Diagnostic Tools for Rapid Eye Movement Sleep Behaviour Disorder



Navin Cooray

St Cross College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2020

Dedicated to
Amma, Thatha, & Malla
for their endless love and support

Abstract

There is clear evidence to support Rapid-Eye-Movement (REM) sleep behaviour disorder (RBD) as an early predictor of Parkinson's disease, dementia with Lewy bodies, and multiple system atrophy. Those diagnosed with RBD enact their dreams and therefore present an abnormal characteristic of movement during REM sleep. This unusual sleep disorder presents an opportunity to better understand neurological degeneration and in-turn develop preventative medicine. However, manual polysomnograph (PSG) analysis is mandatory when diagnosing RBD, which demands complex monitoring equipment, specially trained staff, bed availability, and clinicians to interpret abundant amounts of data. Inevitably this process is time consuming and expensive, creating a demand for automated diagnostic support tools to better utilise the time of clinicians.

The first aim of this thesis was to achieve automated sleep stage classification using signal processing and machine learning techniques. A total of 159 features were calculated using Electroencephalography (EEG), Electrooculography (EOG), and Electromyography (EMG) signals. After extracting the features a Random Forest (RF) algorithm was used to classify five sleep stages. With respect to REM stage classification, this technique achieved an accuracy of 93%, sensitivity of 63%, and specificity of 97% (five stage Cohen kappa of 0.63).

After sleep staging, clinicians must manually identify abnormal REM sleep behaviour. Several methods have been proposed for RBD detection that use EMG recordings to objectively quantify abnormal REM movement. In this work we further develop these proven techniques with additional features that incorporate the relationship of muscle movement between sleep stages and general sleep architecture. Using an RF classifier with established and additional features, the performance of RBD detection was shown to improve upon established metrics (achieving 94% accuracy, 94% sensitivity, and 94% specificity). This technique used in combination with automated sleep staging sustained a high performance in RBD detection, with an accuracy, sensitivity, and specificity of 90%, 86%, and 94%, respectively.

Lastly, this thesis investigated a minimal set of sensors to fully-automate the detection of RBD, integrating automated sleep staging followed by RBD detection without the need for cumbersome EEG sensors. This thesis demonstrated a cost-effective, practical, and simple RBD identification support tool using an EOG and EMG sensor to achieve an accuracy, sensitivity, and specificity of 86%, 82%, and 90%, respectively.

Acknowledgements

Personal

Upon beginning my doctorate, I had only a vague understanding about the task I was about to undertake, but knew I would grow personally and professionally. None of this would have been possible without the guidance and support of many people.

Firstly, my sincerest gratitude to my supervisor, Prof. Maarten De Vos. I want to thank you for your guidance and enthusiasm over the years. Without you I would have never even started this project, which began from a simple meeting and ended up producing my thesis, entailing multiple post-doctoral positions, and motivating a multi-hospital longitudinal wearable sleep study. Beginning my research after working as a professional engineer was awfully daunting at first, but your guidance and mentorship was responsible for transforming my skills into a competent and published academic researcher. I am forever grateful.

I would like thank all those neurologists, specialists, and doctors at the John Radcliffe hospital, your professionalism and clinical expertise were a crucial and vital part of my work. Thank you Dr Zenobia Zaiwalla and Dr Michele Hu, your enthusiasm and guidance have been invaluable throughout the years and enabled the foundations upon which my work could even begin. Thank you to Dr Mkael Symmonds and Dr Christine Lo, your professionalism, determination, and clinical expertise were paramount to obtaining and anonymising the sleep data necessary for my research. Once more, your clinical analysis and knowledge were an inspiration to my work, which benefited greatly from your feedback and continual guidance.

Thank you to my colleagues in the CIBIM research group and those in the institute of biomedical engineering, your support, guidance, and friendship over the years have left an indelible imprint on my work and time at Oxford. Our lunches and coffee breaks have been a highlight of my time here and were a constant source of joy and intellectual stimulation.

To the friends I have made at Oxford and St Cross College, I am truly blessed for the time and memories we have shared. My mental and physical well being have been nourished by the numerous pot-luck dinners, impromptu dancing, exhausting games of squash, epic cricket fixtures, competitive board games, and many shared pints.

To my Australian friends in the UK and back in Australia, your love, support, and friendship have been an inspiration and a source of great strength while working

and living a world away from home. To Paul, Lucy, Pat, and Emma, I am forever grateful to have you in my life, you made me feel at home in the UK while living $16,516$ kilometres from Brisbane.

To the Andradi and Peiris families, thank you for being my family away from family. Living and working away from home was only tolerable because you all embraced me as family. I will always cherish the memories of Christmas parties, family gatherings, and home-cooked Sri-Lankan food.

To my flatmate, Sharinee, thank you for your ongoing support and encouragement through the final stages of the doctorate. Your company and friendship through the COVID-19 lockdown were invaluable to my sanity and well-being. I will always fondly remember our delicious home-cooked meals and nursery of vibrant plants.

Finally, to my family, I am eternally thankful for all your love and support. To my mother and father, thank you for always encouraging and believing in me to achieve what I thought was beyond my grasp. To my brother, thank you for motivating and inspiring me to take risks and push myself. To our dog Fido, thank you for your unconditional love, you truly are a good boy.

Institutional

I would like to thank Rotary International, your charity and organisation have given me a sense of community and enabled me to pursue my dreams. Special thanks to my home district, Jindalee as well as Haddenham and District for their support over the years.

Contents

List of Abbreviations **xiii**

List of Notations **xvii**

1 Introduction **1**
1.1 Motivation . . . 1
1.2 Statement of the problem . . . 3
1.2.1 Sleep . . . 4
1.2.2 Sleep staging rules . . . 7
1.2.3 Incidence and prevalence of RBD . . . 12
1.2.4 RBD as a Predictor . . . 12
1.2.5 Behavioural features of RBD . . . 13
1.2.6 Diagnosing RBD . . . 14
1.2.7 Treatment . . . 14
1.3 Research aims . . . 15
1.4 Thesis Structure . . . 18
1.5 Research publications . . . 19

2 Datasets **21**
2.1 Introduction . . . 21
2.2 Background . . . 22
2.3 Montreal Archive of Sleep Studies . . . 23
2.4 John Radcliffe Hospital . . . 24
2.5 Cyclic Alternating Pattern Sleep Database . . . 25
2.6 Combined Dataset . . . 26
2.7 Conclusion . . . 29

3 Automated Sleep Stage Classification **31**
3.1 Associated Publication . . . 31
3.2 Introduction . . . 32
3.3 Background . . . 33
3.3.1 Automated Sleep Staging . . . 38
3.4 Methods . . . 48

3.4.1 Data . . . 48
3.4.2 Signal Pre-Processing . . . 48
3.4.3 Feature Extraction . . . 49
3.4.4 Electroencephalogram Feature Extraction . . . 50
3.4.5 Electrooculogram Feature Extraction . . . 57
3.4.6 Electromyogram Feature Extraction . . . 61
3.4.7 Miscellaneous Feature Extraction . . . 65
3.4.8 Random Forest Classification . . . 67
3.4.9 Model training and evaluation . . . 70
3.5 Results . . . 73
3.6 Discussion . . . 84
3.7 Conclusion . . . 90

4 Automated Detection of REM Sleep Behaviour Disorder 93
4.1 Associated Publication . . . 93
4.2 Introduction . . . 94
4.3 Background . . . 95
4.3.1 Quantification of Abnormal Muscle Movement . . . 95
4.3.2 RBD Detection . . . 97
4.4 Methods . . . 99
4.4.1 Data . . . 99
4.4.2 Pre-processing . . . 99
4.4.3 Extract Abnormal Muscle Movement Metrics . . . 99
4.4.4 RBD Classification . . . 102
4.4.5 Evaluation . . . 102
4.5 Results . . . 103
4.6 Discussion . . . 107
4.7 Conclusion . . . 112

5 Automated End-to-End RBD Detection 115
5.1 Associated Publication . . . 115
5.2 Introduction . . . 116
5.3 Background . . . 117
5.3.1 Automated Sleep Stage Classification . . . 117
5.3.2 Automated RBD Detection . . . 118
5.4 Methods . . . 118
5.4.1 Data . . . 118
5.4.2 Signal Pre-Processing . . . 119
5.4.3 Feature Extraction . . . 119
5.4.4 Automated Sleep Staging . . . 119

5.4.5 RBD Detection . . . 119
5.4.6 Evaluation . . . 120
5.5 Results . . . 121
5.5.1 RBD Detection . . . 121
5.5.2 Correlation of Automated RBD Metrics . . . 121
5.5.3 Bland-Altman Analysis of Automated RBD Metrics . . . 122
5.5.4 Automatic RBD Classification . . . 124
5.6 Discussion . . . 126
5.7 Conclusion . . . 130

6 Automated RBD Detection with Minimal Sensors 133
6.1 Associated Publication . . . 134
6.2 Introduction . . . 134
6.3 Background . . . 136
6.3.1 The Autonomic Nervous System . . . 136
6.3.2 R-Peak Detection . . . 136
6.3.3 Automated Sleep Staging with Heart Rate . . . 138
6.3.4 Summary . . . 142
6.3.5 Autonomic Dysfunction in RBD . . . 142
6.3.6 Minimal Automated Sleep Staging . . . 143
6.4 Methods . . . 143
6.4.1 Data . . . 143
6.4.2 Pre-processing . . . 144
6.4.3 Electrocardiogram Feature Extraction . . . 144
6.4.4 Minimal Sleep Stage Classification . . . 157
6.4.5 Extract Autonomic Dysfunction Metrics . . . 158
6.4.6 Minimal RBD Detection . . . 159
6.4.7 Evaluation . . . 160
6.5 Results . . . 161
6.5.1 Minimal Automated Sleep Staging . . . 161
6.5.2 Minimal RBD Detection . . . 162
6.5.3 Bland-Altman Analysis on RBD Metrics . . . 165
6.6 Discussion . . . 165
6.7 Conclusion . . . 174

7 Conclusion and Future Work **177**
7.1 Introduction . . . 177
7.2 Summary . . . 178
7.2.1 Automated Sleep Stage Classification . . . 178
7.2.2 Automated RBD Detection . . . 179
7.2.3 End-to-End RBD Detection . . . 180
7.2.4 Minimal End-to-End RBD Detection . . . 181
7.3 Limitations and future work . . . 182
7.3.1 Data . . . 182
7.3.2 Automated Sleep Staging . . . 182
7.3.3 RBD Detection . . . 183
7.3.4 Take-home Settings . . . 184
7.4 Conclusion . . . 185

Appendices

A Additional Algorithms **189**
A.1 Signal Decomposition . . . 189
A.1.1 Discrete Fourier Transform (DFT) . . . 189
A.1.2 Discrete Wavelet Transform (DWT) . . . 190

References **193**

List of Abbreviations

AASM . . . . . American Academy of Sleep Medicine.

AF . . . . . Atrial fibrillation.

AHI . . . . . Apnoea Hypopnea Index.

ANN . . . . . Artificial Neural Network.

ANS . . . . . Autonomic nervous system.

AR . . . . . Autoregressive.

AUC . . . . . Area-Under-the-Curve.

B&A . . . . . Bland-Altman.

BC-SVM . . . Binary Class Support Vector Machine.

BCG . . . . . Ballistocardiography

BPM . . . . . Beats per minute.

CAP . . . . . Cyclic alternating pattern.

CNS . . . . . Central nervous system.

CTRG . . . . . Clinical Trials and Research Governance.

CV . . . . . Cross-validation.

CWD . . . . . Choi-Williams distribution.

CWT . . . . . Continuous wavelet transform.

DFT . . . . . Discrete Fourier Transform.

DWT . . . . . Discrete Wavelet Transform.

ECG . . . . . Electrocardiogram.

EDF . . . . . European Data Format.

EDR . . . . . ECG Derived Respiratory.

EEG . . . . . Electroencephalography

EMG . . . . . Electromyography.

EOG . . . . . Electrooculography.

FFT . . . . . Fast Fourier Transform.

FIR . . . . . Finite Impulse Response.

FN . . . . . False negative.

FP . . . . . False positive.

HC . . . . . Healthy Control.

HF . . . . . High frequency.

HMM . . . . . Hidden Markov Models.

HHT . . . . . Hilbert-Huang Transform.

HRV . . . . . Heart rate variability.

ICSD . . . . . International Classification of Sleep Disorders.

iRBD . . . . . Idiopathic-RBD.

KMCFW . . . K-means clustering based feature weighting.

k-NN . . . . . K-nearest neighbour.

LBD . . . . . Lewy bodies dementia.

LDA . . . . . Linear Discriminant Analysis.

LF . . . . . Low frequency.

LZ . . . . . Lempel-Ziv.

MASS . . . . . Montreal archive of sleep studies.

