POLITECNICO DI TORINO

Master's Degree in ICT FOR SMART SOCIETIES

Master's Degree Thesis

N2 Stage Impact on REM Sleep Behavior Disorders Detection

Supervisors

Prof. Gabriella OLMO

Prof. Guido PAGANA

Dott. Irene RECHICHI

Dott. Gabriele S. GIARRUSSO

Candidate

Seyedeh Sadaf EKRAM

November 2024

Summary

Sleep is a biological, natural process in which the body and mind rest and recover themselves. During sleep, consciousness is suspended, the brain cycles through different stages, and vital physiological processes occur that support physical health, emotional well-being, and cognitive function. The sleep architecture also consists of a number of stages that uniquely contribute to the overall quality of sleep and neurological functions. Further, this cyclic sleep has been divided into Rapid Eye Movement (REM) and non-REM (NREM) stages, including N1, N2, and N3. Amongst these, REM sleep is of particular significance related to dreams and brain health. Disturbances in this stage give rise to specific sleep disorders, one type of parasomnia being the Rapid Eye Movement Sleep Behavior Disorder (RBD). It is characterized by the absence of normal muscle atonia in the REM stage of sleep; therefore, individuals tend to act out their dreams physically. This behavior may include talking, shouting, or other limb movements (can even include violent acts such as punching and kicking). RBD interferes with the quality of sleep for the patient but also carries the risk of injury for themselves and their bed partner. It is therefore an early warning of Parkinson's disease and Lewy body dementia, and for this reason, early detection is very important. Though the REM sleep stage has been directly implicated with RBD, recent works point out that stage N2 may also be an important stage to use in sleep disorder detection. This is because the N2 stage occupies a large part of total sleep time. It is characterized by specific Electroencephalography (EEG) features, such as sleep spindles and K-complexes, implicated in memory consolidation and sensory processing. Understanding how the N2 stage impacts RBD detection is relevant for a number of reasons. First, changes in the pattern of the N2 stage can form early signs of sleep disorders that precede or accompany RBD. A second point is that the inclusion of the N2 stage analysis might provide more diagnostic methodologies that are insightful for RBD, thus allowing earlier treatments. Lastly, the investigation into the relationship existing between the stages N2 and REM will provide better insight into the underlying mechanisms of RBD and its movement in relation to neurodegenerative conditions. The work done was extended to investigate the influence of the N2 sleep stage on RBD detection. We performed a more detailed analysis by first extracting 236 features from the time, frequency, time-frequency, and nonlinear metrics. From these, the 5 most important features were selected by the Minimum Redundancy Maximum Relevance (mRMR) method feature selector. We then used machine learning classifiers with methods such as K-Nearest Neighbors, Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, Kernel Support Vector Machine, and Gaussian Naive Bayes on features extracted from the EEG signals of all three N2, N3, and REM stages of sleep individually and in different concatenations. Our results indicated that overall accuracy is 70% when using only the N2 sleep stage and a Random Forest model, while sensitivity to the class RBD was 70%. Further narrowing down, by using just the REM sleep stage with logistic regression, the overall accuracy increased to 75%, with sensitivity for the detection of RBD reaching 100%. Notably, in combining N2 and REM stages with Logistic Regression, overall accuracy stands at 75%, but the sensitivity decreased somewhat to 90%. This combination, however, improved the specificity from 50%, as derived in using REM alone, up to 60% and greatly enhanced the AUC from 83% up to 93%. These results strongly underpin the inclusion of the N2 stage of sleep along with REM in the detection of RBD. Although sensitivity decreases slightly for the RBD class in adding N2 to REM, improvement in specificity and AUC indicates more balanced and robust diagnostic performance. Overall diagnostic methodologies are strengthened by the incorporation of the N2 stage analysis, thus possibly leading to earlier and more accurate interventions.

Table of Contents

List of Tables

List of Figures

Acronyms

NREM
Non-Rapid Eye Movement

REM
Rapid Eye Movement

EEG
Electroencephalography

TREM
Tonic Rapid Eye Movement

FREM
Phasic Rapid Eye Movement

RBD
REM Sleep Behavior Disorder

R&K
Rechtschaffen & Kales

AASM
American Academy of Sleep Medicine

SVM
Support Vector Machine

K-SVM
Kernel Support Vector Machine

SWS

Slow-Wave Sleep

OSA

Obstructive Sleep Apnea

PSG

Polysomnography

EOG

Electrooculogram

EMG

Electromyogram

ECG

Electrocardiogram

RWA

REM Sleep Without Atonia

AUC

Area Under The Curve

SViT

Spectral Vision Transformer

MLP

Multi-Layer Perceptron

RAI

REM Atonia Index

PD

Parkinson's Disease

DLB

Dementia Lewy Bodies

RF

Random Forest

K-NN

K-Nearest Neighbors

RUSBoost

Random Undersampling Boosted Trees

CNN

Convolutional Neural Network

NFLE

Nocturnal Frontal Lobe Epilepsy

TuSDi

Turin Sleep Disorders Database

LIME

Local Interpretable Model-Agnostic Explanations

SWA

Slow Wave Activity

Chapter 1

Introduction

1.1 Sleep and Its Functions

Sleep is a fundamental biologic process, important for physical and mental health, considering the altered state of consciousness, limited sensory activity, and reduced interaction with the environment. Sleep consists of various stages: non-rapid eye movements (NREM), each contributing in different ways to physiological and cognitive maintenance. NREM sleep, particularly in deeper stages, is important for physiological restoration, including tissue repair and healing, muscle growth, and immune strengthening. Rapid eye movement (REM) sleep is associated with active brain states similar to being awake and is especially important for consolidating memories, regulating emotion, and cognitive processing. The circadian rhythm and the homeostatic sleep drive are two fundamental biological processes of the sleep-wake cycle. While the circadian rhythm controlled by the suprachiasmatic nucleus in the brain times sleep in relation to day and night, The homeostatic process increases the need for sleep based on how long a person has been awake. Poor sleep, either through duration or disorders such as insomnia, seems to cause a poor outcome in health due to increased risks for cardiovascular diseases, metabolic disorders, immunodeficiency, and cognitive impairments. The area of sleep research currently investigates certain specific details at the cellular and molecular levels on the restorative function of sleep regarding brain plasticity and its relatedness to mental health [1].

Sleep needs and patterns change dramatically through the life course, from the highest needs and longest durations in infants and young children. Newborns require up to about 14 to 17 hours of sleep daily, whereas toddlers require about 11-14 hours/day. Throughout the life course, there are changes in sleep needs and patterns due to alterations in physiological factors and changes in hormonal levels, which vary somewhat by sex. As the child grows older, the need for sleep reduces gradually to about 9-11 hours for school-aged children and 8-10 hours for teenagers. Many teenagers, however, get less because of the demands of school and social life. Adult humans generally require about 7-9 hours of sleep, while older adults may require somewhat less, sleeping 6-7 hours, as changes in the sleep architecture lead to lighter and more fragmented sleep [2].

Sex differences in sleep tend to become more pronounced during adolescence onwards. Women, on average, reported a need for slightly more sleep compared with men and experienced more sleep disruptions, particularly those related to hormonal changes associated with menstruation, pregnancy, and menopause. Insomnia is more common among women, while men are at an increased risk of obstructive sleep apnea. Sleep quality and quantity can change across the lifespan depending on hormonal changes. Estrogen and progesterone have opposing effects on sleep cycles and are, therefore, important during times in which hormone levels significantly change [3].

1.2 Sleep Structure

Sleep architecture is generally characterized by the various stages of sleep the body passes through in an overall cyclical pattern. These are split between NREM and REM during sleep. NREM sleep is divided into three stages: N1, N2, and N3. Stage N1 is considered light sleep; it is a transitional period when the body enters through restfulness. Stage N2 is a relatively deeper stage of sleep, with distinctive brain wave patterns that promote the stability of sleep and integration of memories, including sleep spindles and K-complexes. Stage N3, otherwise known as deep sleep or slow-wave sleep (SWS), is critical for bodily restoration, immune function, and release of the growth hormone. After the body proceeds with these stages of NREM sleep, it then goes into REM sleep, which includes an increased brain activity just like during wakefulness and usually has vivid dreams. The body takes on atonia (a brief paralysis of its muscles) during this stage to prevent acting out of dreams. In a normal night's sleep, this cycle of NREM and REM repeats every 90 minutes throughout the night, with the duration of REM increasing and that of deep NREM sleep decreasing as the night progresses [4].

Stage N1, also known as the first stage of NREM sleep, reflects the initial transition in sleep to wake. The first stage of sleep is marked by reduced activity of the brain; heart rate and muscular activity also slow down at this stage, which is associated with relaxation. During N1 sleep, there is a shift in Electroencephalogram (EEG) from alpha waves (apparent representation of a relaxed yet awake condition) to slower theta waves (sleep is starting to occur). It is a very light stage where an individual can wake up rather easily. Importantly, the N1 stage is involved in memory processing and is frequently dominated by hypnagogic hallucinations or

dream-like experience that can incorporate elements from recent wakefulness.
Lacaux et al. [5] highlighted the very notion that sleeping during this stage could predict an increased tendency to forget recently formed memories. Moreover, the development of technologies for automatically staging sleep, such as those reviewed by Sun et al. [6] is improving the ability to accurately detect the N1 stage, which has been particularly difficult to detect considering its very short duration and variability in characteristics.

N2 is a transitional stage of sleep and forms a component of NREM, during which the heart rate slows down, together with muscle relaxation and sleep spindles observed on EEG. It would take about 45 to 55% of an average night's sleep and is considered vital for sleep continuity and memory processing. Specific to N2 sleep are the occurrences of sleep spindles, which are bursts of brain activity linked to memory consolidation, and K-complexes, which respond to external stimuli without waking the sleeper. Increased beta wave in N2 sleep is associated with reduced heart rate variability, suggesting heightened autonomic arousal, according to a study by Migliaccio et al. [7]. In a different study, Ma et al. [8] focused on the "first-night effect" and viewed instability in the stage of N2 as higher due to increased activity of the central nervous system as reflected by variation in heart rate variability. It logically and physically follows that this stage is crucial for restorative sleep or even for cognitive functions.

N3 sleep, also known as SWS or deep sleep, is a vital stage in NREM sleep, during which physical recovery connected to immune activity and memory consolidation takes place. The brain waves in N3 sleep are the slow delta waves that mark deep sleep. This is, therefore, an essential sleep stage in the processes of physical recuperation, such as the repair of tissues, growth of muscles, and strengthening of the immune system. It is well known that stage N3 is sensitive to disturbances in adults, with conditions like sleep apnea greatly reducing the duration of this stage and thus affecting the overall quality and health of an individual's sleep. Tseng et al. [9] showed that sleep apnea is one of the strongest reducing factors that diminish the duration of N3 sleep even more than other disorders, such as chronic tinnitus, pointing out the importance of this stage in restorative sleep. N3 is the most critical stage of embedding memory, where the brain arranges and stores newly accessed information.

REM stage involves intense brain activity and rapid movements of the eyes with vivid dreaming. The brain has patterns during REM sleep similar to those of wakefulness, and this stage is hypothesized to act as an emotional process, cognitive consolidation, and neural reorganization. Smith et al. [10] said that REM sleep is very contributory for motor learning, while NREM sleep, especially N2, contributes in other ways to learning and memory. This theory claims REM helps with the fine motor adjustments necessary to achieve skill acquisition and retention; hence, it is also involved in cognitive and motor readjustments. Brunner et al.

[11] studied the effects of partial sleep deprivation on different stages, including REM sleep. After its deprivation, the duration of this stage of sleep is increased correspondingly to compensate for it, marking the importance of REM in sleep architecture maintenance and physiological function. Feriante et al. [12] further explain that physiological markers of REM have included REM latency, which refers to the time between sleep onset and the first REM period, and REM density, expressing the frequency of eye movements within REM. These measures can signal neural health and responsiveness during the REM state and, in many instances, the intensity of the dream experience. The research of Rechichi et al. [13] deals with REM sleep's dual microstructure. She started the investigation with tonic REM(TREM) and phasic REM (FREM) stages. Her work investigates the nature of these subphases in REM sleep and their role in diagnostics with special attention to neurodegenerative diseases and Sleep Behavior Disorder (RBD). She and her colleagues highlight the different frequency ranges (2-8 Hz for FREM and 7-16 Hz for TREM) that will most likely enrich the feature extraction and classification algorithms on REM sleep and substructure detection, with a view to boosting diagnosis relating to sleep disorders and neurodegenerative conditions [14].

1.3 Sleep Scoring

The two major systems for sleep stage classification based on polysomnography (PSG) data, which encode the different stages of sleep, include the Rechtschaffen & Kales (R&K) and the American Academy of Sleep Medicine (AASM) sleep scoring methods; these differ in their complexity and applications. The classic R&K scoring system from 1968 divided sleep into five stages: REM and NREM stages 1, 2, 3, and 4. The AASM scoring system from 2007 further refines that approach by combining the deep sleep stages 3 and 4 into one stage called N3, updating detection and classification criteria for sleep patterns, especially around REM characteristics and NREM transitions. Each of these scoring methods has different utilities, and the revised guidelines from AASM are often preferred in the clinical setting because they are more straightforward and regularized, whereas the system of R&K is still useful in historical research contexts.

1.4 Sleep Frequency Bands and Their Relation to Sleep Stages

Sleep stages are defined by unique brainwave patterns measured through EEG, with each frequency band (delta, theta, alpha, beta, and gamma waves) serving specific roles during sleep.

Delta waves, the slowest brainwaves (0.5–4 Hz), are prominent during deep NREM sleep, particularly stage N3. These waves are essential for restorative functions such as physical recovery, memory strengthening, and tissue repair. Disruptions in delta wave activity have been linked to negative health outcomes, including cardiovascular diseases and higher mortality risks. Studies suggest that fragmented or reduced delta wave activity could predict long-term issues like coronary heart disease and all-cause mortality [15].

Theta waves, ranging from 4–8 Hz, emerge during the early stages of sleep, notably NREM stages 1 and 2. These waves play a vital role in transitioning from wakefulness to sleep, supporting long-term memory formation and brain synchronization, which organizes neural circuits for complex tasks. Research indicates that theta activity during sleep may reflect brief, localized sleep-like episodes in the brain during wakefulness, potentially impairing reaction times and cognitive processing when under sleep pressure [16].

Alpha waves, with frequencies between 8–12 Hz, are commonly associated with relaxation during wakefulness and the initial transition into sleep. They are most evident in NREM stage 1 but can appear during deeper sleep stages in a phenomenon known as alpha-delta sleep. This pattern, often seen in individuals with chronic pain, insomnia, or fibromyalgia, disrupts deep sleep by overlaying alpha waves onto the delta waves of NREM stage N3. The result is unrefreshing sleep. Studies show that alpha-delta sleep patterns vary across brain regions, with certain areas exhibiting more pronounced alpha activity [17].

Beta waves, faster brainwaves (12–30 Hz), are typically linked to wakeful alertness and active thinking. During sleep, beta activity may signal disturbances or heightened cognitive arousal. Excessive beta waves have been associated with conditions like Parkinson's Disease (PD), where they may disrupt SWS. Research using a primate model of PD found a correlation between increased beta wave activity and reduced SWS, highlighting the role of beta oscillations in sleep disorders associated with early-stage PD [18].

Gamma waves, the fastest brainwaves (30–120 Hz), are involved in attention, memory, and consciousness. These waves are also present during sleep, including SWS, where they align with the peaks of cortical slow waves. This suggests gamma waves contribute to processes like memory consolidation and off-line brain activity during sleep. Intracranial and scalp EEG recordings have revealed that gamma oscillations may support the brain's neural reinforcement during rest [19].

Each of these brainwave patterns contributes to specific aspects of sleep's restorative and cognitive benefits, underscoring the importance of maintaining healthy sleep cycles for overall physical and mental well-being.

1.5 Sleep Disorders

Sleep disorders encompass a variety of conditions that disrupt the quality, timing, or duration of sleep, negatively impacting daytime functioning and overall health. Common sleep disorders include:

- **Insomnia**: Difficulty falling or staying asleep, or non-restorative sleep, despite adequate opportunity. It is the most common sleep complaint, often associated with stress, mental health challenges, and chronic social or environmental factors. Insomnia is more prevalent in women, older adults, and people at social risk. Modern classification combines various causes under the term "insomnia disorder," emphasizing its complex nature. Spielman's model explains chronic insomnia as a combination of predisposing, precipitating, and perpetuating factors, guiding ongoing research for targeted treatments [20].
- **Sleep Apnea**: Particularly obstructive sleep apnea (OSA), characterized by repeated airway blockages during sleep, leading to interrupted breathing and oxygen deprivation. Closely linked to obesity and metabolic disorders, untreated OSA increases the risk of cardiovascular diseases, cognitive impairments, and other health concerns. Awareness and early diagnosis are crucial to mitigate its widespread effects [21].
- **Narcolepsy**: A REM sleep disorder causing excessive daytime sleepiness, sudden sleep attacks, cataplexy (sudden loss of muscle tone), vivid hallucinations, and sleep paralysis. Early diagnosis through PSG and Multiple Sleep Latency Testing (MSLT) is key to reducing its impact and improving outcomes [22].
- **Restless Legs Syndrome (RLS)**: A sensorimotor condition characterized by an irresistible urge to move the legs, often accompanied by discomfort. Symptoms typically occur during rest and are relieved by movement. RLS can lead to sleep-onset and sleep-maintenance insomnia and is linked to brain iron deficiency and dopaminergic abnormalities. ICU patients are particularly prone due to immobility, iron deficiencies, and medication effects [23].
- **Periodic Limb Movement Disorder (PLMD)**: Repetitive limb movements during sleep, primarily in the lower extremities, disrupt sleep quality and contribute to daytime fatigue and excessive sleepiness. Movements can also occur in the upper limbs in some cases [24].
- **Circadian Rhythm Sleep Disorders (CRSD)**: Chronic or recurrent misalignment between the internal sleep-wake cycle and societal schedules. Subtypes include Advanced Sleep Phase Syndrome (ASPS), Delayed Sleep Phase Syndrome (DSPS), Free Running Disorder (FRD), and Irregular Sleep-Wake Rhythm (ISWR). Management often involves behavioral modifications, light therapy, and chronotherapy to realign the sleep schedule [25].

- **Parasomnias**: These involve unusual activities during sleep or transitions between sleep and wakefulness.
    - **NREM Parasomnias**: Include sleepwalking, sleep terrors, and confusional arousals, typically occurring in deep sleep. Sleepwalking may involve simple or complex actions, while sleep terrors are marked by sudden arousal with intense fear and autonomic activation.
    - **REM Parasomnias**: Include RBD, where individuals act out dreams due to the absence of normal muscle atonia, and nightmares, which involve vivid, distressing dreams that awaken the sleeper. Other parasomnias, like sleep paralysis and sleep-related hallucinations, occur at the sleep-wake boundary..

  Parasomnias can stem from genetic predisposition, stress, sleep deprivation, or medication use. Diagnosis often involves detailed sleep history and PSG to monitor episodes. Treatment strategies vary and may include lifestyle changes, safety measures, behavioral therapy, or medication under expert guidance [26].

1.6 RBD

RBD is a parasomnia, an undesired event occurring during sleep, and a condition in which normal atonia during REM sleep is absent and individuals act out their dreams physically. The absence of atonia is manifested through vocalizations, limb movements, and complex behaviors portraying the dream content; these events can lead to injury to the individual or their bed partner [27]. While sleepwalking and night terrors occur during the stage of NREM, RBD is specific to sleep in the stage of REM and is characterized by muscle activity when atonia would be expected to take place, confirmed by PSG [27][28].

People with RBD report having bright dreams associated with behaviors of shouting, talking, punching, kicking, or other violent movements. These tend to happen along with the later stages of sleep when REM stages become more frequent during the night. Injuries range from minor contusions to significant fractures and thus eminently affect the safety and well-being of affected individuals and their partners [29][30].

RBD results from dysfunction in brainstem regions, including the subcoeruleus and magnocellular nuclei, regulating REM muscle atonia. The pons and medulla of the brainstem counteract atonia by means of inhibitory neurotransmitters such as GABA and glycine. Disruptions in these regions result in excessive motor activity during REM sleep that characterizes RBD [27]. Diagnosis is based on PSG to document REM Sleep Without Atonia, RWA, and clinical evaluations of dream enactment behaviors. The PSG usually demonstrates muscle activity during REM sleep, thus confirming the diagnosis. RBD may, however be difficult to diagnose as some symptoms tend to overlap with other parasomnias, and it is possible that PSG will fail to capture episodes during the test period. In addition, mildly affected individuals with subtle symptoms may remain undiagnosed, indicating the need for refined diagnostic criteria or longer monitoring [29][30].

RBD usually constitutes an early warning sign of the development of certain neurodegenerative disorders, such as Parkinson's disease and Dementia with Lewy Bodies, predating other symptoms for several years. Based on this relationship, early intervention and follow-up possibilities are opened. Early recognition of RBD can be very useful for early treatment strategies that may delay or manage the evolution of such neurodegenerative disorders. The management of symptoms of RBD also enhances safety and improves the quality of life of the patients [31][28]. Drugs that could modulate REM sleep mechanisms would allow more specific diagnosis and earlier detection of vulnerable populations should improve treatment outcome. Research into neural pathways and neurotransmitter systems regulating REM sleep might point to novel therapeutic targets for RBD and related conditions. Pharmacological options are a general RBD treatment, and clonazepam has proven very effective in producing a reduction of both frequency and intensity of dream enactments. However, it may cause side effects such as cognitive impairment, particularly in older adults. Melatonin is another common treatment; it is used alone or in combination with clonazepam. Also, changes need to be carried out in environmental conditions such as removal of sharp objects and padding around the bed to reduce risks of injury [31].

Equally important in the management of RBD are lifestyle modifications. Modifications to the sleeping environment, such as padded bed railings, removal of hazardous objects, and sleeping separately when necessary, greatly minimize injury. In providing realistic expectations and reducing anxiety, education regarding RBD to both the patient and partner is also significant. These are particularly helpful for those who have less access to medical treatments [30].

1.7 Polysomnography

Polysomnography, commonly referred to as a sleep study, is a detailed diagnostic test used to identify sleep disorders by monitoring various physiological functions during sleep. These functions include brain activity, oxygen levels, heart rate, breathing patterns, and muscle movements. The test can be conducted in a controlled environment, such as a sleep lab, or at home using portable equipment, and it is invaluable for diagnosing conditions like sleep apnea, insomnia, narcolepsy, restless legs syndrome, and RBD.

PSG evaluates sleep health through multiple components, including:

- **Electroencephalogram (EEG)**: Tracks brain wave activity to determine sleep stages (REM and NREM) and overall sleep architecture.
- **Electrooculogram (EOG)**: Monitors eye movements to identify REM sleep, characterized by rapid eye activity, providing insights into sleep cycles.
- **Electromyogram (EMG)**: Measures muscle activity, particularly in the chin and legs, to assess muscle relaxation (atonia) during REM sleep and detect involuntary movements linked to disorders like restless legs syndrome.
- **Electrocardiogram (ECG)**: Captures heart rate and rhythm to detect irregularities during sleep.
- **Respiratory Measurements**: Sensors placed on the chest, abdomen, and nostrils monitor breathing effort, airflow, and blood oxygen levels, crucial for diagnosing respiratory disorders like obstructive sleep apnea (OSA).
- **Pulse Oximetry**: Measures blood oxygen saturation to identify hypoxemia often associated with breathing disturbances.
- **Snoring Sound Recording**: Identifies snoring patterns that may indicate airflow blockages and potential sleep apnea.
- **Body Position and Movement Tracking**: Evaluates whether specific sleep positions, such as lying on the back, exacerbate symptoms like positional sleep apnea.

