mentioned neurodegenerations, preceding often the characteristic neurological manifestations of such disorders by several years. Early identification of patients

with RBD provides a unique opportunity to spot individuals who are prone to develop α-synucleinopathy neurodegenerative disorders allowing them to join clinical trials of neuroprotective therapies to stop or halt the progression at a point before serious motor/cognitive consequences develop, or to postpone the emergence of these symptoms. Moreover, it would allow contrasting RBD symptoms like injuries by advising environmental measures to ensure bedroom safety or to reduce frequency and severity of RBD episodes through an appropriate administration of medication, overall improving the QoL of patients and their familiar [13].

Although the presence of this correlation seems clear, its nature should be further investigated. In fact, RBD symptoms can rise years before (iRBD) or years after (secondary RBD) such neurodegenerations highlighting two main plausible paths. On one hand, iRBD progresses from RSWA to RBD that hence causing loss of sleep architecture, sleep efficiency, and reduced restorative sleep, bringing consequences like impairing daytime cognitive functions and other pathological problems like a non-correct nocturnal clearance of α-synuclein, and other proteins implicated in pathologies like PD and DLB. On the other hand, RBD can rise as effect of lesioning of brainstem structures responsible for atonia during REM (SubC and VMM) due to spreading α-synuclein pathology, which leads to impaired function in related neuronal networks and impaired daytime cognitive functions [15]. Knowing better these mechanisms may be important since there is evidence that patients who convert from RBD to PD (PDRBD+) have α-synuclein pathology with a likely greater pathological burden with respect to those who present neurodegeneration without RBD prodrome, with the RBDPD+ phenotype characterized by faster progression of cognitive impairment, less response to treatment, depression, anxiety, autonomic dysregulation, and other non-motor symptoms.

1.2.2 Role of Slow Wave Sleep in Neurodegeneration

Based on the idea NREM sleep modulates the clearing of the brain from toxic metabolites accumulated during the wake period and which may contribute to neurodegenerative processes, some studies highlighted an interesting connection between poor SWS quality (which causes a higher neuronal activity) and the accumulation of Amyloid-β protein which if not cleared has proven to be correlated with aging-symptoms of cognitive impairment and brain atrophy [19],[20], and with a higher risk of developing cognitive decline in PD [21]. Two probable simultaneous working mechanisms are suggested for explaining the correlation between sleep and the accumulation of Amyloid-β: on one hand, Amyloid-β is removed from the brain during sleep (glial-lymphatic system) while their creation also declines as a result of decreased synaptic activity. On the other hand, if Amyloid-β is not sufficiently removed from the brain or is created in large amounts as a result of prolonged wakefulness or sleep disruption, it may raise the risk of cognitive impairment.

Long-term studies suggested a negative correlation between Amyloid-β and the SWS frequency below 1Hz (SOs microstructure) and a positive correlation between Amyloid-β and the SWS frequency between 1-4Hz (SWA microstructure), in the sense both a lower activity in SOs and a higher activity in SWA brings to increased metabolite deposition in the long run [19]. These analyses thus highlight the potential role of disturbed-NREM activity in the onset/worsening of some brain disorders and suggest SWS microstructures as predictive biomarkers of some neurodegenerative processes. Moreover, EEG analyses of the distribution of slow wave amplitude performed during NREM sleep highlighted abnormal pattern in RBD phenotype which did not show any reduction in δ and θ band activity as sleep went from the first to the last cycle, in contrast to what occurs in healthy subjects [9]. These results imply that SWS sleep homeostasis may be harmed when RBD is present, suggesting EEG abnormalities during NREM exist and that together with REM ones can be exploited to detect RBD.

1.3 How Sleep is Analyzed

The diagnosis of a certain number of sleep disorders (as well as the plain study of sleep macrostructures and cycles) can be accomplished through a wide number of methodologies, comprising clinical interviews based on history and physical examination, questionnaires for the evaluation of strange and unusual sleep-related behaviors as well as in-lab tests that help to quantify and analyze the presence and the severity of the underlying pathology discriminating it from others which can potentially mimic some features making the diagnosis challenging. History should include details about sleep-wake habits, present or past medical, neurological, and psychiatric illness, substance abuse, and even family history with a particular focus on the frequency, type, and time of symptoms onset. It is important also to interview the patient's bed partner (or caregiver) since frequently sleep disorder-related subjects cannot recall or be aware of abnormal behaviors. Completing questionnaires can be useful to keep track of sleep habits and sleep hygiene, followed by careful physical examination to document the evidence of some medical disorders (respiratory, cardiovascular, endocrinological, or neurological) especially brain and neuromuscular related [2].

1.3.1 Polysomnography

As an extension of the previously-mentioned analysis sequence, hospital/in-lab sleep monitoring should be taken into consideration to highlight the presence of primary conditions causing secondary (or comorbid) sleep disturbances as well as to study the disorder itself. Among the most important monitoring methods it can be possible to mention Multiple Sleep Latency Testing (MLST) which records multiple

naps during the day, Maintenance of Wakefulness Testing (MWT) which determines how long wakefulness can be maintained, and overnight Polysomnography (PSG) which is the most comprehensive, and used laboratory test allowing to obtain simultaneously records of different physiological variables like EEG, EMG, EOG, EKG, airflow (nose and mouth), respiratory effort (Oxygen saturation), body position, and considered the diagnostic gold standard for the most common sleep disorders such as Insomnia, OSAS, RLS, Sleep Epilepsy, and Narcolepsy diagnosis as well as essential for some diseases such as RBD (ICSD-3) [2]. Some sleep disturbances require video recording (vPSG) to be confirmed. The complete clinical guidelines about recommended PSG parameters, event measures, technical specifications, and montage settings are described in detail in the *AASM Manual for the Scoring of Sleep and Associated Events of 2007* [22], updated in 2015.

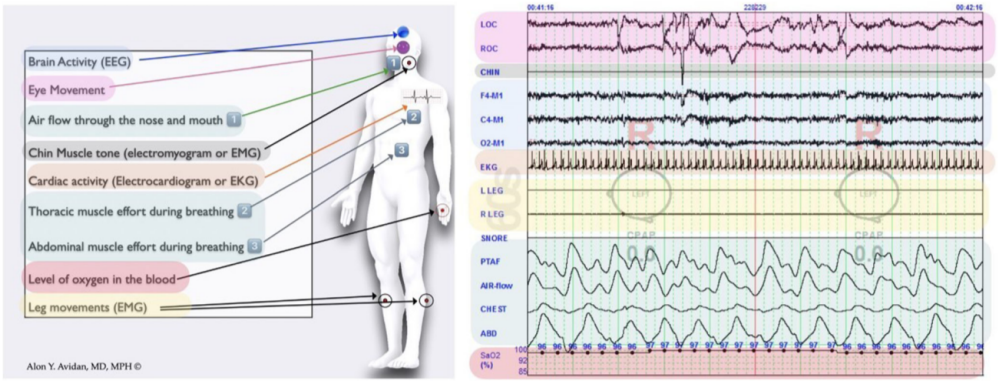


Figure 1.7: Standard Polysomnography configuration [23].
Electroencephalography (EEG), measuring brain activity, Electrooculography (EOG), helping in sleep staging, Electromyography (EMG), recording muscle tone in the chin and limbs, Electrocardiography (EKG), measuring cardiac activity, and respiratory channels (which depict airflow and effort with pulse oximetry, useful for detecting sleep-disordered breathing) make up a standard PSG configuration [23].

Standard PSG analysis consists in reviewing each of the above-mentioned parameters (and related reference values) following the AASM sleep disorders scoring guidelines, on each of the 30s-epochs in which the total overnight record is divided. Although it is usual practice to diagnose a sleep disorder based on a single recording (about eight hours long), some authorities warn that more than one night of recordings may be required so the patient can get used to novel surroundings and sleep more naturally, reducing biases the environment may produce (first-night effect). The most common polysomnographic parameters are Sleep onset latency (normally less than 20 minutes), and Sleep efficiency (normally 85-90%) which respectively indicate the time between lights are turned off and sleep onset, and the ratio between sleep time and total time in bed, objectively assessing sleep quality.

As far as PSG-related (Level I) EEG montage is concerned, the electrodes position is determined according to 10-20 System criteria with a minimum of three derivations (six electrodes), recommended to extract activity from the frontal (F), central (C), and occipital (O) regions. AASM-advised electrodes (Figure 1.8, left) are F4, C4, O2 (exploring), and M2 (reference) coupled forming the channels F4-M1, C4-M1, and O2-M1 while taking into account F3, C3, and O1 referenced to M2 as backups electrodes (F3-M2, C3-M2, and O1-M2). As alternative acceptable derivations (Figure 1.8, right) Fz-Cz, Cz-Oz, and C4-M1 can be set, with backup electrodes placed at Fpz, C3, O1, and M2 which can substitute respectively Fz, Cz, Oz, and M1 in case of malfunction [24].

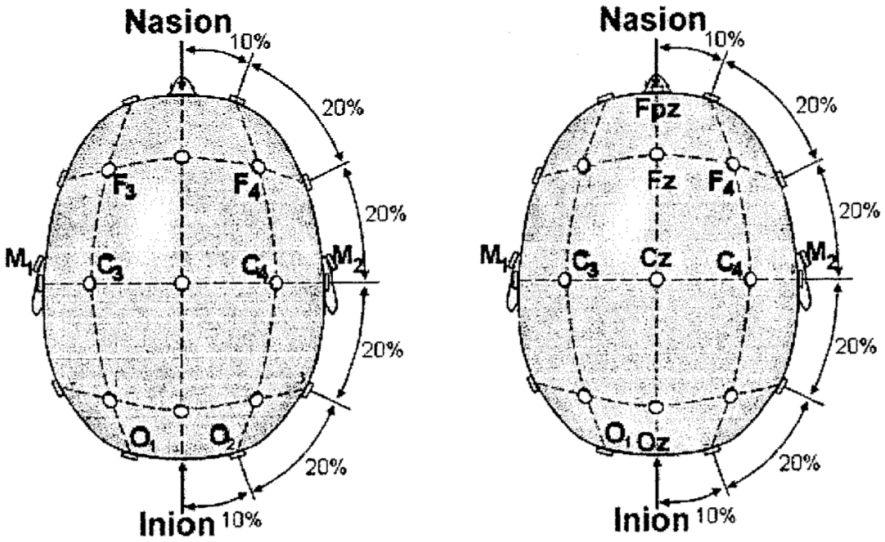


Figure 1.8: AASM electrodes guidelines (10-20 criteria) for standard PSG-related EEG configuration [22].

1.3.2 Emerging Alternative Technologies

The fact patients around the world face significant obstacles (shortage of specialists, heavy financial burden, and lengthy wait times) to receive vPSG when they exhibit symptoms that could be related to sleep disorders, makes it exceedingly difficult for neurologists to examine symptoms and establish a diagnosis in their patients. To slightly alleviate this burden and improve early diagnosis in the at-risk population, in the last years, different alternative methods have been studied and implemented. One of the most common examples is related to home-based PSG (type II) approaches with studies showing their reliability, efficiency, low-failure rate, and robustness (with respect to hospital/in-lab PSG) on respiratory sleep

disorders patients such as OSAS, as well as greater cost-effectiveness ratio. From these studies, it is evident patients experience a more comfortable sleep (as far as quality and quantity are concerned) in their home instead of hospital or laboratory environments.

Some recent research focused on implementing new computer-based technologies which offer lower cost, easier accessibility, and user-friendliness as Actigraphy which exploits movement analysis and Artificial Intelligence (AI) technologies implemented in a watch-styled non-invasive device, able to give a good measurement of sleep patterns continuously over time (days and weeks). AI algorithms estimated sleep parameters (e.g. total time sleep, sleep percentage, wake after sleep onset) are useful for sleep disorder assessment, particularly when concerning sleep-wake disturbances of circadian disorders.

Interesting recent studies compared sleep parameters obtained through smartphone accelerometers (in different settings configurations), commercial wrist accelerometers, and actigraphy against PSG on sleep-disordered breathing subjects, discovering good accordance between the different sources and evidencing there is a need for more studies into alternative technologies for measuring sleep disorders (such as iRBD), as it is possible that combining different sensors (none of which fits the accepted guidelines diagnostic criteria) may yield the ideal PSG substitute [3].

1.4 RBD Diagnosis

The initial suspect of iRBD is usually raised by the patient's bed partner who has noticed some unusual behavior during sleep. This can make the diagnosis problematic since there is the possibility for the subject to be left unattended for many years in the case iRBD symptoms fluctuate, are considered medically irrelevant, or if there is no bed partner at all. Misdiagnoses may occur due to a lack of awareness of iRBD and/or symptoms that mimic other sleep disorders. These factors may also contribute to diagnostic delay with recent studies revealing 31% of the patients do not receive a timely diagnosis of iRBD due to the failure of their specialist to recognize symptoms (mean delay 8.7 years from symptom onset). All these findings suggest thus the importance of clinical assessment when a patient shows iRBD symptomatology since lack of initial recognition can reduce the possibilities for management, neurodegenerative diseases (ND) identification, and treatments [3].

RBD can be diagnosed through the evaluation of its main features like RSWA, vocalization, and dream enactment behavior which represents the more evident symptoms usually triggering the suspicion of RBD. The ICSD-3 produced by AASM in association with the European, Japanese, and Latin American Sleep Research Societies proposes the following RBD diagnostic criteria [25]:"

1. Repeated episodes of sleep-related vocalization and/or complex motor behaviors.

2. These behaviors are documented by polysomnography to occur during REM sleep or, based on clinical history of dream enactment, are presumed to occur during REM sleep.

3. Polysomnographic recording demonstrates REM sleep without Atonia (RSWA).

4. The disturbance is not better explained by another sleep disorder, mental disorder, medication, or substance abuse."

These criteria should be necessarily satisfied and require history and physical examination to verify the disease suspicion, then hospital/in-lab confirmation through PSG (type I) to ensure definitive diagnosis of RBD, also distinguishing it from other conditions which mimic RBD like non-REM parasomnias, OSAS, PLMD, sleep-related-hypermotor epilepsy, or drug-induced RBD.

1.4.1 Screening Questionnaires

Unfortunately accessing PSG is problematic for most clinicians due to expense or to the fact is unavailable in many locations, hence various questionnaires have been developed to try to avoid PSG. Commonly used questionnaires aim at assessing the symptoms which have arisen during the patient's lifetime as well as their occurrence frequency (RBDQ Hong Kong Sensitivity 98% and Specificity 87%, Mayo Sleep Questionnaire Sensitivity 98% and Specificity 74%, RBD Single-Question Screen Sensitivity 98% and Specificity 87%, RBD Screening Questionnaire Sensitivity 96% and Specificity 96%, Innsbruck Questionnaire RBD-I Sensitivity 91% and Specificity 86%) and sometimes can be completed by bed partners, being more efficient in cases where patients do not correctly or completely recall their dreams [16]. Some questionnaires such as RBDSQ have been found with a specificity drop when including patients with other sleep disorders, showing the low reliability of such clinical step in differentiating iRBD from these states [3]. Moreover, recent studies and evidence have highlighted that uncritical and non-expert use of questionnaires may lead to results that in some cases turn out to be incoherent with PSG findings producing high false-positive rates, thus suggesting they just should be used as screening or for diagnosis of "probable RBD" and not as a sole diagnostic criterion [25], emphasizing the irreplaceable importance of PSG evaluation.

1.4.2 RSWA Visual Scoring Methods

PSG can be often accompanied by video recordings (vPSG) in order to identify iRBD by capturing RSWA and/or dream enactment, although evidence shows that

even in severe RBD patients, the majority of motor events are small and involve elementary movements, suggesting the dramatic "violent" behaviors are rare and may be interpreted as the "tip of the iceberg". While video analyses rely upon the occurrences of these unpredictable, not always visible, and possibly rare events, the PSG quantification through EMG has the advantage that RSWA exhibits high night-to-night stability giving the possibility to adequately diagnose RBD even through just a single-PSG night suggesting the promise for such a criterion [25].

RSWA is the polysomnographic hallmark of RBD and it is crucial for its diagnosis but may be also present in subjects without clinical symptoms or overt RBD behavior like dream enactment (isolated RSWA) introducing further difficulties [17]. Although the significance of RSWA alone is uncertain some evidence suggests it is a precursor of RBD representing in some way a prodromal form of REM disorder helping though in predicting future neurodegenerations like α-synucleinopathies. RSWA manifests through constant excessive muscle tone during REM sleep, resulting in either (alone or coupled) REM sleep-related increased sustained (tonic) activity in the chin EMG, intermittent (phasic) excessive activity in the chin or limb EMG, or both [15]. The current in-use visual scoring approaches for PSG-based RSWA quantification mainly exploit the differentiation between tonic and phasic muscle activity although there is no general consensus about the definition of tonic and phasic activity nor about how scores and cutoff values are computed. Following AASM's latest scoring manual [17]:"

- A 30s-epoch of REM sleep is defined as tonic when at least 50% of its duration is related to a chin EMG amplitude greater than the minimum amplitude present in NREM sleep.

- The phasic RSWA components are excessive transient muscle activity bursts (in chin and limb EMG channels) lasting 0.1-5.0s with an amplitude at least 4 times higher with respect to the background EMG. A 30s REM sleep epoch (divided into 3s mini epochs) is considered to contain excessive phasic activity if at least 50% of these 3s mini epochs contain the above-mentioned bursts of transient EMG activity."

While AASM does not provide any established cut-off values for RSWA, other methods like the currently in use in the clinic SINBAR proposed a slightly different and more detailed definition of EMG activity division in tonic and phasic, suggesting submentalis muscle and bilateral Flexor Digitorum Superficialis (FDS) as the best muscle channels combination, as well as defining thresholds (EMG amplitude-based) with the best specificity and sensitivity to identify RBD and distinguish RBD patients from their controls, provided that other diagnostic clinical and videographic criteria are fulfilled.

1.4.3 RSWA Semi-Automatic Scoring Methods

One of the main disadvantages of manual EMG quantification during PSG is that it is time expensive and requires high technical demands on the scorer. For this reason, recent studies have proposed automatic computed-assisted scoring approaches like Ferri's REM Atonia Index (RAI), which works just on the amplitude of the rectified submentalis EMG signal in 1s-mini epochs reaching performances comparable to that of visual methods. This score ranges from 0 (complete loss of atonia) and 1 (total atonia) and the proposed threshold under which RSWA can be defined is 0.8. Despite the promise of simple indices and computer-assisted RSWA quantification methods, stable raw data analysis and precise artifact elimination (instrumentation, or patient movement) are necessary to avoid false positives which makes these approaches semi-automatic rather than automatic, slowing down the process anyway [25]. Like RSWA visual methods, computer-assisted ones exploit the EMG signal amplitude thresholds, which create three main issues to the automation task [26]:

- Exogenous/endogenous noise and interference affect amplitude in a way that is impossible to be totally removed, exacerbated by the long continuative duration of the Polysomnographic recordings. If expert "manual" evaluation succeeds in overcoming these issues, on the other hand, automatic threshold-based can have some difficulty.
- Signal processing applied to remove artifacts changes the amplitude of the EMG signal.
- Thresholds are chosen based on operator-made visual assessments rather than objective quantification. Moreover, surface EMG is believed to show high inter- and intra- subject variability which in turn causes variability of both threshold and scoring criteria.
- Amplitude can be affected by pathological phenomena related to a patient medical condition.

1.5 RBD Treatment

The first measure that can be adopted to manage RBD cases (improving also the bed partner's QoL) concerns the establishment of a safe sleeping environment to reduce or avoid sleep-related injuries possibly caused by dream enactment behaviors and, for patients in whom RBD episodes are significant, pharmacotherapy treatment.

Preferred pharmacological treatments involve the administration of Melatonin (circadian rhythm regulator) or Clonazepam (anxiolytic, anticonvulsant, and muscle relaxant) with the former to be better tolerated in particular among neurodegenerative adult patients [14]. Some retrospective cohort studies report (in both cases)

67% of the patients experience at least a mild improvement of symptoms while 12% had complete resolution of RBD abnormal behavior even if for Clonazepam some additional follow-up studies showed more mixed results [16].

Although the action mechanism is unclear, 3 to 12 mg of Melatonin at bedtime (increased nightly until disruptive and injurious behaviors have ceased) has been proven to augment REM sleep atonia and to improve RBD symptoms both in the frequency of occurrence and severity, even if behaviors tend to come back once Melatonin use is reduced or discontinued, suggesting the need of a lifelong therapy for the patients. Observed common side effects are mild and concern gastrointestinal distress, headache, sleepiness, fatigue, and cognitive alteration even if in general Melatonin tends to be well tolerated at the above-mentioned doses.

Clonazepam is well-recognized for RBD management with nightly administrations consisting of an initial dose of 0.25 mg which increases progressively to a maximum of 2.0 mg [15]. Like Melatonin, the therapeutic mechanisms in RBD are not fully understood even if some studies reported reduced frequency of unpleasant dreams hence decreased possibility of violent dream enactment behavior as benefits. Side effects have been observed including as most common residual morning sleepiness, increased fall risk, memory dysfunction, impotence, and unstable gait. These might be problematic, especially in older adults, and in some cases may limit the treatment's utility since lower doses are suggested in those situations [16].

1.6 A Fully-Automated Diagnostic Tool

As mentioned in the previous sections PSG type-1 is the gold standard for RBD diagnosis but is expensive both in terms of time and money, not always available around the world, invasive for the patients, concerns wait times varying from weeks to more than 12 months [3], and requires specialized technical personnel for the manual score, which significantly hinders the diagnostic process. In the last decade, different alternatives aimed to replace PSG and lighten its load have been studied showing good results in the context of some disorders (e.g. OSAS, circadian rhythms, and sleep-wake disturbances) but encountering criticism when applied to others. Home-based PSG (type II) for example does not routinely include video recording and lacks a series of features which create some dubs about diagnostic powers in sleep disorders such as RBD. In fact type II PSG settings do not comprise a complete array of data montages (lack of EMG) required for iRBD diagnosis. Different studies performed through questionnaires and Actigraphy (compared to vPSG) on H/C and sleep disorders-related subjects, indicate that actigraphy has a high potential as an extra method for detecting iRBD, especially when extensive clinical data is available [3], thanks to the fact in iRBD characteristics sleep-wake patterns exist, which can help clinicians to identify possible diseased patients and

to increase the confidence of diagnosis of "probable iRBD". On the other hand, a diagnosis from this device should be cautious since also other sleep disorders may show up with similar patterns highlighting the limited accuracy of this method in some situations, being not able to catch all the standard PSG parameters needed for a complete and exhaustive RBD diagnosis and producing misclassifications. As far as RSWA scoring methods is concerned, in recent years, different studies promoted semi-automatic diagnosis of RBD through a computer-assisted manual evaluation which speeds up the diagnosis but requires fine artifact removal to produce reliable results.

The current work aims to overcome the cost, time, accuracy, and robustness limits of the to-date available diagnostic tools, proposing a fully-automatic diagnostic approach, which through the usage of a single EEG channel without any particular nor deep pre-processing, reaches in classifying reliably and efficiently healthy patients from RBD ones, exploiting underlying characteristics of PSG-based EEG signal instead of classical EMG one. RSWA is thought to be a dissociative state, characterized by a mismatch between brain (REM) and muscular activity (non REM) which is however linked to the former to some extent [27]. Basing the current strategy on EEG signal instead of EMG would avoid to perform a classical two-steps approach concerning the chain of sleep study and muscular activity evaluation, working directly on the source of the phenomena i.e. the brain. EEG signal-based approach allows thus to simplify the overall pipeline, relying on some recently discovered evidence about EEG abnormalities in RBD patients with respect to H/C ones (mainly concerning differences in δ and β power during phasic and tonic REM, in δ and θ power during the first and the last sleep cycle in NREM, and in Slow-wave amplitude during NREM [9]).

Moreover, the last part of the work concerns the development of a three-stages semi-unsupervised method that can qualify REM Sleep without Atonia (RSWA) as an intermediate pathological state (isolated RSWA) supporting the development of a metric that tries to more finely characterize the neurodegenerative progression from Healthy to RBD.

Chapter 2

Material and Methods

2.1 Databases

2.1.1 CAP

The CAP Sleep Database (`https://physionet.org/content/capslpdb/1.0.0/`) is a public collection of 108 polysomnographic recordings supplied by the Sleep Disorders Center of Ospedale Maggiore of Parma (Italy). Each record, stored in European Data Format (EDF) files, contains at least three monopolar EEG channels (F3, C3, and O1 or F4, C4, and O2, with A1 or A2 as references) and additional bipolar EEG traces, in line with 10-20 system (Fp1-F3, F3-C3, C3-P3, P3-O1 and/or Fp2-F4, F4-C4, C4-P4, P4-O2) as well as two EOG channels, chin and tibial (submentalis muscle and bilateral anterior tibial respectively) EMG, respiratory signals (airflow, abdominal and thoracic respiratory effort, Oxygen saturation SaO2), and ECG data, all complemented by annotations such as scored sleep stages (30s-epoch, R&K rules, W=wake, S1-S4=sleep stages, R=REM) provided by expert neurologists and stored in TXT files [28], [29]. EDF files include information regarding channel labels, sample frequency (F_s), Unit of Measure (UoM), filtering information, and recording duration. They contain PSG recordings from subjects affected by different sleep-related pathologies although just H/C subjects (16 recordings, labeled as n*) and RBD patients (22 recordings, labeled as rbd*) were taken into account, considering only EEG signals.

Given the interest of the current work toward a single-channel EEG approach, only recordings from the central EEG channel were used (less subject to artifacts and with a higher signal-to-noise ratio with respect to other channels). As a result, the C3-A2 channel, or C4-A1 if the former was unavailable, were chosen for the succeeding analysis. Moreover, the work aims to study the effect of REM and SWS as potential RBD biomarkers, so just segments classified as R and N3 (AASM rules) are extracted from the database and used for the following steps. All the

recordings features are resumed in Table 2.1.

Table 2.1: CAP Database technical information.

Filters: LP=Low-Pass, HP=High-Pass, BP=Band-Pass, BS=Band-Stop

ID	Selected channel	UoM	Pre-filter Hz	F_s Hz
n1	C4-A1	uV	LP:30 HP:0.30 BS:50	512
n2	C4-A1	uV	LP:30 HP:0.30 BS:50	512
n3	C4-A1	uV	LP:30 HP:0.30 BS:50	512
n4	C4-A1	uV	...	100
n5	C4-A1	uV	LP:30 HP:0.30 BS:50	512
n6	C3-A2	uV	...	128
n7	C3-A2	uV	...	128
n8	C3-A2	mV	...	100
n9	C3-A2	uV	...	128
n10	C4-A1	uV	LP:30 HP:0.30 BS:50	512
n11	C4-A1	uV	LP:30 HP:0.30 BS:50	512
n12	C3-A2	uV	BP:0.10-100	100
n13	C3-A2	...	...	200
n14	C3-A2	...	...	200
n15	C3-A2	...	...	200
n16	C4-A1	uV	...	100
rbd1-rbd22	C4-A1	uV	LP:30 HP:0.30 BS:50	512

Table 2.2: TuSDi Database technical information.

Filters: LP=Low-Pass, HP=High-Pass, BP=Band-Pass, BS=Band-Stop

ID	Selected channel	UoM	Pre-filter Hz	F_s Hz
n1-n10	C3-A2	uV	LP:50 HP:0.30 BS:50	256
rbd1-rbd10	C3-A2	uV	LP:50 HP:0.30 BS:50	256
rswa1-rswa9	C3-A2	uV	LP:50 HP:0.30 BS:50	256

2.1.2 TuSDi

The Turin Sleep Disorders Database (TuSDi) is a private collection of 29 polysomnographic recordings gathered and provided by the Center for Sleep Disorders at Molinette Hospital (Turin, Italy). The process was carried out in conformity with the Helsinki Declaration and was authorized by the "Ethics Committee of the A.O.U. Città della Salute e della Scienza di Torino (approval No. 00384/2020)"

[27]. Recordings of 10 Healthy individuals (label n*), 10 RBD patients (label, rbd*) as well as 9 subjects with isolated RSWA (label, rswa*), are included. Like CAP Database, waveforms and technical information about recordings (Table 2.2) are stored in EDF file, while TXT files contain annotations related to specific within-recording episodes (e.g. arousal, bruxism, apnea, artifacts, movements) and scored sleep stages (hypnogram) produced by expert neurologists. Also in this case C3-M2 channel (or C4-M1, if the former was unavailable) was chosen for the study and just REM and SWS segments (according to AASM criteria) were extracted.

2.2 RBD and RSWA Quantification Pipeline

The current work aims to use Machine Learning (ML) algorithms to automatically classify new samples into healthy or pathological classes in the context of both binary or multiclass problems. To achieve this purpose, some fundamental characteristics of the EEG signals (features) were characterized from a quantitative point of view (i.e. extracted in the form of a number). This process consists of many different and sequential steps which range from signal numerical preparation to the effective AI application [27].

In this regard the first part of this chapter (developed in MATLAB Version: 9.11.0.2022996, R2021b software) will explore all the pre-processing solutions that were exploited to uniform the available waveforms as far as UoM and F_s are concerned, and to remove unnecessary noise with the help of filtering techniques which are deliberately kept basic in order to stress the robustness of the approach which thus relies on raw EEG data (no spatial filtering nor artefact removal). Consequently, signals were subjected to a multi-domain feature extraction process ending up with matrices containing fundamental parameters providing both physical and medical interpretations of subjects' health conditions, and that in the end were necessary to feed ML models. Thereby, the last part of this section (developed in Python 3.11 software) will focus on the Machine Learning models' training and testing process together with the development of evaluation techniques useful for quantifying the achieved performances, opening a wide range of analysis, comparisons, and comments.

2.2.1 Pre-Processing

This step includes preliminary adjustments of the signals to fit some uniformity requirements, the creation of time series used in the subsequent analysis, and the filtering process. First, the annotations concerning the sleep-stage scoring (of both CAP and TuSDi) were changed from R&K rule to AASM rule, by merging N3 and N4 stages in a unique one called N3 (or SWS). All the signals with UoM different from uV were converted to this order of magnitude, and all recordings sampled at

F_s different with respect to 512Hz were re-sampled. With these two last expedients, all the samples are unified to avoid errors during the division in epochs and to prevent the presence of outliers in the Machine Learning part.

Moreover, (for each subject) only 30s-epochs scored as REM (R) or SWS (N3) were selected to form unique time series (called *REM* and *SWS* respectively). One interesting idea characterizing the current work and inspired by the research of Rechichi et al. (2022) [27], concerns the creation of a third segment-type (called *REM+SWS*) (Table 2.3), built as the combination of REM and SWS ones, and used to potentially enhance the overall classification performances since exploiting at the same time the power of both REM and SWS stages as potential RBD biomarkers (Section 4.1).

Table 2.3: Time-series used in the current Thesis work (for each individual).

R and N3 refer to the sleep-stage scoring (AASM rules) present in .txt annotations included in the Databases and performed by expert neurologists.

Description	Time-Series Name
Built merging all the 30s-epochs classified as R	*REM*
Built merging all the 30s-epochs classified as N3	*SWS*
Built merging all the *REM* and the *SWS* time-series	*REM+SWS*

Following physiological studies on sleep EEG spectral composition, the filtering process was thought and developed to extract the significant band comprised between 0.01 and 40 Hz and to increase the signal-to-noise ratio by removing both low frequency signals that are of non-neural origin (e.g. electrodermal activity, drying, or chemical stability of the electrolyte as a result of heat changes and skin contact) and unnecessary high components which are indeed rare in the REM and NREM activity. In particular, for REM segments the used stopband frequency values are 0.01 Hz (HPF) and 40 Hz (LPF) while for SWS segments 0.01 Hz (HPF) and 30 Hz (LPF) were employed. The choice of these values is based on a couple of main simple facts:

- Although most of the existing literature on EEG proposes (0.5-4 Hz) as the δ band frequency range, the study of Acunzo et al. (2012) suggests zero-phase HPF cut-off frequencies should be kept as low as possible and in no case greater than 0.1Hz. According to the research findings, High Pass filtering an EEG signal is a critical step that "can generate a systematic bias easily leading to misinterpretations of neural activity", since increasing the cut-off frequency may perturb and distort the signal with the presence of offsets and time delays [30] (Figure 2.3, middle). Different trials were made using cut-off frequencies of 0.01 Hz, 0.05 Hz, 0.1 Hz, and 0.5 Hz, with 0.01 Hz providing the best results i.e. almost absent distortion (Figure 2.3, bottom).

- As mentioned in Section 1.1.2, REM sleep is characterized by TREM and FREM microstructures which in general occupy (7-16 Hz) and (2-8 Hz) respectively, with FREM power having relative increase in higher frequencies included in the γ band, which thus makes it reasonable to pick LPF cut-off frequencies up to 40 Hz for REM sleep.