MLP . . . . . Multi-layer perceptrons.

MODWT . . . Maximum overlap discrete wavelet transform.

MSA . . . . . Multiple Systems Atrophy.

NHS . . . . . National Health Service.

OSA . . . . . Obstructive Sleep Apnoea.

OC-SVM . . . One-Class Support Vector Machine.

PCA . . . . . Principle Component Analysis.

PLM . . . . . Periodic Limb Movement.

PD . . . . . Parkinson's disease.

PSG . . . . . Polysomnography.

QMA . . . . . Quantified Motor Activity.

RBF . . . . . Radial Basic Function.

REM . . . . . Rapid eye movement.

RF . . . . . Random Forest.

R&K . . . . . Rachtschaffen and Kales.

RSWA . . . . . REM sleep without atonia.

iRBD . . . . . Idiopathic REM sleep behaviour disorder.

RBD . . . . . REM sleep behaviour disorder.

SEF . . . . . Spectral edge frequency.

SEM . . . . . Slow eye movement.

SEO . . . . . Square Energy Operator.

SPWVD . . . Smoothed pseudo Wigner-Ville distribution.

SVM . . . . . Support Vector Machine.

TKEO . . . . . Teager-Kaiser energy operator.

TN . . . . . True negative.

TP . . . . . True positive.

ZCI . . . . . Zero Crossing Interval.

List of Notations

$\boldsymbol{x}$	Column or row vector $\boldsymbol{x}$.
$\boldsymbol{X}$	Matrix, $\boldsymbol{X}$.
x_n	The n^{th} element of $\boldsymbol{x}$.
$\boldsymbol{x'}$	First derivative of $\boldsymbol{x}$.
$\boldsymbol{x''}$	Second derivative of $\boldsymbol{x}$.
x_n^*	The complex conjugation of the n^{th} element of $\boldsymbol{x}$.
$\in$	Elements of the set.
$\mathbb{R}$	The set of real numbers.
$\mathbb{N}$	The set of natural numbers.
$\mathbb{R}^{1\times N}$	Row vector of real numbers of length N.
$\mathcal{FT}$	Discrete Fourier Transform.
$\lvert . \rvert$	Absolute value.

Come Sleep, O Sleep, the certain knot of peace,
The baiting place of wit, the balm of woe,
The poor man's wealth, the prisoner's release,
The indifferent judge between the high and low...

— Sir Philip Sidney's *Sonnet 39*

1 Introduction

Contents

1.1 Motivation 1
1.2 Statement of the problem 3
1.2.1 Sleep 4
1.2.2 Sleep staging rules 7
1.2.3 Incidence and prevalence of RBD 12
1.2.4 RBD as a Predictor 12
1.2.5 Behavioural features of RBD 13
1.2.6 Diagnosing RBD 14
1.2.7 Treatment 14
1.3 Research aims 15
1.4 Thesis Structure 18
1.5 Research publications 19

1.1 Motivation

Rapid-eye-movement (REM) sleep Behaviour Disorder (RBD) is a parasomnia whereby patients physically enact their dreams during REM sleep. During healthy REM sleep, individuals typically exhibit muscle atonia, where skeletal muscles relax and are inactive. Muscle atonia during REM is the result of active inhibition of motor activity by specific regions of the brain and is understood to be quite complex [1]. It has been shown to be the result of a combination of inactivation

of brain-stem motor inhibitory systems and brain-stem facilitative systems [2, 3]. Those diagnosed or suspected with having RBD no longer exhibit muscle atonia during REM, which can result in movement during this period of sleep. Often dreams can be vivid and involve intense movement that may lead to self-injury or injury to bed partners [1]. RBD can either be idiopathic (iRBD) or secondary, where idiopathic infers the cause is unknown, whereas secondary RBD are associated with a diagnosed degenerative disorder.

As more patients are identified and studied over time, there is mounting evidence indicating RBD as a precursor to Parkinson's disease (PD) that may precede the disease by years, potentially decades [4, 5, 6, 7, 8, 9, 10, 11]. PD is the second most prevalent neurological disorder in the world (after Alzheimer's disease) and is characterised by involuntary shaking, slow movements, and inflexible muscles [12]. Recent research involving autopsies of probable RBD cases also found a strong association between RBD and synucleinopathies [9], thus potentially linking RBD to a wider array of disorders. Synucleinopathies describe a group of neurodegenerative disorders that are defined by an accumulation of alpha-synuclein protein in neurons. This category of disorders are clinically characterised by chronic and progressive decline in motor, cognitive, behavioural, and autonomic functions that include PD, Dementia with Lewy Bodies (LBD), and Multiple Systems Atrophy (MSA) [13, 14]. LBD describes a form of dementia, where protein deposits (called Lewy bodies) develop around the brain regions associated with thinking, memory and motor control [15]. MSA is also a neurological disorder identified by the presence of abnormally folded glial cells (non-neuronal cells found in the brain) [16]. While RBD is not apparent in all PD sufferers it has been known to occur in 40-60% of PD patients [5] and 80% or more of patients with RBD eventually develop PD or LBD [6, 7, 8].

Chen *et al.* (2013) believe that even though this body of research suggests idiopathic-RBD (iRBD) may be isolated to a degeneration of REM atonia, the majority of cases are associated with an underlying pathology that develops over time and research in this field may shed light on neural circuitry, underlying REM

sleep, motor control, and the development of neurodegenerative disorders [10]. Potentially, this very predictive sleep disorder would prove useful in epidemiological studies and in the development of preventative medicine.

Clinically, polysomnography (PSG) is necessary in establishing the diagnosis of RBD, but this process demands monitoring equipment, specially trained staff, bed availability and clinicians who can interpret the complex and abundant amount of data. Inevitably, this process is laborious and expensive [17, 18]. Due to the limited number of sleep clinics and sleep professionals, it is often impractical to perform PSG on large number of patients in epidemiological studies of sleep disorders. This is especially pertinent in the context of RBD, where episodes of dream enactments do not occur every night. Therefore, there is increasing demand to develop take-home automated diagnostic tools to better utilise the time of clinicians and provide objective measurements to help diagnose this disorder. Once more, PSG monitoring itself can be uncomfortable and clunky, as depicted in Figure 1.1 and would benefit from a more ergonomic and minimal solution.

It is the aim of this thesis to develop clinically useful diagnostic tools, which can identify individuals with RBD. This is done by providing a decision support tool that automatically processes electroencephalography (EEG), electrooculography (EOG), and electromyography (EMG) to detect REM sleep. This is the crucial first step towards diagnosing RBD, which is subsequently followed by automatically identifying sleep characteristics associated with RBD. Lastly, this work evaluates the minimal use of PSG sensors to effectively achieve fully-automated RBD detection.

1.2 Statement of the problem

Before introducing the specific research objectives and structure of the thesis, this chapter begins by outlining the basics of clinical sleep analysis, including its challenges and limitations. Next, the prevalence and the relevance of RBD as a predictor are investigated. This is followed by the current methodology in RBD identification, focusing on key challenges in clinical practise. Finally, a set

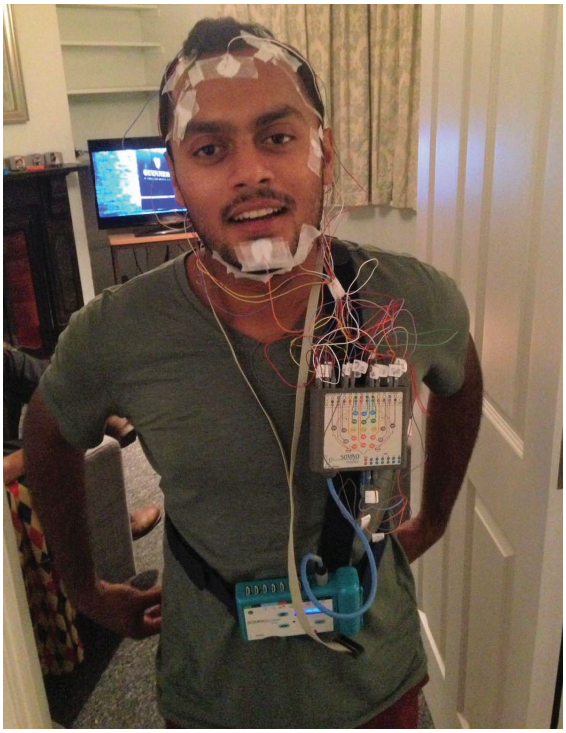


Figure 1.1: An example of a take-home PSG device, which include EEG, EOG, and EMG sensors. These devices can be very uncomfortable and are difficult to administer. Photographed by Dr Brian Chu in 2017.

of research questions and objectives emerging from this analysis are introduced, motivating the structure of the thesis.

1.2.1 Sleep

As a simple behavioural definition, sleep is defined as a "reversible behavioural state of perceptual disengagement from and unresponsive to the environment" [1]. Normal human sleep comprises of two states, REM and Non-Rapid Eye Movement (NREM). NREM itself can be understood to consist of several distinct stages which are defined by physiological and neuronal activity. The onset of sleep begins in NREM and progresses through deeper NREM stages before the first occurrence of REM sleep. This first instance of REM typically occurs 80-100 minutes into sleep [1] and is the state associated with dreaming. The cycle of NREM sleep and REM sleep continue in a periodic fashion.

NREM is typically divided into three stages that are identifiable from EEG recordings, namely N1, N2, and N3. These three stages also depict depth-of-sleep,

EEG Bandwidth	Frequency Range (Hz)
Delta (δ)	$0.5 - 4$
Theta (θ)	$4 - 8$
Alpha (α)	$8 - 12$
Beta (β)	$12 - 30$
Gamma (γ)	>30

Table 1.1: EEG bandwidth activity

with N1 typically having the lowest arousal threshold and N3 with the highest [1]. These three NREM stages have distinct EEG patterns with low activity signatures found in EOG and EMG recordings.

In order to diagnose various sleep disorders clinicians perform sleep studies which rely on visual analysis of PSG recordings. Clinicians score sleep stages based on the visual inspection of EEG, EOG, and EMG signals on the basis of pre-specified epochs (typically 30 seconds) thus obtaining the hypnogram which expresses the sleep stage sequence during the participant's sleep. EEG signals are typically described by their frequency and commonly characterised by the bandwidths shown in Table 1.1. Furthermore EEG events can also be used to characterise EEG activity in distinct transient waveforms such as vertex sharp waves, sleep spindles, and K-complexes. To correctly acquire these neuronal characteristics, EEG electrodes are placed in the manner detailed in Figure 1.2.

Additionally EOG signals are able to illustrate identifying sleep characteristics in the form of eye movement activity (slow and rapid eye movements). REM is characterised by EOG movement patterns that are irregular with sharp peaks that usually last less than 500 milliseconds. While Slow Eye Movement (SEM) are characterised by reasonably regular sinusoidal EOG patterns that usually last longer than 500 milliseconds. To correctly capture EOG characteristics, electrodes are placed in the locations detailed in Figure 1.3.

EMG signals are able to provide insight into the level of muscle tone, which provides vital information regarding muscle activation and relaxation that are regulated in healthy sleep. Healthy REM sleep is associated with muscle atonia and complete muscle relaxation and hence a low amplitude signal. Whereas during

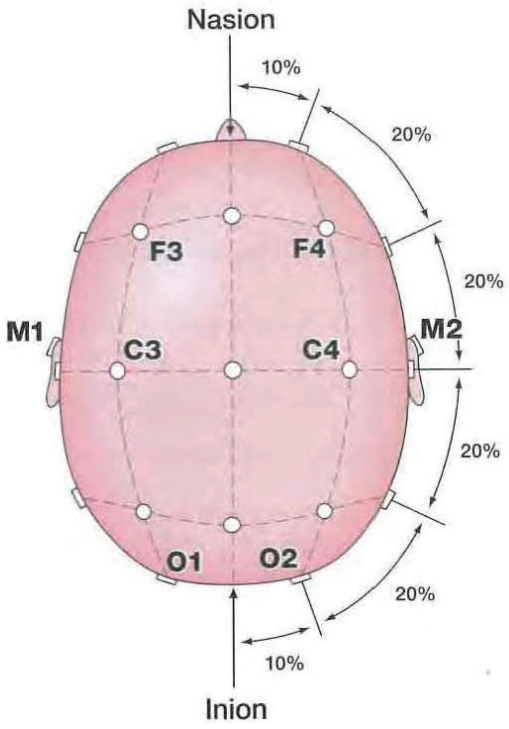


Figure 1.2: These are the recommended EEG electrode placement as specified by the American Academy of Sleep Medicine[1].



Figure 1.3: These are the recommended EOG electrode placement as specified by the American Academy of Sleep Medicine[1].

wakefulness and other NREM sleep EMG is associated with tonic and phasic activity, represented by a constant elevated amplitude or periodic changes in amplitude, respectively. These characteristics can also provide insight into numerous sleep disorders. EMG electrodes are attached in the manner depicted in Figure 1.4.