By providing a comprehensive view of sleep architecture and identifying disruptions, PSG helps clinicians develop tailored treatment plans. Based on the diagnosis, solutions may include CPAP therapy for managing sleep apnea, medications for conditions such as narcolepsy or restless legs syndrome (RLS), or cognitive behavioral therapy (CBT) for insomnia [32][33][34][35].

Chapter 2

Literature Review and Objective

2.1 Literature Review

Traditional RBD diagnosis has been done with clinical questionnaires, analysis of PSG, and behavioral assessments. Although these methods are reliable, they are beyond resource (intensive and unsuitable for large-scale screenings or early detection) in particular in view of the possible role of RBD as a prodromal symptom of neurodegenerative diseases such as PD and dementia combined with DLB.

To this effect, Bugalho et al. [36] investigated the prevalence and phenomenology of RBD in ET patients. In this study, RBD Screening Questionnaire (RBDSQ) was employed in the screening phase and video-PSG amongst those who tested positive. Subjects were categorized into the following groups: ET with RBD, ET without RBD, Parkinson's Disease with RBD (PD-RBD), and Idiopathic RBD (iRBD). This study noted some features of sleep and motor events during sleep that may indicate a link between disorders of ET and alpha-synucleinopathies, and RBD in patients with ET may be characterized by these various changes.

From traditional approaches to recent contributions, automated detection of sleep disorders (but especially RBD) has increasingly become a focal interest in healthcare. Many sleep disorders are still undiagnosed since continuous monitoring is, unfortunately, rather complex and challenging; hence, an automated system that is able to analyze normal sleep behavior over time may provide early detection without intrusive observation and offer crucial advantages for the detection of RBD as an early marker of neurodegenerative diseases. Different studies have used various data modalities and machine learning to generate automated methods of RBD detection. Abdelfattah et al. [37] used regular 2D video data obtained from in-laboratory video-PSG for the diagnosis of RBD. They observed that, through the

analysis of optical flow in video recordings to identify movement patterns during REM sleep, it would be possible to develop a machine learning classifier with sensitivity of 0.921 and specificity of 0.674, with improvements when additional features such as gender and sleep metrics were included to an accuracy of 0.886.
Adaimi et al. [38] also investigated the use of computer vision techniques for detecting RBD from body movements captured during PSG. The study classified sleep behaviors using a Multi-Layer Perceptron model with background subtraction and advanced methods of action recognition. This approach yielded an accuracy of 91.9%, with sensitivity at 78.3% and precision of 100%, highlighting automated video analysis as a potential diagnostic tool.
Meanwhile, an SViT model was proposed in parallel by Gunter et al. [39] for the detection of RBD from PSG data via the transformation of the EEG, EOG, and EMG signals into spectral images. The deep learning model was able to achieve an F1 score of 0.93 outperforming traditional CNN models with validation for EEG and EOG channels.
Focusing on comparative methods, Cesari et al. [40] evaluated various automated approaches that detect RWA, the hallmark of RBD. Comparing metrics such as the REM Atonia Index (RAI) and supra-threshold REM activity, the study assessed effectiveness across configurations and showed that RAI was particularly sensitive for the detection of RBD, but no single method was optimal across all settings. This study underlined the difficulty of automatic diagnosis of RBD and pointed out the need to work on fine-tuning the methods for better diagnostic accuracy.
Cooraya et al. [41] went a step ahead by using automated sleep staging to enhance the detection of RBD automatically using a Random Forest classifier trained with combined EEG, EOG, and EMG signals. The methodology reached an accuracy of up to 96% using manually annotated data and up to 92% when it was part of a fully automated sleep staging algorithm, again pointing to a great value of sleep architecture for differential diagnosis of RBD from healthy conditions.
Further to the above, Papakonstantinou et al. [42] introduced the Ikelos-RWA algorithm as a means of automatic quantification of RWA. Compared to the visual scoring approaches, this algorithm yielded great sensitivity and specificity, allowing for as high as 0.98 AUC to be reached, thus validating its robustness for RBD diagnosis.
Building on automated detection techniques, studies have also turned to single-channel EEG data as a simpler and easily available diagnostic alternative. Rechichi et al. [13] discussed single-channel EEG sleep stage classification, with emphasis on the microstructures of REM sleep. In this study, the authors added several new features used from the phasic and tonic microstates of REM to further improve the sleep stage classification especially between REM and NREM stages. It sets up a high accuracy of the REM detection using machine learning algorithms, including RF, KNN, and RUSBoost methods. The best performance using RF was as high as

92.7%. This simplifies PSG and will further enable at-home sleep monitoring with fewer sensors, thus serving as a promising application for tracking sleep disorders.

Rechichi et al. [14] continued single-channel EEG analysis in a follow-up study focused on the features of EEG from REM and SWS stages, reaching an accuracy of 86% and sensitivity of 91% when using SWS features. These findings show the high potential of non-invasive and low-cost tools as early RBD screening tools, which are quite practical compared to tradition PSG and thus allow for earlier diagnosis.

Giarrusso et al. [43] extended these using machine learning on single-channel EEG segments within a leave-one-out cross-validation framework. In the study, high accuracy in classifying RBD could reach as high as 91%, with sensitivity of up to 94%. A semisupervised method was also proposed in this study, based on the characterization of RWA, to assist early diagnosis and access to preventive therapy.

Buettner et al. [44] introduced a new algorithm that reduced recording times of EEGs; the classification accuracy did not fall below the high range of above 90%. For diagnoses, overnight EEG data is traditionally needed. It succeeded in classifying over 90% in a 10-min snip-pet of EEG. The approach focused on fine-grained EEG frequency bands, namely the delta, theta, and alpha frequency range, thus probably opening a new avenue toward more efficient diagnostics for early neurodegeneration associated with RBD. A systematic review underlined the importance of RBD as a prodromal marker.

Galbiati et al. [45] study synthesized data from longitudinal studies to estimate the conversion rate from RBD into neurodegenerative diseases, mainly synucleinopathies like DLB. The results indicated an impressive conversion rate: more than 90% of patients suffering from RBD developed progressive neurological problems within 14 years. These insights underline the need for early interventions.

Machine learning has also been applied to predict conversion from isolated RBD to neural degenerative conditions [46]. This research analyzed PSG-derived features of EEG and underlined the importance of the EEG features, especially the features indicating EEG slowing in REM sleep, such as increased theta power. Consequently, high accuracy of the prediction allowed new avenues for the intervention strategies.

The literature in other words indicates an emerging emphasis; there's an emerging spotlight on automatic and machine learning-based methodological approaches in the light of multi-modalities like video, PSG, single-channel EEG recordings for the detection and prediction of RBD. These innovations can pave a way for better early diagnosis and early intervention, considering the good association of RBD with neurodegenerative conditions. Thus, the establishment of accessible and effective diagnostic tools is of primary importance for enhancing quality in everyday life for patients as well as for improving our current understanding of prodromal neurodegeneration stages.

2.2 Objective

The core discussion in this work will revolve around how the inclusion of the N2 sleep stage affects RBD detection. By not excluding features related to N2 stage EEG patterns, sleep microstructures, and other physiological signals, we want to enhance the accuracy and robustness of the automated RBD detection methods further. This aims to devise a more powerful non-invasive diagnostic tool that could allow for early detection of RBD and thus timely interventions for the associated neurodegenerative diseases like Parkinson's Disease and Dementia with Lewy Bodies. Current state-of-the-art diagnostic tools for RBD include clinical questionnaires and PSG, which are resource-intensive and not suited for large-scale screening or early detection. While automated methods have been promising in doing so, many of these approaches focus on the data of REM sleep and may underappreciate valuable information that could come from other stages of sleep. The N2 stage constitutes a large portion of the sleep cycle and, if adequately analyzed, might contain subtle cues to RBD, enhancing detection. As RBD is a strong prodromal marker of neurodegenerative diseases-on follow-up studies, over 90% of the patients suffering from RBD have been shown to develop progressive neurological disorders within a period of 14 years-developing better detection capabilities right from an early stage is extremely important. The incorporation of data from the N2 stage in this model may overcome some limitations in the automated detection techniques by:

- **Improvement in Accuracy**: The feature from the N2 stage may enable the machine learning models to better detect healthy sleep patterns from those indicative of RBD.
- **Less Stress on Resources**: Since effective markers could be identified in the N2 stage, diagnostics might be more available because heavy overnight monitoring would not be necessary.
- **Early intervention**: The enabling of early diagnosis using better methods of detection, which may well allow timely therapeutic interventions that slow the disease process.

This will be an important phase in the further development of practical, low-cost diagnostic methods of RBD. This thesis tries to make a worthier contribution not only for clinical practice but also for a greater understanding of sleep disorders as early harbingers of neurodegenerative diseases by venturing into hitherto unexploited potential of the N2 sleep stage.

Chapter 3

Methodology

In this chapter, we will present exhaustive methods to study the effect of different stages of sleep, more importantly, the N2 stage on RBD. Apart from that, we have also carried out binary classification tasks that would distinguish between healthy controls from patients suffering from RBD. Rather than testing our models in isolation, we developed an end-to-end pipeline that considers every step of the process from raw inputs to final validation. Figure 3.1 shows our entire workflow to ensure transparency and ease of reproduction.

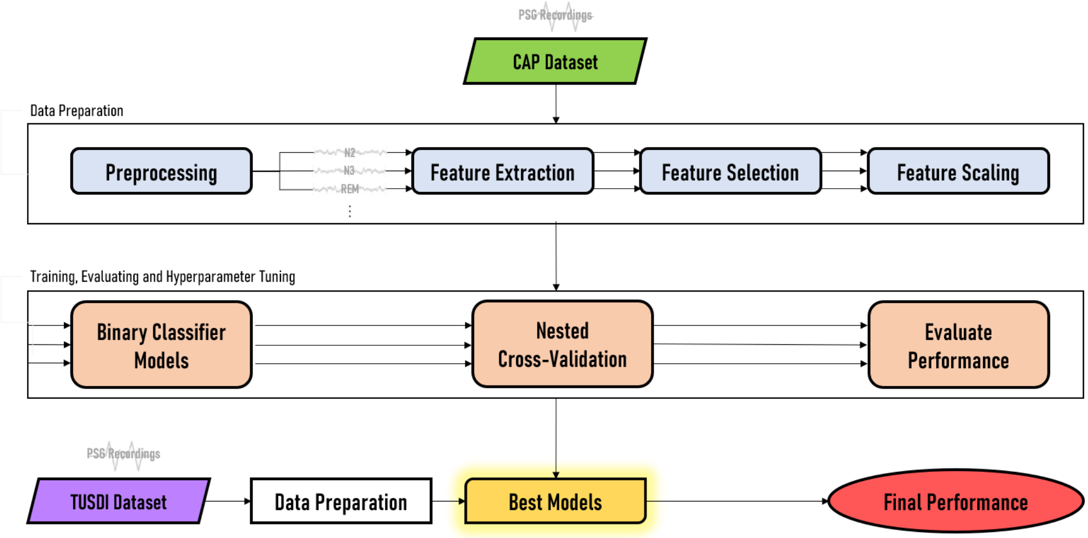


Figure 3.1: Overview of the methodology pipeline.

Our method is organized as follows: preparation of data (PSG recordings of the CAP dataset); cleaning of the signals; selection of an EEG channel; management

of annotations; extraction of meaningful features for different sleep stages (N2, N3, REM, etc.) are described in Sections 3.1, 3.2, 3.3, 3.4, and 3.5. We carefully select the features in order to retain characteristics describing the essential difference between healthy and RBD patients during the various sleep stages, while simultaneously filtering out noise and redundant information.
Next, we develop and refine our binary classification models (Section 3.6). In Section 3.7 we revisit several strategies for model training and performance evaluation. To optimally use our relatively small dataset without overfitting, we apply the technique of nested cross-validation. This provides an avenue to conduct extensive training for different machine learning algorithms and tuning parameters while maintaining that all reviews are done fairly. We present the performance reports of each of our models using metrics, which are explained in great detail in Section 3.8. Finally, we validate our best performing models on the TuSDi dataset. We apply identical preparation steps to this new data in order to ensure comparison on fair terms. This external validation will help us to see how our models will perform in real scenarios and at other centers beyond their training data. In the following methodology section, we shall be addressing theoretical concepts and practical challenges we have hitherto encountered during the analysis.

3.1 Data

In the present work, two different datasets are used. The CAP Sleep Database [47] is an updated collection of 108 PSG recordings acquired at the Sleep Disorders Center of Ospedale Maggiore of Parma, Italy. Each recording represents a variety of physiological signals, including at least three EEG channels: F3/F4, C3/C4, and O1/O2, referenced to A1/A2; two EOG channels; EMG signals from the submentalis muscle; bilateral anterior tibial EMG; respiratory data-airflow, abdominal, and thoracic effort, besides Oxygen Saturation (SaO2); and an ECG channel. Additional EEG bipolar recordings according to the 10-20 international system include Fp1-F3, F3-C3, C3-P3, P3-O1 and/or Fp2-F4, F4-C4, C4-P4, P4-O2. The following 16 records were from healthy individuals without evidence of any neurological disorders and/or intake of any medication influencing the central nervous system. The rest 92 recordings are from patients diagnosed with different sleep disorders: 40 recordings from individuals with Nocturnal Frontal Lobe Epilepsy (NFLE), 22 from patients experiencing RBD, 10 from those with Periodic Limb Movements (PLM), 9 from insomniac patients, 5 from individuals diagnosed with narcolepsy, 4 from patients experiencing Sleep-disordered Breathing (SDB), and 2 recordings from individuals with bruxism. This is a diverse dataset, organized in such a way to enable the study of different sleep conditions and patterns, making it a strong resource for research in sleep disorder diagnosis and analysis. Various recordings exist within

the CAP Sleep Database in a .edf format with detailed annotations, making it a very interesting set of signals for developing and testing algorithms in sleep research.
TuSDi is a sleep disorder database that includes 20 PSG recordings acquired at the Center for Sleep Disorders at Molinette Hospital in Turin, Italy. TuSDi is a representation of great variety in physiological channels. Physiological signals record neural, muscular, and cardiorespiratory functions. In this respect, EEG channels correspond to six derivations, including references to M1 or M2 and are represented as follows: F3-M2, C3-M2, O1-M2, F4-M1, C4-M1, and O2-M1. Additionally, eye movements are recorded by the EOG channels E1-M2 and E2-M2, while the muscular activity is recorded by the left and right eminence. Cardiac activity is picked by the ECG channel. The additional muscular activity in the lower limbs was recorded with four channels for the bilateral tibialis anterior muscles, namely Gamba-L-0, Gamba-R-0, Gamba-L-1, and Gamba-R-1. The respiratory signals are airflow measurement (Fluss), thoracic effort (Tor), abdominal effort (Addo), and carbon dioxide level (Russ) measurements. Additional physiological signals are provided: blood oxygen saturation (SpO2), heart rate frequency (FreqCard), pulse wave (Pleti), and body position (Pos). In total, there are 20 available records: 10 healthy subjects and 10 RBD subjects, in .edf format. This dataset consists of a manifold set of physiological signals supported by annotations in .txt files, providing real value.

3.2 Preprocessing

Preprocessing is a typical process in every EEG analysis cycle. This ensures structuring of the data in a consistent manner, free from unwanted noises, and ensures that accurateness and meaningfulness are achieved in analysis. In this section, we go into detail with respect to the steps taken to preprocess the obtained EEG data: time cropping, channel selection, resampling, epoching, unit standardization, synchronization with sleep stage annotation, and filtering. As a matter of fact, each step in the pipeline is specially set toward preparing the EEG signals for downstream analysis with respect to the detection of RBD.

3.2.1 Time Cropping of Initial Periods

The first segments of each recording contain no EEG data either because of setup procedures or delay in the start of recording and thus must be deleted to avoid any necessity for artifacts introduction into analysis. Times for start and end of each recording are calculated with the help of information about metadata extracted from the file.

3.2.2 Selection of Relevant EEG Channels

The recordings in the EEG usually involve several channels, each of which represents a different part of the head. Only some of the channels were selected for the analysis: C4-A1 or C3-A2; these are among the widely used references in electrode montage in studies on sleep since the Researchers have continuous, unchanging views of the whole sections of this brain activity in sleep.

3.2.3 Resampling of Data

It will be very instrumental in making proper comparisons and spectral analyses if all the EEG recordings have one common sampling frequency. All the recordings in this dataset were acquired under different sampling rates, so a target of 512 Hz was chosen for resampling. This resampling process at a common rate minimizes the disparities across recordings and allows each EEG sample to represent a uniform interval in time.

3.2.4 Epoching and Event Creation

Epoching is the method by which continuous EEG data are divided into discrete, nonoverlapping time windows. It represents a necessary structure in sleep studies, where an epoch length of 30 seconds is considered standard. Each epoch highlights a brief, meaningful aspect of brain activity that can be assigned to an individual sleep stage.

3.2.5 Standardization of Data Units

Since volt, millivolt, or microvolt measurement ranges were applied to the recording, the actual range of each recording was studied. After studying the above fact, all data was unified into µV (microvolts) by using that inference.

3.2.6 Handling Dropped Epochs

Some epochs can be preprocessed as "dropped" either due to excessive noisiness or due to artifact contamination. Dropped epochs are listed in order for better alignment with sleep stage annotations. The program records which epochs are discharged with the goal of maintaining only high-quality epochs that could enhance the overall reliability of results.

3.2.7 Hypnogram Loading and Synchronization with EEG Data

The hypnogram is the sleep stages assigned to every epoch, which becomes quite important in linking the EEG data with its specific stage of sleep. We extracted the stages of sleep from the annotation files. The original annotation is according to the R&K standard. Stage 4 in R&K was converted to be labeled as stage 3 in accordance with modern standards, following AASM scoring guidelines. If the gaps between recorded sleep stages reflected the number of missing epochs, the temporal gaps were used to interpolate them accordingly. For example, if the gap reflected one or two missing epochs (30 or 60 seconds intervals), they were interpolated to align with the corresponding EEG data. The hypnogram was then aligned with the EEG epochs, which were cleaned and prepared previously - that is, all the dropped epochs were removed. This step created a synchronized dataset where each EEG epoch directly mapped to a sleep stage. The result was a tidy array for further analysis.

3.2.8 Filtering of EEG Data

In order to eliminate irrelevant noise and other possible artifacts, a bandpass filter was applied on the EEG data. Low-frequency drifts (artifacts) and high-frequency noise in EEG signals may damage sleep stage classification. In this work, a Chebyshev filter is used, which represents an IIR filter with the most sharp cutoffs and minimal passband ripple:

- High-Pass Filter: A high-pass filter of cutoff 0.3 Hz was applied to remove low-frequency drifts, such as those due to perspiration, movement, or electrical interference that is not a reflection of neural activity.
- Low-Pass Filter: A low-pass filter at 30 Hz cutoff was applied for reducing high-frequency noise. Including muscle artifacts and electrical noise.

First, it removed slow drifts using the high-pass filter, then the low-pass filter to reduce high frequencies; thus, only frequencies within the 0.3–30 Hz range, the bandwidth for sleep-related brain activity, remained in the filtered signal.

3.3 Feature Extraction

The EEG signal is naturally non-stationary and incredibly complex, reflecting the dynamic activity of the brain. The direct performance of analysis without the extraction of features may require very intensive computational procedures and might not express all underlying physiological patterns. Feature extraction is a necessary process in the transformation of raw EEG signals into more interpretative, computationally manageable forms suitable for a particular analytical or predictive task. In this work, 13 polysomnographic features and 236 electroencephalographic features were extracted. For each electroencephalographic feature, 4 statistical measures were computed: mean, standard deviation, 75th percentile, and mode. At the end, that summed up to a total of 957 features.

3.3.1 Polysomnographic Features

PSG features are the quantitative measures by which a comprehensive recording technique adopted to evaluate sleep physiology and disorders is derived. In fact, these reflect the multidimensional composition of sleep architecture and continuity and disruptions, therefore providing information not only about quality sleep but also about underlying mechanisms for sleep life conditions [14].

- **Sleep Onset Latency (SOL)**: The time (in minutes) to transition from wakefulness to the first stage of sleep after lights out.
- **Wake After Sleep Onset (WASO)**: Total time (in minutes) spent awake after initially falling asleep, reflecting fragmented sleep.
- **Total Sleep Time (TST)**: The cumulative duration of all sleep stages (in hours), representing the actual sleep achieved.
- **Time In Bed (TIB)**: Total duration (in hours) from going to bed to final awakening, including periods of wakefulness.
- **Sleep Efficiency (SE)**: A ratio of TST to TIB, expressed as a percentage, assessing how efficiently time in bed is spent sleeping.
- **Arousal Index (ARI)**: Frequency of arousals per hour of sleep, indicating sleep fragmentation and disturbances.
- **REM Sleep Episodes (MREM)**: Total duration (in minutes) spent in REM sleep, a critical stage for memory and mood regulation.
- **Sleep Stage Proportion (SSP)**: Percentage of TST spent in each sleep stage (NREM stages 1-3 and REM), offering insights into sleep architecture.
- **NREM Fragmentation Index (NFI) & REM Fragmentation Index (RFI)**: Metrics assessing interruptions within NREM and REM stages, respectively.
- **Wake Proportion (WP)**: The fraction of time spent awake relative to TIB, highlighting sleep efficiency and disturbances.

- **Sleep Transitions Index (STI)**: The frequency of transitions between different sleep stages or from sleep to wake, reflecting stability of sleep stages.
- **Average Segment Length (ASL)**: Mean duration of continuous periods in a specific sleep stage, providing insights into the consolidation of sleep stages.

3.3.2 Electroencephalographic Features

Electroencephalographic features are extracted from EEG signals of either a normal physiological state or a pathological one. These are features determined through computational and mathematical techniques that analyze electrical activities of the brain as recorded from electrodes placed on the scalp. The main categories of classification of EEG features include the time domain, frequency domain, time-frequency, and nonlinear dynamics.

Time Domain

Time domain features are those statistical and signal-based features that, given a time series signal $x(t)$ $(t = 0, \ldots, N-1$, where N represents the number of samples in an epoch, have been extracted without any necessary transformations in the frequency domain. Typical applications of time domain features include signal processing, machine learning, and several others related to the characterization and pattern or anomaly detection in signals. Basic statistical features are the most fundamental time-domain features, like:

- **Mean**: The average value of the signal, indicating its central tendency.

$$\text{Mean} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{3.1}$$

- **Mode**: The value that appears most frequently in the signal. Useful for understanding common patterns.
- **Median**: : The middle value of the sorted signal, dividing it into two equal halves.