The filtering action was implemented through the cascade of Low-Pass (LP) and High-Pass (HP) filters, instead of a unique Band-Pass (BP) one, to not constrain the stages to the same order or filter type. This allowed to customize the filter mask directly on the specific purpose which required an highly asymmetrical frequency response, as far as roll-off is concerned. At this part of the project, there are no specific requirements about the computational cost, so the entire filtering process was set favoring performances, and only in the second instance lightening offline operations and computational cost, then converging these two points in a unique trade-off.

A crucial goal of EEG-related filters is to guarantee stopband/passband ripples are as low as possible since nothing must happen to the informative part of the signal whose amplitude is usually of the order of uV. Different digital filters were tested (both IIR and FIR) and compared based on some criteria like execution time (speed), order, bandpass/stopband ripples amplitude, and stopband attenuation. The best compromise which optimally fulfilled the above-mentioned requirements was the Butterworth filter-type (Figure 2.1) well-known for its "no-ripple" features, used then for implementing minimum-order zero-phase LPF and HPF to be applied on both REM and SWS segments (MATLAB R2021b). The technical parameters and information about the designed Butterworth filters are grouped in Table 2.4.

Table 2.4: Stopband frequency values.

Ap=Passband ripples, Fp=Passband frequency, Fst=Stopband frequency, Ast=Attenuation in stopband. In Butterworth ("no-ripple" filter) Ap changes the slope in the transition band identifying the attenuation associated to the bandpass frequency

Filter	Order	Cut-off$_{-3dB}$	Ap	Fp	Fst	Ast
HPF	2	0.01 Hz	0.01 dB	0.05 Hz	0.01 Hz	3 dB
LPF	26	26.5 Hz	0.1 dB	25 Hz	30 Hz	25 dB
LPF	35	36.5 Hz	0.1 dB	35 Hz	40 Hz	25 dB

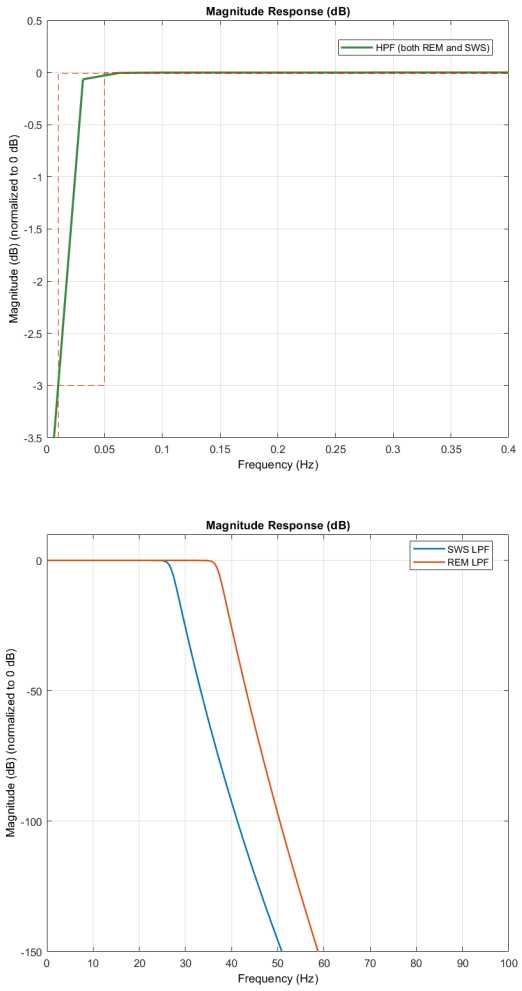


Figure 2.1: Implemented Butterworth filters.

As it is relevant for EEG processing, a zero-phase anti-causal filter was applied to remove (or at least attenuate) the nonlinear time shift (Figure 2.3, top) introduced by the not symmetric nor centered IIR impulse response [30], and to avoid the so-called Edge effects[1] which results in a phenomenon that distorts the boundaries

[1]These almost-unavoidable effects are due to the fact that each time point in the filtered signal is defined as the linear combination (through Kernel coefficients) of previous values of the original signal (FIR), and for IIR also of past samples of the filtered signal. For this reason, the filtered signal can really only begin after a delay of one kernel length with respect to the original time series, since before that, the filter Kernel could not exploit past information. This effect has consequences also on the end of the filtered waveform since it is actually shorter with respect to the original one.

of the filtered signal. The basis of this zero-phase anti-causal filter is the Reflection algorithm, a bidirectional approach (implemented here in a "home-made" fashion exploiting the MATLAB function *filter()*), which involves the following steps [31]:

1. A number of samples of the signal equal to the order of the filter is picked both at the beginning and at the end of the time series.

2. These parts are concatenated (in a backward i.e. mirrored, fashion) before and after the signal in order to avoid discontinuities in the junctions.

3. The new merged signal is filtered. Now the edge effect will be present but on the reflected parts.

4. The reflected parts are cut off, leaving only the true filtered signal.

These steps are graphically shown in Figure 2.2. The overall process produced filtered signals like the one illustrated in Figure 2.3 (bottom), whose effect in frequency domain can be appreciated through Power Spectral Density estimates (Figure 2.4).

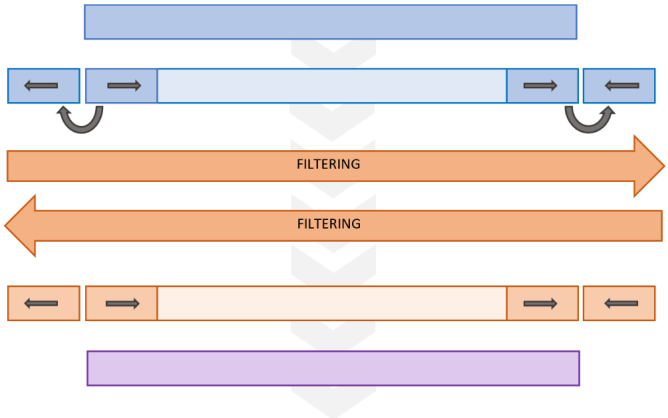


Figure 2.2: Reflection algorithm scheme.

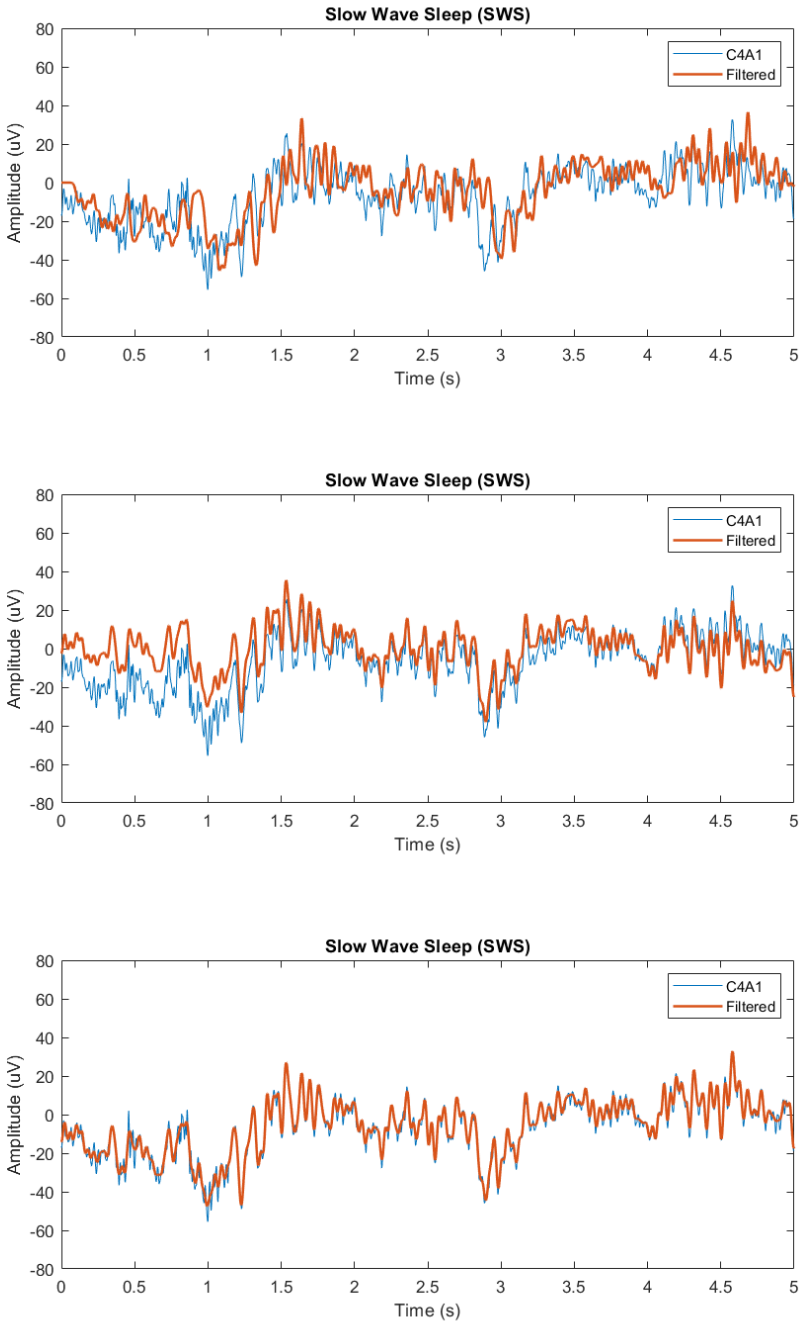


Figure 2.3: Filtered SWS signal.
Patient n1, channel C4-A1. Top: Conventional Filtering, HPF @0.01 Hz.
Middle: Zero-phase Filtering, HPF @0.5 Hz.
Bottom: Zero-phase Filtering, HPF @0.01 Hz (correct).

In Figure 2.3 (middle), they are shown the effect of using a cut-off frequency higher than 0.1 Hz, confirming Acunzo et al. (2012) [30] research work findings, and suggesting the reliability of the proposed guidelines, along with the effectiveness of keeping the influence of High-Pass filter as low as possible to avoid EEG signal distortions (Figure 2.3, bottom).

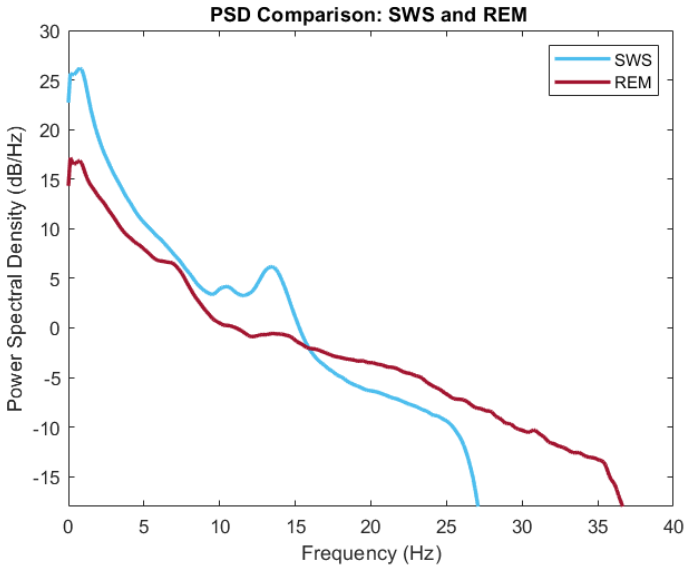


Figure 2.4: Power Spectral Density of filtered REM and SWS segments. Patient n1, channel C4-A1, averaged across all the 2s-epochs.

2.2.2 Feature Extraction

In this section, relevant sleep features are analyzed and manually extracted in MATLAB (R2021b) environment, to quantitatively characterize both the Polysomnographic and Electroencephalographic "information" about the individuals contained in the databases. An important aspect of this step concerns keeping as simple as possible the interpretability of the results by focusing on features that have a clear meaning in the medical field or at least have a tight link with it, to be interpreted by expert clinician personnel. Starting from the works of Rechichi et al. (2021) [8], (2022) [27] and following the existing literature about this topic [32], [33], [34], [35], [36], [37], [38], [39], a review of the most commonly used features in the field of EEG-based sleep study was performed, ending up with an omni-comprehensive list of multi-domain attributes to be used in the context of automatic classification of healthy and sick patients. Features have been divided into Polysomnographic and Electroencephalographic features.

Polysomnographic Features. Concern parameters that were extracted from

the hypnogram, which define the quality, duration, and characteristics of sleep macrostructures [27] (Table 2.5).

Table 2.5: Polysomnographic features, adapted from [27]

Feature	Description
Sleep Onset Latency (SOL)	The amount of time required for transitioning from awake to N1 (min)
Wake After Sleep Onset (WASO)	The amount of time the subject stays awake during the night (min)
Total Sleep Time (TST)	Total number of hours of sleep (h)
Time in Bed (TIB)	Lights-off lights-on interval (h)
Sleep Efficiency (SE)	The percentage of time spent sleeping in bed (%)
Arousal Index (ARI)	The frequency of arousals events[2]
Minutes of REM Sleep (MREM)	Total number of REM epochs (min)
Sleep Stage Proportion (SSP)	Ratio of each sleep stage (N1, N2, SWS, R) to TST (%)
NREM Fragmentation Index (NFI)	The number of transitions from NREM to any other NREM stage (per hour of NREM sleep)
REM Fragmentation Index (RFI)	The number of transition from REM to any other stage (per hour of REM sleep)
Proportion of Wake (WP)	Time spent awake during the night (%)
Sleep Transitions Index (STI)	The count of transitions from REM to NREM (and vice versa) per hour of sleep
Average Length (AL)	The average length of segments classified as the same sleep stage (N1, N2, SWS, R) (min)

[2]Arousals are interruption of the sleep (lasting 3-15s) which consist in abrupt change from SWS (or REM) to light sleep and that can lead to awakening if lasting more than 15s.

Electroencephalographic Features. They were extracted from both REM and SWS segments using the available EEG channel (C3-A2 or C4-A1). The extraction was performed in an epoch-wise fashion according to the division into four domains (Table 2.10):

- Time (TD): focuses mainly on the shape and statistical characteristics of the waveforms.
- Frequency (FD): analyzes the spectral properties of the signals through the estimation of the spectrum and its division in sub-bands.
- Time-Frequency (TFD): analyzes the spectrum properties of the signal accounting for the non-stationary nature of brain processes.
- Nonlinear (NND): focuses on the complexes and nonlinear brain mechanisms.

Given that the AASM criteria evaluate the EEG signal in 30s macro-epochs to score sleep[3], the data in this study (both for REM and SWS segments) were processed in this manner, at least for TD analysis. In FD, TFD, NND features were extracted on mini-epochs of 2s to ensure stationarity for the EEG (which is indeed inherently non-stationary) and allowing to analyze some phenomena more in detail.

A common procedure in this context consists in averaging the feature extracted on these 2s min-epochs across the corresponding 30s macro-epoch, to uniform the overall framework as well as achieve a better spectral frequency resolution. It is worth noticing that averaging requires data on 2s-epochs should be normally distributed along the correspondent 30s epoch, to avoid the mean to bring biased information. Such analysis was performed through a three-step process comprising the Shapiro-Wilk normality test with $alpha = 0.01$ (*Ahmed BenSaïda (2023). Shapiro-Wilk and Shapiro-Francia normality tests, MATLAB Central File Exchange. Retrieved June 26, 2023*) which produces a null hypothesis rejection rate, and the graphical inspection of both PP-plots and frequency distribution histograms (accompanied by normal and kernel distribution fitting) (Figure 2.5) through which the previous results were checked[4] [40].

[3]Initially, the R&K guidelines (Rechtschaffen & Kales 1969) suggested separating the PSG sleep record into 30 s epochs. The 30s interval was chosen because one page equated to 30 seconds of recording at a paper speed of 10 mm/s (ideal speed for viewing alpha and spindles).

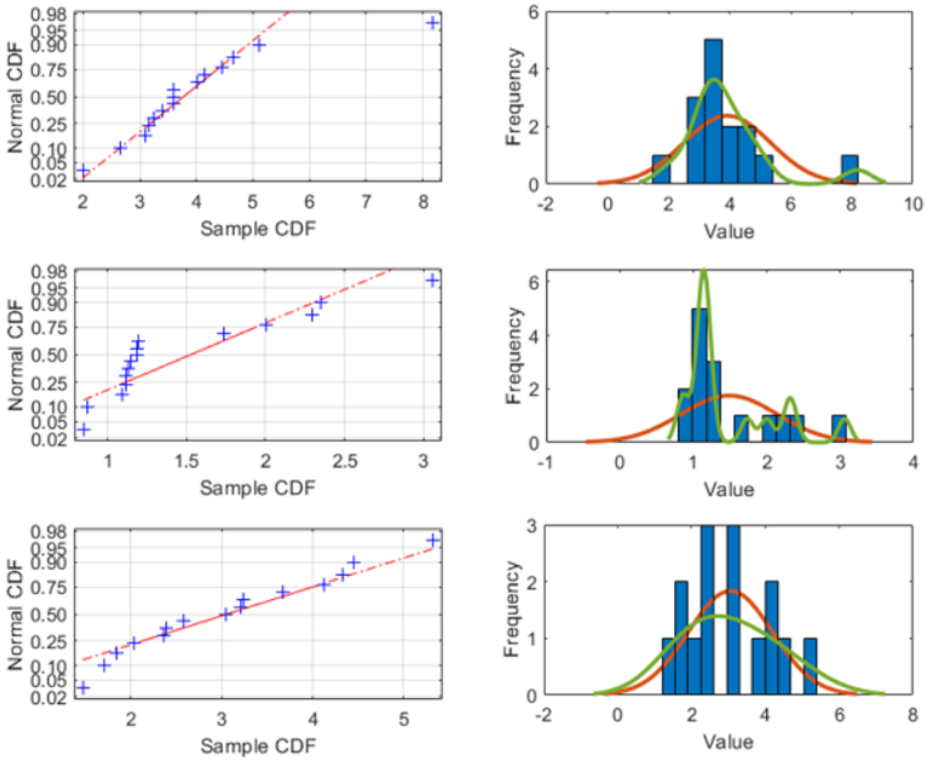


Figure 2.5: Example of normality-test results.
Patient n1, channel C4-A1. The green line on the histogram plots is the kernel fitting curve, while the red one is the normal fitting curve (for comparison purposes). The plots display fifteen 2s epochs features (across the correspondent 30s macro-epoch):
Top: wrongly classified as not-normal by Shapiro-Wilk test.
Middle: correctly classified as not-normal by Shapiro-Wilk test.
Bottom: correctly classified as normal by Shapiro-Wilk test.

From this process it was possible to state not all 2s epoch-data were normal across their correspondent macro-epoch, and for this reasons they were not averaged confirming that in some cases the Mean value was not a good indicator of the underlying distribution.

[4]Shapiro-Wilk test is a statistical test used for assessing if the analyzed data come from a normal distribution. It is based on the null hypothesis (H0) data are sampled from a population following a normal distribution, and produces a significance value (p-value) which if greater than a threshold value (alpha, default 0.05) states there are no enough evidence to reject the null hypothesis, confirming the distribution is not significantly different from normal. PP-plots display the cumulative distribution function (CDF) from the sample data on the x-axis, and the CDF from a normal distribution on the y-axis. If the obtained data points lie on a straight line, the distribution under-analysis is normal. As far as histograms are concerned, normal distributions are displayed through a bell-shaped curve.

Time Domain

This family of features comprises different parameters which work directly on the time-series $x(t)$ ($t = 0, ..., N-1$ with N the epoch length in samples) characterizing the EEG signal from different perspectives, highlighting the underlying shape, amplitude, and amplitude variations, or extracting other information like energy and frequency-related ones.

Concerning statistical measures, it is possible to mention the standard moments like Mean (a measure of the central tendency of a finite set of numbers), Mode (the most common number in a dataset), Median (also called 50^{th}-p, is the value below which the 50% of the population of a certain distribution stays), Standard Deviation (a measure of the dispersion of a set of values with respect to the mean value), Skewness (a measure of the asymmetry of a distribution about its mean), and Kurtosis (a measure of the "tailedness" of a distribution with higher Kurtosis concerning a higher number of outliers) which are graphically depicted in Figure 2.6.

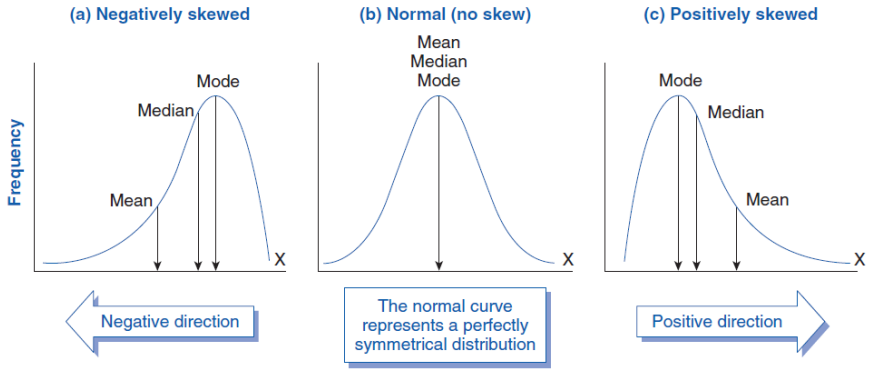


Figure 2.6: Statistical moments.
https://www.biologyforlife.com/skew.html

Zero-Crossing rate (ZCR): is the number of times the signal crosses the x-axis and reflects the frequency of the signal (high ZCR means the signal is dominated by high frequencies), computed as

$$Zero\ Crossing\ Rate = \frac{1}{N-1}\sum_{t=1}^{N} \mathbb{1}\{x(t)x(t-1) < 0\} \tag{2.1}$$

where $\mathbb{1}(X)$ is the indicator function (1 if X true, 0 else).

Hjorth parameters: Activity, mobility, and complexity collectively account for the EEG waveform variability describing its pattern in terms of amplitude, temporal scale, and complexity (Figure 2.7). They are defined in time-domain exploiting the first and second derivative of the EEG signal epoch $x(t)$, and create a sort

of "bridge between the physical time domain interpretation and the conventional frequency domain description" [41], allowing to extract some features about power spectrum i.e. spectral moments, defined as

$$m_n = \int_{-\infty}^{\infty} w^n * S(w)\, dw \tag{2.2}$$

where w is the frequency of interest, and *S(w)* is the correspondent value of the power spectrum.

While Activity represents the variance of the signal amplitude sometimes referred to as the mean power i.e. the surface of the power spectrum of the EEG signal, Mobility provides the average rapidity of the amplitude changes (in time) and it is linked to the variance of the power spectrum along the frequency axis. Complexity gives a measure "of excessive details with reference to the "softest" possible curve shape, the sine wave", which hence corresponds to unity.

$$Activity = var(x(t)) = m_0 \tag{2.3}$$

$$Mobility = \sqrt{\frac{var(\frac{dx(t)}{dt})}{var(x(t))}} = \sqrt{\frac{m_2}{m_0}} \tag{2.4}$$

$$Complexity = \frac{Mobility(\frac{dx(t)}{dt})}{Mobility(x(t))} = \sqrt{\frac{m_4/m_2}{m_2/m_0}} \tag{2.5}$$

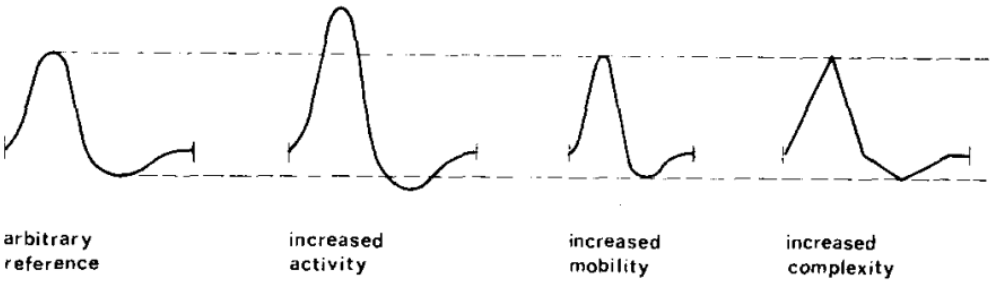


Figure 2.7: Hjorth parameters schematic interpretation [41].

From Hjorth parameters, it was possible to derive additional features concerning the spectral characterization of the EEG signal, defined in the following way

$$Sparseness = \frac{m_0}{\sqrt{(m_0 - m_4)(m_0 - m_2)}} \tag{2.6}$$

$$Irregularity = \frac{m_2}{\sqrt{m_0 * m_4}} \tag{2.7}$$

$$Spectrum\ Bandiwdth = \sqrt{1 - \frac{m_2^2}{m_0 * m_4}} \tag{2.8}$$

$$Average\ Zero\ Up\ Crossing\ Period = 2\pi * \sqrt{\frac{m_0}{m_2}} \tag{2.9}$$

$$Peak\ To\ Peak\ Period = 2\pi * \sqrt{\frac{m_2}{m_4}} \tag{2.10}$$

where Sparseness quantifies how much energy of the overall signal (considered as a vector) is packed into only a few components, and Irregularity (I) defines the ratio between the number of upward zero-crossing and the number of peaks. The Spectrum Bandwidth (B) is a parameter comprised in the range $[0, 1]$ and indicates a broadband (narrowband) signal if B ≥ 0.5 (B ≤ 0.5). Moreover, Average Zero-up Crossing Period (AZCP) and peak-to-peak Period (A3P) quantify the periods between two zero-crossing points and two peaks respectively.

Percentiles: these statistical measures give information about the range of a dataset as well as quantify data denseness. In this project, 25^{th}-p and 75^{th}-p were computed, together with the Interquartile range (IQR) which measures the difference between them characterizing the dispersion of the data with respect to the central value.

Form (FF), Crest (CF) and Impact Factors (IF) are impulsive metrics that were extracted from the EEG records to describe the peak amplitude and the waveform properties exploiting measures of RMS (Root Mean Square).

$$Form\ Factor = \frac{x_{RMS}}{|x|_{mean}} \tag{2.11}$$

$$Crest\ Factor = \frac{x_{peak}}{x_{RMS}} \tag{2.12}$$

$$Impact\ Factor = \frac{x_{peak}}{|x|_{mean}} \tag{2.13}$$

Coastline: is defined as the cumulative length of the waveform (i.e. length of the "stretched" version of the signal) over time, giving simultaneously a measure of the signal amplitude, frequency and duration.

$$Coastline = \sum_{t=1}^{N} |x(t) - x(t-1)| \tag{2.14}$$

Rechichi et al. (2021) [8] proposed envelope-based measures to characterize the morphology and the extrema of the EEG signal by analyzing instantaneously its amplitude changes. This characterization was then performed through three measures namely Number of peaks (NP, number of local extrema), Peaks Prominence (PP, how much a peak stands out with respect to nearby extrema), and Peak Width (PW, distance between peaks defined as the length of the segment created when lines tangent to the peak's left and right inflection points intersect the baseline) (Figure 2.8).

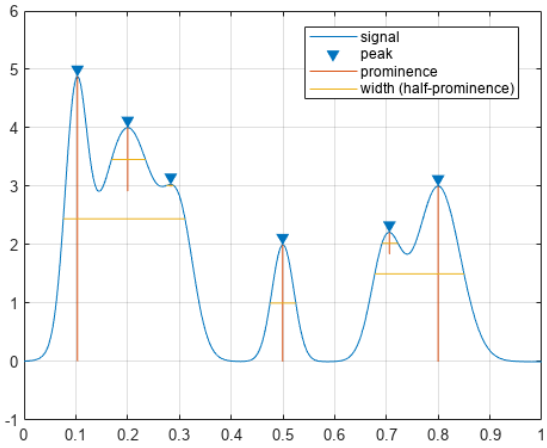


Figure 2.8: Envelope-based parameters.
https://it.mathworks.com/help/signal/ref/findpeaks.html

In the context of this work two novel statistical features were extracted since they have been proved to have a positive impact on the context of sleep stage classification performance [34]. The first feature is "Maximum-minimum Distance (MMD)" which exploits a sub-segmentation of the 30s macro-epoch in λ=3s mini-epochs. The parameter λ is defined by the authors as the "wavelength parameter of the EEG signal" useful for the computation of the second feature called "EnergySis (ESis)".

MMD (Figure 2.9) relies upon the Pythagorean Theorem using the maxima and minima present in each sub-epoch for computing the slope changes in the EEG pattern fluctuation which may be also interpreted as the signal speed changes.

$$d = \sqrt{\Delta t^2 + \Delta a^2} \tag{2.15}$$

where (on each sub-window) Δt indicates the time difference between maxima and minima points while Δa express their amplitude difference. The obtained values

are then averaged across the correspondent 30s epoch

$$Maximum\ Minimum\ Distance = \frac{1}{W} * \sum_{i=1}^{W} |d|_i \tag{2.16}$$

with W referring to the total number of sliding sub-windows in an epoch.

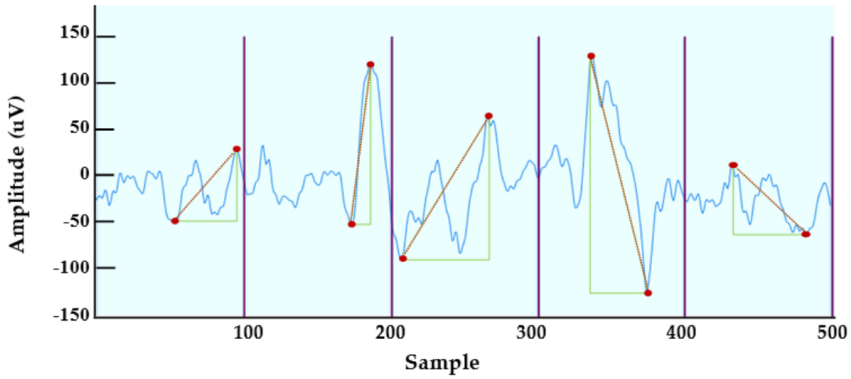


Figure 2.9: Maximum-minimum Distance (MMD) [34].

In signal theory the Energy of a continuous time signal x(t) is defined as $\sum |x(t)|^2$. ESis determines the energy of the EEG signal by considering the formula $\lambda = f/\nu$ which defines the relationship between wavelength (λ), mean frequency (f), and propagation speed (ν) of a radiation, and which allowed to compute the speed of the signal. The feature was then computed as

$$Energy\ Sis = \sum_{t=1}^{N} |x(t)^2| * \nu \tag{2.17}$$

Frequency Domain

The first step to extract features in FD encompasses the estimation of the Power Spectral Density (PSD)[5] which represents the distribution of the power content of a signal with respect to frequency, and that will be the main analysis object. In this regard, it is essential to consider that EEG signal is inherently non-stationary[6] since its source system (the brain) is characterized by activity that is mixed in frequency and which can also change in time based on internal/external stimuli. The mathematical basic foundation of PSD estimation is the Fourier Transform (FT) which does not work well with non-stationary processes since relies on the fundamental idea every signal can be decomposed in a sum of sine waves (which are perfectly stationary) and thus intrinsically assuming that in terms of statistical properties, each frequency component remains as it is permanently [35]. In the case of EEG, this assumption brings to fail some important information about the time location of spectral events which in turn brings to biased results about the Power content.

The solution to overcome this limit involves (as mentioned in the previous section) the segmentation of the signal in mini-epochs before performing FD analysis since here the EEG signal can be considered stationary (WSS) (Figure 2.10). In literature, it exists a widespread consensus about the choice of the values of sleep EEG time windows (T), which usually range from 1s to 5s. To have a good compromise between temporal resolution (T), theoretical frequency resolution ($r_t = 1/T$), and other parameters related to spectral estimation will be discussed in detail later (Table 2.6), $T = 2s$ ($N = 1024$) was chosen which moreover, turned out to be suggested in different topic-related research works [42], [43].

[5]EEG signal is a power signal since its energy is theoretically infinite (practically very big) while its power is finite. This is the reason why spectral analysis relies upon PSD rather than Energy spectrum Density (ESD) estimation. It is a matter of estimate since the EEG exhibits stochasticity (due to properties of the brain) but can be analyzed through finite length single realization only.

[6]A stationary signal is a process whose time period, frequency, and spectral content do not change with time. More specifically the stationarity property requires certain mathematical time-invariance conditions to be satisfied which allow to define two types of stationarities namely strict-sense (SSS) and wide-sense (WSS) stationarity. The second comprises less stringent requirements since requires a time series $x(t)$ to have time-constant mean and variance, and autocorrelation function which depends only on the time difference $\tau = t_2 - t_1$ which means correlation between points of a signal $x(t)$ depends only on how far apart they are in time (τ), not where they are in time.