[1]Extracted from R. Berry et al. *The AASM Manual for Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications.* 2.0. American Academy of Sleep Medicine, 2012.

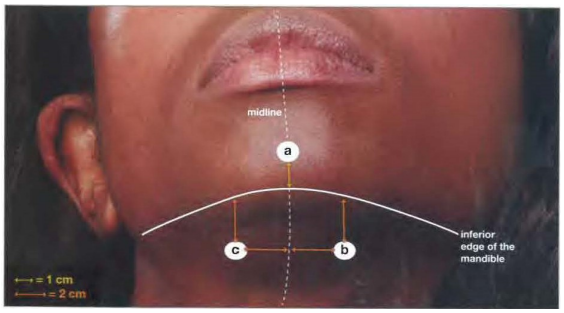


Figure 1.4: These are the recommended EMG electrode placements as specified by the American Academy of Sleep Medicine[1].

1.2.2 Sleep staging rules

Visual scoring remains the gold standard in sleep staging and the following section details the general scoring rules defined by the American Academy of Sleep Medicine (AASM). The initial guidelines to sleep staging were introduced in 1968 by Rachtschaffen and Kales (R&K) [20], where each 30 second recording was labelled either W, S1, S2, S3, S4 or REM. In 2007 the AASM commissioned a new scoring manual that essentially grouped S3 and S4 stages together into a single stage known as N3, since their distinction had no known physiological or clinical basis [19]. Therefore the AASM defined sleep stages as either W, N1, N2, N3 or REM. Additionally the AASM stipulates the use of three EEG electrodes (two were specified in R&K), two EOG electrodes and one EMG electrode when scoring PSG recordings. This thesis has used the AASM terminology and definitions [19].

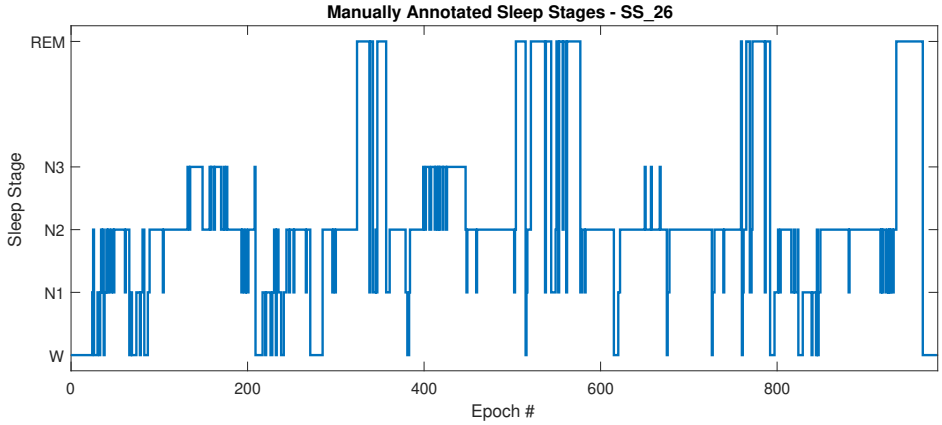


Figure 1.5: An example of manually annotated sleep stages for a healthy control, where every thirty second epoch has been assigned a sleep stage (W, N1, N2, N3 or REM) by a clinical expert.

Wakefulness (W), is defined by a relaxed individual with eyes closed and can be identified by the presence of alpha EEG activity in more than half of the epoch. Within this stage, eye blinks, SEM, and REM are also expected. Interestingly, about 10% of the population does not generate alpha rhythms on eye closure, which results in individuals with low amplitude activity chiefly in the alpha and beta frequencies [1].

Stage N1 has mixed EEG frequencies and is identified by the attenuated alpha rhythm and has low voltage amplitude with no instances of sleep spindles or K-complexes. Eye blinks cease during this stage, although SEM may still be found within EOG signals. In healthy sleep, this stage is the least prevalent stage of sleep.

N2 is identified if either or both a K-complex and train of sleep spindles are seen at the start of an epoch or the last half of the previous epoch. Epochs are continued to be scored as N2 provided EEG signals are low amplitude, contain mixed frequencies, and preceded by an epoch with spindles or K-complexes. In normal sleep, N2 is the most prevalent and typically accounts for 50% of total sleep [1].

Stage N3 is typically associated with slow wave activity with high amplitude and a frequency range between 0.5Hz and 2Hz for more than 20% of the epoch. Sleep spindles may appear within this stage. This stage of sleep accounts for 12.5% to 20% of total normal sleep [1].

Epochs are scored as REM when there is low amplitude, mixed frequency EEG, low chin EMG tone, and rapid eye movements depicted in the EOG. Epochs are continued to be scored as REM even in the absence of rapid eye movement, as long as there is low amplitude, mixed frequency EEG activity without K-complexes or sleep spindles, and EMG tone remains low. In normal sleep, REM stages typically accounts for 20%-25% of total sleep [1].

Inter-rater variability

Although specific rules are defined for sleep staging, there can be a substantial difference in inter/intra-rater variability and is well documented in sleep literature [21, 22]. A popular metric used in sleep literature to measure agreement between

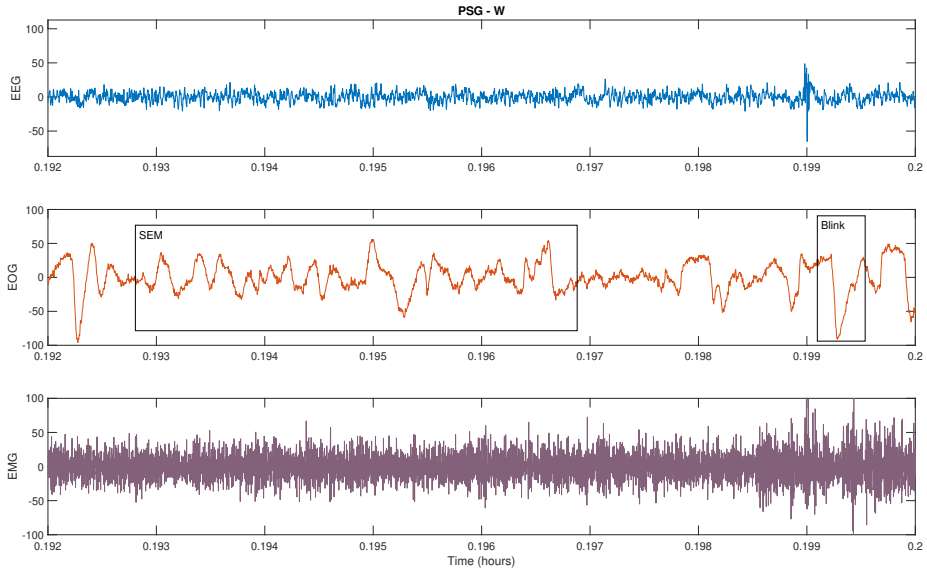


Figure 1.6: A period of wakefulness recorded by PSG, where alpha frequencies are prevalent. Note also the appearance of SEM, REM, and the occurrence of blinks depicted in the EOG signal. During wake, the EMG signal is variable and contains periods of high amplitude, especially compared to other stages.

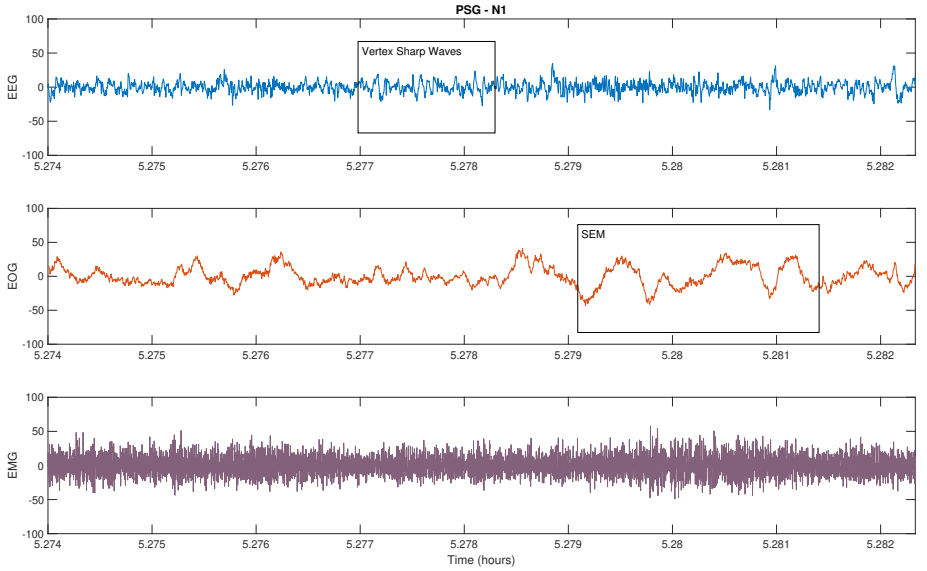


Figure 1.7: An example of N1 sleep, where SEM are visible and EEG has low amplitude and mixed frequencies. N1 stages commence when EEG activity in the range of $4 - 7$Hz, vertex sharp wave or SEM are seen. The EMG during N1 is variable but often lower than during W.

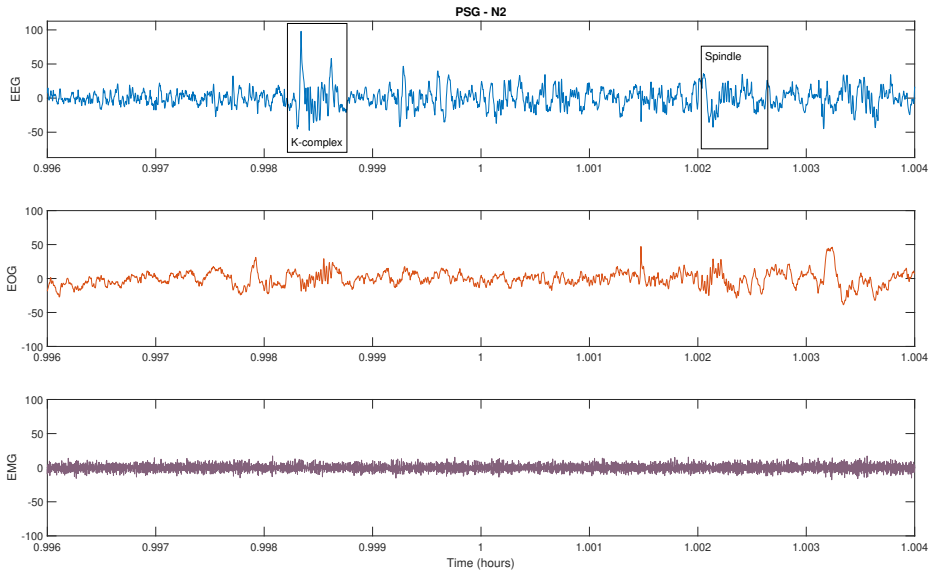


Figure 1.8: N2 sleep is illustrated here, where K complexes are prevalent along with sleep spindles. Note also the decreased amplitude in the EMG signal, compared to W and N1.

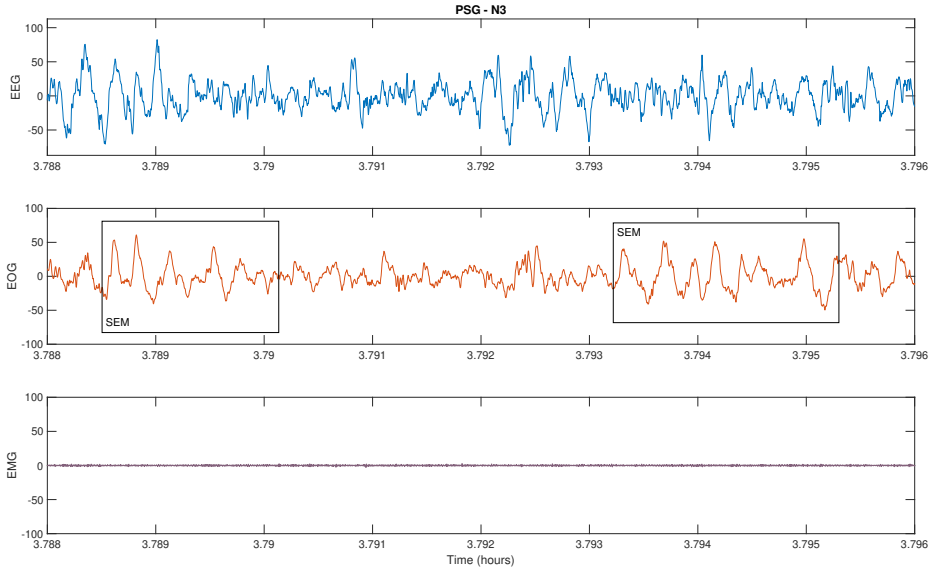


Figure 1.9: Slow wave sleep, also known as N3, is dominated by slow wave activity in EEG with a high amplitude. Note the reduced EMG amplitude compared to W, N1, and N2.