  If N is odd:

$$\text{Median} = x_{\frac{N+1}{2}} \tag{3.2}$$

  If N is even:

$$\text{Median} = \frac{x_{\frac{N}{2}} + x_{\frac{N}{2}+1}}{2} \tag{3.3}$$

- **Variance**: Measures the spread of signal values around the mean.

$$\text{Variance} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \text{Mean})^2 \tag{3.4}$$

$$\text{Activity} = \text{Variance} \tag{3.5}$$

$$\text{Standard Deviation(SD)} = \sqrt{\text{Variance}} \tag{3.6}$$

- **Skewness**: Indicates asymmetry in the signal's distribution.

$$\text{Skewness} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \text{Mean}}{\text{SD}}\right)^3 \tag{3.7}$$

- **Kurtosis**: Measures the "peakedness" or "flatness" of the signal distribution.

$$\text{Kurtosis} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \text{Mean}}{\text{SD}}\right)^4 \tag{3.8}$$

- **Maximum**: The largest value in the signal.

$$\text{Maximum} = \max(x_i) \tag{3.9}$$

- **Minimum**: The smallest value in the signal.

$$\text{Minimum} = \min(x_i) \tag{3.10}$$

- **Range**: The difference between the maximum and minimum values.

$$\text{Range} = \text{Maximum} - \text{Minimum} \tag{3.11}$$

Temporal features describe the signal's dynamics over time, offering insight into its variability and structural complexity, like:

- **Zero-Crossing Rate**: Measures how often the signal crosses zero.

$$\text{ZCR} = \sum_{i=1}^{N-1}\mathbf{1}\left(x_i \cdot x_{i+1} < 0\right) \tag{3.12}$$

- **Mobility**: This is a measure of how quickly the amplitude of a signal changes over time. It gives an indication of the frequency content of the signal.

$$\text{Mobility} = \sqrt{\frac{\text{var}\left(\frac{dx(t)}{dt}\right)}{\text{var}\left(x(t)\right)}} = \sqrt{\frac{m_2}{m_0}} \tag{3.13}$$

- **Complexity**: Describes how detailed or irregular a signal is compared to a simple waveform like a sine wave. It reflects the richness of the signal's structure, considering how its variations differ from smoother forms.

$$\text{Complexity} = \frac{\text{Mobility}\left(\frac{dx(t)}{dt}\right)}{\text{Mobility}(x(t))} = \sqrt{\frac{m_4/m_2}{m_2/m_0}} \tag{3.14}$$

- **Spectral Bandwidth**: This formula uses the variances of the signal and its derivatives to approximate the frequency spread without explicitly transforming to the frequency domain. A higher spectral bandwidth indicates a broader spread of frequencies in the signal.

$$\text{Spectrum Bandwidth} = \sqrt{1 - \frac{m_2^2}{m_0 \cdot m_4}} \tag{3.15}$$

- **Average Zero-Crossing Period**: This is the average time between consecutive zero-crossings of the signal. It provides an indication of the signal's dominant frequency.

$$\text{Average Zero-Crossing Period} = 2\pi \cdot \sqrt{\frac{m_0}{m_2}} \tag{3.16}$$

- **Energy**: The total energy of the signal.

$$\text{Energy} = \sum_{i=1}^{N} x_i^2 \tag{3.17}$$

- **Sliding Integrated Signal Energy**: combines the signal energy with an approximate measure of velocity, derived over a sliding window. It provides a way to evaluate how dynamic or energetic the signal is within a localized time window.

$$\text{Sliding Integrated Signal Energy} = \sum_{t=1}^{N} |x(t)^2| \cdot \nu \tag{3.18}$$

Shape features are another important category, focusing on the geometric properties of the waveform, like:

- **Sparseness**: Quantifies how the signal energy is distributed over its amplitude. It essentially measures whether the signal is concentrated over a few values or spread over a broader range. Higher sparseness indicates that the signal has less spread and may consist of sharper, more distinct peaks.

$$\text{Sparseness} = \frac{m_0}{\sqrt{(m_0 - m_4)(m_0 - m_2)}} \tag{3.19}$$

- **Irregularity**: Measures the degree of unpredictability or variability in the signal structure. It is based on higher-order moments of the signal and provides insight into the chaotic nature of the signal.

$$\text{Irregularity} = \frac{m_2}{\sqrt{m_0 \cdot m_4}} \tag{3.20}$$

- **Peak-to-Peak Period**: Offers a way to quantify the dominant time-scale of a signal using its statistical moments. It helps bridge the time and frequency domains by providing a measure of oscillation frequency using signal variability.

$$\text{Peak-to-Peak Period} = 2\pi \cdot \sqrt{\frac{m_2}{m_4}} \tag{3.21}$$

- **Form Factor**: Helps evaluate how spiky or irregular the waveform is. A higher value indicates more variations (higher peaks relative to the mean).

$$\text{Form Factor} = \frac{x_{\text{RMS}}}{|x|_{\text{mean}}} \tag{3.22}$$

- **Crest Factor**: Indicates the sharpness of the waveform. A high Crest Factor suggests the presence of sharp spikes or high peaks relative to the average power of the signal. Useful in determining transient characteristics or the impulsiveness of a signal.

$$\text{Crest Factor} = \frac{x_{\text{peak}}}{x_{\text{RMS}}} \tag{3.23}$$

- **Impact Factor**: Highlights how extreme the peaks of the signal are compared to the mean signal level.

$$\text{Impact Factor} = \frac{x_{\text{peak}}}{|x|_{\text{mean}}} \tag{3.24}$$

- **Signal Coastline**: Describes the stretched or irregular nature of the signal by measuring its cumulative changes over time. Simultaneously provides insight into the amplitude, frequency, and duration of variations in the signal. A high coastline value indicates a signal with frequent and large changes, suggesting higher complexity or noise.

$$\text{Coastline} = \sum_{t=1}^{N} |x(t) - x(t-1)| \tag{3.25}$$

Percentile features, including the **25th, 75th Percentiles**, and **Interquartile Range** which measures the difference between them highlight key points in the signal's value distribution, revealing how data points are spread across its range.

Envelope features further characterize the signal by examining its amplitude fluctuations [13]. These include metrics such as the **Number of Peaks**, **Peak Prominences** (The degree to which a peak stands out relative to the surrounding extrema), and **Peak Widths**, which help identify significant events or transitions within the signal.

A novel feature in Giarrusso's work [43], called **Maximum-minimum Distance**, divides the standard 30-second macro-epoch into smaller $\lambda = 3$-second mini-epochs. It analyzes slope variations in the EEG signal, which can be interpreted as fluctuations in the signal's speed.

$$d = \sqrt{\Delta t^2 + \Delta a^2} \tag{3.26}$$

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.

$$\text{Maximum Minimum Distance} = \frac{1}{W} \cdot \sum_{i=1}^{W} |d|_i \tag{3.27}$$

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

Frequency Domain

EEG signals are naturally complex, with rich spectral contents in several frequency bands. Frequency domain analyses of these signals could show meaningful insights into the underlying neural dynamics and physiological states. This section includes the methodology followed in feature extracting in the frequency domain from the EEG data: treating the non-stationarity of EEG signals, the segmentation , multitaper approach for PSD estimation, and computation of features for further processing.

The nature of EEG signals is non-stationary, meaning their statistical properties vary with time. The first non-stationarity arises because brain activity is of a dynamic nature itself: neural oscillations induced by cognitive processes, exterior stimulations, and physiological states vary both in amplitude and frequency. Traditional spectral analysis techniques assume stationarity within the analyzed signal segment, which is not satisfied in electric EEG data. It is, therefore, of paramount importance that the methodology that incorporates temporal variability into the EEG signals be followed for accurate modeling of transient events and subtle variation in neural dynamics. Since the EEG signals are non-stationarized,

we further segmented each 30-sec long epoch into further stationary segments, namely micro-epochs, of each length 2 seconds. This is because shorter segments are more likely to fulfill the stationarity assumption upon which spectral analysis is based and make features reflective of signal properties at the time. All these micro-epochs increase the temporal resolution of the analysis, since sudden changes in brain activity can be detected, which would be obscured in longer epochs.
Frequency domain features are extracted to capture several aspects of the spectral content of the EEG signal. Indeed, these features give the quantitative measures of power distribution, spectral shape, and complexity in different frequency bands. Below is an overview of each feature extracted:

- **Absolute Power (AP)**: Represents the total power of the EEG signal within the analyzed frequency range. It is calculated by integrating the PSD over all frequencies. Absolute power reflects the overall signal strength and is sensitive to changes in neural activity levels.

$$\mathrm{AP} = \int_{f_{\min}}^{f_{\max}} \mathrm{PSD}(f)\, df \tag{3.28}$$

- **Mean Power (MP)**: Obtained by normalizing the absolute power by the frequency range. It provides an average power value per unit frequency, offering a standardized measure for comparisons across different frequency ranges.

$$\mathrm{MP} = \frac{\mathrm{AP}}{f_{\max} - f_{\min}} \tag{3.29}$$

- **Spectral Crest (SCr)**: Defined as the ratio of the maximum PSD value within the frequency range to the mean power. The spectral crest measures the prominence of peak frequencies relative to the average power, highlighting dominant oscillatory components in the EEG signal:

$$\mathrm{SCr} = \frac{\mathrm{PSD}_{\max}}{\mathrm{MP}} \tag{3.30}$$

  where $\mathrm{PSD}_{\max}$ is the maximum value of the PSD within the frequency range.

- **Peak Frequency (PKF)**: The frequency at which the PSD reaches its maximum value. The peak frequency identifies the dominant oscillation in the EEG signal, which can be associated with specific neural processes or states:

$$\mathrm{PKF} = f_{\max} \text{ such that } \mathrm{PSD}(f_{\max}) = \max\{\mathrm{PSD}(f)\} \tag{3.31}$$

- **Mean Frequency (MNF)**: Calculated as the centroid of the PSD, weighted by frequency. The mean frequency provides a measure of the average frequency content of the signal and can indicate shifts in spectral power distribution:

$$\mathrm{MNF} = \frac{\int_{f_{\min}}^{f_{\max}} f \cdot \mathrm{PSD}(f)\, df}{\mathrm{AP}} \tag{3.32}$$

- **Spectral Edge Frequencies (SEF)**: Frequencies below which a certain percentage of the total spectral power is contained. SEF provides insights into the distribution of power across the frequency spectrum, indicating the balance between low and high-frequency components. SEF is calculated by finding the frequency $\mathrm{f_p}$ such that:

$$\int_{f_{\min}}^{f_p} \mathrm{PSD}(f)\, df = p \cdot \mathrm{AP} \tag{3.33}$$

  where p is the percentile (e.g., 0.50 for SEF50).

- **Spectral Edge Frequency Differences (SEFd)**: Differences between spectral edge frequencies at different percentiles, quantifying the spread of power distribution e.g., for SEFd_{95_50}:

$$\mathrm{SEFd}_{95\_50} = \mathrm{SEF}_{95} - \mathrm{SEF}_{50} \tag{3.34}$$

- **Shannon Entropy (SEN)**: Measures the randomness or complexity of the power distribution in the frequency domain. High entropy values indicate a more uniform power distribution across frequencies, while lower values suggest concentrated power in specific bands. Shannon Entropy is calculated as:

$$\mathrm{SEN} = -\sum_{j=b_1}^{b_2} P_j \cdot \ln P_j \tag{3.35}$$

  where $P_j = \frac{S_j}{AP}$ is the normalized PSD (probability distribution).

- **Rényi Entropy (REN)**: A generalized entropy measure that emphasizes the contribution of higher power components in the PSD. Rényi entropy of order 2 is given by:

$$\mathrm{REN} = -\ln\left(\sum_{j=b_1}^{b_2} P_j^2\right) \tag{3.36}$$

  Rényi entropy provides additional information about the signal's complexity and the prominence of dominant frequencies.

EEG signals are characterized by distinct frequency bands, each associated with specific neural activities and cognitive states. The features were extracted within the following frequency ranges: Delta (0–4 Hz), Theta (4–8 Hz), Alpha (8–13 Hz), Beta (13–30 Hz), Gamma (30–50 Hz), FREM (2–8 Hz), TREM (7–16 Hz), Slow Oscillations or SOs (0–1 Hz), and Slow Wave Activity (SWA) (1–4 Hz). These frequency bands play a crucial role in understanding brain activity and cognitive processes. For each band, we computed the following features:

- **Maximum PSD Value** : The highest power value within the band, indicating the most dominant frequency component in that band.
- **Relative Power (rp)**: The percentage of the total power (AP) that is contained within the specific band:

$$\text{rp} = \left(\frac{\text{AP}_{\text{band}}}{\text{AP}_{\text{total}}}\right) \times 100\% \tag{3.37}$$

- **Mean Area** : The average power within the band, normalized by the frequency range:

$$\text{meanArea} = \frac{\text{AP}_{\text{band}}}{f_{\text{max}} - f_{\text{min}}} \tag{3.38}$$

- **Area Under the Curve (Area)**: Total power within the band, calculated by integrating the PSD over the band frequencies:

$$\text{Area} = \int_{f_{\text{min}}}^{f_{\text{max}}} \text{PSD}(f)\, df \tag{3.39}$$

- **Spectral Centroid (SCe)**: The weighted mean frequency within the band:

$$\text{SCe} = \frac{\int_{f_{\text{min}}}^{f_{\text{max}}} f \cdot \text{PSD}(f)\, df}{\text{Area}} \tag{3.40}$$

- **Spectral Spread (SSp)**: The standard deviation of the frequencies within the band, weighted by the PSD:

$$\text{SSp} = \sqrt{\frac{\int_{f_{\text{min}}}^{f_{\text{max}}} (f - \text{SCe})^2 \cdot \text{PSD}(f)\, df}{\text{Area}}} \tag{3.41}$$

- **Spectral Skewness (SSk)**: Measures the asymmetry of the power distribution within the band:

$$\text{SSk} = \frac{\int_{f_{\text{min}}}^{f_{\text{max}}} (f - \text{SCe})^3 \cdot \text{PSD}(f)\, df}{(\text{SSp})^3 \cdot \text{Area}} \tag{3.42}$$

- **Spectral Kurtosis (Sk)**: Quantifies the peakedness of the power distribution within the band:

$$\mathrm{Sk} = \frac{\int_{f_{\min}}^{f_{\max}} (f - \mathrm{SCe})^4 \cdot \mathrm{PSD}(f)\, df}{(\mathrm{SSp})^4 \cdot \mathrm{Area}} \tag{3.43}$$

- **Variance of Central Frequency (vcf)**: The variance of the frequencies within the band:

$$\mathrm{vcf} = \frac{\int_{f_{\min}}^{f_{\max}} f^2 \cdot \mathrm{PSD}(f)\, df}{\mathrm{Area}} - (\mathrm{SCe})^2 \tag{3.44}$$

- **Peak Frequency (PKF)**: The frequency within the band where the PSD reaches its maximum.
- **Mean Frequency (mf)**: The average frequency within the band, weighted by the PSD:

$$\mathrm{mf} = \frac{\int_{f_{\min}}^{f_{\max}} f \cdot \mathrm{PSD}(f)\, df}{\int_{f_{\min}}^{f_{\max}} \mathrm{PSD}(f)\, df} \tag{3.45}$$

- **Spectral Edge Frequencies (SEF50,SEF95)**: Frequencies below which 50% and 95% of the band power are contained, respectively.
- **Spectral Edge Frequency Difference (SEFd95-50)**: The difference between SEF95 and SEF50.
- **Shannon Entropy (SEN)**: The entropy of the power distribution within the band.
- **Rényi Entropy (REN)**: Provides an alternative measure of entropy within the band.

Ratios of power between different frequency bands offer additional insights into the balance and interaction of neural oscillations. The following ratios were computed:

- **Theta/Alpha Ratio**:

$$\mathrm{Theta\_Alpha} = \frac{\mathrm{Area}_{\mathrm{Theta}}}{\mathrm{Area}_{\mathrm{Alpha}}} \tag{3.46}$$

- **Beta/Alpha Ratio**:

$$\mathrm{Beta\_Alpha} = \frac{\mathrm{Area}_{\mathrm{Beta}}}{\mathrm{Area}_{\mathrm{Alpha}}} \tag{3.47}$$

- **Theta/Beta Ratio**:

$$\text{Theta\_Beta} = \frac{\text{Area}_{\text{Theta}}}{\text{Area}_{\text{Beta}}} \tag{3.48}$$

- **(Theta + Alpha)/Beta Ratio** :

$$\text{Theta\_Alpha\_Beta} = \frac{\text{Area}_{\text{Theta}} + \text{Area}_{\text{Alpha}}}{\text{Area}_{\text{Beta}}} \tag{3.49}$$

- **(Theta + Alpha)/(Alpha + Beta) Ratio**:

$$\text{Theta\_Alpha\_Alpha\_Beta} = \frac{\text{Area}_{\text{Theta}} + \text{Area}_{\text{Alpha}}}{\text{Area}_{\text{Alpha}} + \text{Area}_{\text{Beta}}} \tag{3.50}$$

- **TREM/FREM Ratio**:

$$\text{TREM\_FREM} = \frac{\text{Area}_{\text{TREM}}}{\text{Area}_{\text{FREM}}} \tag{3.51}$$

- **SWA/SOs Ratio**:

$$\text{SWA\_SOs} = \frac{\text{Area}_{\text{SWA}}}{\text{Area}_{\text{SOs}}} \tag{3.52}$$

Time-Frequency Domain

Analysis of the EEG signals both in time and frequency domains was done using the Discrete Wavelet Transform (DWT). Wavelet transforms are very well adapted to the analysis of non-stationary signals because of their nature: multi-resolution decompositions capture high-frequency/short-duration and low-frequency/long-duration components.

Daubechies 4 wavelet, shortly known as db4, is a member of the family of orthogonal wavelets designed by Ingrid Daubechies in 1988. This wavelet represents one of the most used discrete wavelets in signal processing since it allows both in the time and frequency localizations. It is a wavelet with four vanishing moments, as indicated by "4" in db4, a property that includes a facility to represent polynomial functions of up to degree 3 exactly. It will be helpful in the investigation of sharp transitions; for example, edges in images or transient events in time series data. The wavelet db4 is symmetric and compactly supported. This means computational overhead will remain minimum with this wavelet, without compromising other properties, such as energy concentration. db4 wavelet has a filter length of 8, thereby fixing the number of coefficients in the associated scaling and wavelet filters. These filters are obtained from a mathematical formulation that solves polynomial equations for orthogonality and vanishing moments. db4 wavelet has a shape which is a little

asymmetric, balancing the time-frequency resolution. This property is essential in a variety of applications such as denoising, compression, and feature extraction, where signal fidelity must be preserved. The db4 wavelet has been applied to a wide range of domains: starting from the signal processing domain (it finds wider application in this domain for de-noising signals, detecting singularities, and carrying out audio and image compression) to biomedical engineering, where the analysis of EEG or ECG signals is done with it due to its efficiency in anomaly detection. It also finds applications in numerical solutions to differential equations, as the compact support and orthogonality make computations easier. The scalability of the wavelet, along with its precision, thereby makes it a very powerful tool for analyzing non-stationary signals.
In this paper, the wavelet transform was performed based on the use of Daubechies 4 (db4) wavelet. Decomposition was done up to level 6, resulting in the following components:

- **Approximation Coefficients (A6)**: Represent the low-frequency components of the signal at level 6.
- **Detail Coefficients (D6 to D3)**: Represent the high-frequency components at various levels.

The decomposition levels correspond to specific EEG frequency bands:

- **Level 6 (A6, D6)**: Approximation and detail coefficients capturing Delta band (0.5–4 Hz).
- **Level 5 (D5)**: detail coefficients capturing Theta band (4–8 Hz).
- **Level 4 (D4)**: detail coefficients capturing Alpha band (8–13 Hz).
- **Level 3 (D3)**: detail coefficients capturing Beta band (13–30 Hz).

After obtaining the wavelet coefficients, several statistical features were extracted to characterize the EEG signals within different frequency bands. The features calculated for each band include:

- **Mean Value**
- **Standard Deviation**
- **Coastline Feature**
- **Ratio of Means**: The ratio of the mean value of one frequency band to the mean value of another. It provides insights into the relative dominance of different frequency bands.

Nonlinear Domain

It is relevant to mention that EEG signals are complex dynamics, nonstationarities, and nonlinearities, features which cannot be explained by using traditional linear methods such as a Fourier transform. Nonlinear features give a better understanding of the intrinsic dynamics of EEG signals by extracting the complexity, variability, and fractal properties inherent in them. This will be quite substantial in the field of research for unraveling activities of both the brain and physiological mechanisms, especially the ones related to sleep. In this regard, the nonlinear features extracted within this work are described in detail in this section [48].

- **Detrended Fluctuation Analysis (DFA)**: It basically quantifies the presence of long-range correlations within a time series. The technique have applied DFA to assess the scaling properties of EEG signals by analyzing their fluctuations in several time windows. The obtained scaling exponent reflects fractal properties of the signal.:
  - $\alpha = 0.5$: Represents uncorrelated randomness, akin to white noise.
  - $0.5 < \alpha < 1$: Indicates long-range correlations, often associated with complex physiological processes.
  - $\alpha > 1$: Suggests a transition toward Brownian motion-like dynamics, often observed in deeper sleep stages.
- **Higuchi's Fractal Dimension (HFD)**: This is a technique used to quantify the fractal properties of a signal. In this, the entire EEG signal was divided into small segments and then tested for its self-similarity at different scales. The higher value of HFD, the signal will be more complex, the lower value of HFD means there is more ordered or regular activity. It is useful in capturing the changes between sleep and wake state of the brain.
- **Sample Entropy (SampEn)**: Refers to the irregularity of a time series, thus posing a measure of its respective complexity. It calculates the probability that patterns of a certain length in the signal will continue to be similar when the pattern length increases by one data point. Compared to its predecessor, Approximate Entropy, ApEn, this measure is less dependent on data length. It has been used to assess the degree of unpredictability of EEG signals.:
  - **Higher SampEn**: Suggests increased complexity or variability in brain activity.
  - **Lower SampEn**: Indicates more regular and predictable patterns, often associated with deep sleep or pathological states.

- **Teager-Kaiser Energy Operator (TKEO)**: It computes the instantaneous energy of a signal mapping its amplitude and frequency variations. TKEO measures nonlinear energy fluctuations within the considered EEG signals in a direct way. Features derived from TKEO include:
  - **Mean Energy**
  - **Standard Deviation**
  - **Skewness and Kurtosis**
  - **Maximum Energy**

3.4 Feature Selection

Medical datasets such as ours, often result in extensive number of features after feature extraction to capture the information needed for thorough analysis. In EEG signals analysis, this is especially important since features are often computed across the time, frequency and nonlinear domains (See Section 3.3) to obtain information that is either representative of different aspects of neural activity in different sleep stages [49]. While these features are necessary to represent brain dynamics during sleep, they also add dimensionality and produce a classic data-analysis problem called the curse of dimensionality." In other words, when the number of features is large compared to the size of the dataset, the latter becomes sparse, and any meaningful pattern can be swamped by noise [50]. Such sparsity often results in overfitting, where a model focuses on random noise in the training data rather than learning general patterns that apply to new data. This naturally presents a situation where models that are trained on high-dimensional data can get overly sensitive, doing a good job on training data but failures with regard to new input in real-world scenarios.

The curse of dimensionality also affects the interpretability of the model. In clinical practices, interpretability of models is considered necessary in order to provide insights related to how different features affect the prediction of a model [51]. However, when the number of features is enormous, the interpretability also decreases because it will then become difficult to explain which one among the elements contributes most to the predictive result. This black-box effect discourages the use of such models in a clinical setup since clinicians are very skeptical about predictions given by any model without an interpretable foundation. Dimensionality management is, therefore, also crucial for interpretability and transparency of models if they have to be taken as reliable diagnostic tools.

Given the limitations raised by the curse of dimensionality, selecting a relevant subset of features becomes necessary. Feature selection is a process of selecting a subset of most relevant features available for the target from the total feature space in such a way that it optimizes model performance, computational complexity, and interpretability. Feature selection tries to select a number of features that can best describe the underlying data structure and directly contribute to the prediction task. Traditionally, feature selection methods are categorized into three classes: filter methods, wrapper methods, and embedded methods. Each of them has its strengths and weaknesses. Which of the special approaches is selected depends on several other factors of the size and complexity of the data set, computational resources, properties desired in the model such as interpretability and performance.