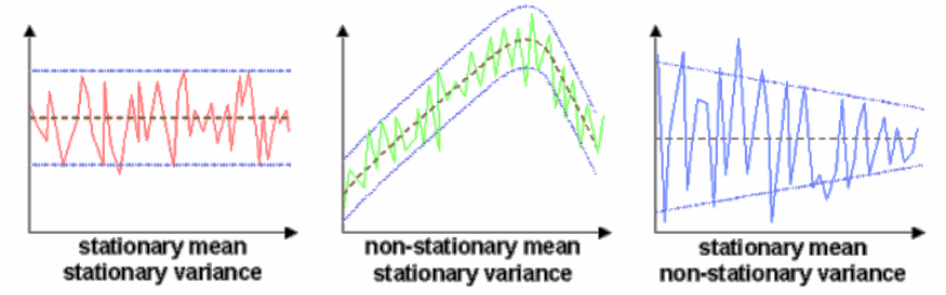


Figure 2.10: Stationarity.

Spectral estimation is a tool introduced to overcome some limitations of TD analysis since breaks down the signal into its constitutive frequency components allowing to discover inner mechanisms and to more accurately distinguish between information of interest from confounding factors such as noise and artifacts [44].

Among the various existing non-parametric techniques to produce spectral estimates, the Multitaper method (Thomson, 1982) was implemented due to its advantages with respect to other common approaches like classical Periodogram and Welch's method, concerning a better trade-off of spectral resolution (narrowband bias[7]), spectral leakage (broadband bias) reduction, theoretical frequency resolution, and variance reduction. Moreover, it results more efficient in dealing with non-stationarity also compared to parametric methods [45] and is particularly suitable for those signals (such as EEG) having power spectra of the form f^{-b} where b is a positive power-low exponent [46].

The Multitaper method produces a power estimate by averaging across K (and not one) modified Periodograms of the same signal $x(t)$, each obtained exploiting a different taper sequence as window function (Figure 2.11):

$$S^{(MT)}(f) = \frac{1}{K} \sum_{k=0}^{K-1} S_k(f) \tag{2.18}$$

$$S_k(f) = |\sum_{t=0}^{N-1} h_k(t) x(t) e^{-j2\pi ft}|^2 \tag{2.19}$$

[7]Theoretically the spectrum of a sinusoid of frequency f is a single vertical peak at that frequency. In practice, any estimate exhibits a main lobe centered @f and is affected by a ringing effect which creates infinite side lobes. The fact central lobe has a bandwidth $[-W, W]$ creates the so-called narrowband bias, which affects spectral resolution since all the frequencies within the small range of its bandwidth are blurred and multiple frequency peaks occurring within this bandwidth appear unified. On the other hand, the side lobes create the so-called broadband bias, due to the fact the power that should be centered @f is transmitted to false sides-frequencies producing then spectral leakage.

where $S^{(MT)}(f)$ is the Multitaper spectrum estimate and $S_k(f)$ denotes the modified periodogram (of the signal epoch $x(t)$ of length N) obtained with the k^{th} taper sequence $h_k(t)$ $(k = 1, ..., K)$.

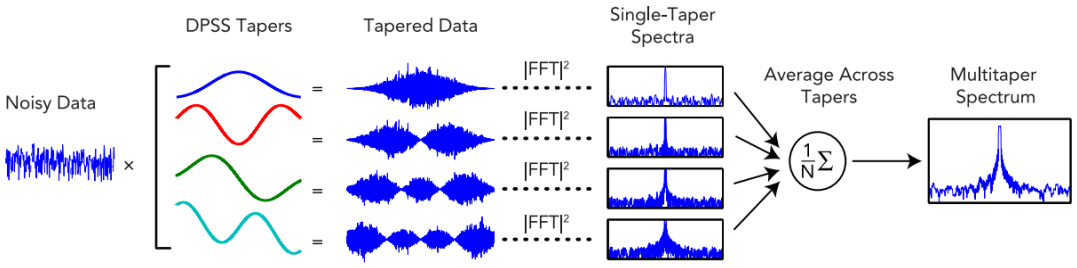


Figure 2.11: Multitaper method [44].

The smooth functions used in the context of this work are the Discrete Prolate Spheroidal Sequences (DPSS) or Slepian tapers (Slepian and Pollak, 1961) which are a particular choice of taper functions, characterized by two main properties that make the real strength of Multitaper method [47], [48]:

1. They have optimal spectral concentration properties since remove the power from the side lobes concentrating it in the main lobe, where is the true frequency. This expedient reduces the broadband bias, at a cost of a slightly larger narrowband one (trade-off).
2. They form an orthogonal set of sequences i.e. the tapers produce estimates of the data which are completely uncorrelated from a statistical point of view, in such a way when the tapered versions of the signal are averaged, the variance is reduced.

Designing a DPSS set means finding the time series $h(t)$ of length N such that the energy of its Fourier transform $H(f)$

$$H(f) = \sum_{t=1}^{N} h(t) e^{-j2\pi ft} \tag{2.20}$$

is maximally localized on a given frequency interval $[-W, W]$.

The optimality condition ends up with a matrix eigenvalue equation for the sequence $h(t)$ whose eigenvectors are the DPSS, and whose eigenvalues

$$\lambda_k(N, W) = \frac{\int_{-W}^{W} ||H(f)||^2 \, df}{\int_{-F_s/2}^{F_s/2} ||H(f)||^2 \, df} \quad \in (0, 1) \tag{2.21}$$

define the spectral concentration (ratio between the power contained in the frequency interval $[-W, W]$ and the power contained in the entire frequency band

$[-F_s/2, F_s/2]$). The remarkable fact is that these sequences do not have equal energy concentration in the frequency range of interest. The first K eigenvalues $\lambda_k(N, W)$ (sorted in descending order) are each almost close to unity, while the others are approximately zero (Figure 2.12).

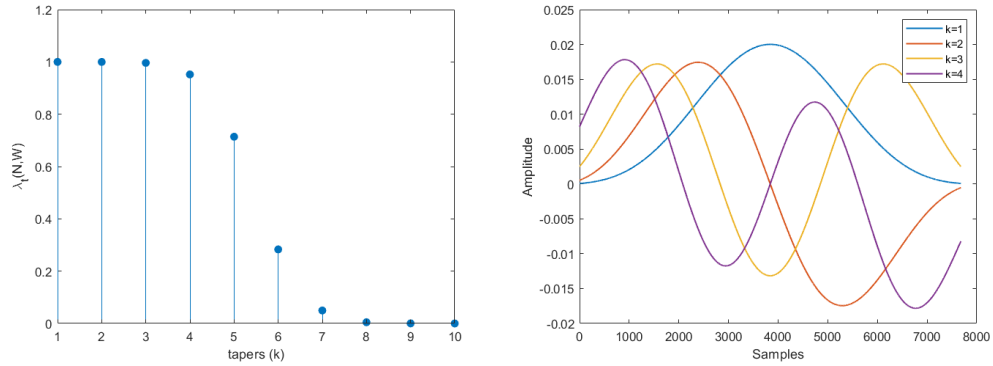


Figure 2.12: Slepian Tapers [47].
The left figure shows the Slepian sequences energy concentration for T=15s, nw=2.5 and K=4. It is possible to notice that the Slepian sequences after the 4^{th} start to concentrate less energy in the interval $[-W, W]$.

The parameter values chosen through the followed design procedure are grouped in the Table 2.6. They were tuned in order to obtain a balanced trade-off of an acceptable broadband bias reduction at a cost of a limited narrowband bias (W=1.5Hz), a good variance decrease ($T = 2s$ and $W = 1.5Hz$ ensure at least a couple of tapers), an adequate temporal resolution ($T = 2s$) as well as satisfactory spectral theoretical ($r_t = 0.5Hz$) and apparent ($r_a = 0.125Hz$) resolutions[8] (Figure 2.13).

[8]The number of points used to compute the DTFT was selected considering the relationship $r_t \geqslant r_a$, where r_t is the theoretical resolution i.e. the max resolution can be obtained and it is uniquely dependent on the length of the mini-epoch ($r_t = 1/T$)), while r_a is the apparent resolution which is related to the fact DTFT requires a certain number of points to represent the spectrum and it is defined as $r_a = F_s/nfft$. Following the first expression, the minimum number of DTFT points that should be used to not have visualization problems is $nfft = N$ ($N = 1024$, $nfft$ chosen as a power of 2). Since $T = 2s$ and $r_t = 0.5Hz$, the apparent resolution was enhanced to $r_a = 0.125Hz$ increasing $nfft$ to 4096.

Table 2.6: Thomson's Multitaper parameters

Parameter		Value	Description
Epoch length	T (N)	2s (1024)	chosen to ensure the stationarity of EEG and to balance the trade-off with the theoretical frequency resolution $r_t = 1/T$.
Resolution bandwidth	W	1.5Hz	width of the Multitaper estimate central lobe i.e. the minimum distance between frequency peaks is expected to be resolved. It is chosen smaller than the minimum frequency range in the signal spectrum (i.e. δ band).
Time-halfbandwidth product	nw	1.5	determines the number of tapers that will be used and hence influences the variance reduction of the estimate. It is defined as $nw = W * T/2$.
Number of Slepian tapers	K	2	number of tapers for which the energy concentration in $[-W, W]$ is close to unity. It is defined as $K = \lfloor 2 * nw \rfloor - 1$

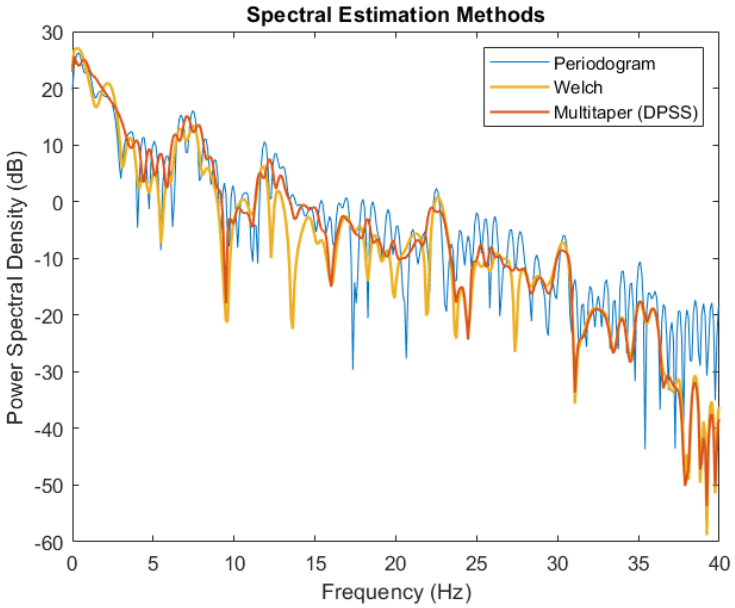


Figure 2.13: Comparison of Spectral estimation methods.
PSD of the first 2s mini-epoch of a REM segment, patient n1, channel C4-A1. Welch's method is performed on the total signal exploiting segmentation of T=2s and 50% of overlapping.

After the power spectra of all REM and SWS 2s epochs were computed (in uV^2/Hz and not in dB in order to work with non-negative power values only), it was possible to enter in more detail with the quantification of some spectral properties which are widely used in the field of EEG-based automatic sleep analysis but also with features commonly exploited in more general biosignals studies.

As mentioned in Table 2.10, spectral features were computed (for both REM and SWS segments) on the total spectrum ($0.01 - 30/40Hz$) but also on each EEG sub-band to highlight those mechanisms that are characteristics of just a reduced portion of the overall spectral activity. Beside classical EEG sub-bands (δ, θ, α, β, γ), additional frequency ranges ($TREM$, $FREM$, SWA, and SOs) were studied since in some research works they exhibited very recognizable characteristics whose anomalies and deviations can be exploited as sleep disorder indicators [8], [9], [19]. Moreover, the choice to extend frequency bands which are characteristics of REM (like TREM and FREM) also to SWS segments, and vice-versa, was inspired by the results achieved by the work of Buettner al. (2020) [49] which states that information with respect to RBD disorder may be hidden in frequency ranges within θ, δ, and α bands, hence suggesting that a finer division of the EEG spectrum during the feature extraction process can improve the performances in healthy-RBD classification problems (Figure 2.14).

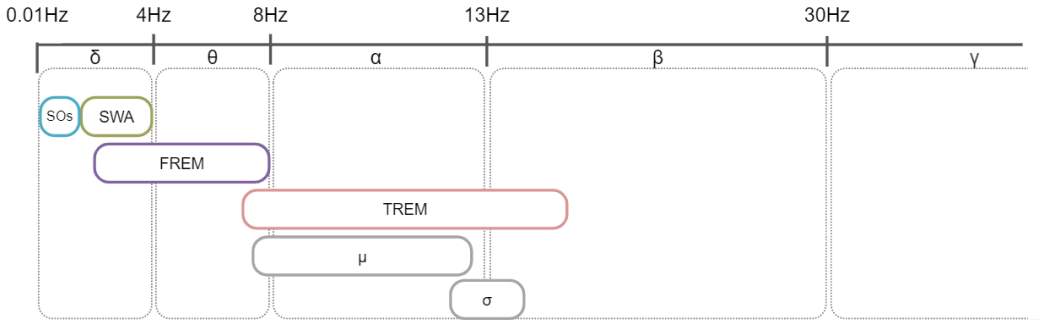


Figure 2.14: EEG spectrum partition.

Absolute Power Spectrum (AP): it describes the power of the signal and it is defined as the area under the PSD curve (or a portion of it) which in this work is computed through numerical integration exploiting the trapezoidal method. It belongs to the family of spectral moments (0^{th} order) and from a conceptual point of view AP is computed as

$$Absolute\ Power = \sum_{j=b1}^{b2} S_j \quad b1, b2 \in [1, M] \tag{2.22}$$

where S_j is the PSD value at the j^{th} bin, and b1 and b2 are the (bin) band edges over which perform the computation (if $b1 = 1$ and $b2 = M$ the power is related to

the total spectrum, since M is the bin corresponding to the last frequency of the PSD).

Mean Power Spectrum (MP): it characterizes the average spectrum content associated to each frequency bin, and it is particularly useful for resuming information about power when frequency bands are not equally occupied.

$$Mean\ Power = \frac{1}{b2 - b1} \sum_{j=b1}^{b2} S_j \tag{2.23}$$

Spectral Crest (SCr): it is an indicator of the peakiness of the spectrum (or a portion of it), since compares its maximum and average values

$$Spectral\ Crest = \frac{\max\{S_{j \in |b1,b2|}\}}{\frac{1}{b2-b1} \sum_{j=b1}^{b2} S_j} \tag{2.24}$$

Peak Frequency (PKF): it is the frequency at which the maximum of the Power spectrum in a certain frequency band occurs.

$$Peak\ Frequency = \max_{j \in [b1,b2]} S_j \tag{2.25}$$

Spectral Centroid (SCe): it is also called Mean Frequency and represents an average frequency. Some works highlighted its reliability in characterizing EEG synchronization during deep sleep, which makes it a good tool in classification contexts [39]. It belongs to the family of spectral moments (1^{st} order).

$$Spectral\ Centroid = \frac{\sum_{j=b1}^{b2} f_j S_j}{\sum_{j=b1}^{b2} S_j} \propto SM1 \tag{2.26}$$

where f_j is the frequency value corresponding to the j^{th} bin.

Median Frequency (MDF or SEF50): it is the frequency below which the 50% of the total power in a certain frequency band is contained, or in other words it is the frequency that divides the power spectrum into two regions with equal amplitude.

$$MDF = f \quad \text{s.t.} \quad \sum_{j=b1}^{MDF} S_j = \sum_{j=MDF}^{b2} S_j = \frac{1}{2} \sum_{j=b1}^{b2} S_j \tag{2.27}$$

In MATLAB this feature was implemented by computing the cumulative sum of PSD values and using it to find (through interpolation) the frequency value which corresponds to half the above-mentioned cumulative sum (query point).

Spectral Edge Frequencies (SEFxx): percentile measurements in the frequency domain, in addition to MDF. They denote the lowest frequency below which the xx% of the total power in a certain frequency band is contained, with xx

typically chosen equal to 25, 75, and 95 (SEF25, SEF75, and SEF95 respectively) (Figure 2.15).

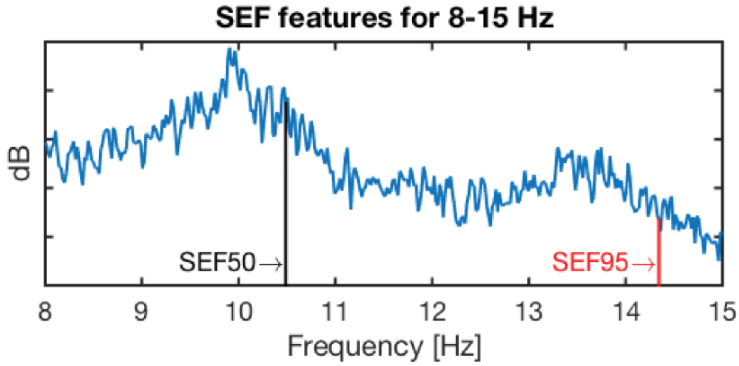


Figure 2.15: Spectral Edge Frequencies [50].

Spectral Edge Frequencies differentials (SEFd): defined as the difference of SEFxx measures. It is a pretty novel features family in the context of sleep studies, used for the first time by Imtiaz et al. (2014) [51], which highlighted its great ability, together with SEF50 and SEF95, in discriminating REM with respect to other sleep stages, especially if applied on TREM (specifically in the band 8-16 Hz), resulting however more powerful compared to SEF.

Following the above-mentioned findings all these edge frequencies metrics were used on REM and SWS segments, and in both total and relative spectral ranges of the EEG in order to exploit their discrimination capabilities in the current healthy/sick classification problem.

Relative Power Spectrum (RP): it is extracted just for the EEG sub-bands and expresses (in percentage) their power content relative to the power of the entire spectrum, indicating how much of the total power is packed in the various relevant EEG frequency ranges.

$$Relative\ Power(\%) = \frac{\sum_{j=b1}^{b2} S_j}{\sum_{j=1}^{M} S_j} * 100 \tag{2.28}$$

Spectral moments (SM): it is an alternative statistical way to extract features from the power spectrum, accounting for its arrangement along the frequency axis. Like in TD, these metrics provide information about the shape (center, dispersion, symmetry, and flatness) of the PSD on average [52], [53]. In the current work these measures were normalized with respect to the Absolute Power (i.e. the 0^{th} order moment SM0) and centered around the Spectral Centroid (called μ_1 in this context) defining:

Spectral Spread (SSp), representing the spread of the power with respect to SCe

$$\mu_2 = \sqrt{\frac{\sum_{j=b1}^{b2}(f_j - \mu_1)^2 S_j}{\sum_{j=b1}^{b2} S_j}} \propto SM2 \tag{2.29}$$

Spectral Skewness (SSk), representing the spectrum symmetry around its SCe

$$\mu_3 = \frac{\sum_{j=b1}^{b2}(f_j - \mu_1)^3 S_j}{(\mu_2)^3 \sum_{j=b1}^{b2} S_j} \propto SM3 \tag{2.30}$$

Spectral Kurtosis (SK), representing the flatness of the spectrum around its SCe

$$\mu_4 = \frac{\sum_{j=b1}^{b2}(f_j - \mu_1)^4 S_j}{(\mu_2)^4 \sum_{j=b1}^{b2} S_j} \propto SM4 \tag{2.31}$$

Exploiting SM0, SM1, and SM2 instead, the Variance of the Central Frequency (VCF) is computed, as

$$VCF = \frac{SM2}{SM0} - (\frac{SM1}{SM0})^2 \tag{2.32}$$

which represents a slightly different way to define the Spectral Spread.

Entropy Measures: in some recent work about the automatic detection of pathological patterns in EEG signals it has been proven the optimality of entropy measures in increasing the performances of the diagnostic system [54] suggesting that diseased states affect the signal complexity. For this reason, a couple of such metrics (namely Spectral Entropy and Renyi's quadratic Entropy) were added to the feature set. Spectral Entropy measures reflect the spectral complexity of the EEG signal i.e. they quantify the peakedness or flatness (density) of the distribution of the EEG power spectrum [55], in such a way higher entropy values correspond to increased pattern irregularity [38]. Different logarithmic bases can be exploited, but natural logarithm was used as suggested in the work of Kristína Šušmáková et al. (2008) [39].

Spectral Entropy (SEN): it is a modified version of the Shannon Entropy in which the normalized Spectrum is treated as a probability distribution

$$Spectral\ Entropy = -\sum_{j=b1}^{b2} P_j \ln P_j \tag{2.33}$$

where P is the PSD normalized with respect to the Absolute Power

$$P_j = \frac{S_j}{\sum_{j=b1}^{b2} S_j} \quad \text{s.t.} \quad \sum_{j=b1}^{b2} P_j = 1 \tag{2.34}$$

Low values of SEN mean the spectrum is dense in some frequency regions since there will be more frequency bins with low power and fewer frequency bins with high power, while high values of SEN mean the spectrum is more or less uniformly distributed. This is the reason why sine wave has SEN=0 (narrow and peaked spectrum), while white noise has SEN=1 (flat spectrum since containing power in all frequencies) [38], [56]. This allows to state that lower SEN values correspond to signals that are more regular and predictable, rather than complex.

Renyi's quadratic Entropy (REN): compared to SEN, the sum is weighted toward lower frequencies which is particularly useful for sleep EEG, which typically exhibits more power in lower frequency bands.

$$Renyi's\ quadratic\ Entropy = -\ln \sum_{j=b1}^{b2} P_j^2 \tag{2.35}$$

Time-Frequency Domain

Until now the EEG signal was segmented in mini epochs to fulfill the stationarity conditions of the FT, however, the obtained results brought in some sense an average aspect of the spectral activity, ignoring remarkable temporal information like the time-location of oscillatory events and time-evolution of the frequency content. A key aspect of biomedical signals is indeed that they are the combination of both transient and diffuse phenomena which may exist in the time-frequency domain. A powerful approach to overcome these limits encompasses the Wavelet Transform (WT) analysis which exploits the property of low-frequency signals to be widespread over time, and of high-frequency bursts occurring during short intervals [57], for implementing a tailored time-frequency localization tool [58] particularly appropriate for processes which show frequencies that vary over time, transients, or slowly varying trends, such as EEG signals.

This method divides the signal under analysis into different frequency components each of which can be studied at a resolution matched to its scale thanks to the suitable variation of the time-frequency aspect ratio (Figure 2.16, left) providing:

- good time resolution and "poor" frequency resolution @ high frequencies.
- good frequency resolution and "poor" time resolution @ low frequencies.

which results in a non-uniform tiling[9].

WT replaces sine waves of FT by translations (by a time factor) and dilation (by a scaling factor) of window functions called wavelets which are finite-energy wave-like oscillations with zero mean [35] (Figure 2.16, right).

[9]FT shows high frequency resolution but zero time resolution, STFT provides a better trade-off between time and frequency resolutions however it is restricted by the Fourier "uncertainty principle", for which there exists a limit to what it is possible to know simultaneously in time and frequency domain since small time windows allow to know more about the location of a frequency event in time, but not on the frequency value itself, and large time windows allow to know about the frequency value and less about the time occurrence. WT leads to the best trade-off since it not only indicates which frequencies are present in a signal, but also when these frequencies occur in time thanks to the fact it provides high frequency resolution and low time resolution for small frequency values as well as low frequency resolution and high time resolution for large frequency values.

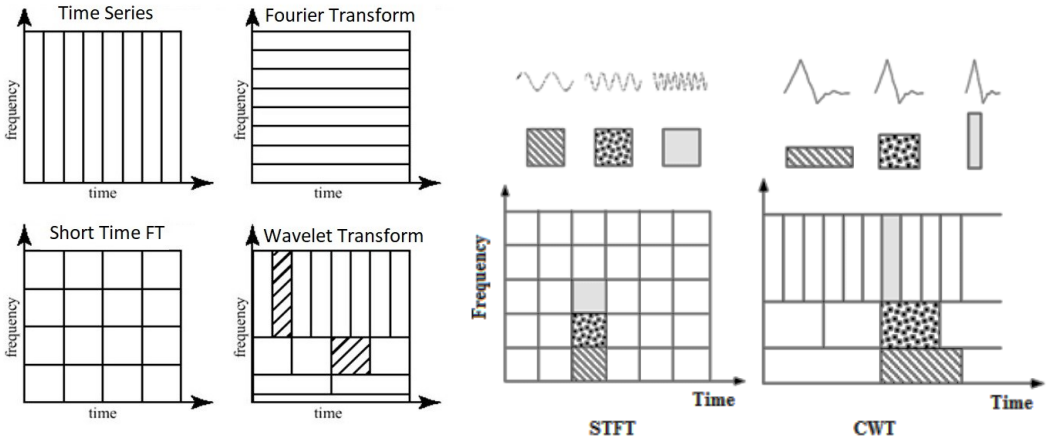


Figure 2.16: Comparison among various time and frequency tiling approaches. *https://ataspinar.com/2018/12/21/a-guide-for-using-the-wavelet-transform-in-machine-learning/ https://www.researchgate.net/.*

The family of wavelets comprises a Father $\Phi(t)$ wavelet and "different" Mother wavelets $\Psi(t)$, whose linear combination (by wavelet coefficients) reconstructs the signal. Each scaled version of the wavelet is shifted along (and correlates with) the entire signal length, and this process is repeated for all the scaled versions creating what are called "levels of decomposition" which (through wavelet coefficients) allow decomposing the signal into a pre-defined number of frequency components (HP and LP) which depends on the number of levels.

The Father wavelet is built by dilating (large scale factor) the mother wavelet and is responsible for the low frequency components i.e. it is good in representing the smooth and low-frequency part of the signal. On the other hand, compressing (small scale factor) the Mother wavelet produces other window functions which are responsible for the high frequency components i.e. they are good at representing the detailed and high-frequency part of the signal.

From a conceptual point of view, the scaling factor s characterizes the frequency action of the wavelet, and it is defined as

$$s \quad \text{s.t.} \quad \Psi(\frac{t}{s}), \quad s > 0 \tag{2.36}$$

the smaller the scaling factor, the more the wavelet will be compressed. A compressed wavelet helps in capturing the abrupt changes of a signal. The reciprocal relationship between the scale and the frequency of the wavelet involves a proportionality constant called "center frequency of the wavelet" which accounts for the fact that, unlike sine waves, wavelets have a bandpass behavior in frequency

domain i.e. they are characterized by a range of frequencies.

$$F_{eq} = \frac{C_f}{s\delta t} \tag{2.37}$$

where s is the scaling factor, C_f is the center frequency of the wavelet, δt is the sampling interval, and F_{eq} is the equivalent frequency [59]. According to this statement, the relationship scale/frequency is defined in Table 2.7.

Table 2.7: Scale/Frequency relationship

Scaling factor	2	4	8	16
Frequency reduction	$F_{eq}/2$	$F_{eq}/4$	$F_{eq}/8$	$F_{eq}/16$

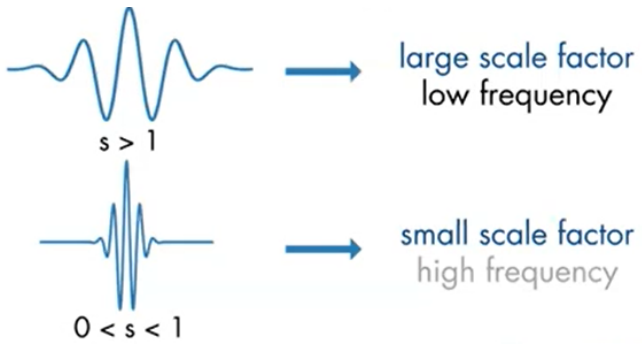


Figure 2.17: Relationship between the scaling factor and the wavelet frequency *https://medium.com/*

On the other hand, the shifting factor k defines the time onset of the wavelet along the length of the signal

$$k \quad \text{s.t.} \quad \Psi(t-k), \quad k > 0 \tag{2.38}$$

In the context of this thesis project, Discrete Wavelet Transform (DWT) was exploited since presenting some advantages with respect to Continuous Wavelet Transform (CWT) [60]:

- it enables a sparser representation in which redundancies are discarded (compression and noise reduction), producing a high quality signal approximation.
- the number of output coefficients is equal to N (epoch length), and at each decomposition stage the signal is downsampled[10] (by 2) thus implying a lower computational burden.

- it does not require an explicit definition of the wavelet functions but exploits just discrete filter banks (LPF and HPF) thus it is equivalent to pass many times (i.e. for many levels of decomposition) the signal through LPF (Father wavelet) and HPF (Mother wavelet). LPF stage produces "approximation coefficients" while HPF produces "detail coefficients".

- the scale parameter $s = 2^j$ is an integer power of 2 and the translation parameter $k = 2^j * m$ is always proportional to it (dyadic scaling and shifting).

$$\Psi(\frac{1}{2^j}(t - 2^j * m)) \quad j, m = 1, 2, 3, ... \tag{2.39}$$

Moreover, among the different types of existing Wavelet functions (Figure 2.18), Daubechies wavelets of order 4 (db4) were used (Figure 2.21) as suggested in the work of Abdulhamit Subasi (2007) [58].

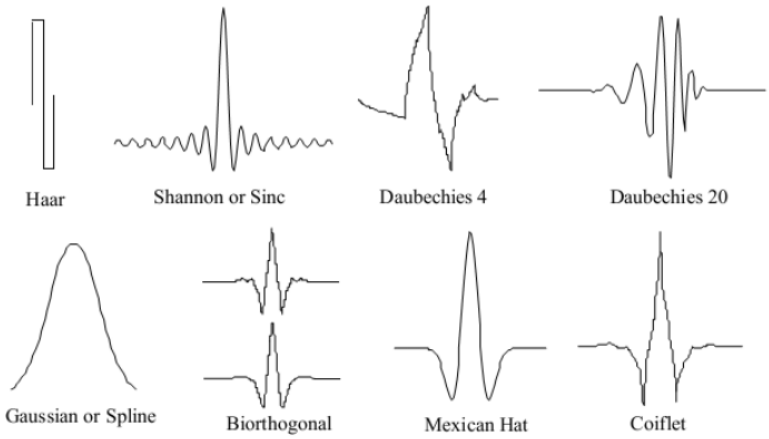


Figure 2.18: Wavelet types.
https://medium.com/

Considering $F_s = 512Hz$, the initial spectrum was defined by the Nyquist frequency (the maximum frequency that may be accurately represented, $\leq F_s/2$) $F_N = 256Hz$. It was then divided by filtering the original signal multiple times, and choosing a level of decomposition equal to 6 an almost exact decomposition of the EEG spectrum into its relevant frequency bands was provided (Figure 2.19).

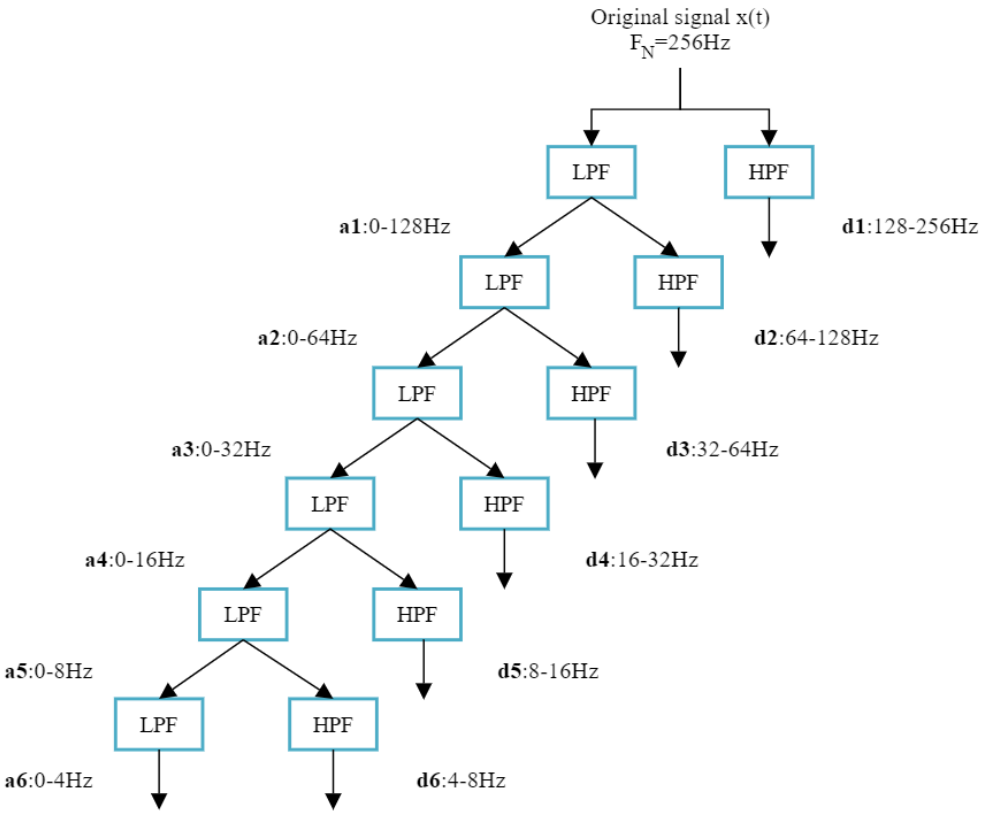


Figure 2.19: Tree structure for EEG discrete wavelet decomposition.