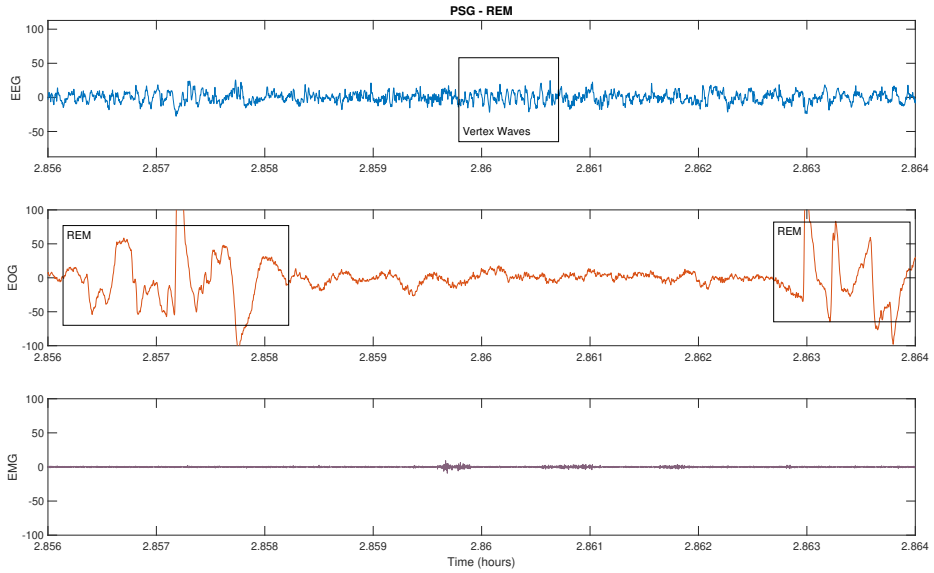


Figure 1.10: REM contains period of rapid eye movement, clear from the EOG signal. Periods between these movements are scored as REM, where there is clear reduction in EMG tone compared to W, N1, N2, and N3. Short irregular bursts of EMG activity (less than 0.25 seconds) can also be observed. Saw-tooth waves are also common in the EEG signal.

Kappa Value	Strength of Agreement
< 0.00	Poor
$0.00 - 0.20$	Slight
$0.21 - 0.40$	Fair
$0.41 - 0.60$	Moderate
$0.61 - 0.80$	Substantial
$0.81 - 1.00$	Almost Perfect

Table 1.2: Useful benchmarks to describe the level of agreement between categorical data, often used in sleep staging literature.

sleep staging is the Cohen-Kappa score [23]. This scores ranges from -1 to 1, with arbitrary benchmarks to indicate level of agreement, as detailed in Table 1.2. Human annotators have a reported inter-rater variability of $\kappa = 0.68$ (for the six stage R&K) [21], which is considered "substantial agreement" [23]. However, the inter-rater variability has been documented to further worsen in cases pertaining to individuals with sleep disorders (κ values of 0.61 for participants with PD) [22]. This inherent variability is the nature of the gold standard by which this study is bounded and assessed.

Sleep and ageing

The structure, duration, and quality of sleep changes with age and gender, however these effects are poorly understood [24]. Luca *et al.* (2015) have shown that spectral power densities, fast spindles, theta/alpha, and beta power in NREM sleep change with age [24]. While in REM sleep, ageing was associated with a decrease in delta waves and an increase in higher frequencies [24]. Documented changes in sleep architecture with healthy ageing also include increased N1 sleep, and reduced slow wave sleep duration (N3) [25, 26]. In addition to sleep architecture, characteristics in EEG features have been observed to alter with age [26]. The occurrences of eye movement during REM sleep are reported to be less frequent with age [27], while EMG activity has been observed to increase [28].

1.2.3 Incidence and prevalence of RBD

The prevalence of RBD within the population is not known, however a phone survey conducted by Ohayon *et al.* (1997) suggests violent behaviour during sleep is more prevalent than it might be presumed [29]. Out of more than 4900 people aged between 15 and 100 years, 2% indicated overall prevalence of violent behaviour during sleep, one quarter of which were likely due to RBD [29]. Another survey estimated the prevalence of RBD to be 0.38% in elderly persons [30]. The typical age of RBD onset is between 40-70 years, but can range from 18-80 years [31]. The male predominance and typically late manifestations of RBD suggest a link with gender differences in brain development with ageing [31]. Once more research has established an association between RBD and neuro-degenerative diseases, such as PD, LBD, and multiple systems atrophy (MSA) [31, 4, 5, 6, 7, 8, 9, 10, 11, 13, 14].

1.2.4 RBD as a Predictor

As more patients are identified and studied over time, there is mounting evidence indicating RBD as a precursor to PD and may precede the disease by years, potentially decades [4]. While RBD is not apparent in all PD patients it has been known to occur in 40-60% of PD patients [5] and 80% or more of patients with RBD

eventually develop PD, PD dementia (PDD), LBD or MSA [6, 7, 8]. The evidence is such that Hogl *et al.* (2018) suggest RBD is no longer merely a parasomnia but a manifestation of early-stage alpha-synucleinopathy [32]. The incredible sensitivity and specificity of RBD as a prodromal stage for these neurodegenerative disorders provide an extraordinary opportunity to evaluate the efficacy of neuroprotective interventions [11].

PD is the second most prevalent neurodegenerative disease worldwide, affecting more than four million people [33]. The chronic nature and growing disability of PD incurs major healthcare expenses that will only continue to escalate in countries with an ageing population [34, 35]. Furthermore studies indicate the diagnosis of PD has a major impact on quality of life and doubles patient mortality [34, 35]. Additionally, the chronic nature and growing disability inherent to the disease contributes towards early retirement and has a considerable impact on patients' families, exacerbating the socioeconomic burden [34, 35]. The cost per patient per year has been estimated to be 17000 euros in Spain [34] and 20095 euros in Germany [35]. The highest expenditure for direct costs include drug treatments and inpatient care (nursing homes) [35, 34]. Indirect costs stem from productivity losses from premature retirement or temporary sickness. In one German study, patients with PD had a mean retirement age of 50.5 years, while the general population was 57.9 years [35]. At this stage, alleviating the symptoms of Parkinson's disease is expensive and places a huge burden on healthcare systems. More work is required to understand the development of this disorder so that preventative measures can be devised. RBD represents one potentially promising early predictor for a large part of PD sufferers, possibly providing a clear avenue to target remedies before the onset of PD.

1.2.5 Behavioural features of RBD

While individuals might display dream-enacting episodes with limb movements on a nightly basis, the majority of episodes are intermittent [1]. It has been documented to occur at a minimum frequency of once in two weeks and up to a maximum of four times a night for 10 consecutive nights [1]. Dream enacting

episodes also include speech, shouting, laughter and at times coherent conversations with bed partners. Patients display a very high threshold for arousal despite these intense movements and activities. In RBD patients, arousal from sleep is rapid and accompanied by complete dream recall (unlike other disorders such as sleepwalking or sleep terrors) [1].

1.2.6 Diagnosing RBD

Since RBD's acceptance as a distinct clinical disorder in 1986 [36], the diagnosis of RBD has evolved over time. A PSG is mandatory when diagnosing RBD, in accordance to the third edition of the International Classification of Sleep Disorders (ICSD-3) [37]. The first criteria requires repeated episodes of behaviour or vocalisations that are either documented by PSG to occur during REM or are presumed to arise from REM based on reports of dream enactment [37]. The second is establishing the lack of atonia during REM sleep (confirmed through PSG), which can be observed by EMG monitoring (an example is given in Figure 1.11). Finally, these dream enactments must not be better explained by another sleep disorder, medical or neurological disorder, mental disorder, medication use, or substance use disorder [37]. When RBD is believed to occur as a result of medication use, a diagnosis of RBD should still be used, provided all above criteria are met [37].

1.2.7 Treatment

Treatment for RBD includes pharmacological agents and practical solutions in an attempt to address self-injury or injury to others [4, 38, 39]. Practical solutions include the removal of dangerous objects from the bedroom, such as projectiles, sharp objects, and obstacles. Also placing cushioning around the bed to soften the impact of falls from dream-enacting behaviour is suggested. There are two primary pharmacologic therapies, which should be used in conjunction with the non-pharmacological practicalities mentioned above. These therapies include Clonazepam and melatonin, which can be taken alone or in combination [40].

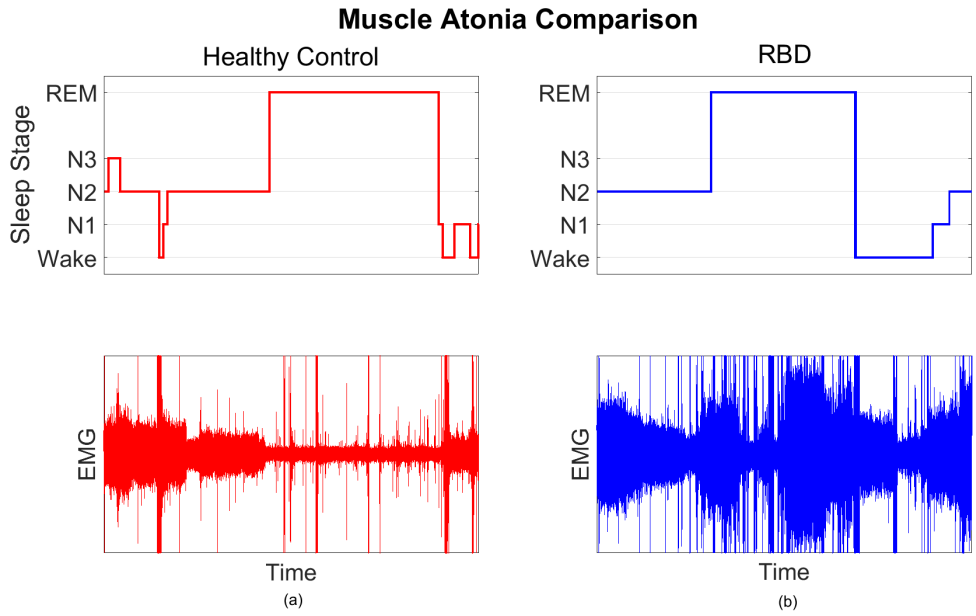


Figure 1.11: An example comparing the difference between atonia levels for (a) a HC and (b) an RBD participant. From the healthy control we can observe the annotated sleep stages and the reducing amplitude as stages progress from N3, N2 and eventually REM (where we observe minimal amplitude due to muscle atonia). Compare this to the RBD participant, where the amplitude increases at REM, often coinciding with movement.

Studies have shown that melatonin has the ability to restore normal muscle tone during REM sleep with rare and mild side-effects [41, 42, 40]. Clonazepam, a sedating benzodiazepine, is effective in promptly alleviating violent dream-enacting sleep in 50%-90% of patients, while also reducing phasic EMG activity [43, 40, 39]. However, it proves ineffective in 10% of patients and a minority of them notice more significant side effects [39, 43].