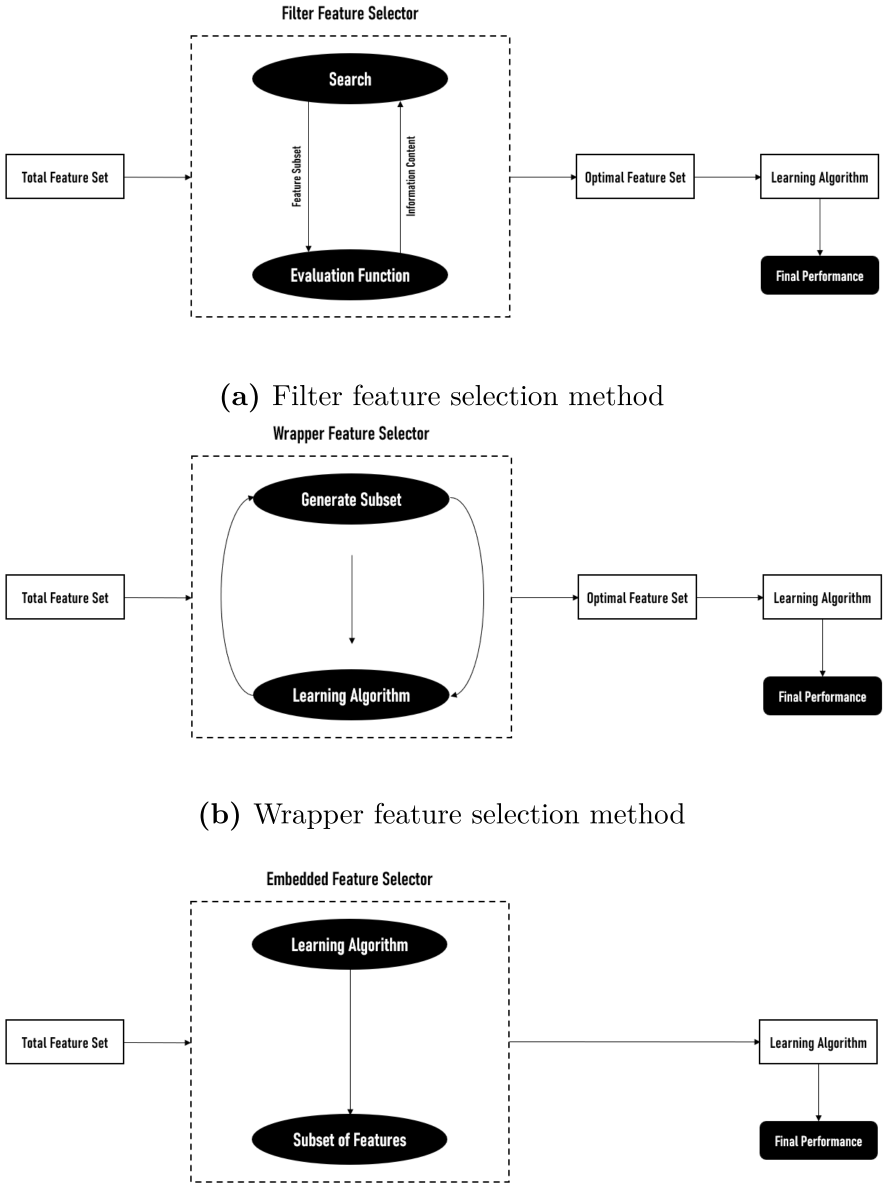
3.4.1 Filter Methods

Filter methods are among the most commonly used techniques due to their simplicity and efficiency. As shown in Figure 3.2a, they assess each feature's relevance independent of any specific machine learning algorithm, using statistical metrics to measure feature importance. Among these, popular methods include the chi-square test, which computes a measure of independence for each feature with respect to the target variable; they also perform well in the case of categorical data [52]. Another popular approach is selection based on correlation, such as Pearson correlation methodology, which selects multicollinearity between features based on a certain threshold of the correlation coefficient. Usually, high feature correlations are removed to avoid redundancy and instead focus efforts on features that provide independent contributions to useful information. Mutual information also serves as another effective score for feature selection that describes linear and nonlinear relationships between features and target variables with the purpose of selecting complex interactions in the data.

These methods have the key advantages of being computationally efficient. They are also very suitable for datasets with high feature counts, since they do not depend on iterative model training. They are algorithm agnostic so, filter methods can be used for a large number of learning model as well. One of the most important cons of such methods is that, every feature is evaluated separately with no interaction between features considered. This would lead to suboptimal results at situations where interactions between features can play an important role in prediction accuracy.

3.4.2 Wrapper Methods

In contrast with filter methods, wrapper methods evaluate feature subsets based on model performance, taking into account feature interactions and dependencies that filter methods overlook. A commonly used wrapper technique is Recursive Feature Elimination (RFE), which iteratively fits a model, ranks features by



(a) Filter feature selection method

(b) Wrapper feature selection method

(c) Embedded feature selection method

Figure 3.2: Overview of supervised feature selection methods

importance, and removes the least relevant features until the desired number of features is achieved. Other wrapper techniques include forward selection, which begins with an empty set of features and progressively adds features based on the model's performance improvements, and backward elimination, which starts with all features and iteratively removes the least significant ones [53]. These techniques are advantageous for capturing interactions between features, often resulting in higher model performance than filter methods.

Nonetheless, wrapper methods are computationally expensive since each model must be trained and evaluated many times for every feature subset. For high-dimensional datasets, unless one have access to massive amounts of computational resources this is impractical. Moreover, since wrapper methods optimize feature selection based on model performance in the training data, they are more likely to overfit and this feature selection may not be optimal for unobserved datapoints in real-world scenarios.

3.4.3 Embedded Methods

Embedded methods, a third category, combine the benefits of filter and wrapper methods by performing feature selection within the model training process itself. These methods are especially efficient because they integrate feature selection with the learning process, making them less computationally intensive than wrapper methods while still capturing feature interactions. Examples include decision trees and other tree-based methods, such as random forests and XGBoost, which provide feature importance scores based on their contributions to reducing impurity at each node, thereby selecting features during model training [54, 55]. Regularization techniques, such as Lasso regression, also serve as embedded methods by introducing an L1 penalty term to the model's loss function, which encourages sparsity in the coefficients and effectively eliminates irrelevant features by setting their coefficients to zero. Embedded methods offer both computational efficiency and high interpretability, as the selected features are integral to the model's training process.

A drawback of embedded methods is that they are model-specific. Thus, the chosen features can be completely different for a learning algorithm used for running a different learning phase. For instance, regularization techniques can struggle in the presence of multicollinearity. It causes a lot of redundant and inefficient feature sets. Even so, embedded methods represent a good compromise between interpretability and computational cost. It makes them them particularly appropriate for many practical applications demanding both performance and interpretability.

3.4.4 Unsupervised Methods

In cases where labeled data is limited or unavailable, unsupervised feature selection methods are utilized. These methods, which include dimensionality reduction techniques like Principal Component Analysis (PCA) and spectral feature selection

(SPEC), are instrumental in reducing the dimensionality of data without relying on a target variable [56]. PCA transforms the data into a set of orthogonal components that capture the maximum variance, while SPEC uses spectral graph theory to select features that retain the data's intrinsic structure [57]. Unsupervised methods are particularly valuable for tasks like biomarker discovery or clustering in high-dimensional, unlabeled datasets.

Unsupervised feature selection methods are used when there is a lack or absence of labeled data. These methods such as Principal Component Analysis (PCA) and spectral feature selection (SPEC) are widely used to reduce the dimension of the data without any dependence on a target variable [56]. While PCA linearly transforms the data into a set of orthogonal transforms to capture the most variance, SPEC applies tools from spectral graph theory to choose features which preserve intrinsic structure of the data [57]. Unsupervised methods especially useful for biomarker discovery or clustering tasks in high dimensional and unlabeled datasets.On the other hand, unsupervised feature selection methods are in general not interpretable because transformed features (like the principal components in PCA) do not have an inherent meaning which makes it harder to interpret model results.

3.4.5 Minimum Redundancy Maximum Relevance (mRMR) Method

The Minimum Redundancy Maximum Relevance (mRMR) method is a powerful and common method that aims to resolve the suffered problem from high-dimensional dataset to select important features which were first proposed in using problems in gene expression microarray data where the challenge was to pick a small number of very informative genes from a pool of thousands of measured genes. mRMR selects features by optimizing two complementary criteria simultaneously: maximizing the relevance of the features with respect to the target variable, minimizing the redundancy among those features.

In the mRMR method, the basic measure to assess relevance and redundancy is based on mutual information. Mutual information quantifies the level of dependency between variables, which allows it to be used for evaluating the relationship between each feature and the target class, and inter-feature correlations. Assuming X_i are features, Y is the target class and S is the subset of features selected in a current iteration. Formally, the relevance of a feature X_i is defined as the mutual information between X_i and Y, which can be expressed as follows:

$$\text{Relevance: } I(X_i;Y) = \sum_{x_i \in X_i} \sum_{y \in Y} p(x_i, y) \log \frac{p(x_i, y)}{p(x_i)p(y)}, \tag{3.53}$$

where $p(x_i, y)$ is the joint probability distribution of X_i and Y, and $p(x_i)$, $p(y)$ are their marginal distributions.

Redundancy is quantified as the average mutual information between each pair of features X_i and X_j in the subset S:

$$\text{Redundancy: } I(X_i; X_j) = \sum_{x_i \in X_i} \sum_{x_j \in X_j} p(x_i, x_j) \log \frac{p(x_i, x_j)}{p(x_i)p(x_j)}. \tag{3.54}$$

The mRMR criterion combines these two objectives which aims to maximize relevance while minimizing redundancy. This is expressed as:

$$\text{mRMR: } \max \left(\frac{1}{|S|} \sum_{X_i \in S} I(X_i; Y) - \frac{1}{|S|^2} \sum_{X_i, X_j \in S} I(X_i; X_j) \right), \tag{3.55}$$

where $|S|$ is the number of features in the subset S.

mRMR is generally an iterative optimization process. Starting from empty subset and we add one element to the feature set at a time according to which one contributes the mRMR objective the most. During each round, the algorithm scores each candidate feature in terms of its relevance and redundancy, and selects the one that optimizes the criterion. To make feature selection more efficient, hybrid schemes like mRMR combined with Genetic Algorithms (GA) or Differential Evolution (DE) have been proposed to operate in large search spaces. Hybrid algorithms combine exploration (scanning broadly across the feature space) and exploitation (refining feature subsets that look promising) to achieve enhanced classification performance. Overall, mRMR provide a balance in the tradeoff between relevance and redundancy, in other words mRMR prevents the selection of features that may explain over with overlapping information to the target with undesirable core of the model.

In our work, we applied the mRMR method to rank the 957 features extracted in the previous Section 3.3. From this ranked list, we selected the top 5 features based solely on the CAP dataset and identified the corresponding feature sets in the TuSDi dataset for consistency in further analysis. While there is no universally pure theory to determine the optimal number of features for classification tasks, we followed the guidelines suggested by Hua et al. [58], which recommend selecting approximately $\sqrt{N}$ features, where N is the total number of samples. This heuristic is empirically grounded in their analysis of feature selection and classification rules, demonstrating that choosing $\sqrt{N}$ features maintain a balance between minimizing overfitting and preserving sufficient information for robust classification. Following this principle, we ensured that our feature subset remains both computationally manageable and theoretically aligned with the characteristics of our data and classification task. The summarized feature ranking results for different sleep analysis datasets are presented in Figures 3.3, 3.4, 3.5, 3.6, 3.7, and 3.8. Each figure displays the top-ranked features, which have been selected for use in the subsequent steps.

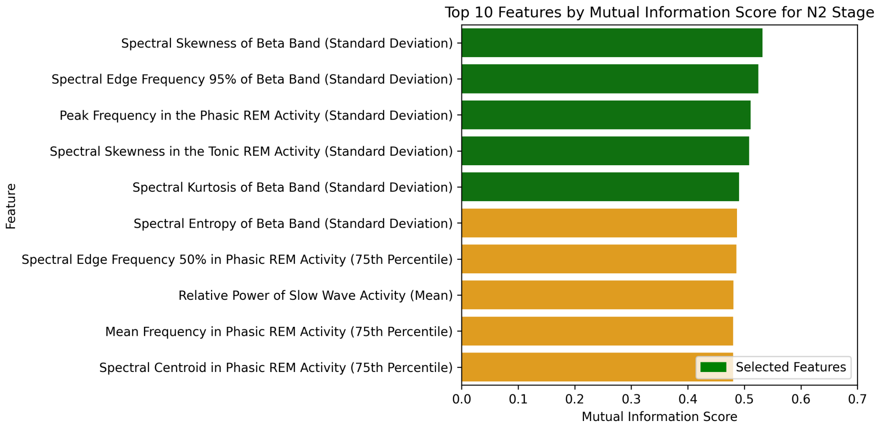


Figure 3.3: Selected features for N2 Stage

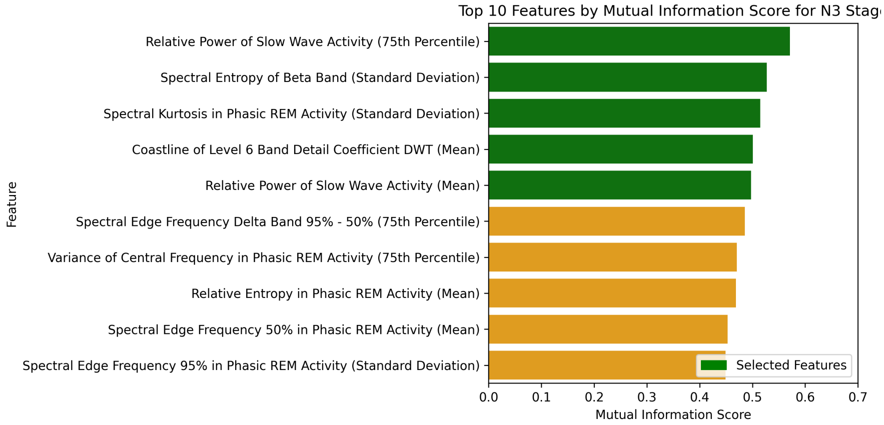


Figure 3.4: Selected features for N3 Stage

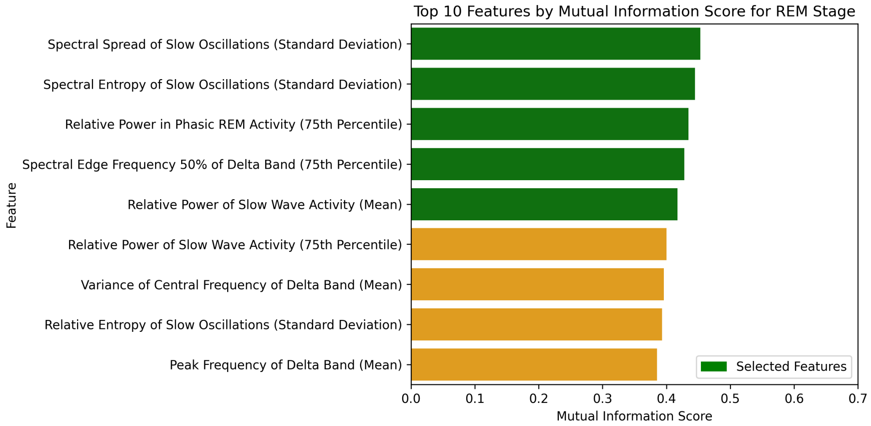


Figure 3.5: Selected features for REM Stage

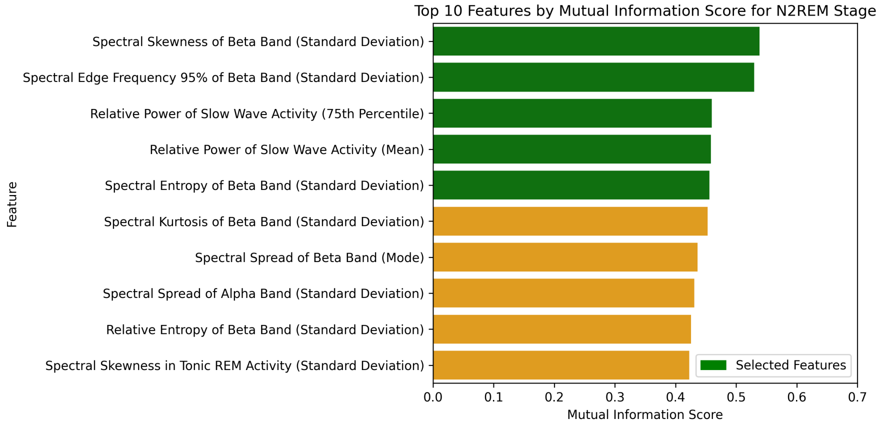


Figure 3.6: Selected features for N2+REM Stages

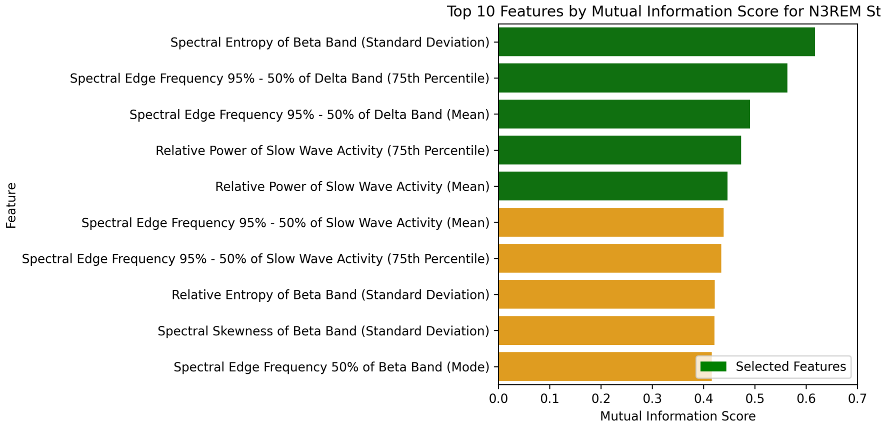


Figure 3.7: Selected features for N3+REM Stages

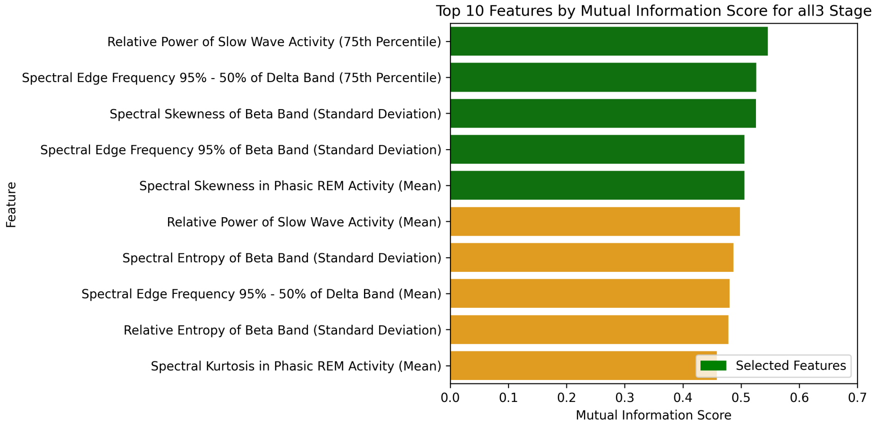
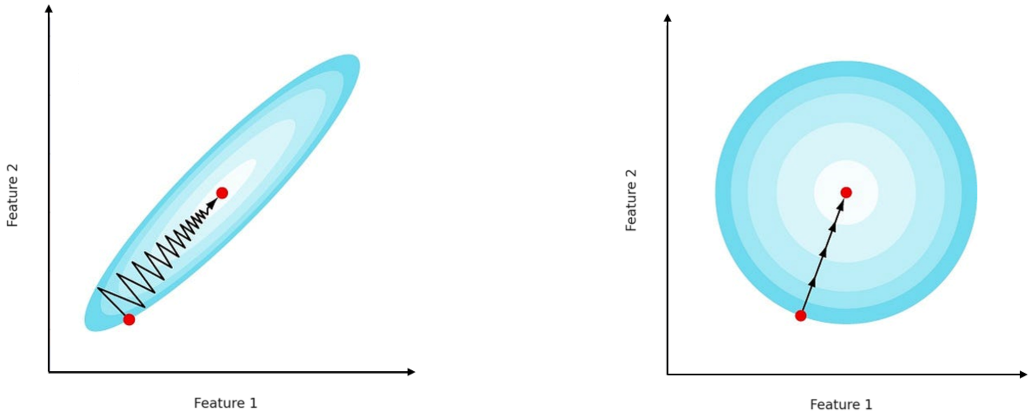


Figure 3.8: Selected features for N2+N3+REM Stages

3.5 Feature Scaling

In our next step, we scaled the features to standardize their ranges; this is done after the feature selection. Feature scaling is an essential pre-processing step of machine learning workflows as the performance of many algorithms such as Logistic Regression that uses gradient descent as the optimization method is sensitive to the magnitude of feature values (See Figure 3.9). Model convergence and stability are enhanced by keeping features on similar scales. [54].

Each feature was scaled, which is done using the mean and standard deviation of each specific feature. In particular, for any feature X, we computed its scaled value X_{scaled} according to the formula:



(a) Without Feature Scaling

(b) With Feature Scaling

Figure 3.9: Comparison of optimization trajectories and convergence with and without feature scaling.

$$X_{\text{scaled}} = \frac{X - \mu}{\sigma}, \tag{3.56}$$

where μ and σ represent the mean and standard deviation of the feature, respectively. This normalization transformes the feature so that it has the mean of zero and a standard deviation of 1, which eliminates the impact of the original scale.

Feature scaling is done to ensure that all selected features equally contribute to the analysis and larger numeric ranges do not dominate the features. This step was thus important for obtaining consistent results in subsequent steps.

3.6 Binary Classification

As discussed earlier, our goal is to distinguish RBD patients from healthy individuals. To achieve this, we employed several algorithms for classification. In this section, we provide a detailed exploration of each algorithm and the corresponding parameters used in the analysis.

3.6.1 K-Nearest Neighbors (KNN) Algorithm

The K-Nearest Neighbors (KNN) is a simple, yet, widely used non-parametric, instance-based learning method for classification and regression tasks. When it comes to classification, the K nearest neighbors algorithm (KNN) assigns a class to a data point given the classes of its k-nearest neighbors in the feature space. The algorithm uses some pre-determined distance metric to calculate similarity between data points that governs how close or far they are from each other. For a given data point x, the KNN classifier predicts the class $\hat{y}$ based on the majority vote among its k-nearest neighbors, defined as:

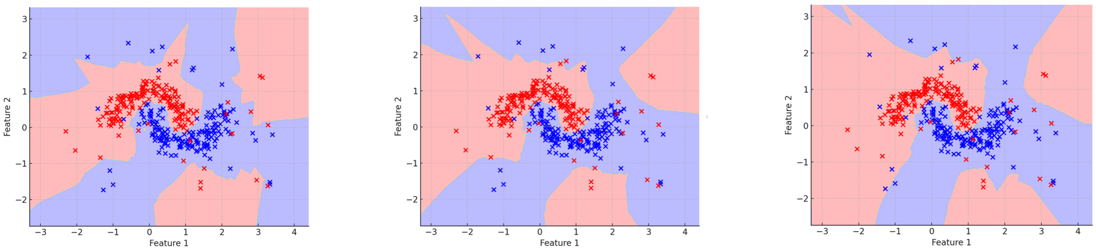
$$\hat{y} = \arg\max_{c \in C} \sum_{i=1}^{k} w_i \cdot \nVdash\{y_i = c\}, \tag{3.57}$$

where C is the set of all possible classes, w_i represents the weight assigned to the i-th neighbor, $\nVdash\{y_i = c\}$ is an indicator function that equals 1 if the neighbor's label y_i matches class c, and 0 otherwise. The weight w_i can be uniform ($w_i = 1$) or distance-based ($w_i = 1/d_i$), where d_i is the distance between x and its i-th neighbor.