The coefficients (Table 2.8 and Figure 2.20) were computed for each 2s mini-epoch and then statistical measures like Mean, Standard Deviation, Coastline (defined in the TD section), and the ratio between successive sub-bands (RASb) were extracted in order to decrease the dimensionality of the obtained vectors [58]. The latter metrics is defined as

$$RASb = \frac{|coef_1|_{mean}}{|coef_2|_{mean}} \tag{2.40}$$

where $coef_1$ and $coef_2$ are couples of successive coefficients.

[10]Assume the original sampling rate of the signal is F_s (thus $F_N = F_s/2$) and that the filters perfectly divide the band into two portions at each stage. After the first LPF the frequency content will range from $\in (0, F_s/4)$ and thus the sampling rate can be reduced to $F_s/2$ maintaining compliance with the Nyquist theorem. For the HPF, the frequency content lies in $\in (F_s/4, F_s/2)$ and there is no content in the range $\in (0, F_s/4)$. For this reason also in this case it is possible to downsample by a factor of 2. With this expedient DWT reduces the overall computational burden.

Table 2.8: Wavelet coefficients

Coefficient	Frequency Range (Hz)	EEG relevant band
d3	32-64	$\approx \gamma$
d4	16-32	$\approx \beta$
d5	8-16	$\approx \alpha$
d6	4-8	$\approx \theta$
a6	0-4	$\approx \delta$

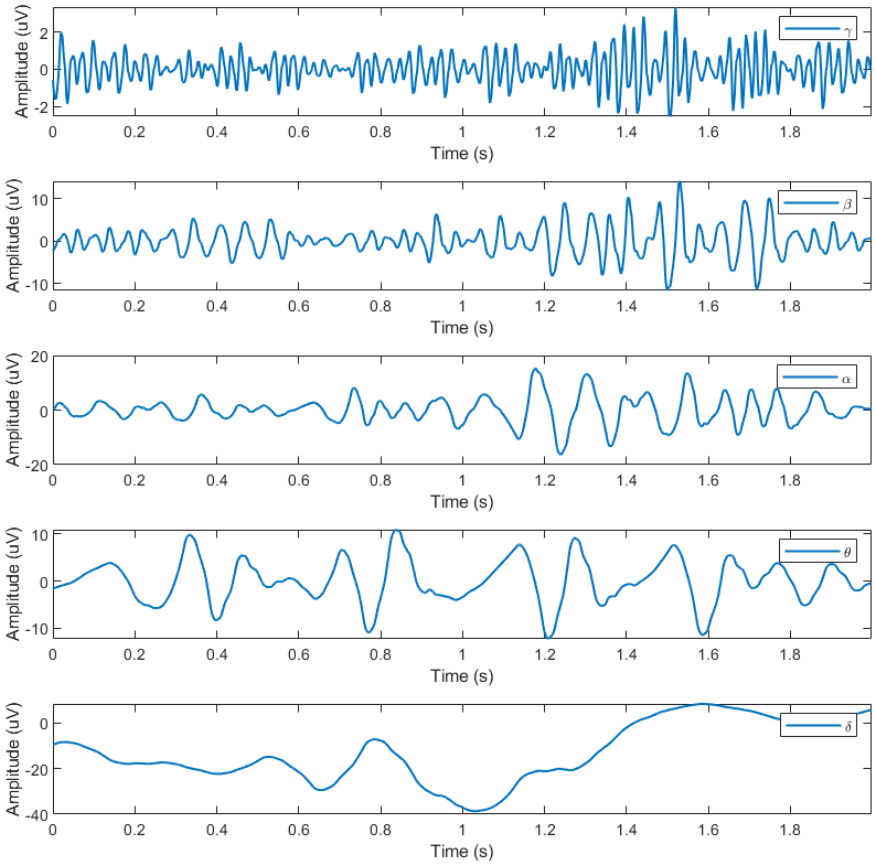


Figure 2.20: Wavelet coefficients evolution in time.
DWT coefficients computed for the first 2s mini-epoch of a SWS segment, patient n1, channel C4-A1

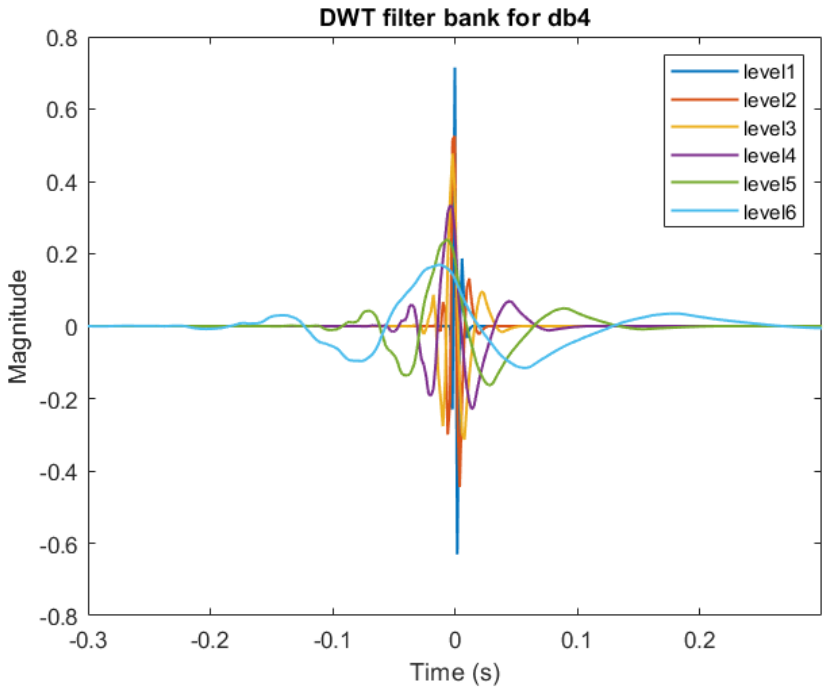


Figure 2.21: Wavelet functions (db4).

Nonlinear Domain

Besides assuming stationarity of the EEG signal, conventional Fourier-based EEG analysis techniques rely on the hypothesis of linearity of the signal under study. This is however in contrast with the well-established knowledge that the EEG is a signal originated by complex desynchronized oscillatory mechanisms related to the combination of its relevant frequency sub-bands. This complexity may arise from interactions among several brain control nodes that operate together at multiple time scales making the brain a system with nonlinear (chaotic) dynamics, source of signals that change amplitude randomly with respect to time [61],[62]. Nonlinear analysis allows to better understand these complex fluctuations of EEG dynamics across the various sleep stages, and results in features which may be efficiently exploited for studying sleep mechanisms. Moreover, they show good diagnostic capabilities, since may evaluate deviations from physiologic conditions highlighting specific sickness states that typically demonstrate a gradual loss of complexity, a symptom of systems' diminished adaptation to stimuli [61]. This power is demonstrated by different works that used nonlinear features in sleep-related H/C-sick classification problems, or automatic sleep staging processes [62], [63].

Among the several ways to estimate signal complexity, mainly divided into two families namely Fractal-based methods and Entropy-based methods, the more commonly used in the context of sleep EEG analysis, which also provide the best results [61] and for this reason exploited in this work, are Detrended Fluctuation Analysis (DFA, Peng et al. 1994), Higuchi Fractal Dimension (HFD, Higuchi, 1990), Sample Entropy (SampEn, Richman and Moorman, 2000), and Taiger-Kaiser Energy Operator (TKEO, Kaiser, 1990 and 1993) which instead does not belong to any of the above-mentioned family, being an energy measure [27].

Fractal-based methods rely upon two main aspects of fractal theory i.e. long-range correlation (linked to self-similarity)[11] and fractional dimensionality.

DFA (*Martin Magris (2023). Detrended fluctuation analysis (DFA), MATLAB Central File Exchange. Retrieved June 26, 2023*): The degree of long-range correlation and self-similarity can be quantified through the Detrended Fluctuation Analysis (DFA) which, by estimation of the so-called scaling exponent α, identifies trends in the signals' variance when analyzed with different time-block lengths [64]. DFA evaluates signal fluctuations across different time window lengths by computing the Root Mean Squared error (RMSe) of linear fits over successively bigger bins (which are overlapped boxes of equal size) of the detrended and integrated time series, allowing the discovery of intrinsic long-range correlations present in it (Figure 2.22).

[11] Long-range correlations refer to the property of some time-series to exhibit correlation between

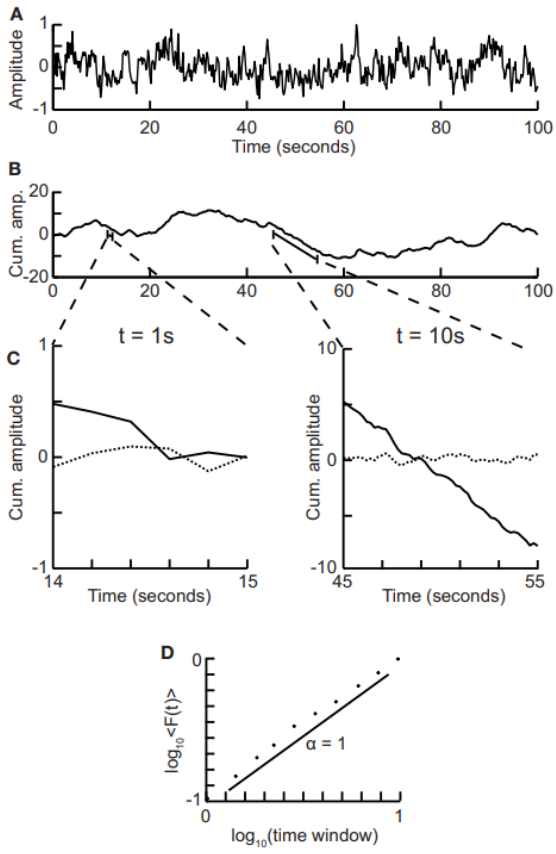


Figure 2.22: Detrended Fluctuation Analysis (DFA) [64].
(A) Time-series $x(t)$. (B) Integration of the signal (cumulative sum, which defines the signal profile $y(t)$) and removal of global trends by subtracting the signal mean. (C) $y(t)$ is divided into sub sequences of equal length n and is deprived of any linear trend to avoid spurious detection of apparent long-range correlations. Detrending the signal profile makes the robustness of DFA since reveals its true scaling properties. (D) The previous step produced $F(n)$ which is the average fluctuation as a function of the box size. $F(n)$ is then computed for different values of n and displayed in a $log(F(n)) - log(n)$ plot whose slope is the scaling exponent α.

The range of window lengths was chosen according to the procedure proposed by Hardstone et al. (2012) [64]. Besides the fact windows should be logarithmically spaced in order to give the same weight to all the time scales (in the log-log

values that do not decay (or decay slowly) as the time lag between them increases (i.e. at different time scales). On the other hand, self-similarity is a fractal property related to the presence of repeated patterns at various temporal scales. In other words, patterns or features that are present in the data at one scale also appear at other scales, and for this reason, self-similar processes are often referred to as scale-free. Both properties can be identified as scaling properties since related to characteristics of the signal that may vary when the time scale of the data is changed.

plot), the lower fitting range was chosen of at least four samples (since it has been demonstrated linear detrending with $n < 4$ produces poor results) while the higher fitting range was picked $n < 10\%$ of the signal size avoiding DFA estimates to be too noisy due to the small number of the available windows. DFA is inherently robust against non-stationarities thus $T = 30s$ is picked. The above-mentioned process results in the range $(8 - 1536)$ samples which means $(0.016 - 2s)$.

The results of DFA can be interpreted as follows [61],[64]:

- If $0 < \alpha < 0.5$ the process has-short range memory thus it not exhibits long-range correlations. This means that "large fluctuations are not followed by similar-sized fluctuations in the future".
- If $0.5 < \alpha < 1$ the process has long-range memory thus it exhibits long-range correlations. This means that "large fluctuations are more likely to be followed by similar-sized fluctuations in the future".
- If $1 < \alpha < 2$ the process is non-stationary.
- If $\alpha = 0.5$ the process is indistinguishable from white noise
- If $\alpha = 1$ the process corresponds to pink noise which exhibits perfect self-similarity.
- If $\alpha = 1.5$ the process corresponds to a Brownian noise.

In general, it is possible to state that complex time series have lower α, while non-complex signals have higher α. Yan Ma et al. (2018) [61], in their work about sleep EEG, demonstrated that healthy subjects are characterized by $\alpha > 1$ when passing from wake to NREM sleep (N3, $\alpha \approx 1.3 - 1.5$) and this values decrease during REM ($\alpha \approx 1 - 1.2$) suggesting that "the dynamics of the sleep EEG is like a Brownian noise process in deeper sleep stages", probably due to the greater neuronal synchronization they are associated with (Figure 2.26).

DFA is more efficient with respect to other classical estimation approaches like R/S analysis (Hurst exponent), being more robust against non-stationarities, not overestimating the scaling exponent, and requiring lower sample size to reduce the estimate variance, what allows it to be applied also on short portions of the signal [65].

HFD (*Jesús Monge-Álvarez (2023). Higuchi and Katz fractal dimension measures, MATLAB Central File Exchange. Retrieved June 26, 2023*): It is another member of the family of Fractal-based approaches, and focuses on the fractional dimensionality estimation (topological dimension of the attractor[12]) as a way to characterize the complexity of the system that generates the signals under study. When using such a nonlinear approach (but also SampEn relies on the same assumption) an irregular EEG waveform is frequently considered as a consequence of the

existence of a strange attractor in the state-space of an underlying system. It is then possible extracting information about this attractor directly from the signal thanks to the embedding theorem (Takens, 1981) which looks at the one-dimensional time series as the compressed information of higher dimensional space (i.e. information from the state-space) [39].

The algorithm builds partial time series from the signal $x(t)$ (N samples) by exploiting two parameters $k = 1, ..., k_{MAX}$ and $m = 1, 2, ..., k$ where the former defines the width of the time window, and the latter refers the point of the original time-series from which the time-window starts. For fixed k and m values, i-differences between k-spaced samples of $x(t)$ are computed, defining the "height" between the extrema of the time windows, and obtaining what is called mean segment length $L_m(k)$ (Figure 2.23)

$$L_m(k) = \frac{1}{k}\left(\sum_{i=1}^{\lfloor (N-m)/k \rfloor} |x(m+i*k) - x(m+(i-1)*k)|\right)\frac{N-1}{\lfloor (N-m)/k \rfloor * k} \quad (2.41)$$

where $\lfloor (N-m)/k \rfloor * k$ is a normalization factor.

These mean segment lengths are then averaged across various values of initial points m obtaining a function of k, called length of the curve $L(k)$

$$L(k) = \frac{1}{k}\left(\sum_{m=1}^{k} L(m,k)\right) \quad (2.42)$$

If $L(k)$ follows a power law of the type $L(k) \propto k^{D_f}$ then the underlying process is fractal-like with dimension D_f. In this case, the straight line fitting $log(L(k))$ as a function of the time delay $log(k)$ is computed and its slope is the HFD [33].

As well as indirectly representing the dimension of the system attractor, HFD can be seen as a measure of the self-similarity of the underlying process, since trying to quantify its property to exhibit pattern at different time scales.

HFD is pretty sensitive to the values of k and N. The parameter k_{MAX} should be estimated from data, and in its original work Higuchi (1989) suggested $k_{MAX} = 2^{11}$ (for $N = 2^{17}$) as the best value, since it performed better when applied on EEG analysis. Starting from the fact there is no general consensus on the choice of such

[12]A dynamical system is characterized by its state (m-dimensional vector in the phase, or state space) and its dynamics (set of equations describing how the state changes over time i.e. defining the system trajectories). Trajectories in some cases can converge to a bounded subspace of the state space called attractor since "attract" them from their initial conditions. Based on the dimension of the resultant geometric object, attractors are grouped into various categories (0-D steady state, 1-D limit cycle, 2-D limit torus, and non-integer dimension strange attractors which are hence fractal objects). Attractors analysis can be useful for characterizing system complexity and stability.

a parameter, Wanliss et al. [66] proposed an efficient and reliable approach for estimating it through the usage of an empirical relationship involving the length of the time series, obtaining $k_{MAX} = 28$ (with $N = 1024$ samples).

$$k_{MAX} = A_1 * sin(B_1 * N + C_1) + A_2 * sin(B_2 * N + C_2) \tag{2.43}$$

Table 2.9: Fitting parameters for k_{MAX} estimation [66]

Parameter	Value
A_1	129.8 ± 3.0
B_1	$(1.292 \pm 0.045) * 10^{-5}$
C_1	0.04488 ± 0.0255
A_2	18.82 ± 2.56
B_2	$(6.488 \pm 0.280) * 10^{-5}$
C_2	1.332 ± 0.220

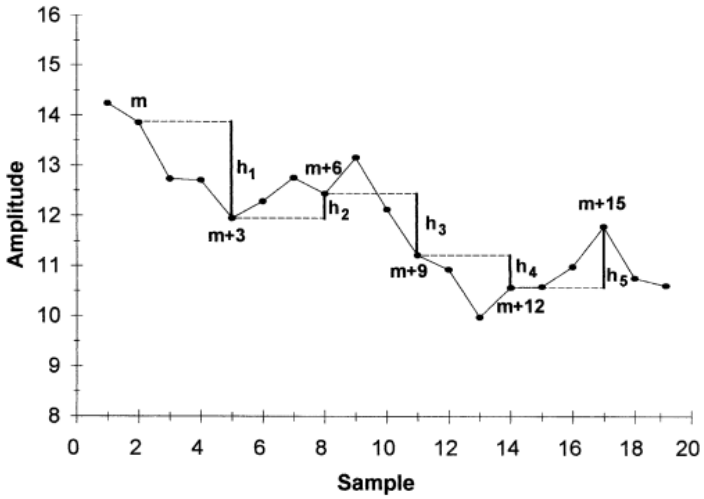


Figure 2.23: Higuchi Fractal Dimension (HFD) [67].
In this graph the name h_i is used to indicate $L_m(k)$.

HFD was used in several studies concerning brain pathologies (like Parkinson's Disease and Epilepsy) [67], in classification problems involving Healthy subjects and depressive patients [68], and for the analysis of other pathological conditions [62] showing as one of the most promising nonlinear features. In sleep-related

studies, it is revealed a complementary behavior with respect to DFA (Figure 2.26), with higher values during REM and lower during N3 [61]. In their work, Accardo et al. (1997) [67] suggested that the lower values of HDF (lower complexity) for Slow Wave Sleep may be due to the higher neural synchronization of this stage, which produces more synergism and less chaos. Similar observations can be found in the work of Yan Ma et al. (2018) [61] which state that a lower fractal dimension (less complexity) is associated with sleep-deprived subjects while higher HFD values are typical of narcoleptic patients with respect to control ones.

SampEn (*Flood, Matthew W., and Bernd Grimm, EntropyHub (2021)*): the other family of nonlinear features comprises Entropy-based methods (called embedding entropies to distinguish them from the spectral entropies computed in Section 2.2.2). Such measures do not result in the computation of the exponent of a power law distribution but in the characterization of the time-series self-similarity properties by assessing the uncertainty regarding its information source offering a measure of the system's complexity, randomness and irregularity [61]. Entropy measurements are based on the assumption that time series with repeated elements are associated with ordered systems (low entropy values), whereas more irregular patterns (arising from less ordered systems) are related to high entropy values [69]. Literature is full of entropy measures, but one of the most effective and simple is the Sample Entropy (SampEn) which thus was implemented in this work.

SampEn quantifies time-series complexity and self-similarity by computing the conditional probability that two sequences of the original signal (template vectors) of a given length m, similar for m data points, remain arbitrarily similar within tolerance r at the $(m+1)^{th}$ point (Figure 2.24).

Considering the following template vector definition

$$u(i) = [x(i) \quad x(i+1) \quad ... \quad x(i+m-1)]^T \tag{2.44}$$

and the conditional probability (self-similarity measure) of the pair $u(i)$ and $u(j)$ as

$$\Phi^m(r) = \sum_{j=0, j\neq i}^{N-m} \sum_{i=0}^{N-m} \mathcal{H}(r - ||u(i) - u(j)||_\infty) \tag{2.45}$$

where $\mathcal{H}(x)$ is the Heaviside function (1 if $x \geq 0$, 0 otherwise), and $||x||_\infty$ is the infinity norm ($max_i|x_i|$), the SampEn can be computed as [33], [70]

$$SampEn(m, r) = -ln(\frac{\Phi^{m+1}(r)}{\Phi^m(r)}) = -ln(\frac{\sum_i A_i}{\sum_i B_i}) = -ln(\frac{\mathbf{A}}{\mathbf{B}}) \tag{2.46}$$

This process is performed for all the possible template vectors pairs $u(i)$ and $u(j)$ and the results are then averaged. If the time series is perfectly ordered, templates that are similar for m points stay similar also for $m+1$ points (i.e. $A = B$) and thus $SampEn = -log(1) = 0$.

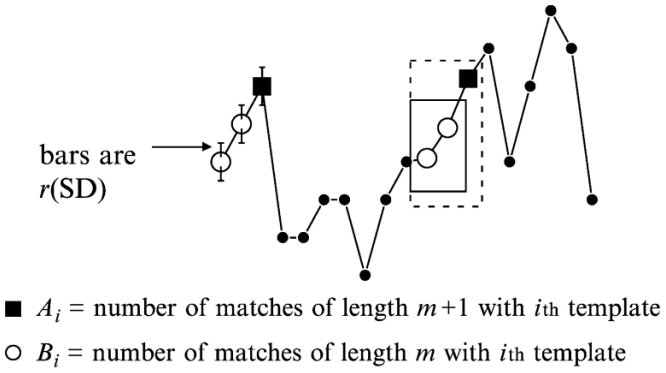


Figure 2.24: Sample Entropy (SampEn) [70].
The template length is $m = 2$, the tolerance for accepting matches is r times the standard deviation (error bars), and the time series begins with the i^{th} template. The template is matched both by the 11^{th} and 12^{th} points (represented by the solid box), and by the $(m+1)^{th}$ points (dashed box), thus B_i and A_i are increased by one.

SampEn shows advantages with respect to other Entropy-based methods like ApEn [54], [71], reason why it was chosen for the current analysis. In fact, SampEn is less sensitive to the choice of the parameters m, and r, does not consider self-matches (producing a lower bias) [70] works well even with short time series, and is robust against noise, producing lower results.

SampEn relies on the choice of three parameters:

- N, the length of the time series. It has not great impact on the final estimate and can be chosen just to satisfy the stationarity condition of the EEG.
- m, the length of the template vectors.
- r, the threshold of similarity between template vectors.

There are no established guidelines for selecting these parameters but most of the works present in literature use $m = 2$, and $r = 0.1-0.2$ times the standard deviation of the original time-series [54], [56], [68]. The estimation method proposed by Castiglioni et al. (2013) [69] was applied in this work, together with the suggestion of Yentes et al. (2013) [71] to avoid (m, r) couples that cause abrupt changes in SampEn results (Figure 2.25).

The estimation process relies upon the idea that escaping vectors (i.e. template vectors that are similar for m points but not for the $m + 1^{th}$) are those effectively contributes to the estimation of SampEn, and that the number of escaping vectors $E(m, r) \propto B - A$ is linked the choice of r:

(a) If r is too small, few vectors have neighbours, so few vectors escape from a neighborhood even if m increase to $m + 1$, and thus $E(m, r)$ approaches 0.

(b) If r increases, the tolerance is too high and then almost all vectors remain similar even if m increases to $m+1$ (there are few escape vectors), and also in this case $E(m,r)$ is close to 0.

The goal is thus to find r such that $E(m,r)$ is maximum i.e. to find r in order to work between two extrema situations (a) and (b). The analysis was performed with $r \in [0.01, 5]$ and $m \in [1, 11]$. The obtained values, $m = 2$ and $r = 0.26 * std(x(t))$, are also coherent with the above-mentioned literature.

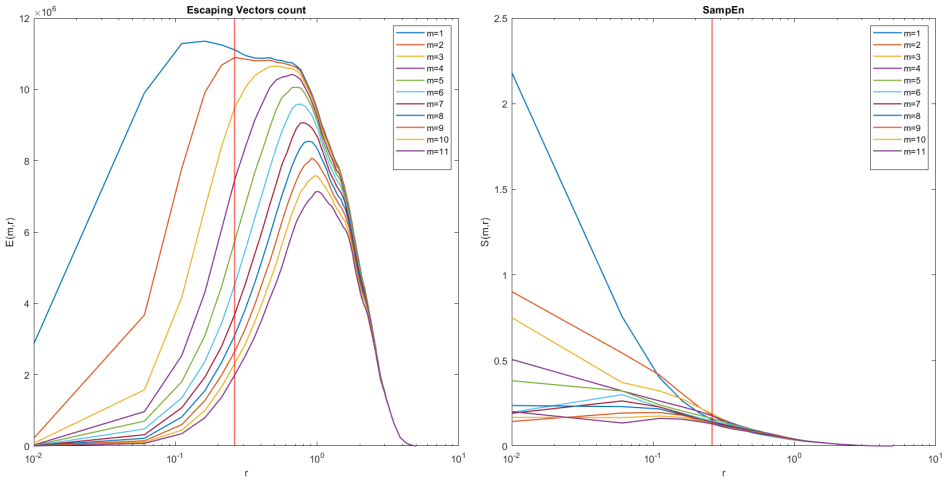


Figure 2.25: E(m,r) curves and SampEn(m,r) curves.
(right) $E(m,r)$ curves, (left) $SampEn(m,r)$ curves for different values of m and r.

Studies computed on different categories of subjects such as adults, children and new-borns [61] reveal that (like HFD) the Entropy measures of physiological EEG signals gradually decrease going from Wake to N1-N3 stages and increase in REM, suggesting that during NREM the brain activity is more regular and coherent with respect to Wake and REM sleep (Figure 2.26).

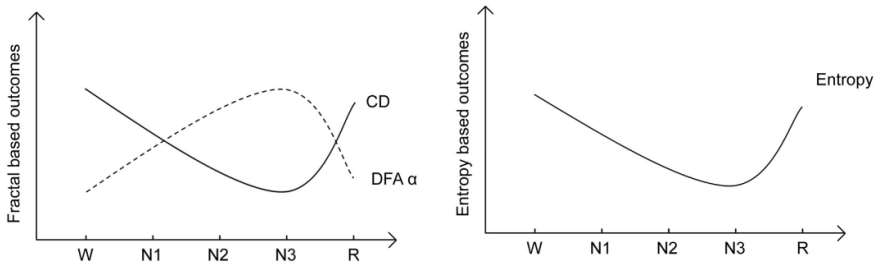


Figure 2.26: Relationship between Nonlinear features and sleep stages [61].
These findings support the prevailing idea that NREM sleep phases entail more organized, coordinated, and regular, or less-involved, brain cell activity. CD stands for Correlation dimension and behaves like HFD.

TKEO (*Hooman Sedghamiz (2023). Teager Keiser Energy Operator Vectorized, MATLAB Central File Exchange. Retrieved June 26, 2023*): Typically, instantaneous energy measures used in signal processing just include amplitude information. However, it has been demonstrated that measures that additionally include frequency details do a better job in assessing the energy required by a system to generate a signal, resulting in a more sensitive analysis. This job is accomplished via the Teiger-Kaiser Energy Operator (TKEO)[13], a frequency-weighted metric often used in EEG research since characterizes the system's "true" energy. Being based on the second derivative of the time series, it is capable to detect instantaneous signal changes such as discontinuities and amplitude and frequency alteration, allowing to better and more finely characterize the energy content of a system.

For discrete signals TKEO is defined as

$$\Psi\left[x(t)\right] = x(t)^2 - x(t+1)x(t-1) \tag{2.47}$$

Some works highlighted that post-processing TKEO can increase the performances in detection problems, bringing to similar results with respect to other methods, while keeping the algorithm simple [72], [73]. For this reason after energy was estimated, the absolute value of TKEO was considered (Figure 2.27) and then statistical measures like Mean, Standard Deviation, Skewness, Kurtosis, and Max were computed, in order to reduce the dimensionality of the obtained objects.

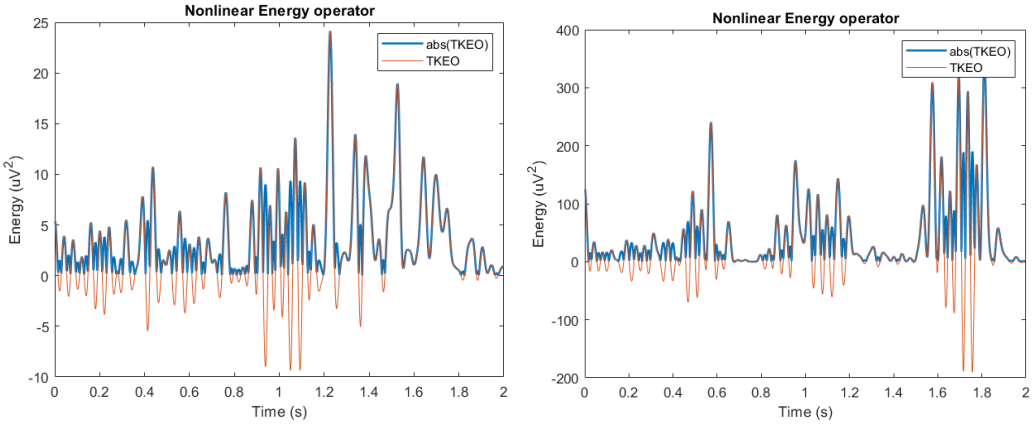


Figure 2.27: Teiger-Kaiser Energy Operator.
TKEO computed for the first 2s mini-epoch of a SWS segment, patient n1 (top) and rbd1 (bottom), channel C4-A1.

[13]For a signal of the type $x(t) = Acos(\omega_0)t + \phi$, the TKEO outcome is proportional to both amplitude and frequency, $\Psi\left[x(t)\right] = A^2\omega_0^2$.

Table 2.10: Electroencephalographic features

Domain	Features
Time	Mean, Mode, Median (50^{th}-p), Standard Deviation, Skewness, Kurtosis, Maximum and Minimum values, Range, Zero-crossing rate, Hjorth Parameters (Activity, Mobility and Complexity), Hjorth-based (Sparseness, Irregularity, Spectrum Bandwidth, Average-Zero-up, Crossing period, Average-peak-to-peak period), Percentiles (25^{th}-p, 75^{th}-p), Interquantile range, Form Factor, Crest Factor, Impact Factor, Coastline, Coefficient Variation (std/mean), Energy of the signal, Envelope-based (Number of peaks, Peaks Prominence, Peak Width), Maximum-minimum-distance, EnergySis.
Frequency (total)	Absolute Power Spectrum, Mean Power Spectrum, Peak Frequency, Mean Frequency (Spectral Centroid), Median Frequency (SEF50), SEF25, SEF75, SEF95, SEFd75-25, SEFd95-25, SEFd95-50, Spectral Crest, Spectral Entropy, Renyi's quadratic Entropy.
Frequency (sub-bands)	Max Value, Relative Power Spectrum, Mean Power Spectrum, Absolute Power Spectrum, Spectral Centroid, Spectral Spread, Spectral Skeweness, Spectral Kurtosis, Variance of Central Frequency, Peak Frequency, Median Frequency (SEF50), SEF95, SEFd95-50, Bandpower-ratios, θ/α, β/α, $(\theta+\alpha)/\beta$, θ/β $(\theta+\alpha)/(\alpha+\beta)$, TREM/FREM, SWA/SOs, Spectral Entropy, Renyi's quadratic Entropy.
Time-Frequency	6 levels Discrete Wavelet Analysis (DWT), db4 (Mean, Standard Deviation, Coastline, Ratio of absolute mean values of adjacent sub-bands).
Nonlinear	Detrended Fluctuation Analysis (DFA), Higuchi Fractal Dimension (HFD), Sample Entropy, Taiger-Kaiser Energy Operator (TKEO) (Mean, Standard Deviation, Skeweness, Kurtosis, Maximum value).

After the feature extraction process, for each variable, the values corresponding to the 30s macro-epochs (for TD) or to the 2s mini-epochs (for the other domains) were collected into a feature array, and then statistics such as Mean, Standard deviation, 75^{th}-p, and Main Mode were computed, to sum up all the information in just a value that catches salient peculiarities about the underlying distributions of the data (Figure 2.28), obtaining 883 features. This process was then repeated for all the patients and segment types, forming three features matrices $\mathbf{U_k}$, $k = 1,2,3$ (with features columns $\mathbf{u_i}$ and patients rows $\mathbf{s_h}$).