1.3 Research aims

This chapter provides the scope and motivation for developing an automated diagnostic support-tool to identify RBD in the hope of better understanding and preventing alpha-synucleinopathies, such as PD, which have huge socioeconomic impact on a global scale. The core elements of an RBD screening support-tool are illustrated in Figure 1.12, which consists of PSG recording, automated sleep stage classification, and automated RBD detection. Visually this can be summarised

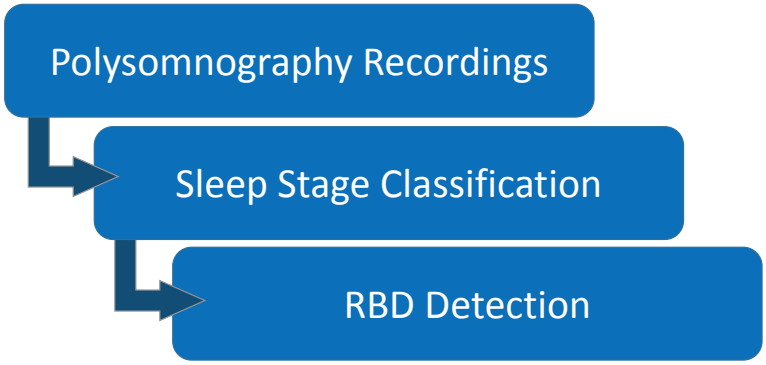
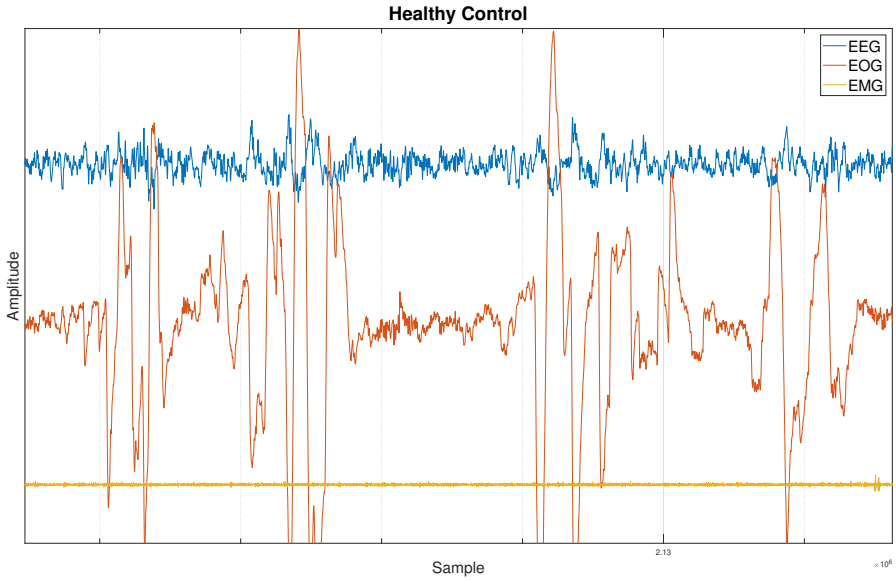


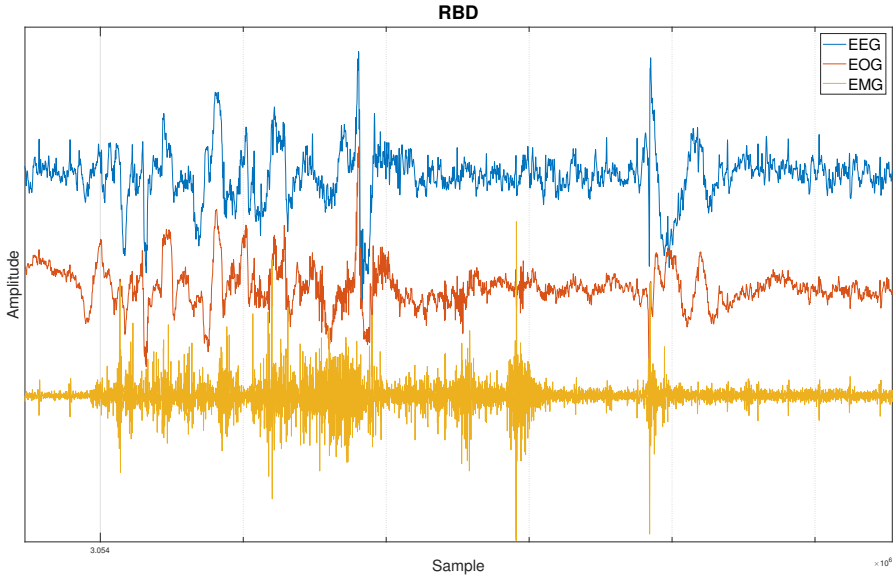
Figure 1.12: To emulate the diagnostic requirements of clinicians, an RBD diagnostic support-tool would require three core elements. The first would be to facilitate an easy self-applied take home PSG recording device. The second would provide automated sleep stage classification in order to identify REM sleep. Lastly, RBD detection would be made based on abnormal REM sleep behaviour.

by Figure 1.13, identifying periods of REM and distinguishing RBD participants from healthy controls (HCs). A successful support tool would need to be highly sensitive and precise to prove effective for epidemiological studies, demanding a limited use of sensors and devices to ensure portability, usability, and economy. Within this context, this thesis has two main objectives: (1) Develop machine learning tools for fusion of a range of PSG signals, to provide evidence based diagnosis of RBD; (2) Develop strategies to facilitate an automated take-home RBD screening device, that would incorporate frugal use of signals and sensors ensuring an effective and efficient tool for epidemiological studies. Consequently, the following research questions are investigated:

- Can machine learning of PSG signals be used to automate sleep staging in patients with RBD?
- Could automated sleep staging be achieved on a mixed cohort of healthy and RBD participants?



(a) Healthy control.



(b) RBD participant.

Figure 1.13: (a) A typical REM period for a healthy control. Note the low amplitude and mixed frequencies in the EEG signal, the rapid eye movement in the EOG signal, and the low amplitude of the EMG signal (indicative of muscle atonia). (b) This is an example from an RBD participants, during an episode where the absence of muscle atonia is observed during REM. The elevated EMG confirms the lack of atonia, the EOG displays rapid eye movement through sharp transient waves, and the EEG signal indicates mixed frequencies and low amplitudes (with some leakage apparent across all channels).

- Are there new predictors and metrics that can be derived from PSG recordings to improve automated detection of RBD?
- Could machine learning be used on a combination of objective metrics to improve RBD identification?
- Could effective RBD detection be achieved from automatically annotated sleep staging?
- What is the minimal combination of PSG sensors to adequately automate the identification of RBD, in order to provide a simple, cost-effective, and practical screening tool for epidemiological sleep studies?

1.4 Thesis Structure

This thesis will investigate the research questions proposed above in the following structure:

- Chapter 2: Introduce existing datasets used for analysis in this thesis, including a detailed description of data limitations, which inform the design and implementation of this thesis.
- Chapter 3: Proposes a machine learning framework for automated sleep staging that functions for both HCs and RBD participants.
- Chapter 4: Proposes a machine learning framework for the quantification and detection of RBD participants.
- Chapter 5: Explores the application of using an end-to-end machine learning framework to identify RBD participants by combining the models presented in chapter 3 and 4.
- Chapter 6: Investigates the possibility of minimal RBD detection using the least number of sensors and devices to facilitate a possible RBD take-home screening tool that would prove practical and cost-effective.

- Chapter 7: A discussion on the findings presented in this thesis, highlighting contributions as well as limitations and the opportunity for improvement in future work.

1.5 Research publications

- N. Cooray, F. Anderotti, C. Lo, M. Symmonds, M. T. Hu, and M. De Vos. "Automating the Detection of REM Sleep Behaviour Disorder". In: *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2018, pp. 1460–3

- H. Phan, F. Andreotti, N. Cooray, O. Y. Chén, and M. De Vos. "Joint Classification and Prediction CNN Framework for Automatic Sleep Stage Classification". In: *IEEE Transactions on Biomedical Engineering* (2018), pp. 1–11

- H. Phan, F. Andreotti, N. Cooray, O. Y. Ch, and D. Vos. "SeqSleepNet : End-to-End Hierarchical Recurrent Neural Network for Sequence-to-Sequence Automatic Sleep Staging". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 27.3 (2019), pp. 400–410

- N. Cooray, F. Andreotti, C. Lo, M. Symmonds, and M. T. M. Hu. "Detection of REM sleep behaviour disorder by automated polysomnography analysis". In: *Clinical Neurophysiology* 130.4 (2019), pp. 505–514

- N. Cooray, F. Andreotti, C. Lo, M. Symmonds, M. T. M. Hu, and M. De Vos. "Screening for REM Sleep Behaviour Disorder with Minimal Sensors (In Review)". In: *Clinical Neurophysiology* (2020), pp. 1–22

There is freedom within, there is freedom without
Try to catch the deluge in a paper cup
There's a battle ahead, many battles are lost
But you'll never see the end of the road
While you're travelling with me.

— Neil Finn's *Don't Dream it's Over*

2 Datasets

Contents

2.1 Introduction 21
2.2 Background 22
2.3 Montreal Archive of Sleep Studies 23
2.4 John Radcliffe Hospital 24
2.5 Cyclic Alternating Pattern Sleep Database 25
2.6 Combined Dataset 26
2.7 Conclusion 29

2.1 Introduction

This chapter will outline the collection of sleep recordings that were investigated in this thesis. A total of three different datasets were analysed, two were freely available online and one provided by the John Radcliffe hospital, Nuffield department of Clinical Neurosciences at the University of Oxford. The Montreal Archive of Sleep Studies (MASS) supplied HC participants, which are described in Section 2.3. RBD participants are supplied by the JR dataset and detailed in Section 2.4. The Cyclic Alternating Pattern (CAP) sleep database provides PSG recordings from both HC and RBD participants, outlined in Section 2.5. This study complied with the requirements of the Department of Health Research Governance Framework for

Health and Social Care 2005 and was approved by the Oxford University hospitals National Health Service (NHS) Trust (HH/RA/PID 11957).

2.2 Background

The data used for this study incorporated recordings from various sources and included specific inclusion and exclusion parameters. This study aimed to have two cohorts of HC and RBD participants who are age-matched. Once more, this study focused on participants who had no concurrent neurological or respiratory disorders.

Concurrent Neurological Disorders

For this work it was important to exclude any concurrent neurological disorders, as the intention is to identify predictors that would be specific only to PD. Participants with PD and secondary RBD were excluded because the motivation of this work was to identify individuals with RBD without PD and observe PD conversion in potential follow-up studies.

Apnoea and Hypopnoea Index

Throughout sleep, breathing related events such as a cessation of breathing (apnoeas) or impeded airflow (hypopnoea) are common and can be observed through PSG recordings. The severity of these events are captured by the apnea and hypopnoea index (AHI), which is defined as the combined average instances of apnoeas and hypopnoeas that occur per hour of sleep. The AASM stipulates that index scores of $5-15$ are mild, $15-30$ are moderate, and >30 are severe [19]. There is no consensus in literature regarding the effect of artefacts caused by snoring and other respiratory related events that impact RBD quantification and detection [49, 50]. For the purposes of this study, participants with high AHI scores were excluded to reduce the potential influence of artefacts in RBD detection evaluation.

2.3 Montreal Archive of Sleep Studies

The MASS database is an open-access collection of 200 PSG recordings from three difference hospital-based sleep laboratories [51]. The motivation behind this database was to provide a common benchmark to compare automatic systems that evaluate sleep recordings, which are freely available through the MASS website (http://www.ceams-carsm.ca/en/MASS). This enables the cross-validation of automated sleep analysis systems, ranging from artefact rejection to sleep staging, as well as the detection of leg movement, apnoeas, micro-arousals, sleep spindles, slow waves, and K-complexes.

The database was subdivided into five homogeneous subsets (SS1 to SS5), which are detailed in Table 2.1. Noticeably, the SS1 cohort are older, which ideally match the cohort of RBD recordings collected for this work. Therefore only the SS1 cohort was analysed and incorporated for this study, while others were excluded. All SS1 recordings provide an EEG (montage of 17-19 channels), EOG (2 channels), EMG (anterior tibialis right, anterior tibialis left, and three chin channels), ECG, and respiratory (thermistance, air flow cannula, oximetry and abdominal/thoracic belt) signals. These recordings are available in European Data Format (EDF++) and the sleep stages have been annotated according to the AASM rules. Only three participants within the SS1 cohort have a AHI greater than 15 (considered moderate), all others were mild and unremarkable. Additionally, all participants in this subset were verified to have no mental disorder or mild cognitive impairment. From the SS1 cohort, the ECG recording of three participants (SS01-01-0001, SS01-01-0010, and SS01-01-0018) were excluded due to their poor ECG signal quality. A summary of the SS1 cohorts included for this work is detailed in Table 2.2

	SS1	SS2	SS3	SS4	SS5	Total
Male	34	8	28	14	13	97
Female	19	11	34	26	13	103
Total	53	19	62	40	26	200
Age (years)	63.6 ± 5.3	23.6 ± 3.7	42.5 ± 18.9	25.3 ± 4.3	25.0 ± 7.4	40.6 ± 19.4

Table 2.1: All participants available from the MASS database.

Cohort	#	Female	Male	Age (years)
Healthy Controls	50	17	33	63.7 ± 5.18

Table 2.2: Included participants from the MASS database.

2.4 John Radcliffe Hospital

The John Radcliffe hospital retains PSG recordings as part of National Health Service (NHS) routine care for individuals suspected of having RBD. This study applied through the Clinical Trials and Research Governance (CTRG) to access anonymised case records for patients who were suspected of having RBD and later confirmed through these recordings. In addition to complete PSG data, these records included: age, sex, diagnosis (recorded by clinical staff) and treatment received at time of recording. PSG recordings were anonymised by those who had authority to access the data. There was no inclusion or exclusion criteria (*i.e.* no lower or upper age limit). In total 36 participants were included in the PSG recordings and are summarised in Table 2.3.