The key parameters of the KNN algorithm include Number of neighbors (k), Weighting Function (*weights*) and Distance Metric (p).

Number of Neighbors

One of the most important hyperparameters of the KNN algorithm is the number of neighbours (or k) that determines the number of nearest neighbours used for classification. Larger values of k look at a larger neighborhood, which smooths the decision boundary, but also introduces bias. On the other hand, smaller values of k used for prediction look at less neighbors to make the prediction and therefore they have a more local decision boundary and can also overfit more easily. Common options of k to examine for this work will be $k = 3$, $k = 5$, and $k = 7$. The KNN performance with different number of neighbors is shown in Figure 3.10.



(a) K=3 **(b)** K=5 **(c)** K=7

Figure 3.10: KNN Decision Boundaries for Different K (Number of Neighbors)

Weighting Function

The weighting function defines how much influence each neighbour has in the classification. We focus on two basic weight schemes:

- **Uniform**: assigns equal weight to all neighbors ($w_i = 1$), regardless of their distance to query point.
- **Distance**: The closer the neighbor to the query point the more weight it has ($w_i = 1/d_i$, in other words, d_i is the distance to the i-th neighbor.) This complements the method of closer neighbors existing frequently and thus leads to better classification in datasets with non uniform density.

Distance Metric

The distance metric (handled by the parameter p) defines how distances between points are calculated:

- $p = 1$: Manhattan distance, given by:

  $$d(x, x_i) = \sum_{j=1}^{m} |x_j - x_{ij}|, \tag{3.58}$$

  where m is the number of features.

- $p = 2$: Euclidean distance, defined as:

  $$d(x, x_i) = \sqrt{\sum_{j=1}^{m} (x_j - x_{ij})^2}. \tag{3.59}$$

The choice of p affects the geometry of the neighborhoods, with Manhattan distance creating box-like neighborhoods and Euclidean distance forming spherical neighborhoods.

3.6.2 Logistic Regression Algorithm

Logistic Regression (LR) is a supervised learning algorithm that frequently applied in classification problems, both binary and multi-class [59].Logistic regression relates input features to the probability of the target class by using a logistic function and thus mapping inputs into probabilities in the range of [0, 1] (Figure 3.11). The model calculates the likelihood of the target class y given the input $\mathbf{x}$ as follows.

$$P(y = 1|\mathbf{x}) = \frac{1}{1 + \exp(-(\mathbf{w}^T\mathbf{x} + b))}, \tag{3.60}$$

where $\mathbf{w}$ represents the weights assigned to the input features, b is the bias term, and $\mathbf{x}$ is the feature vector. For multi-class classification, the algorithm uses the softmax function to generalize probabilities across multiple classes.

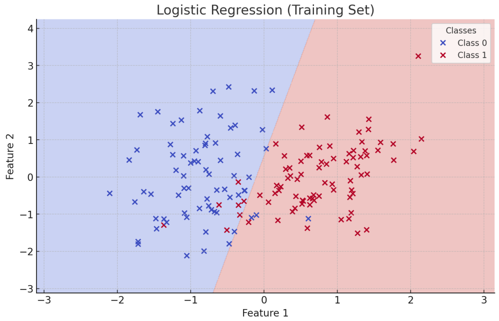


Figure 3.11: Logistic Regression Decision Boundary

To train the Logistic Regression model, the goal is to minimize the regularized log-loss function, which is defined as:

$$\min_{\mathbf{w},b} \frac{1}{N} \sum_{i=1}^{N} [-y_i \log P(y_i|\mathbf{x}_i) - (1 - y_i) \log(1 - P(y_i|\mathbf{x}_i))] + R(\mathbf{w}), \tag{3.61}$$

where N is the number of training samples, y_i is the true label of the i-th sample, $P(y_i|\mathbf{x}_i)$ is the predicted probability, and $R(\mathbf{w})$ is the regularization term. The regularization term $R(\mathbf{w})$ incorporates the influence of key parameters like Penalty, C, and $L1_{ratio}$.

Penalty

The penalty parameter describes the type of regularization being applied to $\mathbf{w}$ which affects the term $R(\mathbf{w})$ Regularization is used for avoiding overfitting by imposing a constraint on the weight magnitudes of the model. Where $R(\mathbf{w})$ is defined, depending on the value of `penalty`, as follows:

- L1 (Lasso Regularization): $R(\mathbf{w}) = \lambda \sum |\mathbf{w}_i|$, which aims to find the sparse solution by driving some weights to zero.

- L2 (Ridge Regularization): $R(\mathbf{w}) = \lambda \sum \mathbf{w}_i^2$, which penalizes large weights as an effort to stabilize.

- ElasticNet: $R(\mathbf{w}) = \lambda[\alpha \sum |\mathbf{w}_i| + (1 - \alpha) \sum \mathbf{w}_i^2]$, combining L1 and L2 regularization, where α is the `l1_ratio`.

- *otw*: No regularization, thus assuming $R(\mathbf{w}) = 0$.

Inverse Regularization Strength

The C parameter regulates the power of regularization, since in $R(\mathbf{w})$ it is being assumed that $\lambda = 1/C$. The regularization effect is high for small values of C, and for large values, the regularization effect decreases. This parameter is useful in managing the complexity versus performance of the model.

Solver

Numerous optimization methods and solvers exist to minimize the loss function, with each incorporating different types of regularization:

- **Liblinear**: Efficient for small datasets; supports L1 and L2 penalties.
- **Saga**: Handles large datasets and supports all penalties, including ElasticNet.
- **Lbfgs**: For multi-class classification; including L2 penalty.

ElasticNet Mixing Parameter

The $L1_{ratio}$ parameter is applicable when the penalty is set to ElasticNet. It determines the ratio between L1 and L2 regularization in $R(\mathbf{w})$:

- $L1_{ratio} = 0$: Equivalent to pure L2 regularization.
- $L1_{ratio} = 1$: Equivalent to pure L1 regularization.
- $0 < L1_{ratio} < 1$: A mix of L1 and L2 regularization.

3.6.3 Gaussian Naive Bayes Algorithm

Gaussian Naive Bayes (GNB) is a probabilistic classifier based on Bayes' theorem, with an assumption of independence among the features conditional on the value of the class label [60]. It works particularly good in case of classification based tasks, which have there features quite Gaussian distributed. We can express the conditional probability over the feature value x_j of class y with:

$$P(x_j|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_j - \mu_y)^2}{2\sigma_y^2}\right), \tag{3.62}$$

where μ_y and σ_y^2 are the mean and variance of the feature x_j for the class y, respectively. Using Bayes' theorem, the posterior probability of a class y given the feature vector $\mathbf{x}$ is computed as:

$$P(y|\mathbf{x}) = \frac{P(y)\prod_{j=1}^{m} P(x_j|y)}{P(\mathbf{x})}, \tag{3.63}$$

where $P(y)$ is the prior probability of the class y, $P(x_j|y)$ is the likelihood of the feature x_j given the class y, and $P(\mathbf{x})$ is the marginal probability of the feature vector. The class label is predicted by selecting the class with the highest posterior probability:

$$\hat{y} = \arg\max_{y} P(y|\mathbf{x}). \tag{3.64}$$

Variance Smoothing

Gaussian Naive Bayes has a single key parameter, which is the variance smoothing parameter, ϵ used to handle the variance estimates.. This is particularly important to prevent numerical instabilities caused by very small variance values. The smoothed variance σ_y^2 is calculated as:

$$\sigma_y^2 = \hat{\sigma}_y^2 + \epsilon \cdot \max(\hat{\sigma}_y^2), \tag{3.65}$$

where $\hat{\sigma}_y^2$ is the observed variance for the class y, and ϵ is a small positive constant added to stabilize the estimates. Typical values explored for ϵ include $\{1\text{e-}09, 1\text{e-}08, 1\text{e-}07\}$.

3.6.4 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised learning algorithm commonly used for classification and regression tasks. It aims to find the optimal hyperplane that separates data points of different classes in the feature space with the maximum margin [61]. The SVM classifier uses a linear decision boundary and optimizes the following objective function:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N} \xi_i, \tag{3.66}$$

subject to the constraints:

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad \forall i,$$

where $\mathbf{w}$ is the weight vector, b is the bias term, ξ_i are slack variables that allow misclassification for non-linearly separable data, C is the regularization parameter, y_i are the true labels, and $\mathbf{x}_i$ are the feature vectors. The hyperplane is determined by the support vectors, which are the data points closest to the boundary. The key parameters for linear SVM include:

Regularization Parameter

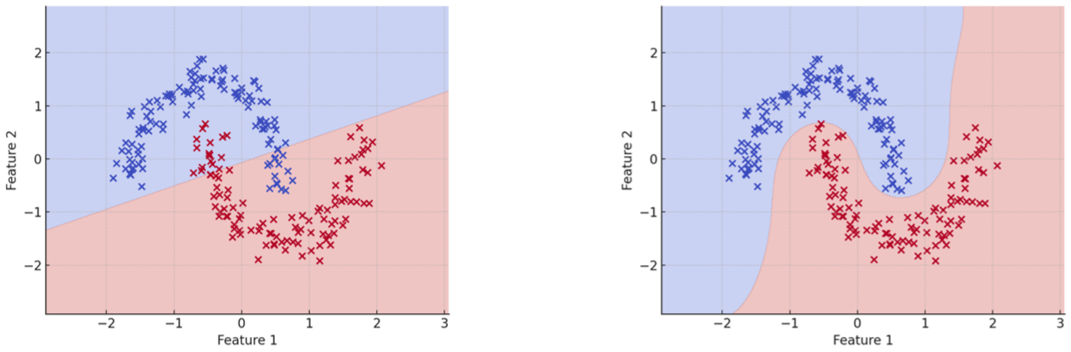
The Regularization parameter C, controls the trade-off between maximizing the margin and minimizing classification errors. A smaller C value allows for a wider margin at the cost of more misclassified points, while a larger C emphasizes correct classification, potentially leading to overfitting. In this study, we explored values of $C = \{0.1, 1, 10, 100\}$ to evaluate the model's performance under different regularization strengths.

Kernel Function

The linear SVM uses a linear kernel function, defined as:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j.$$

This kernel is computationally efficient and works well for linearly separable data.



(a) Linear SVM

(b) RBF Kernel SVM

Figure 3.12: Comparison of decision boundaries for Linear SVM and RBF Kernel SVM

3.6.5 Kernel Support Vector Machine (K-SVM)

Kernel Support Vector Machine (K-SVM) extends the SVM algorithm by employing kernel functions to map input data into a higher-dimensional feature space, enabling the model to handle non-linearly separable data [61]. The optimization objective remains the same as Equation 3.66, but the kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$ replaces the dot product $\mathbf{x}_i^T \mathbf{x}_j$, allowing the model to capture complex relationships as shown in Figure 3.12. The key parameters of K-SVM include:

Regularization Parameter

Similar to linear SVM, the c parameter in K-SVM controls the trade-off between maximizing the margin and minimizing classification errors. Smaller c values encourage a wider margin, while larger values prioritize correct classification. In this study, we evaluated $c = \{0.1, 1, 10, 100\}$.

Kernel Function

K-SVM supports multiple kernel functions, including:

Radial Basis Function (RBF) Kernel: The RBF kernel captures non-linear relationships by considering the distance between data points:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2\right),$$

where γ determines the kernel's sensitivity to data points. Larger values of γ focus on closer neighbors, while smaller values allow broader influence. In this study, we tested $\gamma = \{\texttt{scale, auto}\}$, where:

- $\gamma_{scale} = \frac{1}{\text{number of features}}$,
- $\gamma_{auto} = \frac{1}{\text{number of samples}}$.

Polynomial Kernel: The polynomial kernel captures polynomial relationships of degree d:

$$K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^T \mathbf{x}_j + 1)^d.$$

The degree d controls the complexity of the polynomial decision boundary. Higher degrees capture more complex relationships but may lead to overfitting. In this study, we explored $d = \{3, 5, 7\}$.

Sigmoid Kernel: The sigmoid kernel mimics neural network activation functions:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\alpha \mathbf{x}_i^T \mathbf{x}_j + c_0),$$

where α and c_0 are kernel-specific parameters. Default values were used in this work.

3.6.6 Decision Tree (DT)

The Decision Tree (DT) algorithm is a non-parametric supervised learning method used for classification and regression tasks. It works by recursively splitting the dataset into subsets based on feature values, creating a tree-like structure where each internal node represents a feature split, each branch represents a decision rule, and each leaf node represents an output class [62].

Splitting Criterion (`criterion`)

The algorithm minimizes the impurity of splits using a splitting criterion, such as Gini Impurity or Entropy, and is defined mathematically as:

$$I_{gini} = 1 - \sum_{i=1}^{C} p_i^2, \tag{3.67}$$

$$I_{entropy} = - \sum_{i=1}^{C} p_i \log_2 p_i, \tag{3.68}$$

where p_i is the proportion of samples belonging to class i, and C is the total number of classes.

Maximum Depth

The maximum depth restricts the depth of the tree, controls its complexity and prevents overfitting. A deeper tree may overfit, while a shallow tree may underfit. We evaluated `max_depth` = {None, 3, 10}, where `None` allows the tree to expand until all leaves are pure or contain fewer than the minimum samples.

Splitter

The `splitter` parameter determines the strategy used to split nodes. It can be set to `best`, which selects the optimal split based on the chosen criterion, or `random`, which selects a split randomly among the available options.

3.6.7 Random Forest (RF)

Random Forest (RF) is an ensemble learning algorithm that constructs multiple Decision Trees during training and combines their predictions through averaging (regression) or majority voting (classification) [63]. It reduces overfitting compared to individual Decision Trees by introducing randomness in tree construction, such as selecting random subsets of features for each split.

The overall prediction for classification is given by:

$$\hat{y} = \arg\max_{y \in Y} \sum_{i=1}^{N} \nVdash \{T_i(\mathbf{x}) = y\}, \tag{3.69}$$

where:

- $\hat{y}$ is the predicted class,
- Y is the set of all possible classes,

- $T_i(\mathbf{x})$ is the prediction from the i-th tree,
- N is the total number of trees,

In this study, we explored the following parameters for the Random Forest algorithm:

Number of Estimators

The number of estimators specifies the number of Decision Trees in the forest. A higher number typically improves performance but increases computational cost. We tested `n_estimators` = {100, 200}.

Maximum Depth

The maximum depth parameter controls the maximum depth of each tree, which earlier discussed in Decision Tree parameters. We explored `max_depth` = {None, 3, 5}, where `None` allows the trees to grow until all leaves are pure.

Maximum Features

This parameter defines the number of features to consider when looking for the best split. The available options are `sqrt`, which uses the square root of the total number of features, and `log2`, which uses the base-2 logarithm of the total number of features.

Splitting Criterion

Similar to Decision Trees discussed earlier in Section 3.6.6, The algorithm minimizes the impurity of splits using a splitting criterion, such as Gini Impurity or Entropy, and is defined mathematically as:

$$I_{gini} = 1 - \sum_{i=1}^{C} p_i^2, \tag{3.70}$$

$$I_{entropy} = - \sum_{i=1}^{C} p_i \log_2 p_i, \tag{3.71}$$

where p_i is the proportion of samples belonging to class i, and C is the total number of classes.

3.7 Training, Validation, and Test

In this work, we divide our dataset into two parts: one which contains for the training and validation tasks (CAP Dataset) to search for the optimal models and hyperparameters, and, the other one (TuSDi Dataset) to perform the final testing. This section explores different methods for splitting our tainining and validation data and their suitability for our binary classification task, given that our total data samples are limited.

3.7.1 Basic Train-Validation Split

The train-validation split is the simplest and the most popular evaluation and tuning approach for machine learning models. This method splits the dataset up into two separate data sets; one used to train the model and one to test the model's ability to generalise well to unseen data [54]. A common split ratio can be 80:20 or 70:30 where larger portion is used to train the model and a smaller portion is kept for testing. The choice of ratio may vary depending on the number of samples in the dataset.

The main benefit of this kind of split is its simplicity and computational efficiency. The overall computational cost of this approach is $O(f(N_{\text{train}})) + O(f(N_{\text{test}}))$ for a model with a training complexity of $O(f(N)) = -O(f(N_{\text{train}}))$, where $f(N)$ is the training cost used with N number of training samples. This method is particularly useful in exploratory phases or when computational resources are limited. By separating training and testing datasets, it ensures that the evaluation reflects the model's ability to generalize to unseen data. This separation is a mandatory step in machine learning workflows for identifying overfitting, where a model performs well on training data but poorly on unseen data. However, the method has significant drawbacks, especially for small datasets. The random nature of the split can result in high variance in the evaluation results, as performance may vary substantially depending on which samples are included in the training and testing sets.

3.7.2 Cross-Validation (CV)

Cross-validation is actually one of the best ways to test machine learning models, as it works especially well with low volume datasets. In contrast to a simple train-validation split, when using cross-validation the entire dataset is used iteratively in both training and validation which gives a better estimate of the model generalizes [64]. Cross-validation also maintains a low variance that comes with the single train-test splitting by averaging the different results over these different train-test splits, thus it is also a preferred technique whenever data is limited.

KFold Cross-Validation

K-fold cross-validation involves splitting the dataset into k equally sized folds based on the example of Figure 3.13. In k-fold cross-validation, the model is fitted k times, using $k-1$ folds for training and one fold for validation each time [54]. In each fold, a performance metric M (such as accuracy, precision, or recall) is computed between the true labels y_i and the predicted labels $\hat{y}_i$. Overall performance is calculated as:

$$M_{kfold} = \frac{1}{k}\sum_{i=1}^{k} m(y_i, \hat{y}_i), \tag{3.72}$$

where $m(y_i, \hat{y}_i)$ is the metric computed on the validation set of the i-th fold.

The computational complexity of K-fold cross-validation is $O(k \cdot f(N))$, where $f(N)$ represents the complexity of training the model on N samples. This scaling arises from the need to train the model k times on slightly smaller datasets of size $(1-1/k)N$. This approach is iterative by nature, requiring more computations than a simple train-validation split. This, though, greatly increases the reliability of the performance estimate by reducing variance and effectively using the complete data in both training and validation.

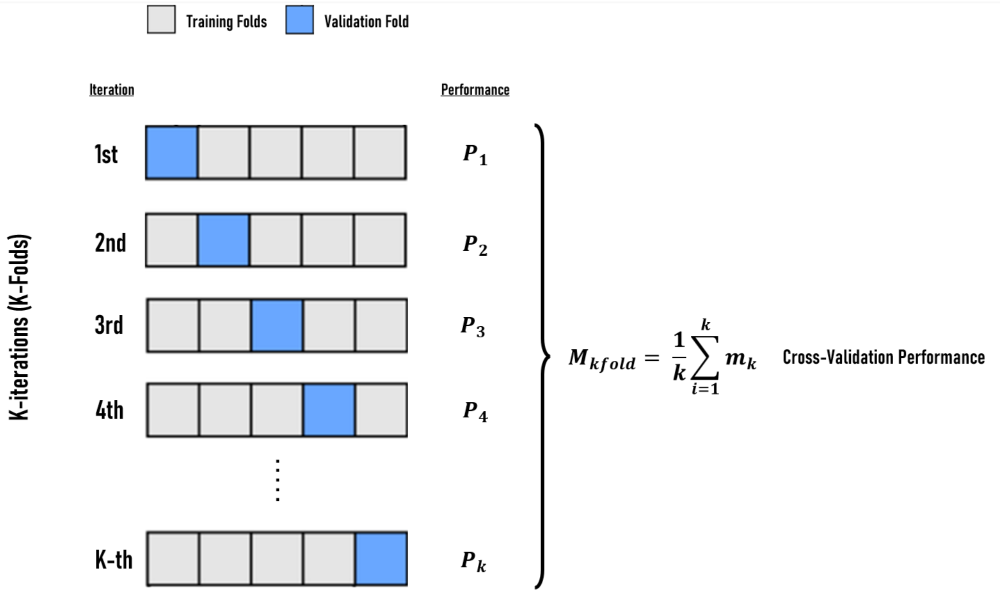


Figure 3.13: K-fold Cross-Validation

Leave-One-Out Cross-Validation (LOO-CV)

Leave-One-Out Cross-Validation (LOO-CV) is a specialized form of cross-validation in which the number of folds equals the total number of samples in the dataset, denoted by N [54]. During each iteration of LOO-CV, a single sample is held out as the validation set, and the model is trained on the remaining $N-1$ samples. This process is repeated N times, ensuring that every sample is used exactly once as the validation point.

The predictions generated during each iteration, denoted as $\hat{y}_i$, correspond to the predicted label for the i-th sample when it is treated as the validation set. These individual predictions are concatenated to form the complete prediction vector $\hat{y} = [\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_N]$. The ground truth labels $y = [y_1, y_2, \ldots, y_N]$ remain fixed throughout the process. The overall performance of the model is then evaluated by comparing y and $\hat{y}$ using a performance metric M, such as accuracy, precision, recall, or mean squared error. Mathematically, the performance metric is expressed as:

$$M_{loo} = m(y, \hat{y}), \tag{3.73}$$

The computational complexity of LOO-CV is $O(N \cdot f(N-1))$, where each training cost is $f(N-1)$. Training the model N times introduces high computational costs for LOO-CV for large and/or complex data and models. The utility of LOO-CV for small datasets, despite its cost, lies in its ability to ensure that the data is fully utilized and also avoids any bias that arises from testing samples in combination. When there is limited data but plenty of computational resources, LOO-CV should be considered the gold standard for performance evaluation.

3.7.3 Training and Evaluation Strategy

In this study, we adopted the LOO-CV approach that maximizes the use of available data by iteratively treating each sample as a test case while training the model on the remaining $N-1$ samples. This method effectively addresses the challenges posed by small datasets.

The development and evaluation pipeline integrated LOO-CV with systematic procedures to mitigate the risk of data leakage. At each iteration of the LOO-CV process, the dataset was partitioned such that one sample served as the test set, while the remaining $N-1$ samples were used for training. Feature scaling was applied exclusively to the training set to avoid incorporating information from the test sample during the scaling process. This scaling transformation, determined on the training set, was then applied to the test sample.

Hyperparameter tuning was performed through a nested cross-validation procedure within the training set. Specifically, a 10-fold cross-validation grid search was conducted, employing the F1-score as the performance metric for selecting the optimal hyperparameters for each binary classifier introduced earlier in Section 3.6. The F1-score was chosen due to its ability to balance precision and recall. The resulting best model was trained on the entire $N-1$ training set using the optimal hyperparameters identified during the grid search. Subsequently, the trained model was evaluated on the isolated test sample, and its prediction was stored.