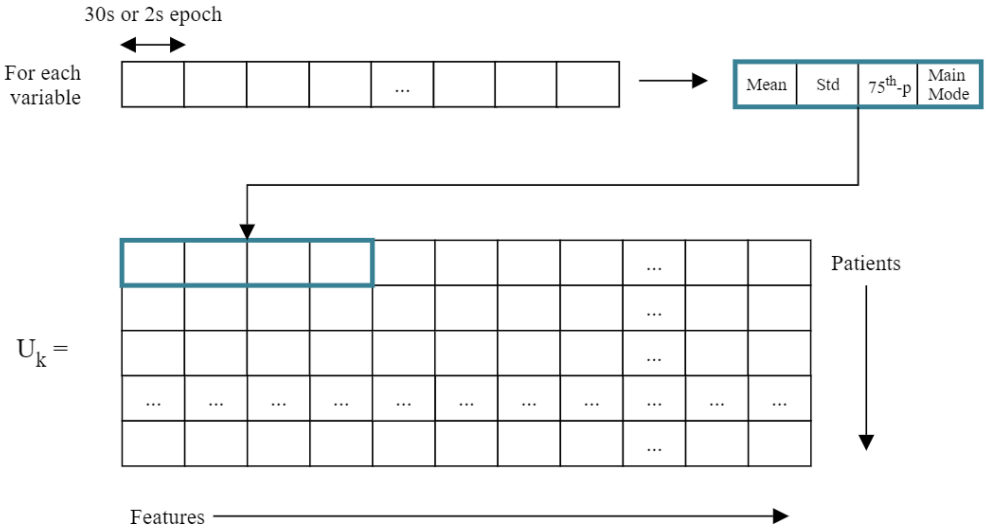


Figure 2.28: Features matrices structure.

At the end of this stage, three feature sets were obtained (Table 2.11).

Table 2.11: Feature Sets employed in this work

FeatSet	Description
$FeatSet_1$	Polysomnographic + REM Sleep Features
$FeatSet_2$	Polysomnographic + SWS Sleep Features
$FeatSet_3$	Polysomnographic + (REM+SWS) Sleep Features

2.3 Binary Classification: H/C and RBD

This section comprises all the steps necessary for implementing an automatic system that through the help of supervised ML models can exploit the available

patients labels to build classification rules that discriminate between healthy and RBD subjects. In this part of the work, H/C and RBD patients from both CAP and TuSDi databases were employed, resulting in three matrices $U_k \in \mathbb{R}^{(58,883)}$ as inputs, and a target vector $\mathbf{y} \in \mathbb{R}^{(58,1)}$ that contains the labels (0 and 1 for H/C and RBD respectively). These $FeatSet_i$ were exploited in a pipeline (Python environment) that aims to standardize the data, reduce the dimensionality of the dataset, optimize the hyperparameters of the ML models used for the analysis, and evaluate the performances through the use of standard supervised ML metrics. All these steps were performed in a Leave-one-out cross-validation framework to overcome the limited number of available patients and decrease the risk of data leakage.

2.3.1 Data Scaling

All the analyzed features come from various domains that are difficult to be compared and have heavily different order of magnitudes that can create imbalanced situations in the models tuning part. Some ML models (e.g. Support Vector Machines, and K-Nearest Neighbors) do not work well with such data since they use distance metrics in the definition of their classification rules. Distance-based algorithms assume in fact, that larger values should be given greater weights bringing them to be more significant with respect to smaller values, and hence playing a more decisive role during the training phase. To overcome this bias source, the data were standardized in such a way each sample contributes approximately proportionally to the final distance, in some cases improving the classification performances and the convergence properties of the algorithms.

There are two big groups of data scaling approaches namely Normalization and Standardization. The latter has different advantages compared to the former e.g. it allows for the presence of outliers more than the normalization family does, making the algorithms less sensitive to anomalies, and it is more robust on new data since relies upon statistical (i.e. global) features rather than local ones (like min and Max). Since features along all the samples (patients) follow a normal distribution, Standardization (also called Z-score normalization) was used (*StandardScaler* in Python framework).

It works independently on each feature along patients, performing the following steps

1. Centering the data by removing the mean.
2. Scaling the data by dividing for the standard deviation.

$$u_{stnd}^{(i)} = \frac{u^{(i)} - \mu_u}{\sigma_u} \tag{2.48}$$

where μ_u is the mean of a certain column of features $\mathbf{u_i}$, and σ_u is the corresponding standard deviation ($u^{(i)}$ is the generic element of the vector $\mathbf{u_i}$). The resulting scaled data follow a normal distribution and have $mean = 0$ and $std = 1$ (not limited to a certain range).

The effect of such procedure on data can be appreciated in Figure 2.29 where two features and their scaled versions were displayed on a 2-dimensional scatter plot. It can be noticed that the relative position of each data point is maintained as far as the presence of outliers, keeping the relationship between features unaffected.

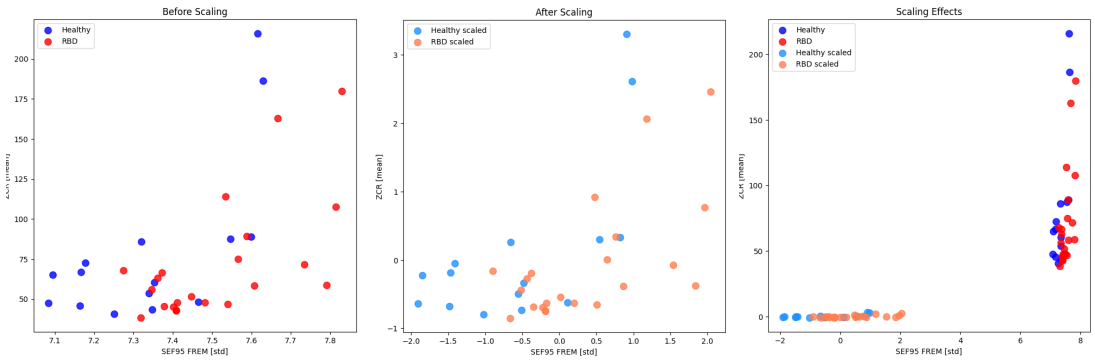


Figure 2.29: Data Scaling effects.
The range of the original features passed from $(7.1, 7.8)$ and $(25.0, 225.0)$ to $(-2.0, 2.0)$ and $(-1.3, 2)$ for the first and the second features respectively.

2.3.2 Feature Selection

Working with a big number of features can increase the complexity of the ML models, rising the chance of overfitting[14] [74]. Selecting a subset S of the overall dataset U with $dim(S) < dim(U)$ through feature selection strategies becomes thus a fundamental step in every classification framework, reducing the computational cost, attenuating the noise brought by non-necessary features, improving the accuracy thanks the more understandable and clear relationship between features and target, and speeding up the training procedure.

[14] A supervised learning pipeline consists in building a model on a portion of the original data (training set) and making predictions on new unseen data (test set). When the model can reach good classification performances on the test set it is said it has good generalization capabilities since it is able to build a rule which is pretty accurate but general enough to fit well with data it does not know. To build such a model, a trade-off between model complexity and performance should be performed, finding a sweet spot between the following situations

- Overfitting. Too complex models (high variance) are too adherent to the training set, performing well on it but failing in generalizing on new data. This condition causes high train accuracy but low test accuracy.

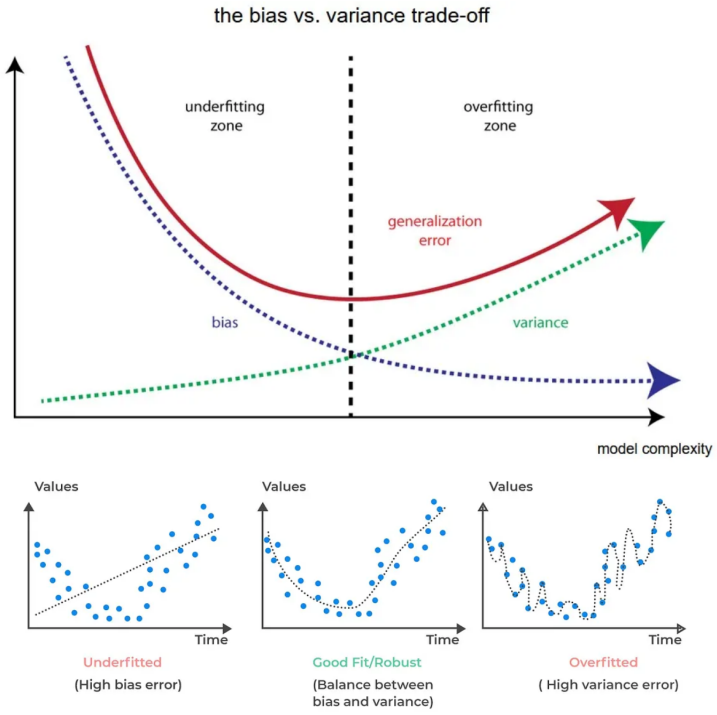


Figure 2.30: Complexity and performances trade-off.
https://towardsdatascience.com, *https://medium.com*

A lot of methods exist, but filter methods were chosen thanks to their speed, computational lightness, and independence from any kind of ML model reducing any potential overfitting problem. Moreover, they shine when the number of features in the original dataset is massive. The key idea they rely upon is to capture the discriminating capability of each feature by statistically quantifying their relationship with the target variable (univariate feature selection) and assigning a score used in heuristic fashion to rank them. Starting from the evidence that "the m best features are not the best m features" [75] Minimum Redundancy Maximum Relevance (mRMR) method was exploited. This approach condenses as much information as possible in a small number of significant features overcoming the

- Underfitting. Too simple models (high bias) are expected to generalize better on new unseen data even if they may fail in catching the variability and the subtle relationship details of the features, being in some sense incomplete and not precise. This condition causes both low train and test accuracy.

Complexity is intimately related to the dimension of the input training dataset (a higher number of samples allow the model to be more complex without overfitting) and with the size of each data point (a higher number of features increase the model complexity bringing to overfitting).

main limitation of a quite large number of basic feature selection methods i.e. they do not account for redundancy of features, which may bring to:

- The poor efficiency through which the resources (features) are used.
- The limited generalization abilities of the resulting feature set. If the features are chosen only on the basis of their individual discriminative power, the resultant subset is not maximally representative of the original space made up of the total number of features.

Being a multivariate approach, mRMR uses more than one statistic, for both establishing the link between features and target as well as taking into consideration the relationship among the features. It belongs to the class of methods that seek the minimal-optimal set, which is defined as the set of all relevant features from which redundant and hence unneeded features are deleted, and having the smallest possible dimension while keeping the highest possible predictive ability [76], [77].

It works in an iterative way choosing at each step the best feature and adding it to the previously selected ones following two criteria, namely Max-Relevance and Min-Redundancy [75].

1. Relevance of each feature with respect to the target. It must be maximized to guarantee the selected features carry important discriminant information (relationship features-target)

$$max \sum_{u_i \in S} I(u_i; y) \tag{2.49}$$

   where y is the target vector, and u_i is the i^{th} feature column.

2. Redundancy with respect to features selected in previous steps. It must be minimized to discard all the irrelevant features (relationship feature-feature)

$$min \sum_{u_i, u_j \in S} I(u_i; u_j) \tag{2.50}$$

   where u_i and u_i are two different feature columns.

$I(a; b)$ represents the Mutual Information[15] (MI) which captures any sort of connection between two random variables. Hence, the correspondent feature-selector belongs to the sub-group of Mutual Information-based Feature Selection (MIFS) methods (implemented in Python through the package `https://github.com/danielhomola/mifs`).

These criteria can be merged in the form of incremental search (mutual information difference, MID)

$$max_{u_j \in U - S_{m-1}} [I(u_j; y) - \frac{1}{m-1} \sum_{u_i \in S_{m-1}} I(u_j; u_i)] \tag{2.51}$$

where m is the number of selected features from the original feature space.

In Equation (2.51) MI between each taken-individually feature and the target is computed, and the feature with the highest MI is placed in the subset S. In each consecutive step a greedy search is performed, picking the features that have the minimum MI with the already selected ones, and the maximum MI with the target.

In their research works, Hanchuan Peng et al. (2005) [75], Ding et al. (2005) [76], and Z. Zhao at al. (2019) [77] proven that, given the same amount m of features, mRMR makes the resultant subset more representative of the target, allowing to use a smaller number of features to effectively cover the same space as the original entire dataset does. Moreover, compared to univariate methods, mRMR improves the performances in terms of classification error, and does not show any kind of model dependency.

After accomplishing the mRMR procedure (with $m = 30$), the following steps and heuristic criteria were performed, ending up with an appropriate small subset of features:

1. For each $FeatSet_i$, the mRMR scores (which are values of $I(u_i; y)$) are normalized to be more comparable and easier to interpret from a numerical point of view.

2. A minimum number of $m_{min} = 5$ is chosen, according to the fact that for all the $FeatSet_i$, the first features carry the majority of MI with the target.

3. When a drastic drop in the value of MI was encountered the m searching was stopped (since the range of the values, also a drop of $0.1 - 0.2$ was considered sufficient).

4. A couple of features were added to include some additional information, remaining above the fixed limit of $m_{MAX} = 10$, necessary for keeping the feature interpretability as high as possible. This method helped to increase the performances and also to have the same dimension for all the $FeatSet_i$.

This process resulted in $m = 8$ top features for each segment type, bringing to three updated $FeatSet_i \in \mathbb{R}^{(58,8)}$ (Table 2.12).

[15]According to information theory, MI derives from Entropy $H(X)$, and conditional Entropy $H(X|Y)$. The former quantifies the uncertainty present in a distribution of the variable X, the latter defines the degree of uncertainty remaining in X after Y is known. Hence MI between two variables X and Y can be computed as the intersection of the above-mentioned sets, representing the amount of uncertainty in X that is removed after Y is known or in other words the amount of shared information between the two data sets, $I(X;Y) = H(X) - H(X|Y)$.

Table 2.12: Updated $FeatSet_i$ with the top-8 ranked features

$FeatSet_1$	$FeatSet_2$	$FeatSet_3$
SEFd(95-50) θ [75^{th}-p]	SCe FREM [Mean]	PKF FREM [Mean]
SK SOs [75^{th}-p]	SSp β [Std]	SEF95 β [Std]
Ssp θ [75^{th}-p]	PP (envelope) [75^{th}-p]	IQR [Std]
$(\theta+\alpha)/(\alpha+\beta)$ [75^{th}-p]	REN2 SOs [75^{th}-p]	SSp θ [Std]
SEFd(95-50) SWA [Mean]	SEFd(95-50) SWA [75^{th}-p]	SEN β [Mean]
SK FREM [75^{th}-p]	SEFd(95-50) β [Mean]	SEFd(95-50) SOs [Std]
Sce Gamma [Std]	PW [Std]	SEF(95-50) FREM [75^{th}-p]
WP	REN2 β [Std]	VCF SWA [75^{th}-p]

Figures 2.31, 2.32, and 2.32 graphically display the feature selection procedure, with a distinction between the discarded attributes (light blue), the ones chosen through a score threshold (blue) and the additional ones (pink).

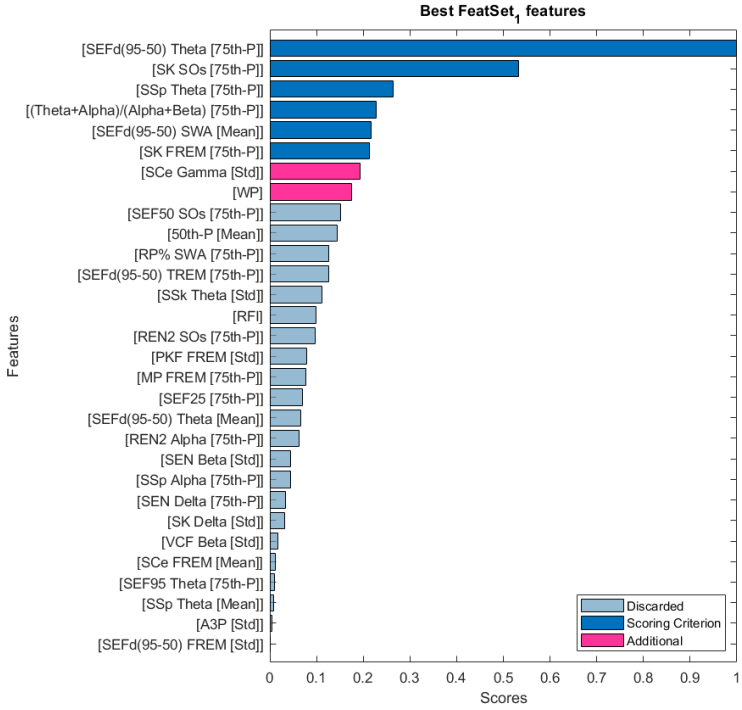


Figure 2.31: Selected Features $FeatSet_1$.

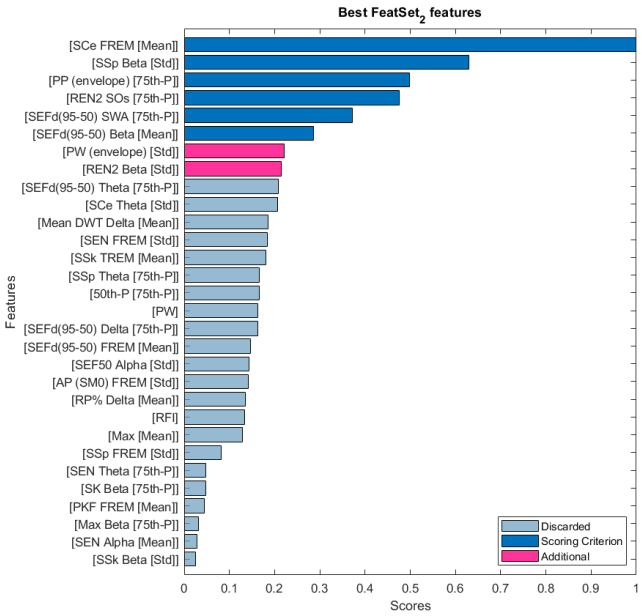


Figure 2.32: Selected Features $FeatSet_2$.

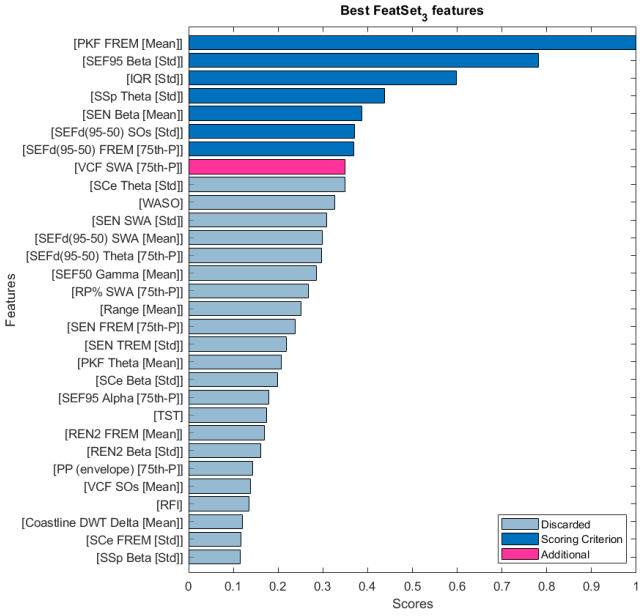


Figure 2.33: Selected Features $FeatSet_3$.

2.3.3 Machine Learning models

The new $FeatSet_i$ were used to train and test eight different supervised binary classification algorithms (sci-kit learn Python library `https://scikit-learn.org/stable/`) whose results were exploited to highlight the segment type that best discriminates between H/C and RBD subjects.

K-Nearest Neighbors (K-NN): It is a non-parametric learning method that classifies data points by computing their distances with respect to near training points (neighbors). According to these distances, a weight is assigned to each observation, and the closest K-top observations are then selected. Any new sample is then classified as the most recurrent label within its neighbors. One of the KNN hyperparameters is the number of neighbors K whose values change the complexity of the resultant model (Figure 2.34). If K is small the decision boundary follows too closely the training data, increasing the complexity and the chance of overfitting. On the other hand, as K increases the boundary becomes smoother until the extrema case in which it equals the number of samples in the dataset (the new point is classified based on the most recurrent label in the whole dataset). The optimal is between the two conditions [74].

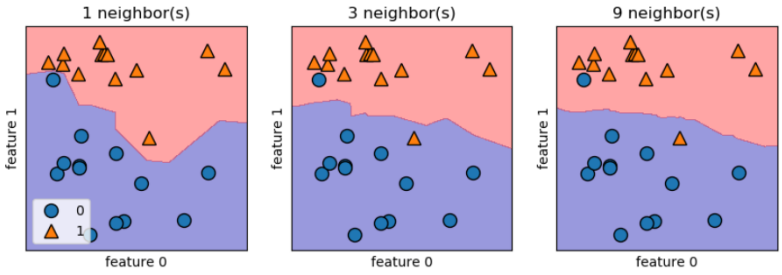


Figure 2.34: Relationship between K and model complexity.
https://medium.com

Another hyperparameter is the distance definition, according to the Table 2.13 (a_i and b_i are the coordinates of the data samples involved in the computation i.e. the patients).

Table 2.13: KNN distances

Distance	Definition
Minkowski	$(\sum_i \lvert a_i - b_i \rvert^p)^{1/p} \quad p \neq 1,2$
Euclidean	$(\sum_i \lvert a_i - b_i \rvert^p)^{1/2}$
Manhattan	$\sum_i \lvert a_i - b_i \rvert$
Chebyshev	$max_i(\lvert a_i - b_i \rvert)$

Support Vector Machines (SVM): it exploits $(m-1)$-dimensional space hyperplanes (where m is the dimension of the feature space) and seeks for the one which optimally separates the features in such a way data points falling on different sides of the hyperplane are attributed to different classes. More specifically SVM seeks the hyperplane with the maximum margins which are the distances between the hyperplane and the so-called support vector points (data points of both classes which are closer to the hyperplane and in some sense the trickiest to be classified, thus basing a decision rule on these points allows to build a robust algorithm). The result is said Maximum Margin classifier. When data are not strictly linearly separable it is possible to relax the problem, creating what is called a Soft Margin Classifier, which allows for misclassifications and accepts points to be inside the margin. The hyperparameter of SVM is the term C which tunes the strength of the above-mentioned regularization introducing a penalty term proportional to its value (Figure 2.35).

- For small C the regularization is heavy, and the decision boundary will allow for misclassifications by imposing big margins and thus making the algorithm more robust and generalizable. Increases the chances of underfitting.
- For big C the regularization is soft and the SVM will try to fit the training data as best as possible imposing smaller margins and thus not allowing misclassified points. Increases the chances of overfitting.

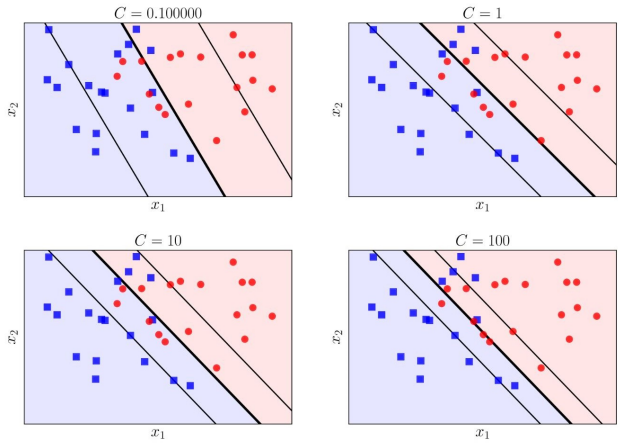


Figure 2.35: Relationship between C and model complexity.
https://dinhanhthi.com/support-vector-machine/

Kernelized Support Vector Machines (k-SVM): There are situations in which samples cannot be linearly separated at all. In these cases may be convenient to bring the data into a higher-dimensional space in which they are linearly separable.

This is done using nonlinear functions of similarity called kernels (Table 2.14), which look for higher-dimensional relationships between the variables, without actually mathematically transforming them (kernel trick).

Table 2.14: SVM kernels

Kernel	Definition
Linear	$\Phi(u_i, u_j) = u_i^T u_j$
Polynomial n-degree	$\Phi(u_i, u_j) = (r + \gamma u_i^T u_j)^d \quad d > 1$
Radial Basis Function (RBF)	$\Phi(u_i, u_j) = exp(-\gamma \|\|u_i - u_j\|\|_2^2)$
Sigmoid	$\Phi(u_i, u_j) = tanh(\gamma u_i^T u_j + r)$

The main two hyperparameters of k-SVM are C (as in the linear SVM) and γ which tunes the kernel action. It determines the number of data points to be taken into consideration for building the hyperplane by scaling the influence each sample has on others. In RBF (Figure 2.36) for high γ the radius of the kernel is small and only points that are strongly near to each other will be considered similar, looking at the similarity concept in a strict sense, and significantly fitting the decision boundary to the training data (high chance of overfitting). In other words, a new observation can be classified in a group only if it is strongly near to the points of that class.

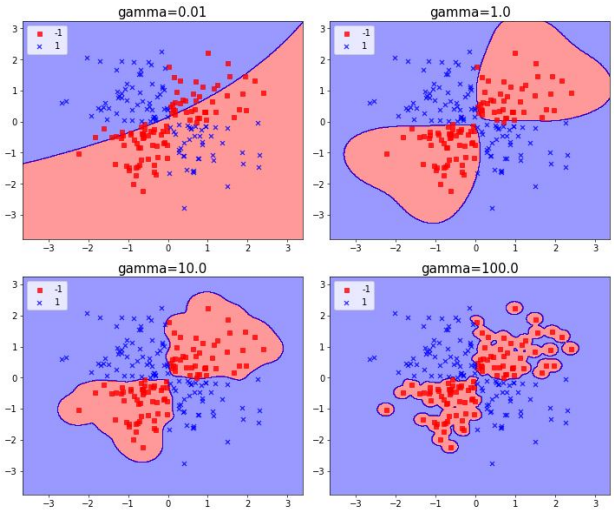


Figure 2.36: Relationship between γ and model complexity.
https://dinhanhthi.com/support-vector-machine/

Naive Bayes Classifiers (NB): It is a probabilistic classifier based on the Naive Bayes Theorem which defines the change in the probability an event happens after

some data evidence is known. This algorithm relies on an optimization problem that seeks to find the outcome y is more probable to be obtained from the model scheme represented by the dataset (features). In practice, the classifier computes the probability distribution of each feature individually (along the training samples of the same class) and makes predictions by comparing the new data points with the statistics of these distributions (in the case of Gaussian NB, the statistics are mean and std) predicting the best matching class. The main disadvantage of this classifier is that it assumes that the features are independent, which is not properly true as demonstrated in Section 2.3.2.

Decision Trees (DT): This supervised method uses a top-down approach to progressively divide the feature space into small groups based on descriptive conditions until at the end each group contains data that predict in a non-ambiguous way the label. In its standard form, DT are made up of nodes (each node represents a feature), branches (representing the decision rule), and leaves (representing the outcome of the decision) which form layers. At each layer, the algorithm splits the feature set in such a way data points falling into the same group are most similar to each other and creating groups that are as different as possible from each other following a within-homogeneity criterion (Figure 2.37).

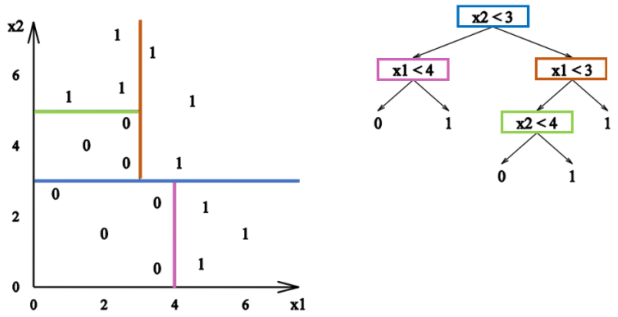


Figure 2.37: Binary bi-dimensional example of DT splitting.
In a bi-dimensional case, the tree splits the feature space into areas in which data points are grouped on the basis of characteristics reflecting the class they belong to. Each node is used as a test condition and each branch as one of the possible answer to that question.

As the number of splits increases, the complexity of the model increases as well as the chance of overfitting. Two common strategies are used to avoid this condition and to keep the DT simple, namely pre-pruning and post-pruning. The first aims to stop the creation of the tree early (e.g. by limiting the maximum depth of the tree or limiting the maximum number of leaves) while the second waits for the completed building and then removes the nodes that contain little information. As all the regularization techniques they decrease the training performances since a smaller tree is like an approximation of the completed one, but bring big advantages

as far as the testing performances are concerned.

In this work, just pre-pruning techniques were employed, and the related limit values (i.e. max depth and max number of leaves) were tuned as hyperparameters. Another hyperparameter is related to the splitting criterion. Among the various possibilities just Gini impurity and Information gain were taken into consideration since they are the most commonly used and considered the most reliable approaches.

Gini impurity is a measure of non-homogeneity and quantifies how fine a split is on the basis of how mixed the classes are within the two groups that originated from the split. In a binary classification problem, it is defined as

$$Gini\ index = p_1(1 - p_1) + p_2(1 - p_2) \tag{2.52}$$

where p_i is the probability of class i. When the classes are equally distributed in the after-split group, $p_1 = p_2 = 0.5$ thus $Gini\ index = 0.5$. On the other hand, if the after-split group contains just one class, one of the two probabilities is 0 and thus $Gini\ index = 0$.

Information Gain derives from Entropy (H) which measures the amount of randomness in a dataset. Perfectly homogeneous groups have $H(y) = 0$, while groups that are perfectly divided into two classes have $H(y) = 1$.

$$IG(y, A) = H(y) - H(y|A) \tag{2.53}$$

If $H(y|A)$ is the homogeneity in the after-split group after knowing the attribute A, IG quantifies the expected entropy reduction caused by splitting according to this feature.

In both cases, the features which are associated to the biggest after-split homogeneity are chosen as nodes for the i^{th} split. The main disadvantage of DT is related to their high variance since they are prone to overfit and their intrinsic instability since an error in the top splits propagates down to all the next steps.

Random Forests (RF): It belongs to the family of ensemble methods since based on the combination of different DT, obtaining performances are better than the ones achieved by each single weak model. They rely on the concept of Bootstrap aggregation (BAG) which creates parallel training procedures, proposing for each rail a random subset of the original data, sampled by means of bootstapping approach[16]. Each collection of data is then used to train a DT ending up with an ensemble of different classifiers. The final prediction is performed as majority voting of the single predictions. (Figure 2.38). Moreover, RF considers random subset of features during the splitting process, and not the overall feature set as DT does.

[16]Bootstrap allows to draw new datasets with the same size of the original one, by randomly sampling the original dataset with replacement (so some points can miss and other can be repeated).

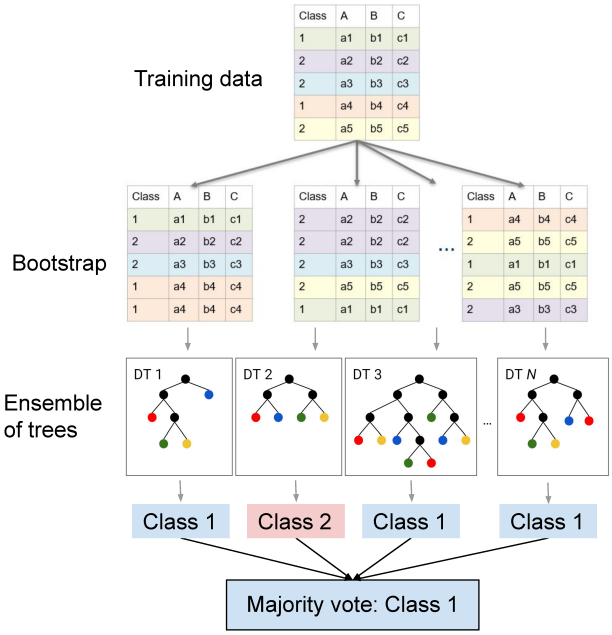


Figure 2.38: Random Forest scheme.

The most important RF hyperparameters to be tuned are the same as in the DT together with:

- The number of features to be randomly selected from the original space. Good rules of thumbs are $sqrt(n.features)$ and $log_2(n.features)$.

- The number of independent estimators to be trained simultaneously. The higher the parameter, the lower the variance.

Boosting Techniques: These strategies work sequentially on different models M_i, training the i^{th} models on a subset of observations that are randomly selected from the original dataset, and then building the $(i+1)^{th}$ by rectifying the mistakes of the previous one. Each observation is associated with a weight that at first is equal for all of them, meaning that at the beginning all the samples have the same probability to be chosen for the training. Before feeding $M(i+1)$, the dataset is updated in such a way that misclassified points weights are increased and correctly classified points weights are decreased. In this way the probability to select a misclassified observation for building the next training set is higher, making $M(i+1)$ more focused on mistakes of the previous steps. This process (Figure 2.39) is repeated for each weak model until some stop condition is verified, and the dataset is predicted as best as possible. All these models create an ensemble such

that when new data are passed, each model gives its classification "opinion" and at the end, the class with the majority vote is chosen.