This dataset provided two nights of full PSG recordings for each participant, but only the second night was used for this work (where available). Some of these participants were taking medication to treat their condition or alleviate movement during REM (briefly detailed in Section 1.2.7), during their PSG recording. This included one participant taking Clonazepam and ten taking anti-depressants (Citalopram, Venlafaxine, Sertraline, Duloxetine, and Lorazepam). For these NHS routine recordings, AHI scores of RBD participants was not measured if obstructive sleep apnoea (OSA) was previously excluded or oxygen saturation monitoring was unremarkable. The measured AHI scores of RBD participants were all considered mild and unremarkable (less than 7.1). However, five RBD participants used a continuous positive airway pressure ventilator during their PSG, indicating they had OSA. Because the ventilator was in use, their AHI was managed and kept low.

Cohort	#	Female	Male	Age (years)
RBD Participants	36	2	34	64.3 ± 7.96

Table 2.3: Included PSG recordings from the JR hospital.

2.5 Cyclic Alternating Pattern Sleep Database

The cyclic alternating pattern (CAP) describes periodic electrocortical activity, distinct from background EEG [52]. These patterns include waveform bursts separated by quiet (low-amplitude) periods [1]. The CAP is thought to provide insight into sleep stage stability and can occur in association with various sleep disorders [52, 1]. This sleep database was compiled to establish an atlas and standardised manual to create a consensus on recording, scoring, and terminology of CAP analysis [52].

The database contains a collection of 108 PSG recordings from the Sleep Disorders Centre of the Ospedale Maggiore of Parma, Italy. These recordings are publicly available online in EDF, including at least three EEG channels (F3 or F4, C3 or C4, and O1 or O2, referenced to A1 or A2), EOG (2 channels), EMG of the submentalis muscle, EMG of the bilateral anterior tibial, respiration signals, and ECG. The database contains 16 HC participants and 92 pathological recordings, detailed in Table 2.4. Trained neurologists annotated every 30 second epoch using the Rechtschaffen and Kales rules [20].

Prefix	**#**	**Pathology**
n	16	Healthy controls
brux	2	Bruxism
ins	9	Insomnia
narco	5	Narcolepsy
nfle	40	Nocturnal frontal lobe epilepsy
plm	10	Periodic leg movements
rbd	22	REM sleep behaviour disorder
sdb	4	Sleep-disordered breathing

Table 2.4: CAP sleep database participants and pathology. Some of the RBD participants were excluded from this study due to concurrent neurological disorders.

Cohort	**#**	**Female**	**Male**	**Age (years)**
Healthy Controls	16	9	7	32.2 ± 5.38
RBD	22	3	19	70.7 ± 6.24

Table 2.5: HC and RBD participants from the CAP sleep database.

For the previous RBD dataset provided from the JR hospital (see Section 2.4), the average age is substantially different from the HCs provided by the CAP sleep dataset. As a result of the purposes of having age-matched participants, HCs from the CAP sleep dataset must be excluded. Most RBD participants from the CAP database have been clinically diagnosed with idiopathic RBD. However, seven participants had secondary RBD, of which five were diagnosed with PD (RBD1, RBD2, RBD3, RBD7, ad RBD9), one with MSA (RBD13), and one with LBD (RBD5). These seven participants were excluded for the purposes of this work. Once more through visual analysis of this dataset, it was identified and confirmed with the clinic that recordings labelled as RBD10 and RBD11 were duplicates and therefore RBD11 was discarded. Of the remaining RBD participants, one was taking Duloxetine and Lorazepam as treatment at time of their PSG recording. For all remaining participants, the AHI does not exceed moderate levels. A summary of the data available from the CAP database is provided in Table 2.6.

Cohort	**#**	**Female**	**Male**	**Age (years)**
RBD	14	2	12	70.2 ± 5.33

Table 2.6: Included participants from the CAP sleep database.

2.6 Combined Dataset

Combining these aforementioned datasets provided a unique collection of recordings to explore and devise sleep staging and RBD detection algorithms. A summary of the collected age-matched data is depicted in Table 2.7. From Table 2.7 there is a clear male bias in the RBD cohort compared to the HCs, which could not be corrected. Table 2.8 details each individual participants used in this work, detailing the source (database), age, sex, and diagnosis (cohort).

Each PSG recording from this combined dataset was annotated by an expert who assigned a sleep stage for every epoch using either R&K or AASM guidelines. Recordings annotated using R&K were converted to AASM, simply by converting S3 and S4 to N3, while S0, S1 and S2 were relabelled as W, N1 and N2, respectively. All recordings provided at the very least the following common PSG channels:

- 1 EEG (either C4-A1, C3-A2 or C1-A1, listed in preferential order)
- 1 EOG (difference of ROC and LOC)
- 1 EMG (chin – submentalis)
- 1 ECG (difference of ECG1 and ECG2)

Cohort	**Dataset**	**#**	**Female**	**Male**	**Age (years)**
Healthy Controls	MASS	50	17	33	63.7 ± 5.18
RBD	JR, CAP	50	4	46	66.0 ± 7.78

Table 2.7: Summary of all participants used in the analysis of this thesis.

Participant	Database	Age	Sex	Cohort (Diagnosis)
01-01-0002	MASS	68	M	HC
01-01-0003	MASS	64	F	HC
01-01-0004	MASS	72	M	HC
01-01-0005	MASS	62	M	HC
01-01-0006	MASS	62	M	HC
01-01-0007	MASS	57	M	HC
01-01-0008	MASS	64	M	HC
01-01-0009	MASS	75	M	HC
01-01-0011	MASS	72	F	HC
01-01-0012	MASS	69	M	HC
01-01-0013	MASS	67	F	HC
01-01-0014	MASS	64	M	HC
01-01-0015	MASS	64	M	HC
01-01-0016	MASS	60	F	HC
01-01-0017	MASS	57	M	HC
01-01-0019	MASS	66	F	HC
01-01-0020	MASS	59	F	HC
01-01-0021	MASS	69	F	HC
01-01-0022	MASS	71	M	HC
01-01-0023	MASS	65	F	HC
01-01-0024	MASS	58	M	HC
01-01-0025	MASS	63	F	HC
01-01-0026	MASS	63	F	HC
01-01-0027	MASS	65	M	HC
01-01-0028	MASS	57	M	HC
01-01-0029	MASS	57	F	HC
01-01-0030	MASS	76	M	HC
01-01-0031	MASS	61	M	HC
01-01-0032	MASS	61	F	HC

Participant	Database	Age	Sex	Cohort (Diagnosis)
01-01-0033	MASS	63	M	HC
01-01-0034	MASS	60	F	HC
01-01-0035	MASS	60	M	HC
01-01-0036	MASS	59	M	HC
01-01-0037	MASS	60	F	HC
01-01-0038	MASS	61	M	HC
01-01-0039	MASS	55	M	HC
01-01-0040	MASS	63	F	HC
01-01-0041	MASS	71	M	HC
01-01-0042	MASS	62	M	HC
01-01-0043	MASS	66	M	HC
01-01-0044	MASS	57	M	HC
01-01-0045	MASS	64	M	HC
01-01-0046	MASS	58	M	HC
01-01-0047	MASS	72	F	HC
01-01-0048	MASS	72	F	HC
01-01-0049	MASS	56	M	HC
01-01-0050	MASS	60	M	HC
01-01-0051	MASS	65	M	HC
01-01-0052	MASS	65	M	HC
01-01-0053	MASS	66	M	HC
N238	JR	56	M	RBD
N239	JR	67	M	RBD
N242	JR	64	M	RBD
N244	JR	64	M	RBD
N245	JR	73	M	RBD
N246	JR	55	M	RBD
N247	JR	69	F	RBD
N248	JR	68	F	RBD
N254	JR	60	M	RBD
N262	JR	53	M	RBD
N265	JR	70	M	RBD
N271	JR	65	M	RBD
N278	JR	59	M	RBD
N280	JR	69	M	RBD
N287	JR	62	M	RBD
N297	JR	73	M	RBD
N312	JR	69	M	RBD
N352	JR	59	M	RBD
N372	JR	48	M	RBD
N403	JR	55	M	RBD
N449	JR	63	M	RBD
N472	JR	57	M	RBD
N482	JR	68	M	RBD

Participant	Database	Age	Sex	Cohort (Diagnosis)
N484	JR	76	M	RBD
N485	JR	57	M	RBD
N487	JR	56	M	RBD
N489	JR	66	M	RBD
N490	JR	74	M	RBD
N491	JR	53	M	RBD
N506	JR	81	M	RBD
N511	JR	65	M	RBD
N513	JR	73	M	RBD
N514	JR	77	M	RBD
N518	JR	52	M	RBD
N522	JR	72	M	RBD
N524	JR	68	M	RBD
RBD4	CAP	70	M	RBD
RBD6	CAP	73	M	RBD
RBD8	CAP	82	M	RBD
RBD10	CAP	73	M	RBD
RBD12	CAP	76	F	RBD
RBD14	CAP	66	M	RBD
RBD15	CAP	69	M	RBD
RBD16	CAP	65	M	RBD
RBD17	CAP	70	M	RBD
RBD18	CAP	59	M	RBD
RBD19	CAP	65	M	RBD
RBD20	CAP	72	M	RBD
RBD21	CAP	73	F	RBD
RBD22	CAP	70	M	RBD

Table 2.8: Participant details for all PSG recordings used in this thesis. The PSG recordings are sourced from the MASS (online), JR hospital, and the CAP sleep database (online), which provided HC and RBD participants.

2.7 Conclusion

This chapter detailed all PSG recordings obtained and collected for the analysis of this thesis. This included freely available (online) datasets that provided HC and RBD participants from the MASS and CAP databases. Additionally, data from NHS routine care was collected and anonymised from the JR hospital, providing 34 RBD participants.

While the focus of this work was to provide a feasibility study into developing

techniques to achieve automated sleep staging and RBD detection using simply two cohorts, this is a major limitation with regard to clinical applications. Once more the heavy male bias in this collection of data must be noted. This is primarily due to a large male predominance of RBD [1], but is also a clear difference between cohorts in the collected data.

As a collection of RBD PSG recordings, this is one of the largest known in literature for automated sleep staging and RBD detection and provided a platform for the rigorous testing described in this thesis. Furthermore, this hybrid collection of data consists of freely available recordings (online), enabling researchers and the public to evaluate this work for independent comparison and validation.

Holy night, you sink down;
dreams, too, float down,
like your moonlight through space,
through the silent hearts of men.
They listen with delight,
crying out when day awakes:
come back, holy night!
Fair dreams, return!

— Matthäus Casimir von Collin's *Nacht und Träume*

3 Automated Sleep Stage Classification

Contents

3.1 Associated Publication 31
3.2 Introduction 32
3.3 Background 33
3.3.1 Automated Sleep Staging 38
3.4 Methods 48
3.4.1 Data 48
3.4.2 Signal Pre-Processing 48
3.4.3 Feature Extraction 49
3.4.4 Electroencephalogram Feature Extraction 50
3.4.5 Electrooculogram Feature Extraction 57
3.4.6 Electromyogram Feature Extraction 61
3.4.7 Miscellaneous Feature Extraction 65
3.4.8 Random Forest Classification 67
3.4.9 Model training and evaluation 70
3.5 Results 73
3.6 Discussion 84
3.7 Conclusion 90

3.1 Associated Publication

- N. Cooray, F. Andreotti, C. Lo, M. Symmonds, and M. T. M. Hu. "Detection of REM sleep behaviour disorder by automated polysomnography analysis". In: *Clinical Neurophysiology* 130.4 (2019), pp. 505–514

3.2 Introduction

As defined by the International Classification of Sleep Disorders (ICSD-3) [37], the first step towards confirming RBD is the visual analysis of REM sleep through a PSG recording. The manual process of identifying REM sleep requires careful annotation of every epoch, which is assigned one of five sleep stages (defined in Chapter 1 Section 1.2.2). This task is time consuming, highly specialised, and subject specific. Therefore any automated RBD screening tool would first need to achieve automated REM classification. However, individual variation represents one of the greatest challenges in achieving automated REM detection, where the nature of sleep patterns and PSG recordings can vary with age and gender [24].

While sleep is relatively well understood for young healthy individuals, the nature of sleep patterns for older individuals and those with sleep related disorders, specifically RBD, have barely been explored. The onset of RBD can vary, but typically occurs after the age of 50 [53, 43, 7]. This study focused on typical onset RBD, where comparatively healthy sleep was defined by HC participants in the same age category. This was the motivation behind the collated PSG recordings from RBD and HC participants as detailed in Chapter 2.

This chapter aims to infer properties of PSG recordings to extract clinically useful information in order to achieve automated sleep staging for a mixed cohort (containing HC and RBD participants, as detailed in Chapter 2). While previous studies that investigated automated sleep staging in RBD participants focused on REM classification (for the purposes of automated RBD detection) [54, 55], this chapter aims to achieve five-stage automated sleep staging with the intention of including sleep architectural features to differentiate RBD individuals from HCs. Once more, this chapter utilised limited PSG sensors in order to emulate the limitations associated with take-home screening devices consisting on a single EEG, EOG, and EMG channel.