This process was repeated for all N samples in the dataset, ensuring that each sample was used once as the test set and $N-1$ times as part of the training set. At the end of the procedure, the prediction vector $\hat{y}$, encompassing predictions for all test samples, was compared against the ground truth labels y to compute overall performance metrics.

This strategy ensures unbiased assessments of model performance while addressing the constraints of limited data samples. Furthermore, the iterative training and validation process increased the reliability of the results, especially when coupled with nested cross-validation for hyperparameter optimization. The resulting models, evaluated using the F1-score, provided a comprehensive understanding of each binary classifier behavior. The optimal model, along with its tuned hyperparameters, will be retrained on the entire train-validation dataset and subsequently utilized for testing on the TuSDi dataset.

3.8 Performance Metrics

Measuring the performance of the machine learning model is important to understand whether or not the model works for the specified classification problem . In general, metrics provide quantitative measures to assess how well a model predicts and offer insight into strengths and weaknesses [65, 66]. These metrics are usually derived from a confusion matrix in binary classification tasks, which summarizes the relationship between the predicted and actual labels. The confusion matrix provides a complete overview of the model performance by allowing several metrics to be computed that assess different dimensions of quality in the predictions. Each of the indicators gives another insight into model performance, and the assessment can be focused on some classification goals [67].

In a binary classification task, the confusion matrix is a 2×2 table that compares the predicted labels with the true labels. It summarizes the model's classification results as follows:

- True Negatives (TN): The number of negative samples correctly classified as negative.
- False Positives (FP): The number of negative samples incorrectly classified as positive.

- False Negatives (FN): The number of positive samples incorrectly classified as negative.

- True Positives (TP): The number of positive samples correctly classified as positive.

The confusion matrix is structured as follows:

$$\text{Confusion Matrix} = \begin{bmatrix} TN & FP \\ FN & TP \end{bmatrix}. \tag{3.74}$$

By default, negative class is represented in first row and positive class is represented in second row (aligning with standard conventions). We can also extract performance metrics from this confusion matrix, expressed mathematically below.

3.8.1 Accuracy

Accuracy measures the proportion of correctly classified samples out of the total number of samples [66]. It is defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{3.75}$$

While accuracy is a widely used metric, it may not be informative for imbalanced datasets, where one class dominates the other, as it treats all errors equally.

3.8.2 Sensitivity (Recall)

Sensitivity, also referred to as recall or the true positive rate, evaluates the model's ability to correctly identify positive samples [67]. It is defined as:

$$\text{Sensitivity (Recall)} = \frac{TP}{TP + FN}. \tag{3.76}$$

A high sensitivity indicates that a high percentage of the positive instances are identified by the model; it does not take into account the share of false positives. Sensitivity has strong medical implications. In scenarios where missing a diagnosis has severe consequences, high sensitivity is prioritized. This metric is crucial for tests that aim to rule out diseases where early detection significantly improves outcomes. However, while high sensitivity reduces the risk of false negatives, it may increase the likelihood of false positives.

3.8.3 Specificity

Specificity, or the true negative rate, measures the model's ability to correctly classify negative samples [67]. It is expressed as:

$$\text{Specificity} = \frac{TN}{TN + FP}. \tag{3.77}$$

Specificity is usually helpful where the distinction of negatives is important, complementing sensitivity through assessing the handling of the negative class. In medical practice, specificity is an assurance that healthy ones are not incorrectly labeled to be suffering from a condition. This is especially important in tests leading to further diagnosis or treatments that may include risks or side effects.

3.8.4 Precision

Precision, or the positive predictive value, calculates the proportion of correctly predicted positive samples out of all samples predicted as positive [68]:

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{3.78}$$

Precision is especially relevant in scenarios where the cost of false positives is significant, ensuring the reliability of positive predictions.

3.8.5 F1-Score

The F1-score balances precision and recall by computing their harmonic mean, offering a single metric that considers both false positives and false negatives [66]:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{3.79}$$

This metric is particularly valuable in cases of imbalanced datasets, where a balance between precision and recall is crucial for meaningful evaluation.

3.8.6 Area Under the Receiver Operating Characteristic Curve (AUC)

The AUC quantifies the overall performance of a binary classifier by measuring the area under the Receiver Operating Characteristic (ROC) curve. The ROC graphically represents the relationship of the true positive rate versus the false positive rate at different thresholds. In fact, the ROC curve presents all the possible realizations on how well the classifier discriminates between positive and negative samples across all possible thresholds. The area under this curve AUC defines the overall performance of the classifier. The AUC is calculated as:

$$\text{AUC} = \int_0^1 \text{Sensitivity}(\text{FPR})\, d(\text{FPR}), \tag{3.80}$$

where the sensitivity is expressed as a function of the false positive rate, and denotes the infinitesimal change in with respect to the variation of the threshold. The AUC sums up these contributions, giving a measure of the ability of the model to balance sensitivity and specificity at any threshold. The performance of the perfect classifier has an AUC value of 1 where the ROC passes through the top-left corner, indicating 100% TPR and 0% FPR. An AUC close to 0.5 is representative of a random guessing model.

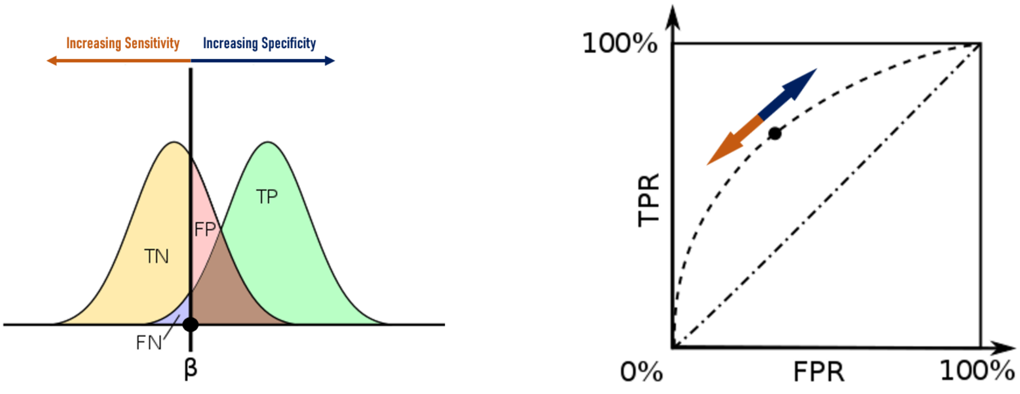
Figure 3.14a reflects the decision boundary between positive and negative classes generated by any threshold value β.

Overlapping regions of the two distributions, positive and negative, make evident the trade-off that is there between sensitivity, or True Positive Rate, and specificity, or True Negative Rate. That balance has been shifted by the threshold β. Lowering β will increase sensitivity in terms of more true positives being caught, but it raises the rate of false positives, therefore lowering the specificity. On the other hand, raising β will decrease the sensitivity while increasing the specificity. Figure 3.14b shows the ROC curve, which is a graph of the balance between the TPR and the FPR as the thresholds change. A ROC plot summarises a classifier performance under all possible operating conditions with regard to a balance of beneficial detections vs false alarms, separately showing both using TPR against FPR at various threshold settings. The measure of overall classifier performance can be given using the AUC.

These plots bring out the fact that AUC is an integral of sensitivity and specificity over all thresholds and is therefore a strong metric when comparing classifiers where the relative importance of these metrics changes with application. By visualizing decision boundary and ROC curve, they emphasize intrinsic trade-offs inherent in binary classification-the need to select an appropriate threshold as required by the application.

3.9 Local Interpretable Model-agnostic Explanations (LIME)

Local Interpretable Model-agnostic Explanations (LIME) is a popular explanation technique of interpretability for any machine learning model; it uses an interpretable model to approximate it locally. LIME relies on the assumption that big, complicated models can be linearized around a single point of prediction where the local



(a) Decision threshold (β)

(b) ROC Curve

Figure 3.14: (A) The relationship between sensitivity and specificity is governed by the choice of the decision threshold (β), illustrated here using distributions of positive and negative classes. (B) The ROC curve captures the classifier's performance across thresholds by plotting TPR against FPR.

surrogates can effectively approximate the decision boundaries. The key idea of LIME is to perturb the input data and observe the changes in the predictions of the model, thus finding the contributions of every feature to the prediction. [69]

LIME produces a model g - usually linear, such that it approximates well the behaviour of the original model f around the neighbourhood of a given instance x. Basically, the explanation model will be constructed by minimising the following objective:

$$\mathcal{L}(f, g, \pi_x) + \Omega(g), \tag{3.81}$$

where:

- $\mathcal{L}(f, g, \pi_x)$ is a loss function that measures the fidelity of g in approximating f locally around x,
- π_x is a locality measure that weights the samples by their proximity to x,
- $\Omega(g)$ represents the complexity of the explanation model, ensuring interpretability by penalizing overly complex models.

To generate perturbed samples, LIME first alters x by adding small random noise and calculates their predictions using f. Then, a weighted linear regression is performed on the perturbed samples using the proximity weight:

$$\pi_x(z) = \exp\left(-\frac{D(x,z)^2}{\sigma^2}\right), \tag{3.82}$$

where $D(x, z)$ is the distance metric between the original sample x and perturbed sample z, and σ controls the scale of locality. The coefficients of the linear model g provide the contribution of each feature in explaining $f(x)$, offering a human-understandable representation of the decision-making process.

LIME is model-agnostic, meaning it can be applied to any predictive model, including neural networks, decision trees, and support vector machines all the models that were discussed earlier in 3.6 that makes it versatile for many applications [69].

Chapter 4

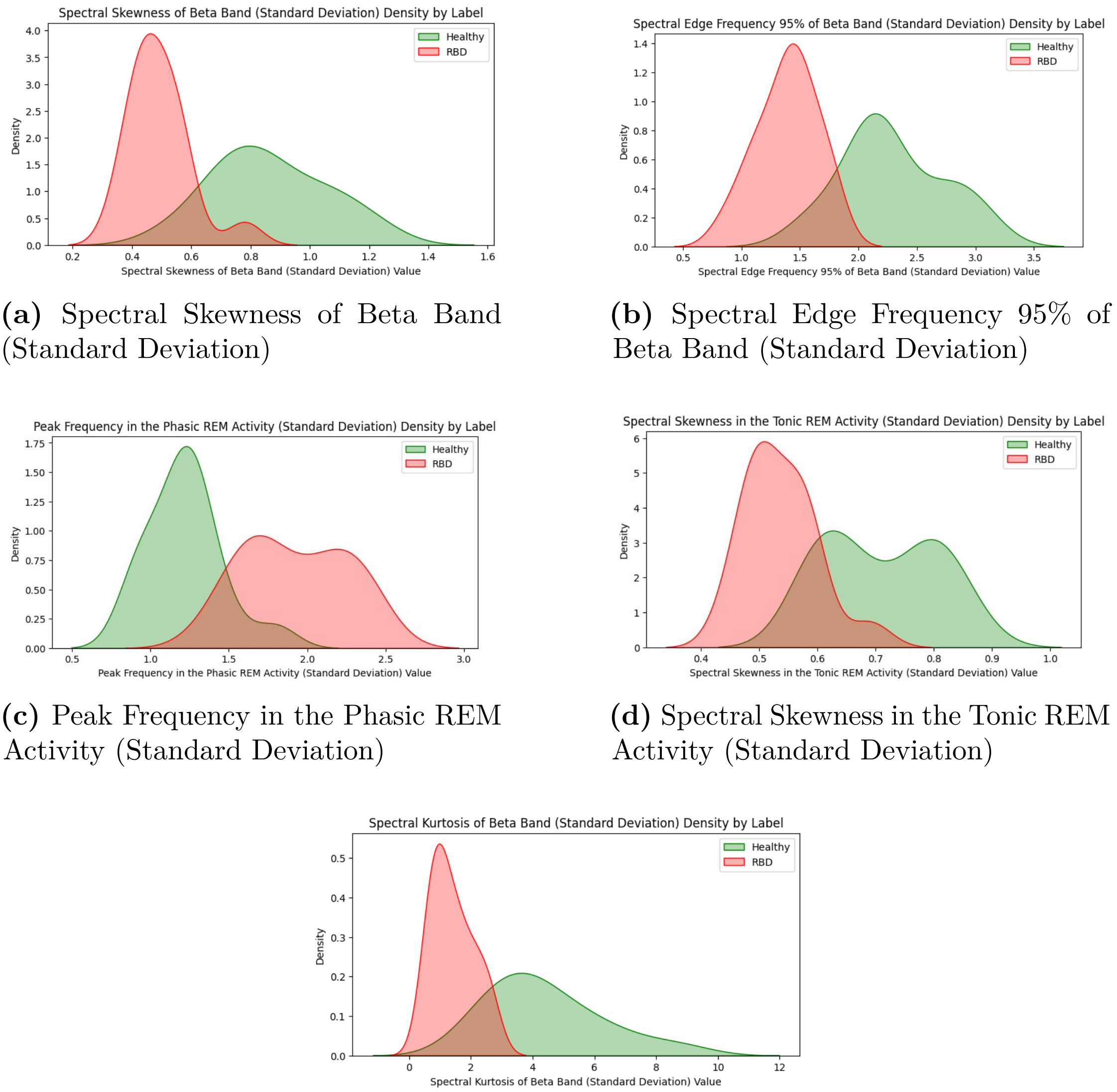
Results

This chapter provides the results of the analyses carried out on the two datasets: the CAP dataset and the TuSDi dataset. AAs highlighted in an earlier section, the main goal pursued in this paper is comparing how the inclusion of the N2 stage would have affected the performance of the classifier regarding the detection of RBD. The first part of the chapter is devoted to interpreting the selected features by means of a Kernel Density Estimation (KDE) plot. A KDE plot is a way of visualizing estimation of the underlying probability density function of a continuous random variable and provides a smooth curve representing the distribution of data points. Section 3.8 introduces the performance evaluation of the developed models using metrics related to accuracy, sensitivity, specificity, precision, F1 score, and AUC. Next, the evaluation of the models' performance is presented using various metrics, including accuracy, sensitivity, specificity, precision, F1 score, and AUC which were introduced in Section 3.8. Finally, the predictions of the machine learning models are explained using LIME-local interpretable model-agnostic explanations. This scheme explains in more understandable terms for non-technical clinicians so that they can convey proper interpretation to the results.

4.1 Selected Features Interpretations

4.1.1 N2 Stage

The higher standard deviation of a healthy group in Figure 4.1a suggests that there is more fluctuation in the skewness of the beta band, which perhaps reflects the normal variability in brain activity and stability of sleep. This can include natural transitions between deeper and lighter stages of N2 sleep, as well as minor arousals affecting beta power distribution. Although the lower standard deviation could indicate more skewedness in the beta band, it really means less fluctuation.



(a) Spectral Skewness of Beta Band (Standard Deviation)

(b) Spectral Edge Frequency 95% of Beta Band (Standard Deviation)

(c) Peak Frequency in the Phasic REM Activity (Standard Deviation)

(d) Spectral Skewness in the Tonic REM Activity (Standard Deviation)

(e) Spectral Kurtosis of Beta Band (Standard Deviation)

Figure 4.1: KDE plots of 5 top most important features for N2 stage.

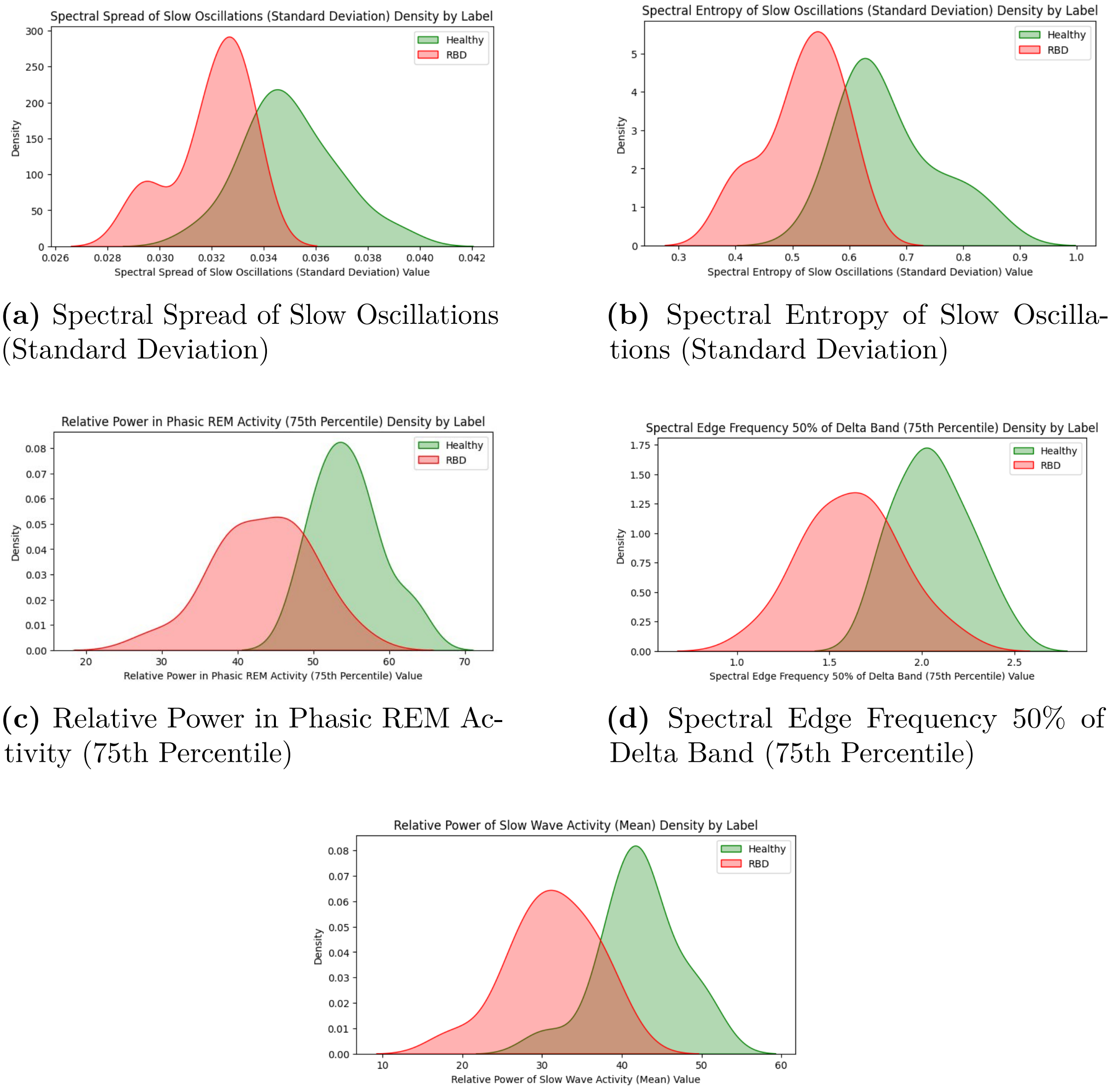
Such stability might suggest a disrupted sleep architecture with fewer transitions or changes in beta activity. In RBD, reduced variability may indicate a pattern of sustained arousal or uniform muscle activity that influences the spectral properties in less dynamic fashion.

In the healthy group, the spectral edge frequency values are greater, with more dispersion, centered at 2.2 in Figure 4.1b; this has suggested a dynamic beta power distribution during N2 sleep, reflecting normal fluctuations such as micro-arousals, muscular twitches, and characteristic physiological features of healthy sleep. In the RBD group, values are lower, peaking around 1.4, indicating reduced beta power variability. This stability may reflect abnormal motor or neural activity, perhaps causally related to disrupted muscle atonia, and might serve as an early marker of RBD.

The plot for the healthy group in Figure 4.1c is concentrated around lower values peaking around 1.0. That means that peak frequencies in FREM for the healthy group are more uniform. Such consistency is generally expected, as it reflects the unchanging rhythmic activity and typical sleep architecture without major disruptions. The RBD group is further spread out and shifted right toward higher standard deviation values, peaking around 1.5 and extending well into the range of 2.5. This would suggest great variability in the RBD group, which may reflect irregular shifts of peak frequency within the FREM. Such variation could reflect perturbations in sleep that coincided with irregular cortical activity or disturbed sleep architecture, perhaps indicative of RBD.

Figure 4.1d shows that the distribution for healthy individuals peaks at a higher range for standard deviation compared to the RBD group. This is indicative that, in healthy individuals, the skewness of spectral power is generally more variable in the TREM band. Variability in skewness could suggest a flexible but stable boundary between sleep and wake states with transitions into and out of each state occurring without disruption. This distribution for RBD individuals peaks at a lower standard deviation value, reflecting less variability in spectral skewness. This reduced variability could indicate that the power distribution in the TREM band is more rigid or always asymmetrical in some sense, which would reflect instabilities in the sleep-wake transition. In the case of RBD, such a lack of variability may underpin abnormal or shallower boundaries of sleep stages that could result in incomplete or disrupted transitions from one sleep stage into another.

The healthy subjects are more greatly dispersed across the higher values of spectral kurtosis in Figure 4.1e. This characterizes a higher dispersion in the standard deviation of spectral kurtosis for the healthy group within the beta band. Such a spread suggests that the beta activity in healthy individuals is less peaked and more variable, reflecting a more consistent distributed concentration of beta power across frequencies. While the distribution of RBD group values is sharply peaked around lower spectral kurtosis values, this would mean less dispersion in spectral kurtosis; more precisely, less standard deviation in spectral kurtosis. This would support, if interpreted accordingly, that the beta band activity in RBD patients has lesser variability and may be more concentrated in specific frequency peaks, leading to a higher sharpness in the power spectral density distribution.



(a) Spectral Spread of Slow Oscillations (Standard Deviation)

(b) Spectral Entropy of Slow Oscillations (Standard Deviation)

(c) Relative Power in Phasic REM Activity (75th Percentile)

(d) Spectral Edge Frequency 50% of Delta Band (75th Percentile)

(e) Relative Power of Slow Wave Activity (Mean)

Figure 4.2: KDE plots of 5 top most important features for REM stage.

4.1.2 REM Stage

Compared to the healthy subjects, the density curve in Figure 4.2a has shifted to higher values for healthy subjects peaking around 0.035. This reflects a wider distribution of low-frequency power due to greater variability in spectral spread in the slow oscillations band among healthy individuals. This would indicate that during REM sleep, healthy subjects have more distributed, even though low, slow oscillatory activity and would be expected from the minimal power in the SOs band during normal REM. The red density curve of the RBD group is more concentrated and peaks at a lower value at about 0.032. This reduced variability suggests a more narrow spread of power in the SOs band, possibly therefore indicating more focused or less variable low-frequency activity during REM.

Figure 4.2b shows that, for the healthy group, the curve is centered around higher values of standard deviation. This points to this healthy group having more variability of entropy within the SOs band as a reflection of the more dispersed or balanced power across frequencies. In general, expected from normal REM sleep with its sparse and less concentrated low-frequency oscillations. This lower entropy variability in the RBD group therefore suggests that the power in the SOs band is less uniformly distributed, tending to become concentrated in a few low frequencies. The source of such concentration might well be disruptions in expected REM architecture, either through arousal-like events or even muscle movements typical of RBD.

Figure 4.2c would suggest that in the frontal region, during REM, there is low-frequency activity; it is balanced with other frequencies so as to represent a relatively stable REM structure that has not been disrupted and without any motor activities. This balance fits with the characteristics one would expect from normal REM sleep: low-frequency dominance is limited in order not to interfere with the lighter, dream-active state of REM. The RBD group has a peak around a lower percentile, around the 40th percentiles, reflecting the fact that their 75th percentile of relative power in the low-frequency FREM band is reduced. Any such lower concentration may result from an overabundance of lower frequency power that could result from muscle twitches or other atypical motor activity during REM. This dominance of low-frequency power represents an increase over baseline and thus reflects a disruption of normal REM architecture, these subjects experiencing more disruptions or motor activity during REM, as is known to occur in RBD.