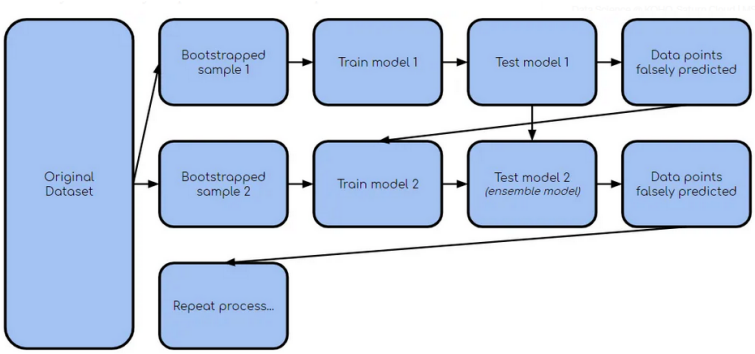


Figure 2.39: Boosting technique scheme.

The core of boosting techniques can be enriched by additional details, defining various algorithms which mainly differ on the way the weights are updated, and on the way model's classification opinions are exploited to produce predictions. In this works these approaches were used with DT weak models (boostDT).

- **Adaptive Boosting DT** (ADAboostDT): works with weak learners that are particular DT called stumps, which have only one node and two leaves. The first stump is built exploiting the attribute with the lower Gini index. Each i^{th} built model is associated to a metric called "Amount of say" α, which is a weight that will influence the final voting part

$$\alpha = \frac{1}{2} ln(\frac{1 - Total\ Error}{Total\ Error}) \tag{2.54}$$

  where the Total Error is the sum of the weights of misclassified points and ranges between 0 (perfect stump) and 1 (terrible stump). This metric is also used to update the sample weights according to the adaptive strategy

$$new\ weight = old\ weight^{\pm\alpha} \tag{2.55}$$

  $+\alpha$ is used for misclassified points, while $-\alpha$ is used for correctly classifications. Each $(i+1)^{th}$ stump is trained on a new dataset built by randomly selecting observations from the dataset favouring those that received higher weights in the previous steps (also duplicates may be present), and giving them the same weights to restart the process. At the end a forest of different stumps is built, each of whom has different "Amount to say" values. When a new data is put in this serial algorithm, it will be classified on the basis of the weighted (α) majority votes of all the stumps.

- **Gradient Boosting DT** (GboostDT): In this case each weak learner (full DT) is fitted on the residual error made by the previous DT (and not on the data). At the first iteration a baseline prediction is built through the log of the odds of the target features

$$log(odds) = log(\frac{n.\ of\ samples\ in\ class1}{n.\ of\ samples\ in\ class0}) \tag{2.56}$$

then converted to a probability value c. If $c > 0.5$ then every instance is predicted as belonging to class 1 (otherwise to class 0). For every instance the algorithm computes the residual $observed - predicted$ and each $(i+1)^{th}$ DT does predictions as

$$prediction = F_0 + lr * residual \tag{2.57}$$

where F_0 is the baseline of $log(odds)$, and lr is the learning rate. It is like each new DT is fit on the gradient of the error to introduce correction in the opposite direction with respect to the previous done mistakes.

The hyperparameters to be tuned are the number of estimators, and the learning rate which quantifies the correction strength each i^{th} DT is allowed to impose to solve the mistakes of the $(i-1)^{th}$ tree. There is a strong relationship between these parameters since if the learning rate is low, then more trees are needed to achieve good performances. In contrast with RF, for which the more estimators the better the performances, for boostDT too high estimators may increase the chance of overfitting. Moreover, since GboostDT uses full DT, also pre-pruning parameters can be tuned.

2.3.4 Pipeline

As mentioned at the beginning of Section 2.3, each model classification pipeline (encompassing all the bricks described in the current section) was inserted in a Leave-One Out Cross Validation (LOO-CV) framework.

CV is an expedient exploited to assess the generalization capabilities of a model by training it on just a portion of the data (training set) and evaluating it on the remaining part (test set) in such a way the latter contains new and never seen data. Evaluating the models' performances on data that have already been used for other steps (like training or hyperparameters optimization), may in fact bring to the phenomenon of data leakage which in turn can produce too optimal results. The most common form of CV is the k-fold CV where the dataset is divided into k-folds. At each iteration $k-1$ parts are used for training and the remaining one is used for testing, in such a way data are more effective with respect to the single split e.g. in a $75-25$ single split, the model uses 75% of the data for training and

25% for testing, while in a 10-CV at each iteration the model uses $9/10 = 90\%$ of the data for training, which results in a more accurate model.

LOO-CV is a special case of CV particularly suitable for tiny datasets, like the ones often encountered in the medical field, which exploits a finer division of the dataset treating each sample as a fold. It thus performs CV for n iterations, each time taking $n-1$ patients as the training set, and leaving 1 for testing purposes (where n is the number of samples, in this work $n = 58$, 38 form CAP and 20 from TuSDi databases).

This approach was particularly beneficial coupled with the hyperparameters[17] optimization step, since dividing the original dataset multiple times (one part for training, one for testing, and one for hyperparameters optimization) to avoid data leakage, would have brought to poor results, due to the limited number of samples.

For each features set ($FeatSet_i$, with $i = 1,2,3$) the overall pipeline was developed in an iterative fashion, encompassing the following steps:

1. Data are split in $n-1$ training samples and 1 testing sample.
2. Data are scaled on the training set.
3. Hyperparameters are optimized through a 10-CV grid search on the training set, using as choice-criterion the F1-score.
4. The best parameters are found and the corresponding best model is fit on the training set.
5. The test sample is used to make a single prediction, stored in a vector y_{pred}.
6. 1-5 are repeated for all the samples ($n = 58$), and for all the algorithms (8).

For each $FeatSet_i$, this procedure ended up with 58 different best models per algorithm type (one for each LOO-CV fold). Nevertheless, each classifier was summarized through an unique representative model (Section 3.1), chosen as the one producing the higher F1-score among all the LOO-CV folds best models.

2.3.5 Performance Metrics

The vectors y_{pred}, containing $n = 58$ predictions, were used then to evaluate the models through standard Biomedic-related ML metrics. Considering a generic confusion matrix

[17] This process exploits a grid search approach which consists in trying all the possible combinations of pre-selected parameter vectors in a CV framework in such a way at the end the parameters corresponding to the best models are chosen on the basis of the best value of some predetermined score metric. It is a time-consuming process but if properly done it allows to significantly improve the performances of all the models reducing the chances of overfitting.

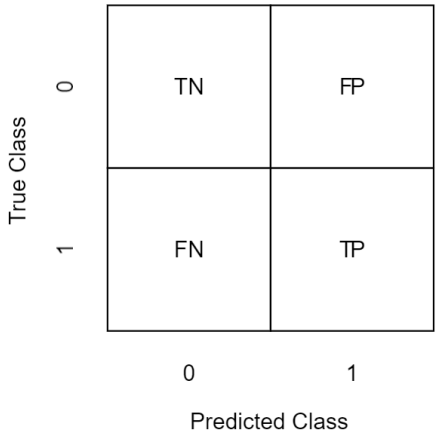


Figure 2.40: Confusion matrix scheme.

the following performances indicator were computed:

- Accuracy: It is a measure of the classification error

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.58}$$

- Sensitivity (Recall, TPR)[18]: Defines the test's ability to correctly identify ill patients out of healthy ones

$$TPR = \frac{TP}{TP + FN} = 1 - FNR \tag{2.59}$$

- Specificity (TNR): Defines the test's ability to correctly reject healthy subjects which do not present the condition

$$TPR = \frac{TN}{TN + FP} = 1 - FPR \tag{2.60}$$

[18]A test which reliably detects the presence of a condition (i.e. positive patients) resulting in high TP and low FN has high sensitivity while a test which efficiently excludes subjects who do not have the condition (i.e. negative patients) resulting in high TN and low FP has high specificity. The weight of these two metrics is different on the basis of the diagnostic application. In all the cases in which failing to treat a condition is serious, high sensitivity is preferred, while situations in which patients classified as positive are subject to more testing, expense, or anxiety, a high specificity is favourite. In the under study case of early diagnostic of neurodegeneration high sensitivity is a must.

- ROC curve and AUC: Represent the graphical ways to display the Sensitivity-Specificity trade-off. Making predictions can be seen as thresholding the degree of uncertainty associated to each outcome (Figure 2.41), whenever the computed predicted probability is under the threshold, the samples are classified as negative and vice-versa. For balanced dataset this threshold $\gamma = 0.5$. In order to define the threshold such that the working point of the classification algorithm reflects the above-mentioned trade-off basing on the specific application needs, it is possible to plot the FPR *i.e.* $1 - TNR$, against the TPR for all the possible decision thresholds. The resulting curve is said Receiving Operating Characteristic (ROC) and when associated to a reliable classifier, it stays above a straight 45° line which represents a model that does not perform better of a coin toss. ROC quality can be summarized in an unique metrics called Area Under the Curve (AUC).

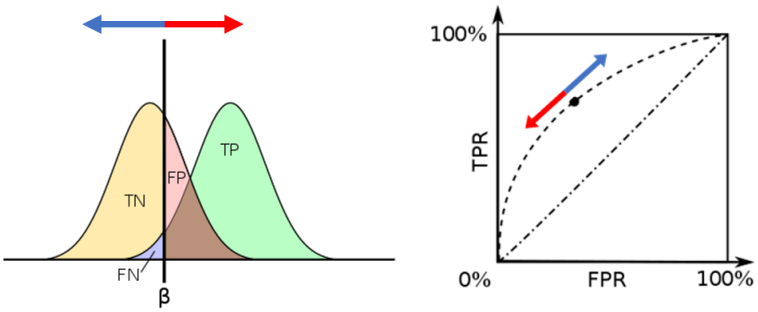


Figure 2.41: ROC and confusion matrix.

- Precision: It is the rate of the actual positive samples

$$P = \frac{TP}{TP + FP} \tag{2.61}$$

- F1-score: It is the harmonic mean of recall and precision

$$F1 = \frac{2TP}{2TP + FP + FN} \tag{2.62}$$

2.4 Three-Stages Classification: H/C, RSWA, and RBD

To date, the majority of works about the automatic score of sleep EEG signals are conducted exploiting supervised machine learning approaches (Section 2.3) which have a series of advantages since by simply employing a sufficient quantity of manually labeled EEG samples, they allow to train models able to make accurate predictions on new unlabelled data. On the other hand, unsupervised methods like clustering are less performant on classification tasks since in general do not exploit any pre-knowledge of data and require an appropriate feature selection processing to be reliable, which makes the training part tedious and complex.

However, the performances of the supervised methods are strictly linked to the quality of the training dataset, since their classification capabilities depend on what they learn in the training phase. They in fact may perform worse whenever fed (test) with new samples containing structures that are slightly different compared to the training data (e.g. if a supervised model is trained on healthy subjects, its performances can degrade as soon as it is fed with test data from a clinical population since the inherent structure of such patients is out of the pattern the model has learned to recognize). In such situations, to obtain acceptable performances, there is the need to re-training the models.

This still works well if the goal is to mimic the human expert identification (Section 2.3) where supervised approaches outperform unsupervised ones, although it may be not the right way e.g. for unknown-data-patterns discovering purposes.

The main advantage of unsupervised techniques is their focus on grouping similar samples from a data-driven point of view, which makes them independent with respect to any particular dataset, and thus pretty suitable for signal exploring of clinical population and for the detection of anomalies. This fact provides these methods with the ability to recognize signal patterns and characteristics which overcome any strict prior knowledge, but that rise from the true inherent and underlying structure of the samples, looking at patterns hidden in data and potentially discovering novel ones in a way that is independent of the subject's health status, age, recording device goodness, and other blurring parameters.

According to the rationale followed by Alireza Kazemi et al. (2022) [78] in this research a semi-supervised-based approach was exploited. This method aims to pick the best of both worlds by using the prior knowledge contained in labeled data for boosting the performances of a clustering framework, able then to generate more stable centroids that can help any type of future unlabelled classification or data-exploring task. The current study can be seen as the first step of a more complex process that aims to qualify REM Sleep without Atonia (RSWA) as an intermediate pathological state, supporting a finer characterization of the neurodegenerative

progression from Healthy to RBD (Section 1.4.2). The cluster foundation set in this work will avoid the need for annotations in data, speeding up the analysis and reducing the overall costs.

2.4.1 Pipeline

The semi-supervised procedure, implemented in a Python framework, used the features extracted in Section 2.2 from both CAP (16 healthy, and 22 RBD) and TuSDi (10 healthy, 10 RBD, and 9 RSWA) databases just for the feature set obtained by the union of REM and SWS segments ($FeatSet_3$), resulting in $n = 67$ samples labeled as 0 (H/C), 1 (RBD) and 2 (RSWA). It encompasses two macro stages (fig. 2.42):

1. A supervised pre-processing stage exploiting available data labels for data scaling (as in Section 2.3.1), and dimensionality reduction (feature projection, and feature selection) in such a way data are compliant with Clustering needs.

2. An unsupervised clustering task which exploits the information obtained in (1) to study how samples group themselves, and to quantify the classification capabilities of the trained model through unsupervised metrics. Moreover, the stability of the clusters centroids was tested in a LOO-CV framework, aiming to establish the robustness of the approach.

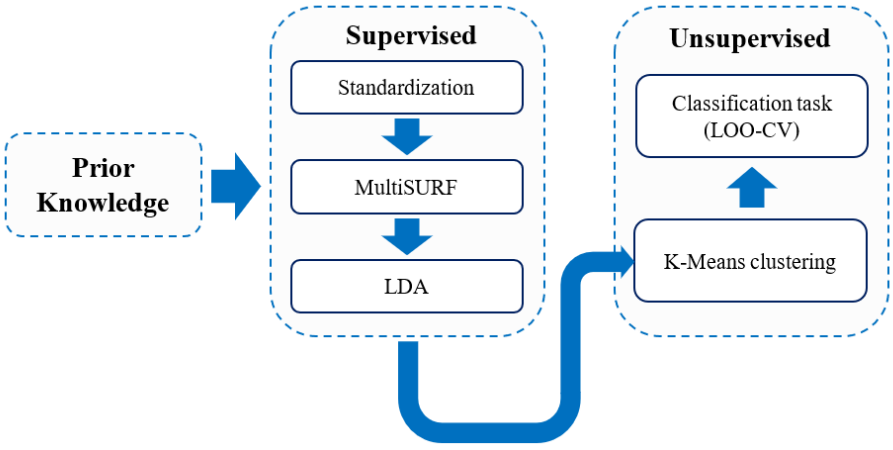


Figure 2.42: Semi-supervised-inspired pipeline scheme.

2.4.2 Feature Selection

The feature selection step is of paramount importance in clustering frameworks, since different feature sets can lead to different grouping results and hence different

performances. Moreover, irrelevant features add noise that can blur or make difficult any clusters revealing.

A comparison among some filter-based univariate selectors was performed, in order to highlight the method that, coupled with LDA (Section 2.4.3), would have allowed to retain the most information about the data structure using the smallest number of features m. It was performed by running, for each method, K-Means clustering (Section 2.4.4) with $K = 3$ for different numbers of features ranging in $(1, 100]$, then reduced to two LDA components, and computing the ARi score. The best results were achieved by MultiSURF especially in the range $m \in [28, 34]$, which is an acceptable features number range to keep high the data interpretability, while ensuring optimal performances. After $m = 50$, all the methods behaved in the same manner.

MultiSURF is one of the core feature selection methods belonging to the family of Relief-based algorithms (RBAs), a bunch of filter strategies that are computationally efficient, and sensitive to the interaction between features avoiding mistaken eliminations of relevant attributes (sci-kit Rebate Library `https://epistasislab.github.io/scikit-rebate/` in Python framework). In their work, Ryan J. Urbanowicz et al. [79] suggest MultiSURF as able to catch multi-dimensional interactions between features (up to 3-way complex pattern associations between attributes) while determining the feature importance. Moreover, it turns out to be the most general and flexible among the RBAs algorithms, with no need of hyperparameters optimization.

The algorithm works iteratively through b random training instances R_i that are picked from the dataset (b chosen equal to the number of samples n) and a features, defining the weight associated to each feature A as $W[A]$. At each cycle R_i is referred as the target instance and the distances between this point and the other $b - 1$ ones are computed. On the basis of these distances, MultiSURF defines two types of nearest neighbours of the target sample:

- One belonging to a different class, called nearest missing M.
- One belonging to the same class, called nearest hit H.

The weight updating strategy is based on the idea that useful attributes should discriminate between samples from different classes and should assume the "same" value for instances of the same class. For this reason if the A value differs between R_i and M, $W[A]$ is increased by a term $+1/b$ (supporting the hypothesis that A discriminates well between classes). On the contrary, if A differs between R_i and H, $W[A]$ is decreased by a factor $-1/b$ (Figure 2.43). This process is repeated until a pre-determined number of features to be selected is reached, each accompanied by its weight (relevance score) which ranges from -1 to 1.

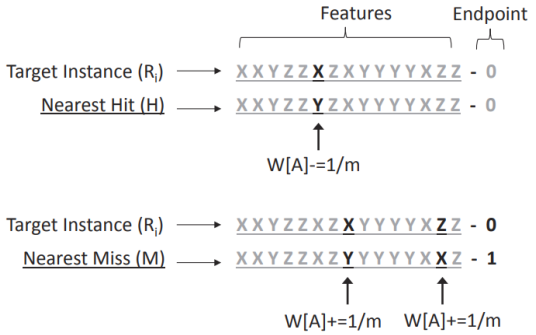


Figure 2.43: MultiSURF weight updating [80].
"Relief updating W[A] for a given target instance when it is compared to its nearest miss and hit. In this example, features are discrete with possible values of X, Y, or Z, and endpoint is binary with a value of 0 or 1."

The algorithm determines the instances that can be considered neighbours using threshold parameters T which creates a hyper-circle in the a-dimensional feature space around each "target" within which all the instances are said “near instances”. Moreover, it defines a dead-band zone which extends to the internal side of the hyper-circle and excludes all the instances in the neighbour (near) which are ambiguously far from the target (Figure 2.44).

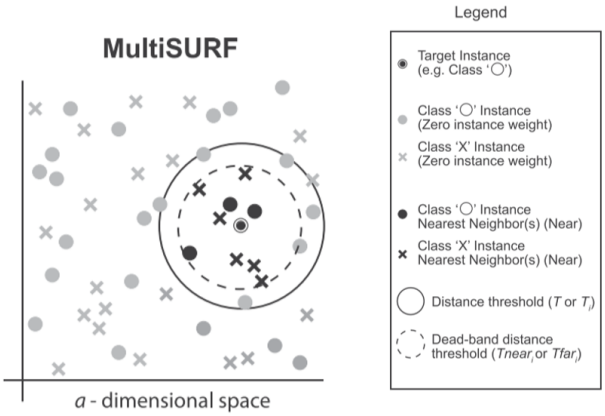


Figure 2.44: MultiSURF neighbours definition [79].

In order to define the optimal number of features m to be selected, the feature selection process was coupled with the optimization of the hyperparameter K (number of clusters). The search was performed by exploiting an unsupervised hyperparameter optimization approach named Elbow method which runs K-Means for different values of K, and reveals the optimal number of clusters as the one related to the elbow point of the objective function curve. All the values of

$K \in [1, 10]$ were considered to not influence this step with any prior-knowledge about the existent classes (H/C, RBD, and RSWA), and to highlight potential underline abnormal patterns. This process was performed for different number of features $m \in (1, 100]$ (then reduced through LDA) to spot the optimal number of clusters in the majority of cases (Figure 2.45, left). The obtained result $K = 3$ was confirmed also from ARi values computed for different number of features and different number of clusters[19] (Figure 2.45, right).

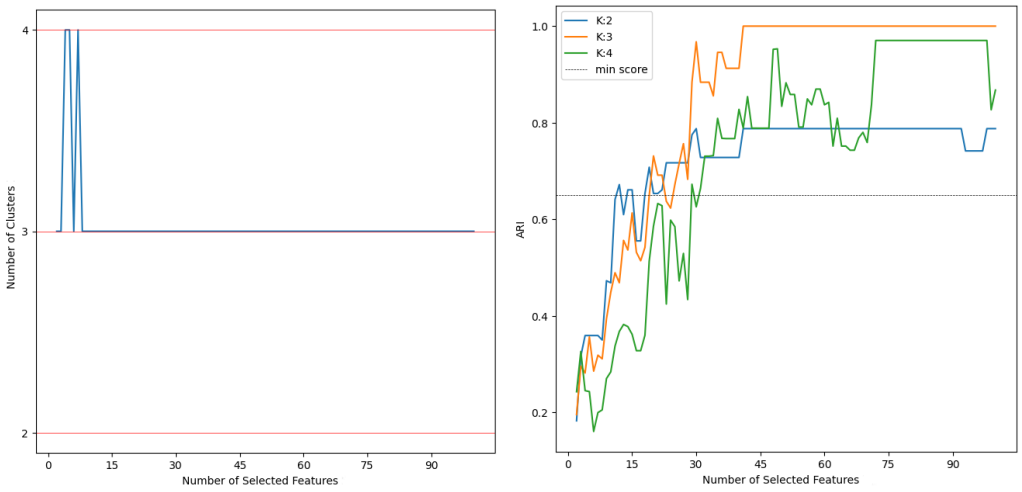


Figure 2.45: Optimal number of clusters.

The optimal number of features was chosen as the minimum value of m for which it was possible to reach the best trade-off between both high ARi score and data interpretability. Already at $m = 45$ (Figure 2.45, orange line) the score index reaches its maximum but pretty similar results can be found also at $m = 30$ that hence seems to be the best compromise.

The optimal results are resumed in Table 2.15, while the best m features scored through MultiSURF selection method are listed in Table 2.16.

Table 2.15: Optimal values for the number of features and clusters

Optimal m	Optimal K
30	3

[19]The small number of $K = 4$ initial cases ($k = 4, 5, 7$ in Figure 2.45, left) were discarded due to the fact that a number of features smaller than k=10 was not sufficient to efficiently grasp the structure of the data, resulting in a not satisfying clustering result.

Table 2.16: Selected Features for REM+SWS segment (updated $FeatSet_3$)

Feature	Score
PKF FREM [MM]	0.4048
SSk FREM [Mean]	0.3556
SSk β [Std]	0.3546
PKF FREM [Std]	0.3542
REN2 β [Std]	0.3534
SCe FREM [Mean]	0.3506
SEF50 FREM [Mean]	0.3487
SEN β [Std]	0.3485
SEFd(95-50) δ [Mean]	0.3414
SSk FREM [75th-P]	0.3317
SCe FREM [75th-P]	0.3249
SEFd(95-50) θ [Mean]	0.3243
SEFd(95-50) θ [75th-P]	0.3223
SEFd(95-50) δ [75th-P]	0.3175
PKF FREM [Mean]	0.3137
SEF50 FREM [Std]	0.3128
SK β [Std]	0.3116
REN2 FREM [Std]	0.3060
SEFd(95-50) FREM [Mean]	0.3016
SK FREM [Mean]	0.3011
SEF50 FREM [75th-P]	0.2976
SK FREM [Std]	0.2908
SK FREM [75th-P]	0.2890
SEFd(95-50) FREM [Std]	0.2855
SK FREM [MM]	0.2838
SEN FREM [Std]	0.2798
PKF θ [75th-P]	0.2794
SEN FREM [Mean]	0.2791
Coastline DWT δ [Std]	0.2777
PKF FREM [75th-P]	0.2771

2.4.3 Feature Projection

In order to further manage the curse of the dimensionality problem, and to graphically visualize the effects of the clustering action, the number of features was reduced to be smaller or equal than three. It was demonstrated that reducing the dimensionality just by exploiting a feature selection strategy was not sufficient, since the resulting number of attributes required to both guarantee good classes separability and high data interpretability, was still too high. To resume the largest information in the smallest number of components, Linear Discriminant Analysis (LDA) was exploited then.

This supervised linear transformation technique removes redundant attributes and noise by weighting the features and linearly combining them into new components which contain a lot of the original information in a smaller dimension framework, while maximizing the linear separability among classes. LDA extracts information from data to create new low-dimensional axes (linear discriminants) onto which project the data in such a way, on those directions, the classes are maximally separated. The maximum number of allowed new axes is $C-1$, with C the number of classes. Since in this work $C=3$, at most two LDA components could be exploited [81].

The transformation of the original axes is based on the ratio of two criteria that should be simultaneously true once data are projected (Fisher's criterion):

- Maximization of the distance between the classes means (between-class variance).
- Minimization of the variance (or scatter) within each class (within-class variance).

The maximization of Fisher's criterion brings to an eigenvalues problem[20] whose c eigenvectors (corresponding to the largest c eigenvalues) are the projection axes i.e. the columns of an optimal projection matrix $\Theta^* \in \mathbb{R}^{m,c}$ used to transform the original features space as following

$$LDAs = (\Theta^*)^T U^T \tag{2.63}$$

where $U \in \mathbb{R}^{n,m}$ is the features matrix, $LDAs \in \mathbb{R}^{c,n}$ are the $C-1$ projections, n is the number of samples, m is the number of features, and c is the number of classes minus one.

[20]The eigenvectors determine the direction of the axes of the new feature subspace i.e. the directions where the maximum class separation is permitted, while the associated eigenvalues tell how "informative" the new axes are or, in other words, they give the amount of between-class variance that each LDA components bring.

Figure 2.46 resumes the overall dimensionality reduction process, which comprises the cascade of MultiSURF and LDA.

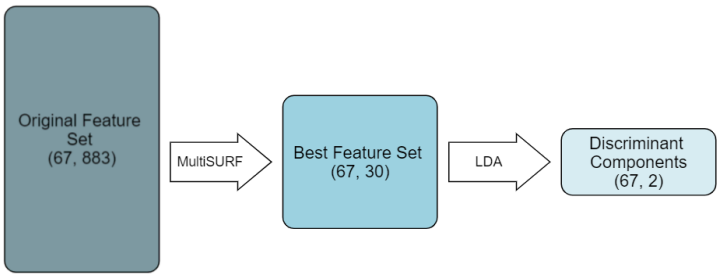


Figure 2.46: Dimensionality reduction cascade.

By checking the explained variance (that is the ratio between the c^{th} eigenvalue and the sum of all the eigenvalues), it was possible to quantify how much information about the variability between the categories was captured by each LDA component (Table 2.17).

Table 2.17: LDA components explained variance

Component	Explained variance
LDA_1	84.55%
LDA_2	15.45%

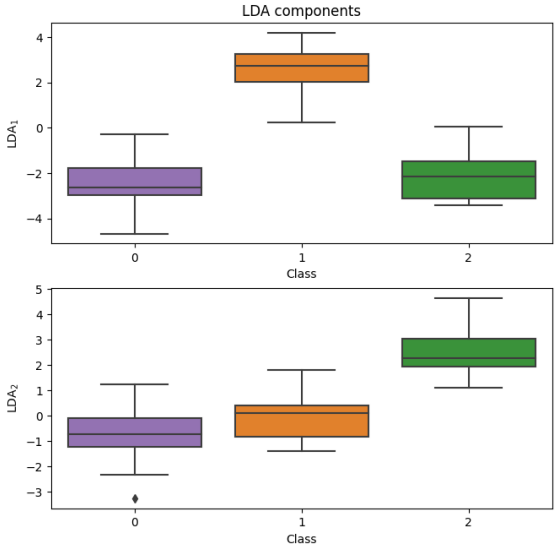


Figure 2.47: Box plots of data on LDA components.

From these results, it is possible to learn just $c = C - 1$ eigenvectors bring most of the information about the separability of the data. The first component is by far the most informative one, and it is possible to not lose much information by forming a 1D-feature spaced based on this eigenpair. Nevertheless, the information brought by the second component has proved to be almost crucial in terms of discriminability power. Hence, starting from the idea that a clustering procedure would have surely benefited of both components in terms of performance, the number of LDA axes was chosen equal to $c = 2$ (Figure 2.47).

2.4.4 K-Means Clustering

Clustering algorithms aim to partition a dataset into groups according to some similarity criterion. This family comprises several techniques which differ in the way the data similarity is quantified (e.g. distance, density, hierarchical grouping). In this work, K-Means clustering was developed since considered the most simple and widespread clustering algorithm (as well as the one that provided the best results). Moreover, it has high interpretability and requires the optimization of just a hyperparameter (Section 2.4.2), speeding up the overall approach.

K-Means algorithm clusters data separating samples in groups with the same variance, simultaneously fulfilling intra-class similarity and inter-class dissimilarity criteria, which result in the definition of clusters' centers (or centroids) that are somehow representative of certain regions of the data. The process concerns the minimization of the so-called within-cluster sum-of-square criterion (inertia criterion) defined as follows

$$\sum_{h=0}^{n} min_{\mu_j \in C}(||s_h - \mu_j||^2) \tag{2.64}$$

where n is the number of samples contained in the feature matrix, s_h is the h^{th} patient row, C is the number of distinct not overlapping clusters, and μ_j is the centroid computed as the mean of the data contained in the j^{th} cluster.

The inertia is a measure of the internal coherence of cluster, and works well with isotropic clusters (like the ones related to this work). The main disadvantage of this criterion is that, since based on Euclidean norm, it can become inflated when applied to high-dimensional data (curse of dimensionality), which however was solved in Sections 2.4.3, and 2.4.2.

The algorithms relies upon an initialization procedure (1) and a looping between two steps $(2 - 3)$:

1. Initial definition of the centroids through basic methods like random choice of C samples from the dataset.

2. Allocation of each data point in its closest centroid cluster.

3. Update of centroids computed as the mean values of the samples contained in each previously defined cluster.

This process is repeated until the centroids stop to change in a significant manner (convergence).

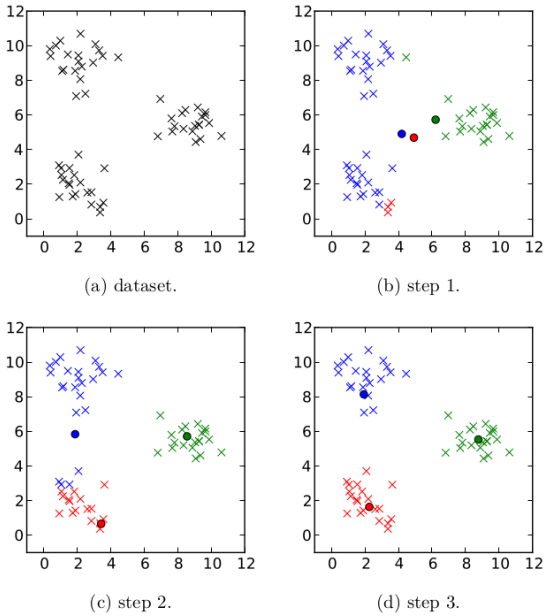


Figure 2.48: K-Means scheme.
https://ludovicarnold.com/teaching/optimization-machine-learning/unsupervised-example-clustering-k-means/

K-Means suffers from the random initialization of the centroids, in terms of convergence speed and reliability of the results. To solve this aspect two strategies were adopted i.e. initializing the centroids in such a way they start far from each other (K-Means++ in sci-kit learn Python environment), and running the algorithm several times with different initialization of centers, collecting the performances as $mean \pm std$. The centroids corresponding to the best run i.e. the one related to the highest Adjusted Rand Index score (ARi)(Section 2.4.5), were selected as representative results of the clustering part.

2.4.5 Performance Metrics

The obtained clustering results were used to quantify the performance of the current approach. The result of such a technique is not an estimation of the class to which the point is intended to belong, but rather the labeling of each sample based on its cluster membership, which may or may not agree with the labels used as target

variables (ground truth) and may change from one run to the next. For this reason, the performance scores should focus on aspects like the shape and the coherence of the final clusters, and the separation quality of the space. They can be divided into two main categories namely External criteria, which exploit label prior-knowledge, and Internal criteria, which do not.

External Criteria:

- Adjusted Rand-Index (ARi): It measures the similarity between the ground-truth and the clustering assignments, ignoring permutations. It is proportional to the number of sample pairs with identical labels in both ground truth and clustering assignments, with values ranging from -1 (poor agreeing labels) to 1 (perfect matching), where $ARi = 0$ defines a random matching.

$$Ri = \frac{a+b}{C_2^n} \tag{2.65}$$

$$ARi = \frac{Ri - E[Ri]}{max(Ri) - E[Ri]} \tag{2.66}$$

  where C is the ground truth class assignment, K is the clustering, a (b) is the number of pairs of elements that are in the same (different) C and in the same (different) K, C_2^n it the number of possible pairs in the dataset (with n samples), and $E[Ri]$ is the expected Ri of random labelings.

- V-Measure (V): Is the harmonic mean of homogeneity (the cluster contains members of a single class) and completeness (all members of a given class are assigned to the same cluster). V $\in [0, 1]$ and is defined as

$$V = \frac{2 * homogeneity * completeness}{\beta * homogeneity + completeness} \tag{2.67}$$

- Confusion Matrix: Displays the true class on the rows and the clustering assignments on the columns, to spot dispersion along different clusters of data belonging to the same class.