The principal contributions of this chapter to the field of sleep stage classification are:

- A rigorous implementation of a random forest classifier for automated sleep staging on age-matched HC and RBD diagnosed participants. This PSG collection of RBD participants is one of the largest in known literature.
- The exploration of features to aid in automated sleep staging for both elderly HCs and RBD participants.
- Identifying the complexities and difficulties that RBD participants introduce to the problem of automated sleep staging.

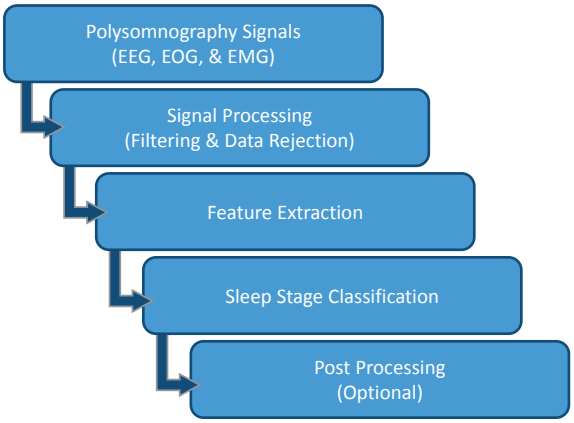


Figure 3.1: Traditional automatic sleep stage classification methodology.

3.3 Background

The problem of automatic sleep staging has been extensively studied over the last few decades where numerous signal processing approaches and various classification techniques have been implemented. There are numerous methods in the literature that utilise PSG recordings to classify REM stages either specifically or as part of multi-sleep-stage classification (see Table 3.1). This section provides a literature review on automated sleep staging techniques with a specific focus on REM stage

classification with a summary provided in Table 3.1. It must be noted that there is great difficulty and uncertainty in comparing automated sleep staging results from studies that utilised different datasets.

Many of these automated sleep scoring algorithms produce results that are comparable to expert annotation of PSG recordings from young HCs. However, very few of those validated automated scoring algorithms were designed for older individuals, let alone those who suffer from sleep disorders, which can exhibit very different EEG characteristics [19, 24]. Nonetheless, manual sleep scoring remains the clinical gold-standard to date and is used to assess the performance of automated sleep stage classification.

This section begins with a review of classifiers with a focus on REM detection. The established methodology adopted in automated sleep staging classification is detailed in Figure 3.1 that begins with pre-processing, followed by feature extraction, and ends with sleep stage classification.

Study	#	♀	♂	RBD	Age	Sleep Staging	PSG (#)	Model	Kappa	Other	REM Score
[56]	12	3	9		42.3±16.4	R&K (6)	EEG (2) EOG (2) EMG	K-means			Sen: 0.725 Spe: 0.876
[54]	8	5	3		60.8±9.4	REM/NREM	EEG	k-NN			Sen: 0.94
	8	3	5	8	60.6±8.5		EOG				Spe: 0.96
[57]	5		5		36-56	AASM	EEG	k-NN			Sen: 0.724
[55]	10	6	4		28.6±5.2	W/REM/NREM	EEG(3)	SVM			Sen: 0.88
	10	8	2		55.2±8.2		EOG				Spe: 0.97
	10	5	5		58.6±10.3						
	10	2	8	10	61.7±8						
[58]	40	14	26		54.35±16.37	AASM	EEG (6) EOG EMG	SVM		Acc: 0.9204	Sen: 0.793 Spe: 0.947 Acc: 0.870
[59]	28				35-56	R&K (6)	EEG	SVM		Acc: 0.964	Sen: 0.943, Spe: 0.736 Sen: 0.851, Spe: 0.968 Sen: 0.832, Spe: 0.961
[60]	8				21-35	R&K (5)	EEG (1)	SVM		Acc: 0.929	Acc: 0.925
[61]	8				21-35		EEG (1)	SVM		Acc: 0.889	Sen: 0.762
[62]	14				55.7±16.7		EEG (6) EOG (2)	SVM			Balanced Acc: 0.93
[63]	15				29.2±8		EEG (2)	SVM			Sen: 0.97

						EOG (2) EMG				Spe: 1.0
[64]	21			22-69	AASM	EEG (1)	Decision Tree	0.61		Sen: 81.1, PPV: 0.743
	23				AASM	EEG (1)		0.62		Sen: 66.5, PPV: 0.815
	44				AASM	EEG (1)		0.61		Sen: 0.732, PPV: 0.782
[65]	20	8	12	21.2±1.1		EEG (1)	Decision Tree	0.79	Acc: 0.867	Sen: 0.905
						EOG				Sen: 0.955
						EMG				PPV: 0.85
[66]	265	26	239	43.5	W/REM/SWS/LS	EOG	Decision Tree			
[67]	33	12	21	67.0±8.60	AASM	EEG	Fuzzy logic	0.76±0.06		Sen: 0.87±0.11
						EOG				Spe: 0.95±0.04
						EMG				
[68]	20	10	10	50±21.5	R&K (6)	EEG	Rules		Acc: 0.81	Sen: 0.862
						EMG				
						EOG				
						ECG				
[69]	16			30-75	AASM	EEG	Random Forest	0.76	Acc: 0.83	
[70]	8			21-35	R&K (6)		Random Forest		Acc: 0.953	Sen: 0.922, Spe: 0.991
	25	4	21	28-68	R&K (6)		Random Forest			Sen: 0.876, Spe: 0.986

	20	10	10	28.7±2.9	R&K (6)		Random Forest		
[71]	25	4	21	50±10	AASM		Random Forest	Acc: 0.980	Sen: 0.95 Spe: 1.0
[72]	8			21-35		EOG	SVM/Random Forest	Acc: 0.917	Sen: 0.842±0.008
	10								Sen: 0.882±0.013
	20	10	10	28.7±2.9					Sen: 0.761±0.09
[73]	20	10	10	28.7±2.9	R&K (5)	EEG (1)	Random Forest	0.83	Acc: 0.915	Sen: 0.691
[74]	8			21-35	R&K (5)	EEG (1)	Random Forest		Acc: 0.91	Sen: 0.81, Spe: 0.97
					R&K (6)		Random Forest		Acc: 0.87	Sen: 0.78, Spe: 0.96
[75]	8				R&K (5)	EEG (1)	RUSBoost	0.864	Acc: 0.980	Sen: 0.83, Spe: 0.97
	25	4	21	50±10				0.818	Acc: 0.812	
[76]	16	5	11	25.3±2.5	R&K (5)	EOG	LDA	0.77	Sen: 0.826	Sen: 0.941
[77]	62			42.5±18.9	AASM	EEG (1)	CNN/RNN	0.8	Acc: 0.862	Sen: 0.902
										F1: 0.893
									K=0.80	

Table 3.1: A list of automated sleep staging literature, with a focus on older age groups, sleep disorders, and limited PSG signals.

3.3.1 Automated Sleep Staging

Identifying REM stage of sleep is of critical importance in diagnosing RBD and the distinct definition of this sleep stage (Chapter 1 Section 1.2.2) provides an opportunity for automation to assist sleep experts. There are numerous methods that utilise PSG recordings to classify REM stages either specifically or part of multi-sleep-stage classification, such as decision trees, Support Vector Machines (SVMs), Artificial Neural Networks (ANNs), and others. A few studies simply focus on automatic REM stage detection [54, 78, 79, 80, 55, 81]. Any successful screening tool for RBD would need to be able to detect REM independent of age and gender. The ability to detect REM independent of age has not been comprehensively explored through literature except for a few isolated papers [82], let alone studies that include RBD participants [54, 55]. As a result, the available literature on automated sleep staging on populations with sleep disorders is also limited [54, 55, 56, 59, 83, 82]. This section details automated sleep staging and REM classification techniques and specifically focused on studies that used relatively older participants reflective of the expected onset of RBD.

K-mean clustering

Agarwal *et al.* (2001) presented a computer assisted method to sleep staging using k-means clustering [56]. Using two EEG, two EOG and one EMG signal this technique extracts time domain features (amplitudes and non-linear operators) and frequency domain features (frequency power ratios of clinically relevant bands) [56]. Once features are extracted from each segment they are organised into homogeneous clusters. These clusters are then manually assigned to sleep stages by human reviewers that sample a small number of representative epochs (up to five) for each cluster. Once clusters are assigned sleep stages, all epochs can be classified and hypnograms can be generated. The hypnograms can then undergo optional smoothing with post-processing, using rules that emulate R&K scoring rules. The performance of this algorithm was assessed with 12 participants aged between 17 and 62 years and achieved 72.5% sensitivity and 87.6% specificity, with respect to

REM sleep. However, this technique still requires manual human interaction to assign clusters a sleep stage. Nonetheless, this technique demonstrated with limited PSG sensors automated classification was feasible for healthy controls.

K-Nearest Neighbours

Kempfner *et al.* (2012) explored k-nearest neighbours (k-NN) in order to classify REM from NREM sleep [54]. This relatively simple machine learning technique works by determining the k-nearest neighbours within the feature space of the training set for an unknown epoch. The most represented sleeping stage in the k-nearest neighbours becomes the assigned stage. The computation and memory space of this algorithm is highly dependent on the number of features used. Kempfner *et al.* (2012) had 8 HCs (aged 60.8 ± 9.4) and iRBD participants (aged 60.6 ± 8.5) [54]. The interquartile difference (75^{th} percentile and 25^{th} percentile) in EOG and EEG signals were then determined within each clinically relevant frequency band. Additionally, REM was characterised by evaluating the normalised cross-correlation coefficient of the right and left EOG probes. After feature scaling, k-NN classification was achieved using the Euclidean distance. Post processing was achieved using a normalised Blackman window. The final performance with respect to REM detection achieved an accuracy, sensitivity and specificity of 96%, 94%, and 96%, respectively. This study demonstrated REM sleep classification was possible on a combined cohort of HCs and RBD participants.

Gunes *et al.* (2010) presented a method using a combination of techniques, where k-means clustering based feature weighting (KMCFW) was used as a pre-processing routine, followed by k-NN classification for all sleep stages [57]. Using Welch Fast Fourier Transform (FFT) 129 features are derived and reduced to four features comprised of the statistical minimum value, maximum value, standard deviation, and mean value. KMCFW begins by determining clusters using k-means, which each have a corresponding centre. The method then calculates the ratio of means of features to their centres, which was then multiplied with each feature. Classification was then achieved using k-NN at a performance of 81% sensitivity. This study only

used five male participants aged between 31 and 56, which means its application on older individuals remains relatively unknown.

Support Vectors Machines

SVMs are a popular machine learning technique and has been applied to automated sleep staging [60, 58, 63, 62, 55, 59, 61]. Generally SVMs show reliable classification with noisy inputs, however this technique does require several parameters to be tuned.

Koley and Dey (2012) explored automatic sleep classification using a single EEG signal on a dataset of 28 participants (aged $35 - 56$) using an SVM [59]. A total of 39 features were extracted and through a recursive feature elimination technique, the optimum numbers of features were used for sleep stage classification using a binary SVM (five in total, for each sleep stage). Using a one-against-all strategy meant the SVM classifier with the highest output, was deemed the classified stage. This dataset included participants with apnoea (13) and results were evaluated for both cohorts. REM classification achieved a sensitivity and specificity of 0.851 and 0.968, respectively for non-apnea participants. Whereas for apnea participants REM classification sensitivity and specificity was 0.832 and 0.961, respectively. This study demonstrated the feasibility of SVMs in automated sleep staging, but also illustrated how sleep disordered breathing could negatively impact performance.

Kempfner *et al.* (2013) developed an automatic sleep staging algorithm that uses a multi-class SVM on young adults to elderly patients [84]. This technique used two EOG channels and three EEG channels. Features were extracted from three second mini epochs, which were derived from a three second sliding window. Features included simple time-domain (median amplitude) and statistical (EOG co-variance) properties. Classification was then achieved using a multi-class SVM, using a one-vs-rest approach, combined with a Radial Basis Function (RBF) as a kernel. Three SVM classifiers were created and denoted as SVMw (W vs. rest), SVMr (REM vs. rest) and SVMn (NREM vs. rest). Hypnograms were up-sampled from one score per epoch to 10 scores per epoch (one score per mini-epoch). Once classified

the MC-SVM output was post-processed by converting the output to smoothed "posterior-like probabilities". This was achieved by a normalised Blackman window that was applied to each of the three SVM classes. The class with the highest "posterior-like probability" in each mini-epoch was the final prediction. In terms of REM classification this techniques achieved an accuracy of 91%, sensitivity of 88%, and specificity of 97%. The primary cause of misclassification was due to the transition stages or when sleep was highly fragmented. This technique was particularly susceptible to short periods of wakefulness, due to the smoothing effect in post-processing [84]. This study detailed the ability of SVMs in achieving automated sleep staging for a diverse age group of healthy controls.