Figure 4.2d also reveals that in healthy subjects the delta band power is distributed at slightly higher frequencies within the 0-4 Hz range. The similarity between these SEF50 values indicates the expected low frequency activity in the delta band during REM where the delta power is at a minimum due to a lack of large, low-frequency oscillations. This stable pattern is consistent with normal REM sleep architecture in healthy subjects, characterized by atonia with low levels of low-frequency interruptions. The RBD group reaches its peak at a lower value of around 1.5 and spreads to even lower frequencies. Such a lower concentration of values suggests increased low-frequency activity within the delta band during REM sleep in the RBD group. This pattern may arise from disruptions in REM atonia-natural muscle relaxation during REM-or due to abnormal muscle movements that do indeed characterize RBD. This may then introduce low-frequency power atypical in REM sleep.

Similarly, the green curve for the healthy group in Figure 4.2e peaks at a higher mean value at approximately 45. That would say for healthy people the relative power in the SWA band during REM sleep lower in general, that is what we would expect. During REM sleep activity in the low frequency band is usually low because this band is highly presented in the deep stages of NREM. This higher mean in the healthy group would suggest a consistent REM sleep structure with minimal interference from low-frequency power. In contrast, the red curve for the RBD group peaks at about 30 for a correspondingly lower mean value and is shifted toward higher relative power in the SWA band. The emergence of higher SWA power during REM may reflect an abnormal emergence of low-frequency activity due to glsNREM, like transitions or micro-arousals. This low-frequency increase during REM in the RBD group may relate to muscle movements or disruptions.

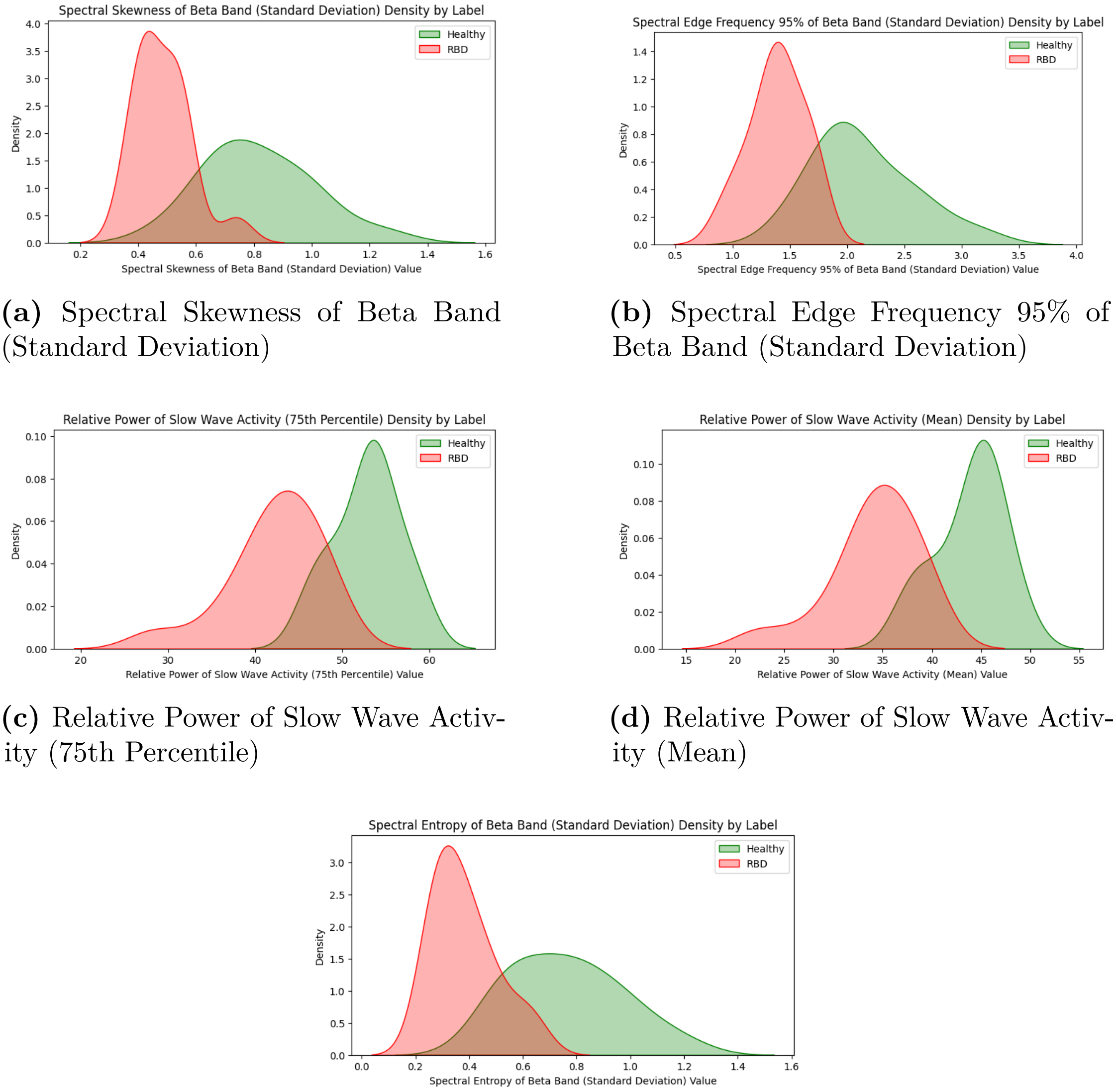
4.1.3 N2+REM Stages

As can be seen from Figure 4.3 , the combined use of N2 and REM stages, the N2 stage captures reduced variability and stability in RBD patients, namely, spectral skewness of beta band and spectral edge frequency 95% of beta band, while REM underlines muscle atonia disruption and low-frequency power, namely relative power of slow wave activity. Figure 4.3e represents the sixth most important feature ranking of stage N2. Although it is not considered during the N2 stage process, combined with the REM stage, it tends to show a more significant effect.

4.2 Binary classification Model Performance

4.2.1 CAP Dataset

Cross-validation was carried out on the CAP dataset to test the generalization of the models. Results are presented, grouped according to sleep stage, as N2, N3, REM, N2+REM, N3+REM, and N2+N3+REM. Overall performance of various models on sleep stages is summarized in Table 4.1 Confusion matrices of each class are summarized in Figure 4.4, where the number of correct and incorrect classifications made by the best models is depicted. Among these, the highest reported accuracy is 96.88% in the stages N3, N3+REM, and N2+N3+REM. The sensitivity of 100% has been found in many stages from the K-SVM and LR models. Moreover, the AUC also reached as high as 100% in the stages N3, N3+REM, and N2+N3+REM, which could hopefully provide very effective models during these stages.



(a) Spectral Skewness of Beta Band (Standard Deviation)

(b) Spectral Edge Frequency 95% of Beta Band (Standard Deviation)

(c) Relative Power of Slow Wave Activity (75th Percentile)

(d) Relative Power of Slow Wave Activity (Mean)

(e) Spectral Entropy of Beta Band (Standard Deviation)

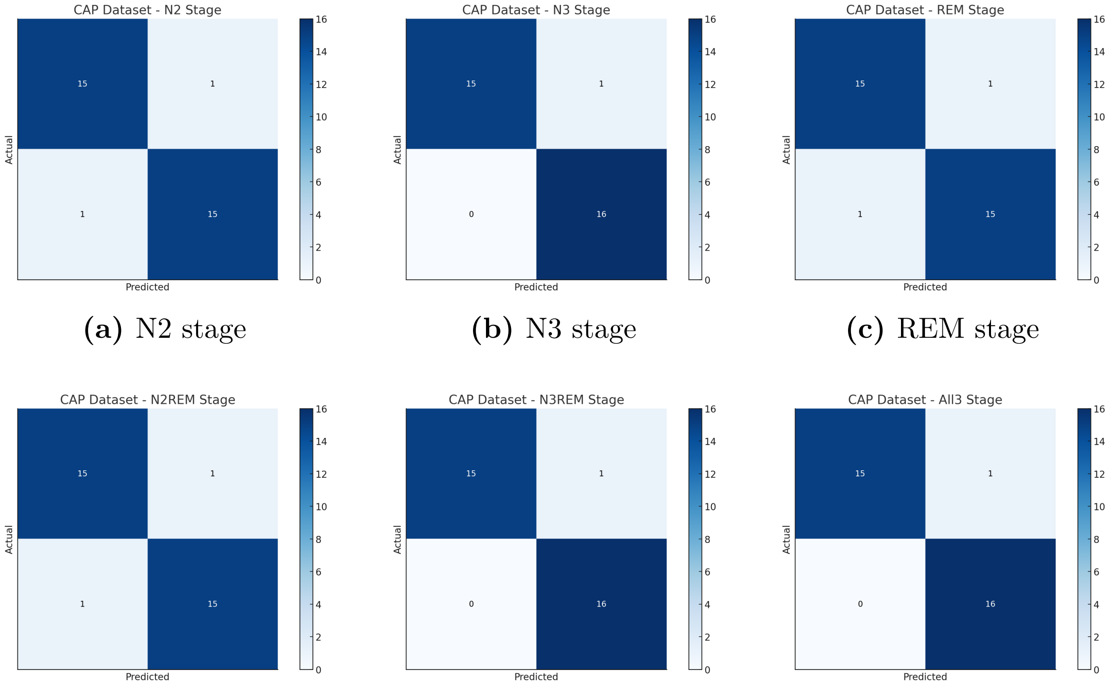
Figure 4.3: KDE plots of 5 top most imprtant features for N2+REM stages.

4.2.2 TuSDi Dataset

The models developed with the CAP dataset were tested with the TuSDi dataset to evaluate their performance on data they had never seen. Results of the test on TuSDi are summarized in Table 4.2 and the confusion matrices of each sleep stage are shown in Figure 4.5. The evaluation revealed that the highest performance, which was 85.00%, was achieved for the stages N2+N3+REM by the LR model. Sensitivity

Stage	Model	Accuracy	Sensitivity	Specificity	Precision	F1	AUC
N2	RF	93.75	93.75	93.75	93.75	93.75	96.50
N3	K-SVM	96.88	100.00	93.75	94.12	96.97	100.00
REM	LR	93.75	93.75	93.75	93.75	93.75	98.80
N2+REM	LR	93.75	93.75	93.75	93.75	93.75	99.20
N3+REM	SVM	96.88	100.00	93.75	94.12	96.97	100.00
N2+N3+REM	LR	96.88	100.00	93.75	94.12	96.97	100.00

Table 4.1: CAP Dataset Cross-validation Best Models Per Stage



(a) N2 stage

(b) N3 stage

(c) REM stage

(d) N2+REM stages

(e) N3+REM stages

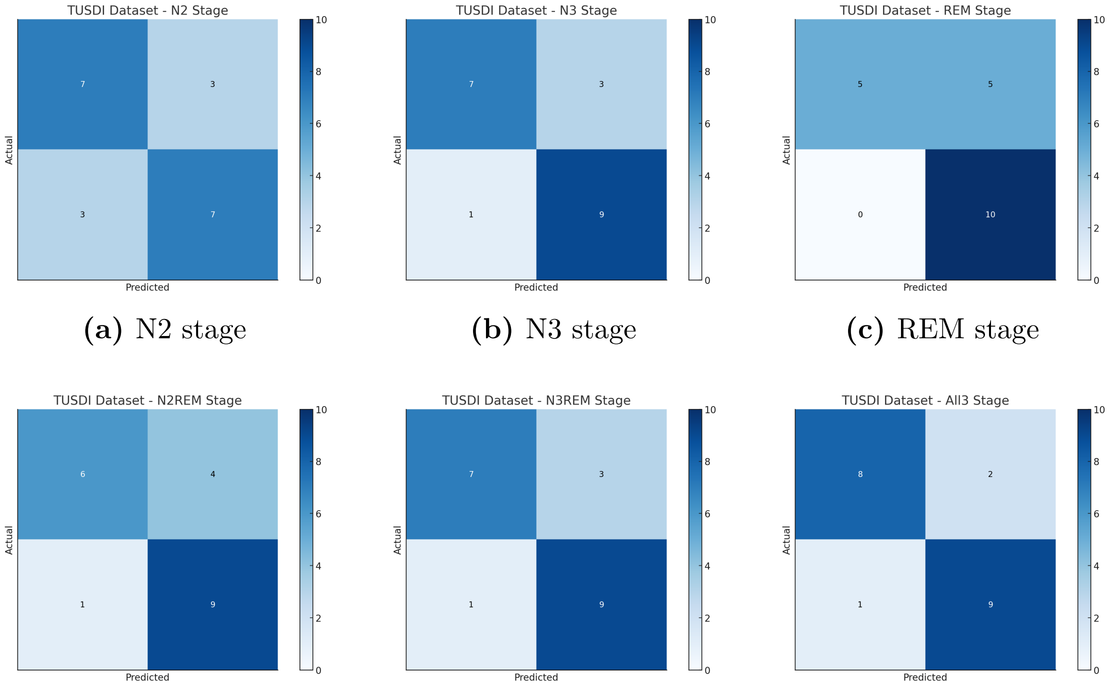
(f) N2+N3+REM stages

Figure 4.4: Confusion matrices for the CAP dataset across different stages.

was highest in the REM stage at 100%, showing how well the model performed in correctly identifying RBD subjects. Specificity had different maximum values across the stages: 80.00% in the N2+N3+REM stages. The F1 score, standing for a great balance between precision and recall, had the highest value in the N2+N3+REM stages of 85.71%. Besides, the high performance of the models demonstrated AUC values; the highest AUC value was 93.00% in the stages N2+REM.

Stage	Model	Accuracy	Sensitivity	Specificity	Precision	F1	AUC
N2	RF	70.00	70.00	70.00	70.00	70.00	84.00
N3	K-SVM	80.00	90.00	70.00	75.00	81.82	89.00
REM	LR	75.00	100.00	50.00	66.67	80.00	83.00
N2+REM	LR	75.00	90.00	60.00	69.23	78.26	93.00
N3+REM	SVM	80.00	90.00	70.00	75.00	81.82	90.00
N2+N3+REM	LR	85.00	90.00	80.00	81.82	85.71	87.00

Table 4.2: TuSDi Test Results Based on Best Models



(a) N2 stage **(b)** N3 stage **(c)** REM stage

(d) N2+REM stages **(e)** N3+REM stages **(f)** N2+N3+REM stage

Figure 4.5: Confusion matrices for the TuSDi dataset across different stages.

4.3 LIME Interpretation

LIME is a very powerful technique for explaining the contribution of individual features to specific predictions, as described earlier. In this section, we follow the case of subject one in an attempt to show and explain the results that LIME produced.

At stage N2, the model predicts the patient being healthy with 85% confidence. This decision was reached based on the key features identified: Spectral Skewness in the TREM Activity Standard Deviation and Spectral Skewness of Beta Band, standard deviation, each positively contributing to classifying it as "Healthy." However, the Peak Frequency in the Phasic REM Activity standard deviation became a contributing feature toward the "RBD" class, although of low influence. Insights from this stage are captured in Figure 4.6.

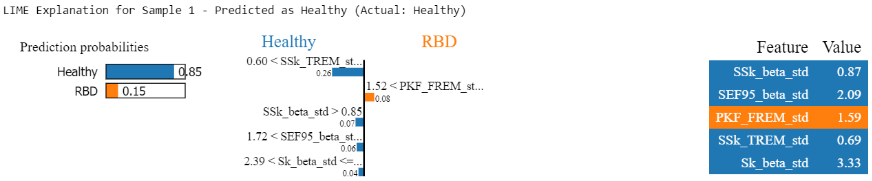


Figure 4.6: LIME explanation for N2 stage: Healthy classification.

When analyzing the N3 stage, the model was 98% confident to predict a patient as healthy. Features such as Coastline of Level 6 Band Detail Coefficient DWT (Mean) and Spectral Entropy of Beta Band (Standard Deviation) played an important role in this classification prediction. This stage proved very reliable in distinguishing the patients as being healthy. The results are illustrated in Figure 4.7.

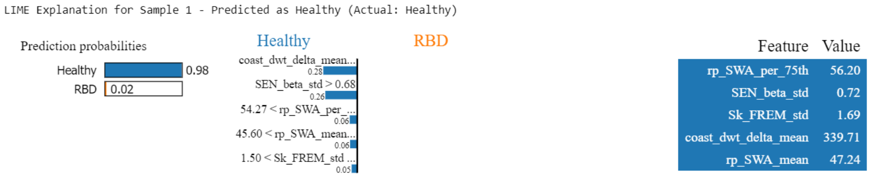


Figure 4.7: LIME explanation for N3 stage: Healthy classification.

In the REM stage, the model labeled the case as "RBD" with 56% confidence, thus showing a weak reliability of this stage in healthy case detection. Features such as pectral Edge Frequency 50% of Delta Band (75th Percentile) and Spectral Spread of Slow Oscillations (Standard Deviation) contributed towards the classification as "RBD," while other features provided mixed signals and did not help the decision-making. The LIME explanation for this stage is shown in Figure 4.8.

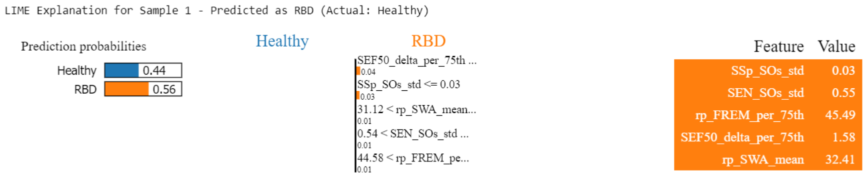
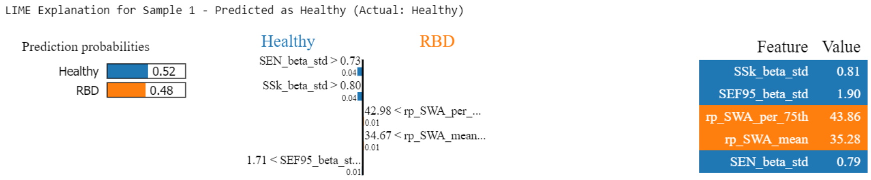


Figure 4.8: LIME explanation for REM stage: Healthy classification.

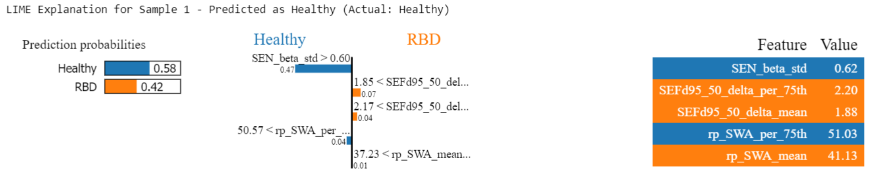
The combination of N2+REM stages provides a balanced evaluation of the model's ability to classify the sample. In the given case, the model predicted the patient to be healthy with a moderate degree of confidence. The features contributing to the prediction, such as Spectral Entropy of Beta Band (Standard Deviation) and Spectral Skewness of Beta Band (Standard Deviation), influenced the healthy classification, while Relative Power of Slow Wave Activity (Mean) leaned towards the "RBD" classification. This combination highlights the interplay between the features derived from N2 and REM stages, emphasizing the importance of incorporating multiple sleep stages for a comprehensive assessment.



Feature	Value
SSk_beta_std	0.81
SEF95_beta_std	1.90
rp_SWA_per_75th	43.86
rp_SWA_mean	35.28
SEN_beta_std	0.79

Figure 4.9: LIME explanation for N2+REM stages: Healthy classification.

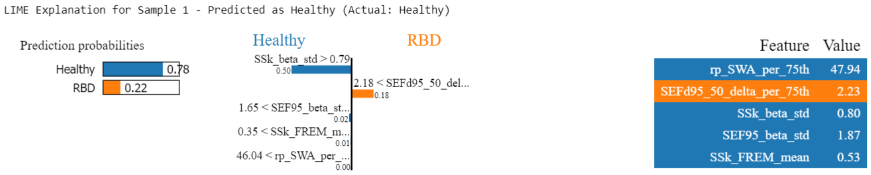
When combining the N3+REM stages, the model seemed to have much more confidence in tagging the sample as healthy. The most relevant feature for the "Healthy" classification was Spectral Entropy of Beta Band (Standard Deviation), while for "RBD", important feature was Spectral Edge Frequency 95% - 50% of Delta Band. Again, a stage combination that emphasizes the strong diagnostic power of including N3 with REM data is highlighted, which reflects the findings in Figure 4.10.



Feature	Value
SEN_beta_std	0.62
SEFd95_50_delta_per_75th	2.20
SEFd95_50_delta_mean	1.88
rp_SWA_per_75th	51.03
rp_SWA_mean	41.13

Figure 4.10: LIME explanation for N3+REM stages: Healthy classification.

The model classified the patient as healthy with high confidence in the combination of N2+N3+REM stages, supported by feature such as Spectral Skewness of Beta Band (Standard Deviation) strongly contributed to making the outcome lean towards being healthy. Feature like Spectral Edge Frequency 95% - 50% of Delta Band (75th Percentile) contributed a little towards the RBD classification, indicating small ambiguities. This all-inclusive approach leverages the strengths of all three stages to deliver a robust and well-rounded classification, as illustrated in Figure 4.11.



Feature	Value
rp_SWA_per_75th	47.94
SEFd95_50_delta_per_75th	2.23
SSk_beta_std	0.80
SEF95_beta_std	1.87
SSk_FREM_mean	0.53

Figure 4.11: LIME explanation for N2+N3+REM stages: Healthy classification.

Such information can provide clinicians with more insight into which characteristics of which part of sleep contribute to RBD. Additionally, in some cases, their diagnoses may conflict with model results, and these findings could help them understand why.

Chapter 5

Conclusion and Future Works

In this paper, we reported a detailed analysis of PSG data using machine learning techniques, showing how the N2 stage particularly affects the detection of RBD. It was found from the results that the N2 stage, traditionally a less emphasized stage as compared to REM sleep, improves the accuracy of RBD detection. It has also been pointed out that the N2 stage is highly critical in maintaining sleep stability and, as such, disruption in this stage may only be an early indication of neurological disorders. Application of machine learning models using Random Forest and Logistic Regression classified the healthy individuals from those suffering from RBD by features extracted across time, frequency, and nonlinear domains. They also found that the addition of the N2 stage data to the REM data enhanced the detection reliability for the RBDs and believed that this may indicate a possible benefit of the use of the analysis in the N2 stage for either earlier or more valid diagnoses. Machine learning was particularly useful in this context due to its extensive capability of capturing and classifying complex patterns of interest present in the PSG data, hence aiding the diagnosis for the RBD. The results showed that the REM stage of sleep is itself very sensitive in the detection of RBD, though it lacks specificity between normal individuals and those with RBD. The reason is that REM analysis alone may highlight features from RBD but cannot segregate them from normal sleep variations in healthy subjects. Again, the model was seen to show increased specificity with the incorporation of the N2 sleep stage data alongside REM data, whereby it could distinctly differentiate between healthy and RBD patients. Adding features from the deep restorative sleep stage N3 again improved this model. Finally, we found that the combined consideration of stage N2, stage N3, and REM yielded the best performance in terms of sensitivity and specificity, underlining that possibly a full-spectrum consideration of sleep architecture might be important in the case of accurate detection and diagnosis of RBD.