- Purity (P): Each cluster ω_k is assigned to the class c_j which is most frequent within it. The related accuracy is evaluated by counting the number of correct assignments (i.e. the number of samples with ground truth c_j contained in the cluster associated with that class), and dividing the obtained value for the number of samples in the dataset. It can be derived directly from the confusion matrix by dividing the sum of the maximum values of each row for the total number of samples in the dataset. Perfect clustering has $P = 1$ (each cluster contains only elements of a single class).

Internal Criteria:

- Silhouette Coefficient: It ranges between -1 and 1, where higher scores are associated with a better cluster definition.

$$s = \frac{b - a}{max(a, b)} \tag{2.68}$$

  where a (b) is the mean distance between a sample and all other points in the same class (next nearest cluster). The metric s is computed for each sample and then averaged.

- Davies-Bouldin Index (DBi): This score quantifies the average similarity between clusters by comparing the distance between clusters and the distance within each cluster. Defining the similarity between two clusters through the measure R_{ij}

$$R_{ij} = \frac{s_i + s_j}{d_{ij}} \tag{2.69}$$

  where s_i is the average distance between each point of cluster i and its centroid, and d_{ij} is the distance between cluster centroids i and j, then DBi may be computed as

$$DBi = \frac{1}{K} \sum_{i=1}^{K} max_{i \neq j} R_{ij} \tag{2.70}$$

  where K is the number of clusters. Since $max_{i \neq j} R_{ij}$ is the measure of the similarity between the i^{th} cluster and the j^{th} cluster most similar to it, the aim is to keep it as low as possible to ensure the best inter-cluster discrimination, and for this reason, good clustering is associated to lower values of DBi.

2.4.6 Centroids Stability

To test the stability of the cluster centroids against different training sets and initialization (and thus to quantify the reliability of the current semi-supervised-based approach), the extracted LDA components were used in a LOO-CV framework, and the obtained clusters were evaluated through Accuracy and ARi scores. This was performed starting from the assumption that if clusters are well formed and stable, then the classification results obtained by using each time a slightly different training set (differing just for one sample at a time) for classifying a new sample, should not dramatically change with respect those achieved without LOO-CV.

As mentioned in Section 2.4.5, each sample is labeled based on its cluster membership, which may or may not agree with the labels used as target variables (ground truth), and that may change from one run to the next. For this reason, the Accuracy score required a suitable label mapping to force the coherence between the values used as targets (0 H/C, 1 RBD, and 2 RSWA) and clustering assignment.

Chapter 3

Results

As previously stated, the models described in Sections 2.3, and 2.4 were used in a binary and three-stages classification assignment respectively, to distinguish healthy from diseased (RBD, and RSWA) participants (CAP Sleep Database and TuSDi Database). The first part of this section displays the categorization performance for each $FeatSet_i$ evaluated (Table 2.11), while the second part resumes the clustering capabilities of the data belonging to $FeatSet_3$.

3.1 Binary problem

Each $FeatSet_i$ was used to train, test, and evaluate eight different supervised models, following the pipeline described in Section 2.3.4. All the models were then compared focusing on the trade-off between high sensitivity score and overall balance. This was done to reduce false negatives as much as possible (of paramount importance in the diagnostic field) without neglecting other metrics that still bring important information.

3.1.1 REM Features

This feature set comprised EEG features extracted from REM segments and PSG features. Table 3.1 shows the classification performance of the eight models examined; the results refer to a LOO-CV framework (one iteration). The macro-averaged accuracy computed across all classifiers evaluated is $86.85\% \pm 0.03$. The k-SVM classifier reached the best overall performance (accuracy: 89.66%, sensitivity: 93.75%). As far as the model optimization is concerned, the employed hyperparameters were C (regularization term), γ (kernel action) both with search range $[0.0001, 1000]$, and the kernel type among Polynomial, RBF, and Sigmoid.

The grid search resulted in the optimal representative model *k-SVM(C:* 2.5, γ*:* 0.01*, kernel: Sigmoid).*

Table 3.1: $FeatSet_1$(PSG+REM) classification performances (%)

Score	KNN	SVM	k-SVM	DT	NB	RF	GbD	ADAb
Accuracy	89.66	87.93	89.66	86.21	87.93	89.66	79.31	84.48
Sensitivity	90.63	87.50	93.75	84.38	93.75	90.63	81.25	84.38
Specificity	88.46	88.46	84.62	88.46	80.77	88.46	76.92	84.62
Precision	90.63	90.32	88.24	90.00	85.71	90.63	81.25	87.10
F1	90.63	88.89	90.91	87.10	89.55	90.63	81.25	85.71
AUC	0.94	0.93	0.91	0.75	0.96	0.94	0.85	0.92

3.1.2 SWS Features

In the second feature set, the EEG features were extracted from SWS segments and were exploited (together with PSG features) in order to quantify the capability of SWS as an RBD biomarker (Section 1.2.2), and to compare it with REM sleep. Table 3.2 shows the classification performance of the eight models examined; the results refer to a LOO-CV framework (one iteration). The macro-averaged accuracy computed across all classifiers evaluated is $79.09\% \pm 0.02$. The RF classifier reached the best overall performance (accuracy: 81.03%, sensitivity: 87.50%). As far as the model optimization is concerned, the employed hyperparameters were:

- The number of estimators with search values $[10, 50, 100, 150]$.
- The splitting criterion among Gini, Entropy, and log-loss.
- The maximum depth of each weak DT, with search range $[1, 10]$.
- The number of randomly selected features to be exploited at each split, between $sqrt(n.features)$, and $log_2(n.features)$.

The grid search resulted in the optimal representative model *RF(n.estimators: 10, criterion: Gini, max depth: 5, features:* log_2*).*

Table 3.2: $FeatSet_2$ (PSG+SWS) classification performances (%)

Score	KNN	SVM	k-SVM	DT	NB	RF	GbD	ADAb
Accuracy	79.31	81.03	81.03	77.59	79.31	81.03	77.59	75.86
Sensitivity	84.38	90.63	84.38	84.38	87.50	87.50	81.25	78.13
Specificity	73.08	69.23	76.92	69.23	69.23	73.08	73.08	73.08
Precision	79.41	78.38	81.82	77.14	77.78	80.00	78.79	78.13
F1	81,82	84.06	83.08	80.60	82.35	83.58	80.00	78.13
AUC	0,84	0.76	0.84	0.67	0.85	0.86	0.79	0.79

3.1.3 REM+SWS Features

The last feature set encompassed PSG features and EEG features that were extracted from the new segment type built by merging REM and SWS epochs. This expedient allowed to slightly improve the classification performance of some of the analyzed models, although the results are, overall, quite comparable with those of $FeatSet_1$. Table 3.3 shows the classification performance of the eight models examined; the results refer to a LOO-CV framework (one iteration). The macro-averaged accuracy computed across all classifiers evaluated is $86.21\% \pm 0.03$. The KNN classifier reached the best overall performance (accuracy: 91.38%, sensitivity: 90.63%). As far as the model optimization is concerned, the employed hyperparameters were K (number of neighbors) with search range $[1, 30]$, and the distance metric among Minkowski, Euclidean, Manhattan, and Chebyshev.

The grid search resulted in the optimal representative model *KNN(K:15, metric: Manhattan)*.

Table 3.3: $FeatSet_3$ (PSG+[REM+SWS]) classification performances (%)

Score	KNN	SVM	k-SVM	DT	NB	RF	GbD	ADAb
Accuracy	91.38	87.93	87.93	79.31	87.93	82.76	86.21	86.21
Sensitivity	90.63	84.38	90.63	87.50	87.50	84.38	81.25	81.25
Specificity	92.31	92.31	84.62	69.23	88.46	80.77	92.31	92.31
Precision	93.55	93.10	87.88	77.78	90.32	84.38	92.86	92.86
F1	92.06	88.53	89.23	82.35	88.89	84.38	86.67	86.67
AUC	0.91	0.92	0.93	0.79	0.89	0.90	0.89	0.93

In Figures 3.1, 3.2, 3.3 they are displayed the ROC curves (with the corresponding operating point), and the confusion matrices for each $Feat_i$ best model.

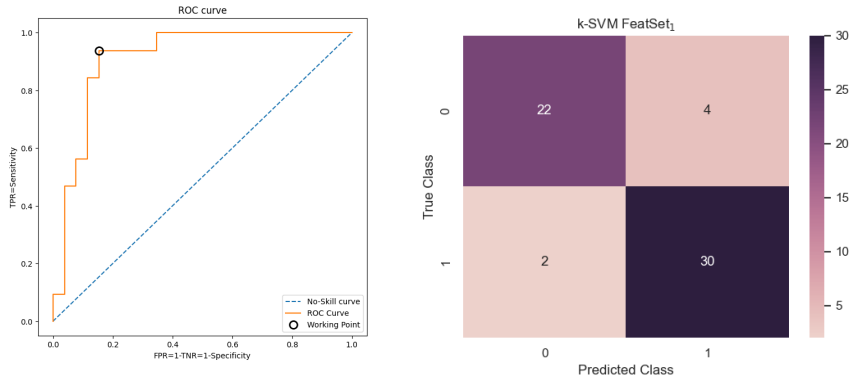


Figure 3.1: $FeatSet_1$ best model's ROC curve and Confusion Matrix.

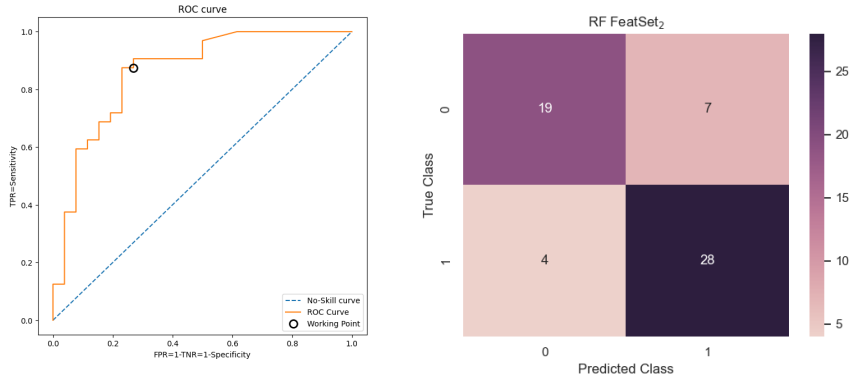


Figure 3.2: $FeatSet_2$ best model's ROC curve and Confusion Matrix.

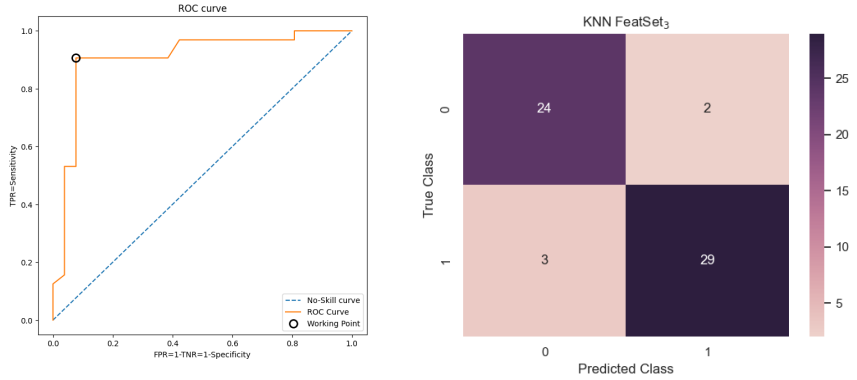


Figure 3.3: $FeatSet_3$ best model's ROC curve and Confusion Matrix.

3.2 Three-stages problem

The two LDA components enclosing the best 30 features from the REM+SWS segment ($FeatSet_3$) obtained following the pipeline described in Section 2.4.1, were used to define the space within which performing data samples grouping.

As mentioned in the related section, K-Means was run multiple times on the dataset, in order to decrease the effect of the initial centroids initialization. The average results obtained from these executions are resumed in Table 3.4.

Table 3.4: $FeatSet_3$ average clustering performances

Score	Value
Adjusted Rand Index (ARi)	0.962 ± 0.021
Silhouette score	0.554 ± 0.001
V-measure	0.937 ± 0.025
Purity	0.983 ± 0.008
Davies-Bouldin Index (DBi)	0.696 ± 0.011

The ground truth points and the clusters obtained from the best K-Means run are displayed in Figure 3.4, together with the Elbow method procedure to define the optimal number of clusters (Figure 3.5).

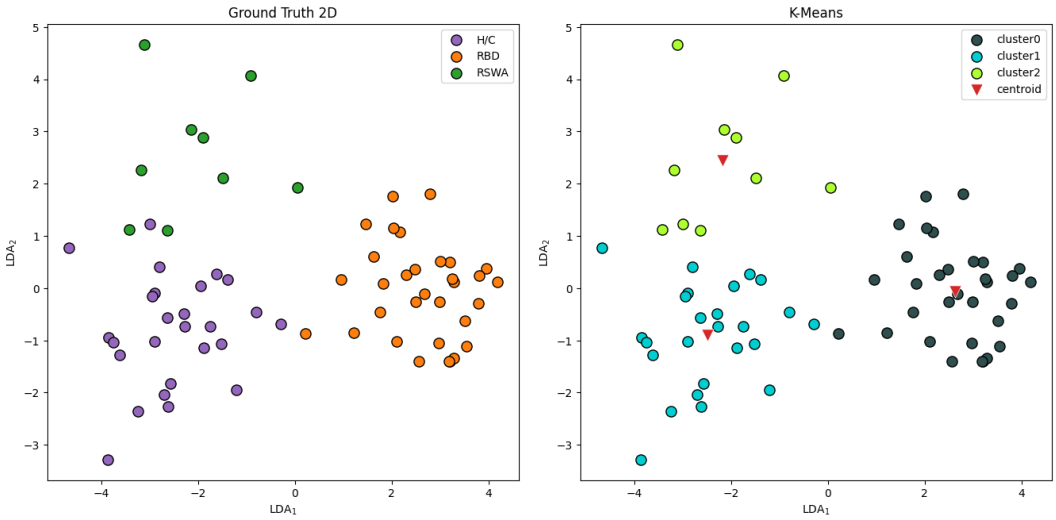


Figure 3.4: Ground truth data (left), clustered data (right) with $m = 30$, and $K = 3$. The red triangles represent the centroids obtained through the clustering model.

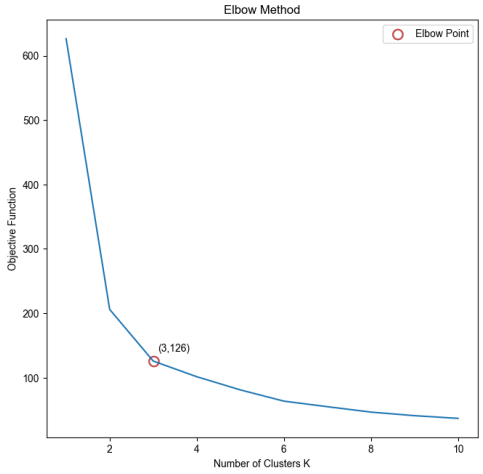


Figure 3.5: Elbow method definition of the optimal number of clusters $K = 3$.

The performances reached by this model and the related centroids coordinated are highlighted in Table 3.5, and Table 3.6 respectively.

Table 3.5: $FeatSet_3$ clustering performances

Score	Value
Adjusted Rand Index (ARi)	0.967
Silhouette score	0.554
V-measure	0.944
Purity	0.985
Davies-Bouldin Index (DBi)	0.693

Table 3.6: Best model's centroids

Cluster	Centroids Coordinates
Cluster 0	$+2.619, -0.059$
Cluster 1	$-2.482, -0.899$
Cluster 2	$-2.175, +2.437$

It is possible to further assess the quality of the obtained clusters by exploiting a confusion matrix (Figure 3.6), and a bar graph which plots the silhouette coefficients

for each data sample contained in the dataset, referenced to the average silhouette coefficient (Figure 3.7).

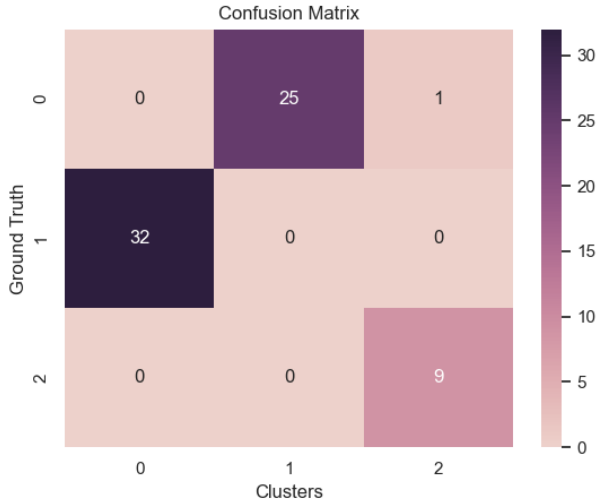


Figure 3.6: Clustering confusion matrix.

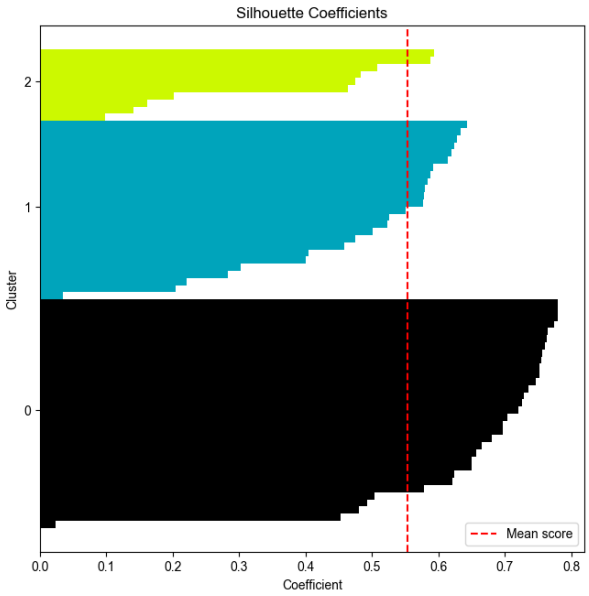


Figure 3.7: Silhouette coefficients for each sample.

The last step of the current pipeline, comprised the evaluation of the clusters'

centroids' stability in a LOO-CV framework, in order to prove the reliability of the current semi-supervised-based strategy. After the labels mapping mentioned in Section 2.4.6 was performed, the following performance values were obtained

Table 3.7: Clusters centroids stability

Score	Value
Accuracy	0.985
Adjusted Rand Index (ARi)	0.967

fulfilling the main assumption this expedient was based on. In fact, the classification error (red circle) made by applying this procedure was the same obtained in Figure 3.4 (right).

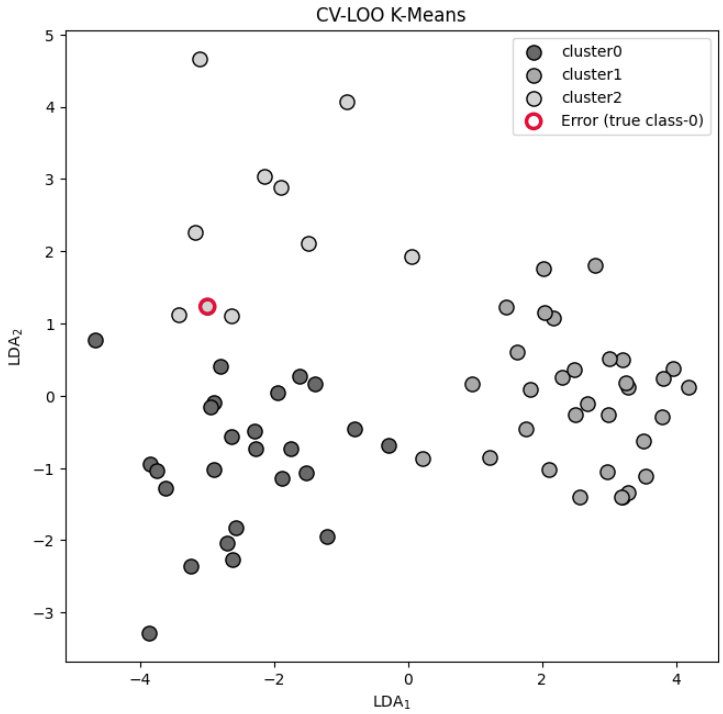


Figure 3.8: LOO-CV clustering error.

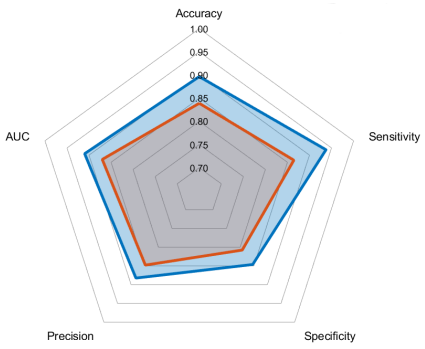
Chapter 4

Discussion

4.1 Binary problem

From the results obtained in Section 3.1 it is possible to appreciate that the developed classifiers generally achieved good overall performances, with macro-averaged Accuracy of 80% and 87%, and Sensitivity of 85% and 88% when using $FeatSet_1$ and $FeatSet_2$ respectively, supporting the belief about the capabilities of both REM and SWS sleep as potential RBD biomarkers [9], [49], [19].

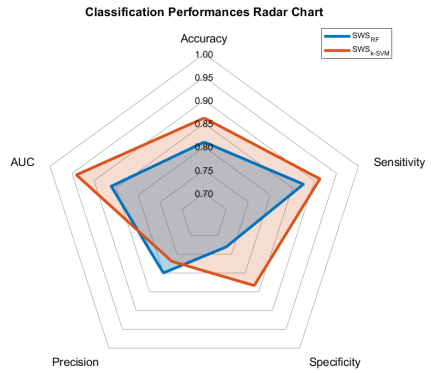
The findings related to Section 3.1 were compared with those obtained by Rechichi and colleagues (2022) [27] (which exploited the same databases used in the current analysis), and then used to highlight differences and discrepancies between the studies.



Metric	New	Old
Accuracy	89.66	83.91
Sensitivity	93.75	86.46
Specificity	84.62	80.77
Precision	88.24	84.78
F1-score	90.91	85.55
AUC	0.91	0.87

Figure 4.1 & Table 4.1: $FeatSet_1$ comparison with previous work [27]

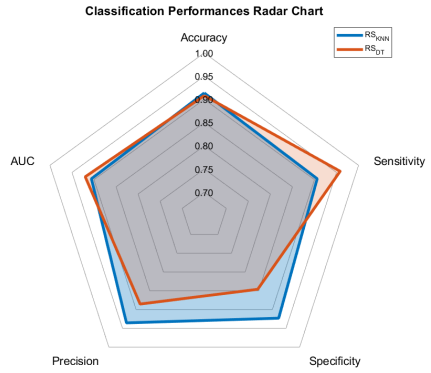
with macro-averaged Accuracy values of 86.85%±0.03 of the new $FeatSet_1$-based best model against 80.29% ± 0.03 of the old one.



Metric	New	Old
Accuracy	81.03	86.21
Sensitivity	87.50	91.23
Specificity	73.08	83.36
Precision	80.00	76.92
F1-score	83.58	83.36
AUC	0.86	0.94

Figure 4.2 & Table 4.2: $FeatSet_2$ comparison with previous work [27]

Although the new $FeatSet_2$-based best model capabilities are generally worse with respect to those of the previous work, it is worth noting that the macro-averaged Accuracy value related to the new $FeatSet_2$ ($79.09\% \pm 0.02$) is pretty near to the $81.10\% \pm 0.03$ associated with the old one, suggesting that the higher performances linked to the latter may be related to a finer tuning of the model rather than a concrete increased features' power.



Metric	New	Old
Accuracy	91.38	90.80
Sensitivity	90.63	95.83
Specificity	92.31	84.62
Precision	94.55	88.54
F1-score	92.06	91.99
AUC	0.91	0.92

Figure 4.3 & Table 4.3: $FeatSet_3$ comparison with previous work [27]

with macro-averaged Accuracy values of $86.21\% \pm 0.03$ of the new $FeatSet_3$-based best model against $85.70\% \pm 0.04$ of the old one.

In contrast with the previous work, REM features outperformed SWS ones with averaged-accuracy improvement of 8%, +4% on averaged-Sensitivity, and +13% on averaged-Specificity (Figure 4.4). Nevertheless, the SWS-based best model defined in Section 3.1 proved to be effective in dealing with the binary classification job, reaching values of Accuracy and Sensitivity above 80%.

As noticed in the above-cited comparison, the performances achieved by $FeatSet_1$ have experienced a sharp increase with respect to those obtained on the same subset in the previous work (Figure 4.1), diminishing the discriminatory power gap with respect to $FeatSet_3$. For this reason, the global performances obtained with $FeatSet_1$ and $FeatSet_3$ are now almost comparable. Despite these findings, and that the highest Sensitivity value is associated with the best REM-based model, KNN fed with REM+SWS data has proved to be the most balanced among all the best classifiers (Figure 4.4, pink), confirming that REM and SWS segments may bring complementary information about the characterization of RBD and healthy differences, allowing to smoothly boost the approach performances when coupled. Moreover, KNN fed with REM+SWS data reached a value $k = 82.64\%$ of Cohen's Kappa, showing optimal agreement with the manual H/C-RBD scores.

$$k = \frac{2(TP * TN - FN * FP)}{(TP + FP) * (FP + TN) + (TP + FN) * (FN + TN)} \tag{4.1}$$

with k ranging from 0 (no agreement) to 1 (perfect agreement).

The evaluation results as well as the comparison between the presented best models can be graphically appreciated in Figure 4.4

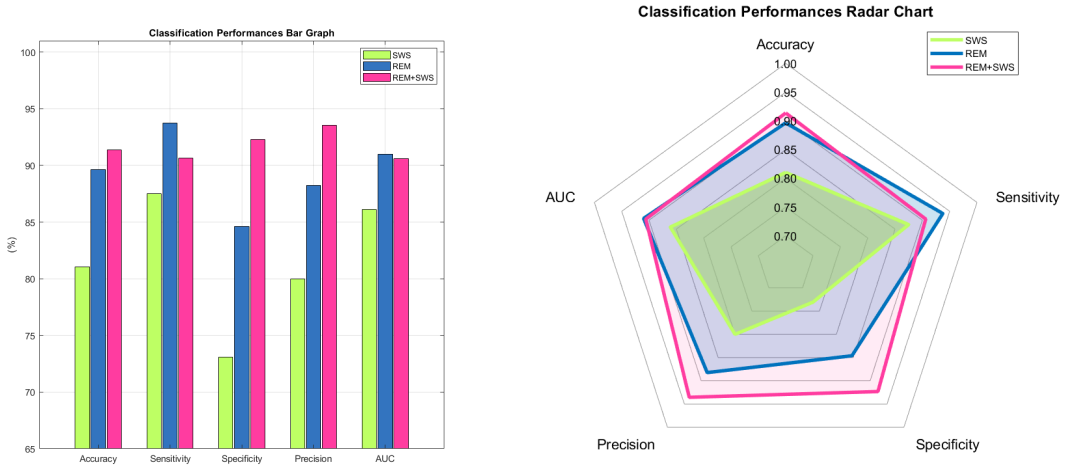


Figure 4.4: Binary classifiers comparison plots.

The fact that the new $FeatSet_3$ allowed to obtain results that were similar or slightly better (in terms of balance) with respect

- Those achieved with the old one (Figure 4.3).
- Those reached by exploiting $FeatSet_{1,2}$ (Figure 4.4).

highlights a potentially crucial aspect of the idea of the current thesis work, strictly linked with the decision to define $FeatSet_3$ starting from the union of REM and SWS segments, rather than from the union of the features sets found from REM and SWS segments taken alone [27]. The achieved higher performances in fact create the chance to distinguish between H/C and RBD subjects without the need of further segmenting a PSG-related EEG signal in N3 and REM sleep stages, hence reducing the in-lab pre-processing steps required for the classification task, as well as both the computational and time burden.

The obtained results point to the potential of an EEG-based, low-cost, automatic RBD detection system that can be used in early diagnosis of neurodegeneration in order to spot prone individuals, and allow them to join clinical trials of neuroprotective therapies to halt or at least delay the progression. Moreover, they are more or less aligned (and show coherence) with the ones found in previous works on the same topic [27], [49], suggesting the reliability and the robustness of the current strategy.

4.2 Three-stages problem

The results obtained in Section 3.2 seem to suggest the ability of the developed approach to extract useful unlabeled data pattern information, and to exploit them for grouping samples according to their membership to a certain clinic class. Despite the small dataset size, the clusters are pretty dense and allow the almost univocal discrimination between the three classes performing just one misassignment, as shown by the confusion matrix in Figure 3.6, and reaching a value of Accuracy of 99% even when the training and testing sets slightly changed (Figure 3.8).

It is interesting to note that the existence of a small subset of features able to clearly discriminate samples in a clustering framework is not trivial nor obvious and increases the confidence level about the reliability of the followed semi-supervised-based approach. The main factors which contribute to achieve these results may be related to:

1. The true inherent discriminatory power of the exploited features (Table 2.16) in underlying and highlighting the differences in the data structure of healthy and clinical populations, as well as in spotting the subtle characteristics of intermediate conditions such as RSWA.

2. The power of LDA in amplifying the capabilities of such features while removing all the blurring aspects that may confound the final results, creating highly discriminatory components.

By reaching satisfactory technical properties, the developed pipeline allows qualifying the presence of RSWA as a separate condition with respect to H/C and RBD.

This seems coherent with the results of some studies which suggest isolated RSWA as the starting point for the iRBD progression, hence representing a prodromal form of such REM sleep-related disorder able to predict future neurodegenerations like those belonging to the α-synucleinopathies family [15]. Overall, these findings, may highlight RSWA as the intermediate stage of a neurodegenerative path that can be thus finely characterized through the implementation of a continuous metric that, ranging from 0 (Healthy) to 1 (RBD) and passing through milder bricks (like RSWA, 0.5) [82], characterizes the whole "spectrum" of the above-mentioned cognitive impairment process, allowing to discover even earlier prone subjects, and increasing the chances of improving their QoL.

4.3 Selected Features

The top-m features selected in both Section 2.3 and Section 2.4 may technically assist medical personnel in their studies and evaluation, illustrating which EEG attribute clinicians should look for in order to diagnose the disease, and therefore it is extremely important to assess their consistency with the existent physiology and clinic literature about the EEG abnormalities in RBD subjects.

The obtained subsets encompass different features in FREM ($\delta + \theta$), SOs (low δ), SWA (high δ), and θ bands which are consistent with the work of Buettner and colleagues (2020) [49] performed on the same dataset (CAP, REM sleep), which state that information with respect to RBD disorder may be hidden in frequency ranges within θ, and δ bands with a relative importance that decreases as the Hz range becomes higher, findings that also inspired the idea of performing a more granular division of the EEG spectrum (Figure 2.14). Similar results can be found in the research of Ngo et al. (2020) [19] which related an irregular Amyloid-β deposition with non-physiologic SOs and SWA activities, and Valomon et al. (2021)[9] which spotted abnormal RBD-related patterns in δ and θ bands (N3 sleep), and FREM band (REM sleep). The obtained FREM band classification capabilities are also in line with what was suggested by the studies of Rechichi et al. (2021) [8].

Features contained in $FeatSet_2$ associated with β activity, can be somehow related to Sleep Spindles. These wave patterns mainly occur during N2 sleep, even though it cannot be excluded that both imperfections in sleep staging and the existence of activity residues from one stage to the next, have led to the appearance of such phenomena also during slow wave sleep. Moreover, the reliability of such features seems to be supported by the works of O'Reilly et al (2015) [83] which discovered that sleep spindle density is changed in RBD patients, with clinical cases showing densities considerably lower for fast spindles and higher for slow spindles, compared to H/C one.

The effectiveness of RBD-related anomalies in δ, θ, and β frequency bands during N3 sleep is supported also by the research work of Kazemi et al. (2022) [78], which proposes novel EEG characteristics in subjects with RBD concerning a much reduced power in the δ band and significantly higher power in the α, β, and θ bands compared to typical N2 and N3 stages.

The presence of SEFd(95-50), SEF50, and SEF95 in the above-mentioned subsets is perfectly consistent with what is stated in topic-related works, confirming the great capability of these features in discriminating between Healthy and RBD/RSWA states [8], [27], and more generally in detecting any substantial deviation from normal REM sleep pattern [51]. Moreover, spectral moments (SCe, SSp, SSk, and SK) have been proven to be reliable in the current classification tasks, suggesting that remarkable changes in EEG spectrum shape are present in RBD and RSWA compared to H/C [9], [49].

It can be interesting to notice that among all the updated subsets (both for binary and clustering problems), just a single feature came from PSG metrics (Wake Proportion, WP), highlighting the power and effectiveness of the exploited EEG-based approach as a valuable substitute of classic PSG-based analysis.

The discriminatory power of the top-1 best features of each $FeatSet_i$ in the binary problem, and of the LDA components in the three-stage study, can be appreciated by analyzing their probability density functions (Gaussian fit) computed along all the patients coherently with class labels, and displayed in Figure 4.5 and Figure 4.6 respectively.

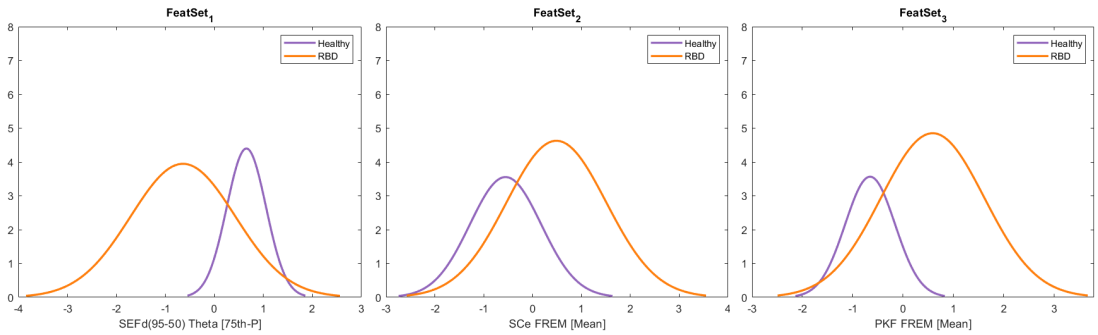


Figure 4.5: Best features discriminatory power.
(left) $FeatSet_1$, (middle) $FeatSet_2$, (right) $FeatSet_3$

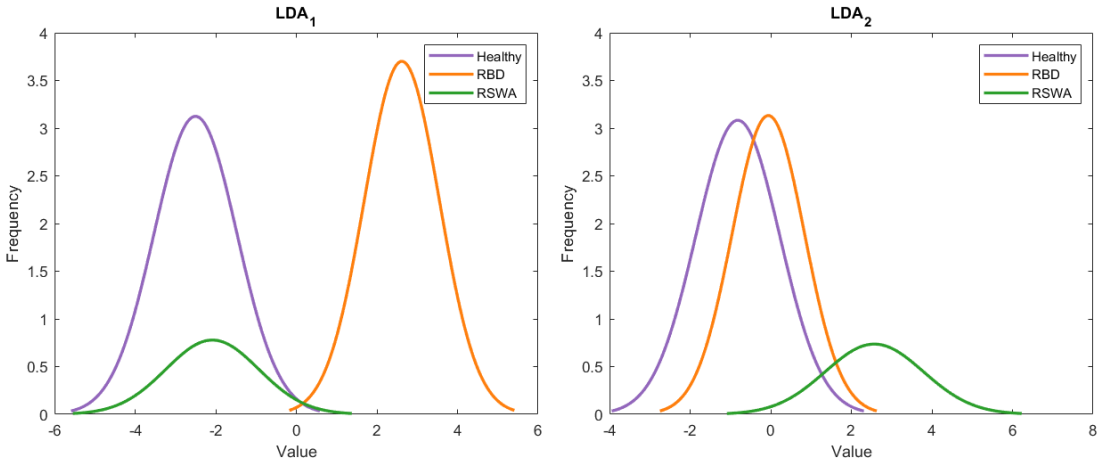


Figure 4.6: LDA components discriminatory power.

As mentioned in Section 2.4.3 the first LDA component separates the samples belonging to H/C and RBD classes, while the second component is crucial to discriminate between H/C-RBD and the RSWA class.

Chapter 5

Conclusions and Future Works

Overnight PSG type-1 is the gold standard for RBD diagnosis but shows a certain number of limitations concerning availability, time and money burden, number of used sensors, discomfort for the patient which in turn may create biases in recordings, and the need for specialized health care staff for the manual scoring that can be laborious and which may significantly hinder the diagnostic process. Moreover, it is usually coupled with additional screening questionnaires and medical history interviews which further increases the expected work load. Given the well-established knowledge about RBD as early biomarker for the onset of α-synucleinopathies (up to 14 years), accelerating the entire prediction process through alternatives to classical manual or semi-automatic RSWA scoring methods becomes crucial.

The current work aimed to overcome the cost, time, accuracy, and robustness restrictions of the to-date available diagnostic tools, proposing a fully-automatic diagnostic approach, which through the usage of a single EEG channel, reaches in classifying reliably and efficiently healthy patients from diseased ones, exploiting underlying characteristics of PSG-based EEG signal instead of classical EMG one. RSWA is thought to be a dissociative state, characterized by a mismatch between brain (REM) and muscular activity (non REM) which is however linked to the former to some extent. Basing the current strategy on EEG signal instead of EMG would avoid performing a classical two-steps approach concerning the chain of sleep study and muscular activity evaluation, working directly on the source of the phenomena i.e. the brain.

Data were taken from REM and SWS Sleep recordings. The trained ML methods (both supervised-binary and unsupervised-multiclass) achieved high performances which are also in line with other topic-related works, supporting the reliability

and robustness of the single-channel EEG approach to accomplish healthy-diseased classification tasks. For the two-stages problem, the feature set derived from the segment REM+SWS outperformed the others, reaching high overall balance, values of Accuracy and Sensitivity above 90%, and optimal inter-rater reliability with manual scoring (Cohen's Kappa= 83%) indicating its relevance to the investigation of RBD and the associated neurodegenerative process. Moreover, exploiting EEG-based data in a semi-supervised-based framework, allowed to reveal some inherent patterns of such pathological conditions, useful for more finely characterizing the neurodegeneration process from healthy to RBD.

The obtained results point to the potential of an EEG-based, low-cost, automatic RBD/RSWA detection system that can be used in early diagnosis of neurodegeneration in order to spot prone individuals in advance and make follow-up procedures easier, allowing patients to join clinical trials of neuroprotective therapies to halt or at least delay the progression, thus leading to an improvement in the QoL of these subjects and their loved ones.

Besides the increase in the size of the exploited dataset in order to generalize and validate the obtained results, further works may encompass the lightening and optimization of the overall approach in terms of computational efficiency, and its real-time implementation on a wearable user-friendly device able to facilitate diagnostics in hospital and in home environments, reducing the level of invasiveness for the patients and decreasing the number of undiagnosed cases.

Bibliography

[1] Velayudhan Mohan Kumar. *Sleep and sleep disorders The effects of maternal sleep deprivation on cognition in the offspring in an animal model View project.* 2008. URL: https://www.researchgate.net/publication/5241469 (cit. on pp. 1, 7).

[2] S Chokroverty. *Overview of sleep and sleep disorders.* 2010, pp. 126–140. URL: http://journals.lww.com/ijmr (cit. on pp. 1, 4, 13, 14).

[3] Samantha Bramich, Anna King, Maneesh Kuruvilla, Sharon L. Naismith, Alastair Noyce, and Jane Alty. «Isolated REM sleep behaviour disorder: current diagnostic procedures and emerging new technologies». In: *Journal of Neurology* 269 (9 Sept. 2022), pp. 4684–4695. ISSN: 14321459. DOI: 10.1007/s00415-022-11213-9 (cit. on pp. 2, 16, 17, 20).

[4] Aakash K Patel, Vamsi Reddy, Karlie R Shumway, and John F Araujo. «Physiology, Sleep Stages». en. In: *StatPearls.* Treasure Island (FL): StatPearls Publishing, Jan. 2023 (cit. on pp. 3, 4).

[5] T.of Encyclopaedia Britannica. it. In: *electroencephalography. Encyclopedia Britannica.* May 26, 2023. URL: https://www.britannica.com/science/electroencephalography (cit. on p. 3).

[6] Sara Mahvash Mohammadi, Shirin Enshaeifar, Mohammad Ghavami, and Saeid Sanei. *Classification of Awake, REM, and NREM from EEG via Singular Spectrum Analysis* (cit. on p. 4).

[7] Péter Simor, Gwen van der Wijk, Lino Nobili, and Philippe Peigneux. «The microstructure of REM sleep: Why phasic and tonic?» In: *Sleep Medicine Reviews* 52 (Aug. 2020). ISSN: 15322955. DOI: 10.1016/j.smrv.2020.101305 (cit. on pp. 4, 5).

[8] Irene Rechichi, Maurizio Zibetti, Luigi Borzì, Gabriella Olmo, and Leonardo Lopiano. «Single-channel EEG classification of sleep stages based on REM microstructure». In: *Healthcare Technology Letters* 8 (3 June 2021), pp. 58–65. ISSN: 20533713. DOI: 10.1049/htl2.12007 (cit. on pp. 5, 30, 37, 44, 109, 110).

[9] Amandine Valomon, Brady A. Riedner, Stephanie G. Jones, Keith P. Nakamura, Giulio Tononi, David T. Plante, Ruth M. Benca, and Melanie Boly. «A high-density electroencephalography study reveals abnormal sleep homeostasis in patients with rapid eye movement sleep behavior disorder». In: *Scientific Reports* 11 (1 Dec. 2021). ISSN: 20452322. DOI: 10.1038/s41598-021-83980-w (cit. on pp. 5, 10, 13, 21, 44, 105, 109, 110).

[10] Michael J. Sateia. «International classification of sleep disorders-third edition highlights and modifications». In: *Chest* 146 (5 Nov. 2014), pp. 1387–1394. ISSN: 19313543. DOI: 10.1378/chest.14-0970 (cit. on p. 6).

[11] Michael J. Thorpy. «Classification of Sleep Disorders». In: *Neurotherapeutics* 9 (4 Oct. 2012), pp. 687–701. ISSN: 19337213. DOI: 10.1007/s13311-012-0145-6 (cit. on p. 6).

[12] Andrea Galbiati, Laura Verga, Enrico Giora, Marco Zucconi, and Luigi Ferini-Strambi. «The risk of neurodegeneration in REM sleep behavior disorder: A systematic review and meta-analysis of longitudinal studies». In: *Sleep Medicine Reviews* 43 (Feb. 2019), pp. 37–46. ISSN: 15322955. DOI: 10.1016/j.smrv.2018.09.008 (cit. on pp. 9, 11).

[13] Stuart J. McCarter, Erik K. St. Louis, and Bradley F. Boeve. «REM sleep behavior disorder and REM sleep without atonia as an early manifestation of degenerative neurological disease». In: *Current Neurology and Neuroscience Reports* 12 (2 2012), pp. 182–192. ISSN: 15346293. DOI: 10.1007/s11910-012-0253-z (cit. on pp. 9, 12).

[14] Christopher K Pham, Abdulghani Sankari, and Jennifer M Slowik. *Rapid Eye Movement Sleep Behavior Disorder Continuing Education Activity*. URL: https://www.ncbi.nlm.nih.gov/books/NBK555928/#:~:text=Rapid (cit. on pp. 9, 10, 19).

[15] Michele T. Hu. «REM sleep behavior disorder (RBD)». In: *Neurobiology of Disease* 143 (Sept. 2020). ISSN: 1095953X. DOI: 10.1016/j.nbd.2020.104996 (cit. on pp. 9, 11, 12, 18, 20, 109).

[16] Michael Howell and Carlos H Schenck. *Official reprint from UpToDate Rapid eye movement sleep behavior disorder*. 2023. URL: www.uptodate.com (cit. on pp. 9, 11, 17, 20).

[17] Michela Figorilli, Giuseppe Lanza, Patrizia Congiu, Rosamaria Lecca, Elisa Casaglia, Maria P. Mogavero, Monica Puligheddu, and Raffaele Ferri. «Neurophysiological Aspects of REM Sleep Behavior Disorder (RBD): A Narrative Review». In: *Brain Sciences* 11 (12 Dec. 2021). ISSN: 20763425. DOI: 10.3390/brainsci11121588 (cit. on pp. 10, 18).

[18] Jimmy J. Fraigne, Zoltan A. Torontali, Matthew B. Snow, and John H. Peever. «REM sleep at its core - Circuits, neurotransmitters, and pathophysiology». In: *Frontiers in Neurology* 6 (MAY 2015). ISSN: 16642295. DOI: 10.3389/fneur.2015.00123 (cit. on p. 10).

[19] Hong-Viet V. Ngo, Jurgen Claassen, and Martin Dresler. «Sleep: Slow Wave Activity Predicts Amyloid-beta Accumulation». In: *Current Biology* 30.22 (Nov. 2020), R1371–R1373. ISSN: 0960-9822. DOI: https://doi.org/10.1016/j.cub.2020.09.058. URL: https://www.sciencedirect.com/science/article/pii/S0960982220314287 (cit. on pp. 12, 13, 44, 105, 109).

[20] Yo-El S Ju et al. «Slow wave sleep disruption increases cerebrospinal fluid amyloid-beta levels». In: *Brain* 140.8 (July 2017), pp. 2104–2111. DOI: 10.1093/brain/awx148. URL: https://doi.org/10.1093/brain/awx148 (cit. on p. 12).

[21] Ee Wei Lim, Dag Aarsland, Dominic Ffytche, Raquel Natalia Taddei, Daniel J. van Wamelen, Yi-Min Wan, Eng King Tan, Kallol Ray Chaudhuri, and Kings Parcog groupMDS Nonmotor study group. «Amyloid-β and Parkinson's disease». In: *Journal of Neurology* 266.11 (Nov. 2019), pp. 2605–2619. ISSN: 1432-1459. DOI: 10.1007/s00415-018-9100-8. URL: https://doi.org/10.1007/s00415-018-9100-8 (cit. on p. 12).

[22] Conrad Iber and American Academy of Sleep Medicine. *The AASM manual for the scoring of sleep and associated events : rules, terminology and technical specifications*. eng. Westchester, IL: American Academy of Sleep Medicine, 2007 (cit. on pp. 14, 15).

[23] Leslie C. Markun and Ajay Sampat. «Clinician-Focused Overview and Developments in Polysomnography». In: *Current Sleep Medicine Reports* 6.4 (Dec. 2020), pp. 309–321. ISSN: 2198-6401. DOI: 10.1007/s40675-020-00197-5. URL: https://doi.org/10.1007/s40675-020-00197-5 (cit. on p. 14).

[24] Manjari Tripathi. «Technical notes for digital polysomnography recording in sleep medicine practice». en. In: *Ann. Indian Acad. Neurol.* 11.2 (Apr. 2008), pp. 129–138 (cit. on p. 15).

[25] Birgit Högl and Ambra Stefani. «REM sleep behavior disorder (RBD): Update on diagnosis and treatment». en. In: *Somnologie (Berl.)* 21.Suppl 1 (2017), pp. 1–8 (cit. on pp. 16–19).

[26] Giulia Masi. «Automating REM Sleep Without Atonia Scoring Methods». Tesi di laurea. Politecnico di Torino, 2021 (cit. on p. 19).

[27] Irene Rechichi, Federica Amato, Alessandro Cicolin, and Gabriella Olmo. «Single-Channel EEG Detection of REM Sleep Behaviour Disorder: The Influence of REM and Slow Wave Sleep». In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 13346 LNBI (2022), pp. 381–394. ISSN: 16113349. DOI: 10.1007/978-3-031-07704-3_31 (cit. on pp. 21, 24, 25, 30, 31, 56, 105, 106, 108, 110).

[28] M G Terzano et al. «Atlas, rules, and recording techniques for the scoring of cyclic alternating pattern (CAP) in human sleep». en. In: *Sleep Med.* 2.6 (Nov. 2001), pp. 537–553 (cit. on p. 22).

[29] A L Goldberger et al. «PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals». en. In: *Circulation* 101.23 (June 2000), E215–20 (cit. on p. 22).

[30] David J Acunzo, Graham Mackenzie, and Mark C W van Rossum. «Systematic biases in early ERP and ERF components as a result of high-pass filtering». en. In: *J. Neurosci. Methods* 209.1 (July 2012), pp. 212–218 (cit. on pp. 25, 27, 30).

[31] Mike X Cohen. *Analyzing neural time series data: Theory and practice*. Issues in Clinical and Cognitive Neuropsychology. Mit Press, Jan. 2014 (cit. on p. 28).

[32] Ashima Khosla, Padmavati Khandnor, and Trilok Chand. «A comparative analysis of signal processing and classification methods for different applications based on EEG signals». In: *Biocybernetics and Biomedical Engineering* 40.2 (2020), pp. 649–690. ISSN: 0208-5216. DOI: https://doi.org/10.1016/j.bbe.2020.02.002. URL: https://www.sciencedirect.com/science/article/pii/S0208521620300231 (cit. on p. 30).

[33] Poomipat Boonyakitanont, Apiwat Lek-uthai, Krisnachai Chomtho, and Jitkomut Songsiri. «A review of feature extraction and performance evaluation in epileptic seizure detection using EEG». en. In: *Biomed. Signal Process. Control* 57.101702 (Mar. 2020), p. 101702 (cit. on pp. 30, 59, 61).

[34] Khald Aboalayon, Miad Faezipour, Wafaa Almuhammadi, and Saeid Moslehpour. «Sleep stage classification using EEG signal analysis: A comprehensive survey and new investigation». In: *Entropy (Basel)* 18.9 (Aug. 2016), p. 272 (cit. on pp. 30, 37, 38).

[35] D Puthankattil Subha, Paul K Joseph, Rajendra Acharya U, and Choo Min Lim. «EEG signal analysis: a survey». In: *J. Med. Syst.* () (cit. on pp. 30, 39, 49).

[36] Igor Stancin, Mario Cifrek, and Alan Jovic. «A review of EEG signal features and their application in driver drowsiness detection systems». en. In: *Sensors (Basel)* 21.11 (May 2021), p. 3786 (cit. on p. 30).

[37] Shayan Motamedi-Fakhr, Mohamed Moshrefi-Torbati, Martyn Hill, Catherine M. Hill, and Paul R. White. «Signal processing techniques applied to human sleep EEG signals—A review». In: *Biomedical Signal Processing and Control* 10 (2014), pp. 21–33. ISSN: 1746-8094. DOI: https://doi.org/10.1016/j.bspc.2013.12.003. URL: https://www.sciencedirect.com/science/article/pii/S174680941300178X (cit. on p. 30).

[38] Tarek Lajnef, Sahbi Chaibi, Perrine Ruby, Pierre-Emmanuel Aguera, Jean-Baptiste Eichenlaub, Mounir Samet, Abdennaceur Kachouri, and Karim Jerbi. «Learning machines and sleeping brains: Automatic sleep stage classification using decision-tree multi-class support vector machines». In: *Journal of Neuroscience Methods* 250 (2015). Cutting-edge EEG Methods, pp. 94–105. ISSN: 0165-0270. DOI: https://doi.org/10.1016/j.jneumeth.2015.01.022. URL: https://www.sciencedirect.com/science/article/pii/S0165027015000230 (cit. on pp. 30, 47, 48).

[39] Kristína Šušmáková and Anna Krakovská. «Discrimination ability of individual measures used in sleep stages classification». In: *Artificial Intelligence in Medicine* 44.3 (2008), pp. 261–277. ISSN: 0933-3657. DOI: https://doi.org/10.1016/j.artmed.2008.07.005. URL: https://www.sciencedirect.com/science/article/pii/S0933365708000924 (cit. on pp. 30, 45, 47, 59).

[40] Asghar Ghasemi and Saleh Zahediasl. «Normality tests for statistical analysis: a guide for non-statisticians». en. In: *Int. J. Endocrinol. Metab.* 10.2 (Apr. 2012), pp. 486–489 (cit. on p. 32).

[41] Bo Hjorth. «EEG analysis based on time domain properties». In: *Electroencephalography and Clinical Neurophysiology* 29.3 (1970), pp. 306–310. ISSN: 0013-4694. DOI: https://doi.org/10.1016/0013-4694(70)90143-4. URL: https://www.sciencedirect.com/science/article/pii/0013469470901434 (cit. on p. 35).

[42] Fahrettin F Gonen and Gleb V Tcheslavski. «Techniques to assess stationarity and gaussianity of EEG: An overview». In: *International Journal Bioautomation* 16.2 (2012), p. 135 (cit. on p. 39).

[43] H Sugimoto, N Ishii, A Iwata, N Suzumura, and T Tomita. «On the stationarity and normality of the electroencephalographic data during sleep stages». en. In: *Comput. Programs Biomed.* 8.3-4 (Sept. 1978), pp. 224–234 (cit. on p. 39).

[44] Michael J Prerau, Ritchie E Brown, Matt T Bianchi, Jeffrey M Ellenbogen, and Patrick L Purdon. «Sleep neurophysiological dynamics through the lens of multitaper spectral analysis». en. In: *Physiology (Bethesda)* 32.1 (Jan. 2017), pp. 60–92 (cit. on pp. 40, 41).

[45] Behtash Babadi and Emery N Brown. «A review of multitaper spectral analysis». en. In: *IEEE Trans. Biomed. Eng.* 61.5 (May 2014), pp. 1555–1564 (cit. on p. 40).

[46] E J McCoy, A T Walden, and D B Percival. «Multitaper spectral estimation of power law processes». In: *IEEE Trans. Signal Process.* 46.3 (Mar. 1998), pp. 655–668 (cit. on p. 40).

[47] The MathWorks Inc. *Multitaper power spectral density estimate*. Natick, Massachusetts, United States. URL: `https://it.mathworks.com/help/signal/ref/pmtm.html#btvap5t` (cit. on pp. 41, 42).

[48] P.P. Mitra and B. Pesaran. «Analysis of Dynamic Brain Imaging Data». In: *Biophysical Journal* 76.2 (Feb. 1999), pp. 691–708. DOI: `10.1016/s0006-3495(99)77236-x`. URL: `https://doi.org/10.1016%5C%2Fs0006-3495%5C%2899%5C%2977236-x` (cit. on p. 41).

[49] Ricardo Buettner, Annika Grimmeisen, and Anne Gotschlich. «High- performance diagnosis of sleep disorders: A novel, accurate and fast machine learning approach using electroencephalographic data». In: *Proceedings of the Annual Hawaii International Conference on System Sciences*. Hawaii International Conference on System Sciences, 2020 (cit. on pp. 44, 105, 108–110).

[50] Takashi Nakamura, Valentin Goverdovsky, Mary J Morrell, and Danilo P Mandic. «Automatic sleep monitoring using ear-EEG». en. In: *IEEE J. Transl. Eng. Health Med.* 5 (June 2017), p. 2800108 (cit. on p. 46).

[51] Syed Anas Imtiaz and Esther Rodriguez-Villegas. «A low computational cost algorithm for REM sleep detection using single channel EEG». en. In: *Ann. Biomed. Eng.* 42.11 (Nov. 2014), pp. 2344–2359 (cit. on pp. 46, 110).

[52] Jérôme Antoni. «The spectral kurtosis: a useful tool for characterising non-stationary signals». en. In: *Mech. Syst. Signal Process.* 20.2 (Feb. 2006), pp. 282–307 (cit. on p. 46).

[53] Cuong Q Ngo, Rifai Chai, Timothy W Jones, and Hung T Nguyen. «The effect of hypoglycemia on spectral moments in EEG epochs of different durations in type 1 diabetes patients». en. In: *IEEE J. Biomed. Health Inform.* 25.8 (Aug. 2021), pp. 2857–2865 (cit. on p. 46).

[54] U. Rajendra Acharya, H. Fujita, Vidya K. Sudarshan, Shreya Bhat, and Joel E.W. Koh. «Application of entropies for automated diagnosis of epilepsy using EEG signals: A review». In: *Knowledge-Based Systems* 88 (2015), pp. 85–96. ISSN: 0950-7051. DOI: https://doi.org/10.1016/j.knosys.2015.08.004. URL: https://www.sciencedirect.com/science/article/pii/S0950705115003081 (cit. on pp. 47, 62).

[55] T. Inouye, K. Shinosaki, H. Sakamoto, S. Toi, S. Ukai, A. Iyama, Y. Katsuda, and M. Hirano. «Quantification of EEG irregularity by use of the entropy of the power spectrum». In: *Electroencephalography and Clinical Neurophysiology* 79.3 (1991), pp. 204–210. ISSN: 0013-4694. DOI: https://doi.org/10.1016/0013-4694(91)90138-T. URL: https://www.sciencedirect.com/science/article/pii/001346949190138T (cit. on p. 47).

[56] N Kannathal, Min Lim Choo, U Rajendra Acharya, and P K Sadasivan. «Entropies for detection of epilepsy in EEG». en. In: *Comput. Methods Programs Biomed.* 80.3 (Dec. 2005), pp. 187–194 (cit. on pp. 48, 62).

[57] Nikesh Bajaj. «Wavelets for EEG analysis». In: *Wavelet Theory*. IntechOpen, Feb. 2021 (cit. on p. 49).

[58] A Subasi. «EEG signal classification using wavelet feature extraction and a mixture of expert model». en. In: *Expert Syst. Appl.* 32.4 (May 2007), pp. 1084–1093 (cit. on pp. 49, 52, 53).

[59] The MathWorks Inc. *Trasformata wavelet continua e analisi basata sulla scala*. Natick, Massachusetts, United States. URL: https://it.mathworks.com/help/wavelet/gs/continuous-wavelet-transform-and-scale-based-analysis.html (cit. on p. 51).

[60] The MathWorks Inc. *Continuous and Discrete Wavelet Transforms*. Natick, Massachusetts, United States. URL: https://it.mathworks.com/help/wavelet/gs/continuous-and-discrete-wavelet-transforms.html?lang=en (cit. on p. 51).

[61] Yan Ma, Wenbin Shi, Chung-Kang Peng, and Albert C. Yang. «Nonlinear dynamical analysis of sleep electroencephalography using fractal and entropy approaches». In: *Sleep Medicine Reviews* 37 (2018), pp. 85–93. ISSN: 1087-0792. DOI: https://doi.org/10.1016/j.smrv.2017.01.003. URL: https://www.sciencedirect.com/science/article/pii/S1087079217300187 (cit. on pp. 56, 58, 61, 63).

[62] Pedro Garc and Callegari Pe. *Analysis of EEG Signals using Nonlinear Dynamics and Chaos: A review*. 2015 (cit. on pp. 56, 60).

[63] Jürgen Fell, Joachim Röschke, Klaus Mann, and Cornelius Schäffner. «Discrimination of sleep stages: a comparison between spectral and nonlinear EEG measures». In: *Electroencephalography and Clinical Neurophysiology* 98.5 (1996), pp. 401–410. ISSN: 0013-4694. DOI: https://doi.org/10.1016/0013-4694(96)95636-9. URL: https://www.sciencedirect.com/science/article/pii/0013469496956369 (cit. on p. 56).

[64] Richard Hardstone, Simon-Shlomo Poil, Giuseppina Schiavone, Rick Jansen, Vadim V Nikulin, Huibert D Mansvelder, and Klaus Linkenkaer-Hansen. «Detrended fluctuation analysis: a scale-free view on neuronal oscillations». en. In: *Front. Physiol.* 3 (Nov. 2012), p. 450 (cit. on pp. 56–58).

[65] Rafał Weron. «Estimating long-range dependence: finite sample properties and confidence intervals». en. In: *Physica A* 312.1-2 (Sept. 2002), pp. 285–299 (cit. on p. 58).

[66] J A Wanliss and Grace E Wanliss. «Efficient calculation of fractal properties via the Higuchi method». In: *Nonlinear Dynamics* 109.4 (Sept. 2022), pp. 2893–2904 (cit. on p. 60).

[67] A Accardo, M Affinito, M Carrozzi, and F Bouquet. «Use of the fractal dimension for the analysis of electroencephalographic time series». en. In: *Biol. Cybern.* 77.5 (Nov. 1997), pp. 339–350 (cit. on pp. 60, 61).

[68] Milena Čukić, Miodrag Stokić, Slobodan Simić, and Dragoljub Pokrajac. «EEG machine learning with Higuchi's fractal dimension and Sample Entropy as features for successful detection of depression». In: *Cognitive Neurodynamics* 14.4 (Mar. 2020), pp. 443–455. DOI: 10.1007/s11571-020-09581-x. URL: https://doi.org/10.1007%5C%2Fs11571-020-09581-x (cit. on pp. 60, 62).

[69] Paolo Castiglioni, Sebastian Zurek, Jaroslaw Piskorski, Marcin Kosmider, Przemyslaw Guzik, Emiliano Ce, Susanna Rampichini, and Giampiero Merati. «Assessing Sample Entropy of physiological signals by the norm component matrix algorithm: Application on muscular signals during isometric contraction». In: *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. Osaka: IEEE, July 2013 (cit. on pp. 61, 62).

[70] Joshua S. Richman, Douglas E. Lake, and J.Randall Moorman. «Sample Entropy». In: *Numerical Computer Methods, Part E*. Vol. 384. Methods in Enzymology. Academic Press, 2004, pp. 172–184. DOI: https://doi.org/10.1016/S0076-6879(04)84011-4. URL: https://www.sciencedirect.com/science/article/pii/S0076687904840114 (cit. on pp. 61, 62).

[71] Jennifer M Yentes, Nathaniel Hunt, Kendra K Schmid, Jeffrey P Kaipust, Denise McGrath, and Nicholas Stergiou. «The appropriate use of approximate entropy and sample entropy with short data sets». en. In: *Ann. Biomed. Eng.* 41.2 (Feb. 2013), pp. 349–365 (cit. on p. 62).

[72] Aykut Erdamar, Fazıl Duman, and Sinan Yetkin. «A wavelet and teager energy operator based method for automatic detection of K-Complex in sleep EEG». en. In: *Expert Syst. Appl.* 39.1 (Jan. 2012), pp. 1284–1290 (cit. on p. 64).

[73] John M O'Toole, Andriy Temko, and Nathan Stevenson. «Assessing instantaneous energy in the EEG: A non-negative, frequency-weighted energy operator». In: *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. Chicago, IL: IEEE, Aug. 2014 (cit. on p. 64).

[74] Andreas C Müller and Sarah Guido. *Introduction to machine learning with Python : a guide for data scientists*. First edition. Sebastopol, CA: O'Reilly Media, Inc, 2017 (cit. on pp. 68, 74).

[75] Hanchuan Peng, Fuhui Long, and C. Ding. «Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy». In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27.8 (Aug. 2005), pp. 1226–1238 (cit. on pp. 69–71).

[76] Chris Ding and Hanchuan Peng. «Minimum redundancy feature selection from microarray gene expression data». en. In: *J. Bioinform. Comput. Biol.* 3.2 (Apr. 2005), pp. 185–205 (cit. on pp. 70, 71).

[77] *Maximum Relevance and Minimum Redundancy Feature Selection Methods for a Marketing Machine Learning Platform*. 2019 (cit. on pp. 70, 71).

[78] Alireza Kazemi, Martin J. McKeown, and Maryam S. Mirian. «Sleep staging using semi-unsupervised clustering of EEG: Application to REM sleep behavior disorder». In: *Biomedical Signal Processing and Control* 75 (2022), p. 103539. ISSN: 1746-8094. DOI: https://doi.org/10.1016/j.bspc.2022.103539. URL: https://www.sciencedirect.com/science/article/pii/S1746809422000611 (cit. on pp. 85, 110).

[79] Ryan J. Urbanowicz, Randal S. Olson, Peter Schmitt, Melissa Meeker, and Jason H. Moore. «Benchmarking relief-based feature selection methods for bioinformatics data mining». In: *Journal of Biomedical Informatics* 85 (2018), pp. 168–188. ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2018.07.015. URL: https://www.sciencedirect.com/science/article/pii/S1532046418301412 (cit. on pp. 87, 88).

[80] Ryan J. Urbanowicz, Melissa Meeker, William La Cava, Randal S. Olson, and Jason H. Moore. «Relief-based feature selection: Introduction and review». In: *Journal of Biomedical Informatics* 85 (2018), pp. 189–203. ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2018.07.014. URL: https://www.sciencedirect.com/science/article/pii/S1532046418301400 (cit. on p. 88).

[81] Cheng Li. «Fisher Linear Discriminant Analysis». In: 2014 (cit. on p. 91).

[82] Irene Rechichi, Antonella Iadarola, Maurizio Zibetti, Alessandro Cicolin, and Gabriella Olmo. «Assessing REM Sleep Behaviour Disorder: From Machine Learning Classification to the Definition of a Continuous Dissociation Index». In: *International Journal of Environmental Research and Public Health* 19.1 (2022). ISSN: 1660-4601. DOI: 10.3390/ijerph19010248. URL: https://www.mdpi.com/1660-4601/19/1/248 (cit. on p. 109).

[83] Christian O'Reilly, Isabelle Godin, Jacques Montplaisir, and Tore Nielsen. «REM sleep behaviour disorder is associated with lower fast and higher slow sleep spindle densities». In: *Journal of Sleep Research* 24.6 (June 2015), pp. 593–601. DOI: 10.1111/jsr.12309. URL: https://doi.org/10.1111/jsr.12309 (cit. on p. 109).