In the future, we want to expand the dataset used in this study to further validate the robustness of the proposed methodology on a wider population across the spectrum of different age groups and diverse backgrounds. Adding more data into the dataset makes it not only more diverse but also adds enough samples for the training of advanced algorithms that can find better patterns. There may be much better ways, especially for the current deep learning models like Convolutional Neural Networks (CNNs) and Spectral Vision Transformers (SViT), to improve sleep stage classification performance and supplement diagnosis accuracy. The development of portable, less-invasive gadgets for home-based monitoring is further suggested in future studies to make this work more practical. Single-channel EEG systems are a potentially low-cost, user-friendly solution that may enable early-stage RBD detection. Further, studies based on the correlation of characteristics in the N2 stage with neurodegenerative markers may lead to insights into neurological disorders such as Parkinson's disease and Lewy body dementia at an early stage, thus contributing to timely interventions with better efficiency.

Chapter 6

Appendix

Model	Accuracy	Sensitivity	Specificity	Precision	F1	AUC
KNN	87.50	87.50	87.50	87.50	87.50	94.34
RF	93.75	93.75	93.75	93.75	93.75	96.48
DT	90.63	93.75	87.50	88.24	90.91	90.63
NB	90.63	93.75	87.50	88.24	90.91	97.27
LR	90.63	93.75	87.50	88.24	90.91	92.97
SVM	93.75	97.80	87.50	88.89	93.12	95.66
K-SVM	84.38	81.25	87.50	86.67	83.87	89.45

Table 6.1: CAP Dataset Cross-validation Results for N2 Stage

Model	Accuracy	Sensitivity	Specificity	Precision	F1	AUC
KNN	96.88	100.00	93.75	94.12	96.97	99.61
RF	90.63	87.50	93.75	93.33	90.32	98.24
DT	84.38	75.00	93.75	92.31	82.76	86.52
NB	96.88	100.00	93.75	94.12	96.97	99.61
LR	96.88	100.00	93.75	94.12	96.97	100.00
SVM	96.88	100.00	93.75	94.12	96.97	98.83
K-SVM	96.88	100.00	93.75	94.12	96.97	100.00

Table 6.2: CAP Dataset Cross-validation Results for N3 Stage

Model	Accuracy	Sensitivity	Specificity	Precision	F1	AUC
KNN	90.63	87.50	93.75	93.33	90.32	92.38
RF	90.63	93.75	87.50	88.24	90.91	97.66
DT	81.25	81.25	81.25	81.25	81.25	78.32
NB	90.63	93.75	87.50	88.24	90.91	98.05
LR	93.75	93.75	93.75	93.75	93.75	98.22
SVM	90.63	87.50	93.75	93.33	90.32	98.44
K-SVM	90.63	93.75	87.50	88.24	90.91	98.83

Table 6.3: CAP Dataset Cross-validation Results for REM Stage

Model	Accuracy	Sensitivity	Specificity	Precision	F1	AUC
KNN	87.50	87.50	87.50	87.50	87.50	95.51
RF	90.63	87.50	93.75	93.33	90.32	96.44
DT	84.38	81.25	87.50	86.67	83.87	87.89
NB	93.75	93.75	93.75	93.75	93.75	98.22
LR	93.75	93.75	93.75	93.75	93.75	98.22
SVM	93.75	93.75	93.75	93.75	93.75	98.83
K-SVM	87.50	87.50	87.50	87.50	87.50	94.53

Table 6.4: CAP Dataset Cross-validation Results for N2+REM Stage

Model	Accuracy	Sensitivity	Specificity	Precision	F1	AUC
KNN	96.88	100.00	93.75	94.12	96.97	99.80
RF	90.63	87.50	93.75	93.33	90.32	98.24
DT	78.13	75.00	81.25	80.00	77.42	85.55
NB	96.88	100.00	93.75	94.12	96.97	99.22
LR	93.75	100.00	87.50	88.89	94.12	99.22
SVM	96.88	100.00	93.75	94.12	96.97	99.22
K-SVM	96.88	100.00	93.75	94.12	96.97	100.00

Table 6.5: CAP Dataset Cross-validation Results for N3+REM Stage

Model	Accuracy	Sensitivity	Specificity	Precision	F1	AUC
KNN	96.88	100.00	93.75	94.12	96.97	96.29
RF	90.63	87.50	93.75	93.33	90.32	98.83
DT	81.25	81.25	81.25	81.25	81.25	83.79
NB	93.75	93.75	93.75	93.75	93.75	99.22
LR	96.88	100.00	93.75	94.12	96.97	100.00
SVM	96.88	100.00	93.75	94.12	96.97	99.61
K-SVM	90.63	87.50	93.75	93.33	90.32	99.61

Table 6.6: CAP Dataset Cross-validation Results for N2+N3+REM Stage

Bibliography

[1] Matthew Walker. «The Role of Sleep in Cognition and Emotion». In: *Annals of the New York Academy of Sciences* 1156 (Apr. 2009), pp. 168–197. DOI: 10.1111/j.1749-6632.2009.04416.x (cit. on p. 1).

[2] Oskar G Jenni and Mary A Carskadon. «Normal human sleep at different ages: Infants to adolescents». In: *SRS basics of sleep guide* (2005), pp. 11–19 (cit. on p. 2).

[3] Elisa Perger, Rosalia Silvestri, Enrica Bonanni, Maria Caterina Di Perri, Mariana Fernandes, Federica Provini, Giovanna Zoccoli, and Carolina Lombardi. «Gender medicine and sleep disorders: from basic science to clinical research». In: *Frontiers in Neurology* 15 (July 2024). DOI: 10.3389/fneur.2024.1392489 (cit. on p. 2).

[4] Vasili Kharchenko and Irina V. Zhdanova. *Structure of sleep explained by wave mechanics.* bioRxiv preprint. 2023. DOI: 10.1101/2023.01.19.524817 (cit. on p. 2).

[5] Célia Lacaux, Thomas Andrillon, Isabelle Arnulf, and Delphine Oudiette. «Memory loss at sleep onset». In: *Cerebral Cortex Communications* 3.4 (2022). DOI: 10.1093/texcom/tgac042 (cit. on p. 3).

[6] Jinbo Sun, Yi ni She, Xiao Zeng, Liming Lu, Chun-Zhi Tang, Nenggui Xu, Badong Chen, and Wei Qin. «A two-branch trade-off neural network for balanced scoring sleep stages on multiple cohorts». In: *Frontiers in Neuroscience* 17 (2023). DOI: 10.3389/fnins.2023.1176551 (cit. on p. 3).

[7] Gianna C. L. Migliaccio, Ian M. Greenlund, Jeremy A. Bigalke, Jennifer Nicevski, Anne L. Tikkanen, and Jason R. Carter. «Beta intrusion during N2 sleep is negatively associated with nocturnal heart rate variability in healthy adults». In: *Physiology* 38.S1 (2023). DOI: 10.1152/physiol.2023.38.s1.5731406 (cit. on p. 3).

[8] Ning Ma, Qian Ning, Mingzhu Li, and Chao Hao. «The First-Night Effect on the Instability of Stage N2: Evidence from the Activity of the Central and Autonomic Nervous Systems». In: *Brain Sciences* 13.4 (2023), p. 667. DOI: `10.3390/brainsci13040667` (cit. on p. 3).

[9] Hsin-Hao Tseng, Sheng-Wei Hwang, Shang-Rung Hwang, and Juen-Haur Hwang. «Sleep apnea plays a more important role on sleep N3 stage than chronic tinnitus in adults». In: *Medicine* 101 (2022). DOI: `10.1097/MD.0000000000030089` (cit. on p. 3).

[10] Carlyle Smith, Jocelyn Aubrey, and Kevin Peters. «Different Roles for REM and Stage 2 Sleep In Motor Learning: A Proposed Model». In: *Psychologica Belgica. Special Issue: Cognition in Slumberland. Mechanisms of Information Processing in the Sleep-Wake Cycle.* Sleep Research Society, 2004, pp. 81–104. DOI: `10.4324/9780203307991_chapter_5` (cit. on p. 3).

[11] Daniel P. Brunner, Derk-Jan Dijk, Irene Tobler, and Alexander A. Borbély. «Effect of partial sleep deprivation on sleep stages and EEG power spectra: evidence for non-REM and REM sleep homeostasis». In: *Electroencephalography and Clinical Neurophysiology* 75.6 (1990), pp. 492–499. ISSN: 0013-4694. DOI: `10.1016/0013-4694(90)90136-8`. URL: `https://www.sciencedirect.com/science/article/pii/0013469490901368` (cit. on p. 4).

[12] Joshua Feriante and John F. Araujo. *Physiology, REM Sleep.* Treasure Island (FL): StatPearls Publishing, 2023. URL: `http://europepmc.org/books/NBK531454` (cit. on p. 4).

[13] 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 (2021), pp. 58–65. DOI: `10.1049/htl2.12007`. eprint: `https://ietresearch.onlinelibrary.wiley.com/doi/pdf/10.1049/htl2.12007`. URL: `https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/htl2.12007` (cit. on pp. 4, 11, 24).

[14] 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: *Bioinformatics and Biomedical Engineering.* Ed. by Ignacio Rojas, Olga Valenzuela, Fernando Rojas, Luis Javier Herrera, and Francisco Ortuño. Cham: Springer International Publishing, 2022, pp. 381–394. ISBN: 978-3-031-07704-3 (cit. on pp. 4, 12, 19).

[15] Sizhi Ai, Shuo Ye, Guohua Li, Yue Leng, Katie L. Stone, Min Zhang, Y. Wing, Jihui Zhang, and Yan Liang. «Association of Disrupted Delta Wave Activity During Sleep With Long-Term Cardiovascular Disease and Mortality». In:

Journal of the American College of Cardiology (2024). DOI: `10.1016/j.jacc.2024.02.040` (cit. on p. 5).

[16] Sara Fattinger, Salome Kurth, Maya Ringli, Oskar G. Jenni, and Reto Huber. «Theta waves in children's waking electroencephalogram resemble local aspects of sleep during wakefulness». In: *Scientific Reports* 7.11187 (2017). DOI: `10.1038/s41598-017-11577-3` (cit. on p. 5).

[17] Aygun Asgarli, Matthew Banks, and Kirill Nourski. «0047 Intracranial EEG Reveals Human Cerebral Cortical Regions That Are Spared from Alpha-delta Sleep». In: *Sleep* (2024). DOI: `10.1093/sleep/zsae067.0047` (cit. on p. 5).

[18] Ajay K. Verma et al. «Slow-wave sleep dysfunction in mild parkinsonism is associated with excessive beta and reduced delta oscillations in motor cortex». In: *Frontiers in Neuroscience* 18 (2024). DOI: `10.3389/fnins.2024.1338624` (cit. on p. 5).

[19] Mario Valderrama et al. «Human gamma oscillations during slow wave sleep». In: *PLOS ONE* 7.4 (2012). DOI: `10.1371/journal.pone.0033477` (cit. on p. 5).

[20] Israel Soares Pompeu de Sousa Brasil and Renatha El Rafihi-Ferreira. «Insomnia». In: *Sleep Disorders: Diagnosis and Therapeutics*. 2024, pp. 31–41. DOI: `10.1007/978-3-031-50710-6_4` (cit. on p. 6).

[21] Urszula Guderska, Agnieszka Urbanek, Jacek Kurzeja, D. Maciejewska, Adrianna Rasmus-Czternasta, Magdalena Bartczak, and Filip Czternasty. «Sleep apnea and comorbid diseases - a review». In: *Journal of Education, Health and Sport* (2024). DOI: `10.12775/jehs.2024.58.009` (cit. on p. 6).

[22] J. T. Srikanta and K. M. Chandan Kumar. «Narcolepsy - A Rare but Underrecognized Problem in Children». In: *Indian Pediatrics* 56.2 (2019), p. 147 (cit. on p. 6).

[23] Salvio Serrano Ortega, M.a Amparo Sangil, Domingo Cañizo, and Patricio Pérez. «Iron deficiency and Restless Legs Syndrome». In: 25.100 (2023). DOI: `10.60147/f62a1452` (cit. on p. 6).

[24] Basheer Khassawneh. «Periodic Limb Movement Disorder». In: *Sleep Disorders and Neurological Disease*. John Wiley and Sons, Inc., 2005, pp. 483–486. DOI: `10.1002/0471751723.ch62` (cit. on p. 6).

[25] John Harrington and Teofilo Lee-Chiong. «Circadian Rhythm Sleep Disorders». In: *Sleep Medicine*. 2013, pp. 60–70. DOI: `10.3920/9789086867639_005` (cit. on p. 6).

[26] Michael S. Aldrich. «Parasomnias». In: *Sleep Medicine*. 1999, pp. 260–287. DOI: `10.1093/oso/9780195129571.003.0015` (cit. on p. 7).

[27] Naoko Tachibana. «REM Sleep Behavior Disorder». In: *Sleep Medicine Clinics* 6.4 (2011), pp. 459–468. DOI: `10.1016/j.jsmc.2011.08.009` (cit. on p. 7).

[28] R. Malhotra. «REM Sleep Behavior Disorder». In: *Encyclopedia of Sleep*. Elsevier Inc., 2013, pp. 209–213. DOI: `10.1016/B978-0-12-378610-4.00427-7` (cit. on pp. 7, 8).

[29] Carlos H. Schenck. «REM Sleep Behavior Disorder». In: *Handbook of Clinical Neurology*. 2014, pp. 12–14. DOI: `10.1016/B978-0-12-385157-4.00569-8` (cit. on pp. 7, 8).

[30] Imran Khawaja, Benjamin C. Spurling, and Shantanu Singh. *REM Sleep Behavior Disorder*. StatPearls [Internet]. 2020 (cit. on pp. 7, 8).

[31] Alex Iranzo de Riquer. «REM Sleep Behavior Disorder». In: *Handbook of Clinical Neurology*. Springer, Vienna, 2017, pp. 361–366. DOI: `10.1007/978-3-7091-1628-9_35` (cit. on p. 8).

[32] Thiago Barral F. Reis, Michel P. Tcheou, and Felipe Da Rocha Henriques. «Detecting Sleep Disorders in Polysomnography Data». In: *Proceedings of the IEEE Latin American Symposium on Circuits and Systems (LASCAS)* (2024), pp. 1–5. DOI: `10.1109/lascas60203.2024.10506123` (cit. on p. 9).

[33] Beatrice Go and Erica R. Thaler. «Home Sleep Testing versus Traditional Polysomnography». In: *Otolaryngologic Clinics of North America* 57.3 (2024), pp. 363–369. DOI: `10.1016/j.otc.2023.11.003` (cit. on p. 9).

[34] Simon Frenkel, Ashen Amaranayake, Charlotte Molesworth, Liliana Orellana, and Anne Marie Southcott. «Allowing ad libitum sleep during overnight polysomnography impacts Multiple Sleep Latency Test results in patients being assessed for hypersomnolence». In: *Journal of Clinical Sleep Medicine* (2024). DOI: `10.5664/jcsm.11302` (cit. on p. 9).

[35] «Polysomnography». In: *Encyclopedia of Sleep Medicine*. Elsevier eBooks, 2023, pp. 374–387. DOI: `10.1016/b978-0-12-822963-7.00174-2` (cit. on p. 9).

[36] Paulo Bugalho, Manuel Salavisa, Cláudia Borbinha, Marco Fernandes, Bruna Meira, Raquel Barbosa, and Maressa Mendonça. «REM sleep behaviour disorder in essential tremor: A polysomnographic study». In: *Journal of Sleep Research* 30 (Apr. 2020). DOI: `10.1111/jsr.13050` (cit. on p. 10).

[37] Mohamed Abdelfattah, Oliver Sum-Ping, Joanna Galati, S. Marwaha, Alexandre Alahi, and Emmanuel During. «0689 Automated Detection of Isolated REM Sleep Behaviour Disorder Using Computer Vision». In: *Sleep* (2024). DOI: `10.1093/sleep/zsae067.0689` (cit. on p. 10).

[38] George Adaimi, Niraj Gupta, Ali Mottaghi, Serena Yeung, Emmanuel Mignot, Alexandre Alahi, and Emmanuel During. «0641 Automated Detection of Isolated REM Sleep Behavior Disorder (iRBD) During Single Night In-Lab Video-Polysomnography (PSG) Using Computer Vision». In: *Sleep* 45 (May 2022), A282–A282. DOI: 10.1093/sleep/zsac079.638 (cit. on p. 11).

[39] Katarina Mary Gunter, Andreas Brink-Kjær, Emmanuel Mignot, Helge Bjarup Dissing Sørensen, Emmanuel H. During, and Poul Jennum. «SViT: a Spectral Vision Transformer for the Detection of REM Sleep Behavior Disorder». In: *IEEE Journal of Biomedical and Health Informatics* PP (2023). DOI: 10.1109/JBHI.2023.3292231 (cit. on p. 11).

[40] Matteo Cesari et al. «Comparison of computerized methods for rapid eye movement sleep without atonia detection». In: *Sleep* 41.10 (2018). DOI: 10.1093/sleep/zsy133 (cit. on p. 11).

[41] Navin Cooray, Fernando Andreotti, Christine Lo, Mikael Symmonds, Michele Hu, and Maarten de Vos. «Detection of REM Sleep Behaviour Disorder by Automated Polysomnography Analysis». In: *Clinical Neurophysiology* 130 (Feb. 2019). DOI: 10.1016/j.clinph.2019.01.011 (cit. on p. 11).

[42] Alexandra Papakonstantinou, Jannis Klemming, Martin Haberecht, Dieter Kunz, and Frederik Bes. «Ikelos-RWA. Validation of an Automatic Tool to Quantify REM Sleep Without Atonia». In: *Clinical EEG and Neuroscience* 55 (May 2023), p. 155005942311753. DOI: 10.1177/15500594231175320 (cit. on p. 11).

[43] Gabriele S. Giarrusso. «Machine Learning Strategies for Single-Channel EEG Automatic Detection of REM Sleep Behavior Disorder: A Model Based on REM and Slow Wave Sleep». Master's Thesis. Politecnico di Torino, 2023 (cit. on pp. 12, 24).

[44] 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: Jan. 2020. DOI: 10.24251/HICSS.2020.396 (cit. on p. 12).

[45] 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 (2019), pp. 37–46. DOI: 10.1016/j.smrv.2018.09.008 (cit. on p. 12).

[46] Matteo Cesari et al. «Machine Learning Predicts Phenoconversion from Polysomnography in Isolated REM Sleep Behavior Disorder». In: *Brain Sciences* 14.9 (2024). ISSN: 2076-3425. DOI: 10.3390/brainsci14090871. URL: https://www.mdpi.com/2076-3425/14/9/871 (cit. on p. 12).

[47] M.G. Terzano, Liborio Parrino, and Arianna Smerieri. «Atlas, rules, and recording techniques for the scoring of cyclic alternating pattern (CAP) in human sleep». In: *Sleep Med* 3 (Jan. 2001), pp. 187–199 (cit. on p. 15).

[48] 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 p. 31).

[49] Q. Zhu, Y. Wang, and T. Smith. «High-Dimensional EEG Analysis for Sleep Stage Characterization». In: *Medical Data Analysis* 25 (2018), pp. 112–121 (cit. on p. 32).

[50] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006 (cit. on p. 32).

[51] F. Li and M. Zhou. «Interpretability in High-Dimensional Medical Models». In: *Journal of Clinical Data* 34 (2020), pp. 10–22 (cit. on p. 32).

[52] I. Guyon and A. Elisseeff. «An Introduction to Variable and Feature Selection». In: *Journal of Machine Learning Research* 3 (2003), pp. 1157–1182 (cit. on p. 33).

[53] R. Kohavi and G. John. «Wrappers for Feature Subset Selection». In: *Artificial Intelligence* 97 (1997), pp. 273–324 (cit. on p. 35).

[54] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2009 (cit. on pp. 35, 40, 51–53).

[55] Mostafa Karami. «Machine Learning Algorithms for Radiogenomics: Application to Prediction of the MGMT promoter methylation status in mpMRI scans». PhD thesis. Politecnico di Torino, 2022 (cit. on p. 35).

[56] I.T. Jolliffe and J. Cadima. «Principal Component Analysis and its Applications». In: *Philosophical Transactions of the Royal Society A* 374 (2016), p. 20150202 (cit. on p. 36).

[57] Z. Zhao and H. Liu. «Spectral Feature Selection for Unsupervised Learning». In: *Proceedings of the 24th International Conference on Machine Learning* (2007), pp. 1151–1157 (cit. on p. 36).

[58] Jianping Hua, Zixiang Xiong, James Lowey, Edward Suh, and Edward R. Dougherty. «Optimal number of features as a function of sample size for various classification rules». In: *Bioinformatics* 21.8 (Nov. 2004), pp. 1509–1515. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bti171. eprint: https://academic.oup.com/bioinformatics/article-pdf/21/8/1509/48973153/bioinformatics_21_8_1509.pdf. URL: https://doi.org/10.1093/bioinformatics/bti171 (cit. on p. 37).

[59] David Roxbee Cox. «The regression analysis of binary sequences». In: *Journal of the Royal Statistical Society: Series B (Methodological)* 20.2 (1958), pp. 215–232 (cit. on p. 43).

[60] George H John and Pat Langley. «Estimating continuous distributions in Bayesian classifiers». In: *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc. 1995, pp. 338–345 (cit. on p. 45).

[61] Corinna Cortes and Vladimir Vapnik. «Support-vector networks». In: *Machine Learning* 20.3 (1995), pp. 273–297 (cit. on pp. 46, 47).

[62] J Ross Quinlan. «Induction of decision trees». In: *Machine learning* 1.1 (1986), pp. 81–106 (cit. on p. 48).

[63] Leo Breiman. «Random forests». In: *Machine learning* 45.1 (2001), pp. 5–32 (cit. on p. 49).

[64] Sylvain Arlot and Alain Celisse. «A Survey of Cross Validation Procedures for Model Selection». In: *Statistics Surveys* 4 (July 2009). DOI: 10.1214/09-SS054 (cit. on p. 51).

[65] V. J. Raja, Dhanamalar M, Gautam Solaimalai, D. L. Rani, P. Deepa, and R. G. Vidhya. «Machine Learning Revolutionizing Performance Evaluation: Recent Developments and Breakthroughs». In: *Proceedings Article* (2024). DOI: 10.1109/icscss60660.2024.10625103 (cit. on p. 54).

[66] Shu Geng. «Analysis of the Different Statistical Metrics in Machine Learning». In: *Highlights in Science Engineering and Technology* (2024). DOI: 10.54097/jhq3tv19 (cit. on pp. 54–56).

[67] Abdulvahap Pinar, Cemil ÇOLAK, and Esra Gültürk. «Evaluation of Performance Metrics in Heart Disease by Machine Learning Techniques». In: *The Journal of Cognitive Systems* 8 (June 2023). DOI: 10.52876/jcs.1276688 (cit. on pp. 54–56).

[68] Oona Rainio, Jarmo Teuho, and Riku Klén. «Evaluation metrics and statistical tests for machine learning». In: *Dental Science Reports* (2024). DOI: 10.1038/s41598-024-56706-x (cit. on p. 56).

[69] Marco Ribeiro, Sameer Singh, and Carlos Guestrin. «"Why Should I Trust You?": Explaining the Predictions of Any Classifier». In: Aug. 2016, pp. 1135–1144. DOI: `10.1145/2939672.2939778` (cit. on pp. 58, 59).