SVMs were also used by Khalighi *et al.* (2013) for sleep stage classification using six EEG and two EOG signals [58]. A total of 570 features were extracted using frequency and temporal features in addition to the maximum overlap discrete wavelet transform (MODWT). This study deemed MODWT and harmonic features to be the most parsimonious and the most discriminative channels appear to be the central probes (C3 and C4). This technique achieved 92% average accuracy for all sleep stages and 79% sensitivity and 95% specificity in terms of REM classification. These results were procured from 40 participants with a mean age of 54.4 years and a standard deviation of 16.4 years.

Bajaj *et al.* (2013) used SVMs in conjunction with features derived from a histogram of segmented time-frequency images generated from an EEG signal [60]. The smoothed pseudo Wigner-Ville distribution (SPWVD) based time-frequency representation of an EEG signal was used to generate the image, which was segmented based on clinically relevant frequency bands. Using eight participants aged between 21 and 35 years old, this classifier achieved more than 80% sensitivity in W, S2, S4 and REM stages.

Sousa *et al.* (2015) proposed a two-step classifier where a SVM classifier was used at each node of a decision tree [62]. Furthermore a post-processing step was used to re-classify "dubious" epochs. These "dubious" epochs are identified and corrected by understanding commonly misclassified stages in automated sleep

staging algorithms and incorporating AASM sleep scoring rules. Features were extracted from six EEG channels and two EOG channels, which included MODWT-based features, relative power and harmonic parameters. Based on 14 participants (mean age 55.7 ± 16.7 years) with various diagnosed sleep issues (obstructive sleep apnoea, periodic limb movement (PLM), effective disorder amongst others) this algorithm achieved a REM balanced accuracy (the average between sensitivity and specificity) of 93%. This performance is particularly impressive due to the wide age span and the variety of sleeping disorders.

Lajnef *et al.* (2015) used a similar combined classifier approach using a wide range of time and frequency domain features extracted from EEG, EOG, and EMG signals [63]. From a dataset of 15 participants (mean age 29.2 ± 8 years) this technique achieved a REM sensitivity of 97% and specificity of 100%. These impressive results were achieved using participants from a narrow age span but their application to a wider population remains unknown.

An SVM was also used by Zhu *et al.* (2014) utilising a single channel EEG and graph domain features [61]. Using a small dataset of eight HCs, this technique achieved a five-stage accuracy of 0.889 and a 6 stage kappa of 0.81. REM classification achieved an accuracy of 0.889 and a sensitivity of 0.762. These results are promising given that only nine features were derived using visibility graphs avoiding time and frequency domain features, detailed in [61]. Once more, the interpretive value of these features might not be easily conveyed to clinical specialists.

Decision Trees & Random Forests

By using a single channel EOG, Virkkala *et al.* (2008) used a decision tree classifier as an easy alternative to the gold standard [85]. This study involved a comprehensive 132 participants for training (aged between 26 and 61 years) and 131 participants for testing (aged 28 to 60 years). Only two features were used that depicted the peak-to-peak amplitude in the $0.5-6$Hz band and the beta power between $18-30$Hz. A simple decision tree based on threshold values classified epochs into wakefulness, light sleep (N1 and N2), N3 and REM, which achieved an overall accuracy of 73.8%.

With respect to REM detection this methods achieved 61.7% sensitivity. These results describe the importance of EEG features in REM detection.

A rule based automatic sleep staging classifier was proposed by Liang *et al.* (2012) that used 12 features derived from EEG, EOG, and EMG signals [65]. The decision tree comprised of 13 decision nodes, which at times used a combination of feature thresholds. Contextual smoothing was then applied after the decision tree, which was a set of rules that ensured "continuity in classification using temporal contextual information" [65]. With 20 participants (mean age was 21 ± 1.1 years) this technique achieved an overall accuracy of 87%, with a REM sensitivity of 88%. It appears that W was often misclassified for N2, followed by REM misclassification of N2. Unfortunately this technique was evaluated on a very specific young age group and therefore its' applicability to elderly HCs remains unknown.

Decision tree logic has been used by authors to develop simple thresholds and rules in an aim to replicate AASM scoring rules [19]. However these rules do not generalise well due to the high variability in sleeping patterns amongst people. An ensemble of decision trees are called a random forest and are able to compensate for this high variability by creating many simple models and was explored by Fraiwan *et al.* 2012 [69]. This study extracted features from a single EEG channel using time-frequency analysis and entropy measures. These included Choi-Williams Distribution (CWD), Continuous Wavelet Transform (CWT), Hilbert-Huang Transform (HHT), and Renyi's entropy. Classification was achieved using a random forest classifier. A total of 16 PSG recordings from participants that were aged between 30 and 75 years were evaluated. This particular study was motivated by exploiting benefits of time-frequency analysis on non-stationary EEG signals. The best performance was attained using CWT features, which achieved an accuracy of 83%. The classification of N1 was the poorest and was attributed to the low number of N1 instances from the available sleep recordings.

Álvarez-Estévez *et al.* (2013) used fuzzy logic inference on a set of features extracted from EEG, EOG, and EMG signals from a dataset of 33 HC participants [67]. These fuzzy rules are structured into five independent set of rules for each sleep

stage and the total numbers of rules implemented was 111, which were derived from medical knowledge and the author's experience. This technique yielded a multi-stage kappa score of 0.76 ± 0.06 and REM sensitivity and specificity of 0.87 ± 0.11 and 0.95 ± 0.04 (mean and standard deviation), respectively. While these rules provide a transparent and easily understandable "human language" structure, they are quite numerous and would need to be expanded much further to incorporate additional sleep disorders and a wider age group.

Using a single EEG and EOG signal, Stepnowsky *et al.* (2013) demonstrated with a simple decision tree, automated sleep staging classification that was in agreement with manual scoring [64]. This study included PSG recordings of 44 healthy participants and achieved a multi-stage kappa of 0.61 and a REM sensitivity and positive predictive value of 0.732 and 0.782, respectively. Interestingly, participants in this study with OSA provided improved REM classification results, which might be attributed to more prominent and visually apparent REM sleep characteristics, which might also be the case for RBD participants.

Şen *et al.* (2014) conducted a comparative study of automated sleep staging using numerous classification techniques, including a random forest, feed-forward neural network, decision tree, SVM, and radial basis function neural network [71]. This study included PSG recordings from 25 healthy participants and extracted a total of 41 features from a single EEG signal. This study produced their best results with a random forest classifier using a hybrid approach to feature selection, achieving an accuracy of 0.98 and a REM sensitivity and specificity of 0.95 and 1.0. While this study highlights the benefit of automated sleep staging using a random forest, the marginal improvement given by feature selection was marginal and might not generalise well for participants with sleep disorders, such as RBD.

Da Silveira *et al.* (2017) also demonstrated the superior performance of random forests in automated sleep staging with a single EEG signal, when compared to a multilayer perceptron, Naïve Bayes network, AdaBoost M1, and 1-nearest neighbours [73]. This technique used 18 features derived using DWT and achieved

a multi-stage kappa score of 0.83 and a REM sensitivity of 0.691 on a dataset of 20 healthy participants [73].

Using a smaller dataset of eight healthy participants, Hassan and Bhuiyan (2016) came to a similar conclusion, where ensemble learning provided consistent and superior performance for automated sleep staging [74]. Using a single EEG signal, complete ensemble empirical mode decomposition with adaptive noise and bagging achieved a multi-stage kappa of 0.862 and accuracy of 0.907. Continuing from this work, Hassan and Bhuiyan (2017) used a random under sampling boosting classification technique using a total of three PSG datasets [75]. The performance varied from a multi-stage kappa of 0.77 to 0.893. These studies demonstrated the value of ensemble learning, but once again only on HCs. These authors then demonstrated automated sleep staging using only an EOG signal to derive DWT features through various statistical features such as spectral entropy, moment-based measures, refined composite multiscale dispersion entropy and autoregressive model coefficient [72]. A random forest classifier was assessed using three datasets, achieving a REM sensitivity of 0.842 ± 0.008, 0.882 ± 0.013, and 0.761 ± 0.09 (mean and standard deviation) on datasets with 8, 10, and 20 healthy participants, respectively. This study was able to verify the sole use of EOG signals for multi-stage sleep classification.

Memar and Faradji (2018) extracted 104 features from a single EEG signal on three datasets containing 8, 10, and 20 healthy participants [70]. Using a statistical significance test in combination with a minimum redundancy and maximum relevance feature selection algorithm, this method then used 30 features in a random forest algorithm to achieve five-state automated sleep staging. Evaluated for each dataset, this technique achieved REM sensitivity of 0.922, 0.876, and 0.832 (while REM specificity was 0.991, 0.986, and 0.961) [70]. These datasets demonstrated that there can be a large variation in the performance of REM sensitivity, even when they are sourced from young HCs [70]. While REM sensitivity was the lowest performing sleep stage in this study, this might be attributed to the lack of information available through a solitary EEG signal.

Linear Discriminant Analysis

Using EEG, EOG, EMG, and ECG signals to extract 74 features, Krakovská and Mezeiová (2011) demonstrated that a carefully selected set of 4 to 14 features were able to achieve sufficient sleep staging results on a dataset of 20 healthy participants [68]. Using quadratic discriminant analysis (an extension of linear discriminant analysis), this technique achieved an overall multi-stage accuracy of 0.81 and a REM sensitivity of 0.862.

Liang *et al.* designed an eye-mask device to record EOG signals, which were compared to PSG recordings to evaluate automated sleep staging (five states) using linear discriminant analysis [76]. Multi-scale entropy and autoregressive coefficients were derived from the EOG signals recorded from eight young healthy participants (25.3 ± 2.5 years) and achieved REM sensitivity of 0.9414 and a five-state kappa score of 0.77 [76]. This study demonstrated the feasibility of a portable sleep recording device that could achieve automated sleep staging on a small set of HCs. This work would benefit from additional data that might include a wider age distribution. Once more the ability of this technique on participants with RBD remains unknown.

Artificial Neural Networks

Due to the high volume of sleep data available and the processing capabilities of modern computers, artificial neural networks (ANNs) have been explored for automatic sleep stage scoring purposes. Charbonnier *et al.* (2011) used a two-part process, where artefact rejection preceded the selection of one of four classifiers [86]. A total of 33 time and frequency domain features were extracted from a single EEG, EOG, and EMG signal for every epoch, which are then processed such that they do not contain artefacts. Artefact-free features were presented to one of four classifiers. Each of the four classifiers was designed based on which signals were available after artefact rejection. Each classifier was a multi-layer perceptron (MLP). This algorithm achieved an overall accuracy of 85.5% and REM sensitivity was 80% from a dataset consisting of 13 healthy participants (aged

between 19 to 47 years). This technique benefited from having EMG features that were able to better distinguish REM sleep.

Ronzhina *et al.* (2012) extracted relative spectral power features from a single EEG channel and used numerous ANN configurations to arrive at an optimal performance [87]. Using a database of eight participants (aged between 23 and 35 years old) their highest overall agreement was 77% with 30 input nodes, 11 hidden nodes and 6 output nodes. REM agreement sensitivity was 82%.

Tian and Liu (2005) proposed a hybrid system comprising a neural network and a fuzzy rule-based reasoning system [88]. This technique extracted temporal and spectral features from a single EEG and EMG signal from a dataset of 16 adults aged between 19 and 45 years. Features were used initially to roughly classify epochs using a self-organising feature map into wakefulness, light or deep sleep. This was further refined using heuristic rules, which were then presented to a process based on fuzzy rules for classification of sleep stages. The output from the fuzzy classification was also followed by another rule based correction strategy to further improve accuracy. The algorithm achieved an overall accuracy of 85% for all sleep stages and REM stages had an agreement rate of 84%.

A well designed neural network with a relatively large number of features has been shown to produce high classification accuracy. However, from a clinical perspective is it difficult to justify and derive meaning from specifying parameter weightings of neural networks.

Summary

This section has reviewed numerous classification techniques and features to demonstrate the feasibility of automatic sleep staging. Many of these automated sleep scoring algorithms produce results that are akin to an expert scorer but are often limited by their focus on healthy individuals with little variation in age. Only two studies demonstrated the ability to achieve automated sleep staging in RBD participants [54, 55]. Comparing numerous automated sleep staging techniques was difficult when studies used numerous performance metrics with

vastly different datasets. The studies above demonstrated that a vast array of features are able to achieve automated sleep staging. A few studies even illustrated the importance of some features over others, but no clear overarching set of superior features for multi-stage sleep classification. This chapter uses these studies as a precedent to further demonstrate automated sleep staging on a mixed cohort of HCs and RBD participants (including some publicly available data) for the purposes of RBD detection. This aim of this chapter was not to identify the best classifier for automated sleep staging, but to use an established technique and collate popular and validated features to evaluate the performance on a new large dataset (described in Chapter 2).

3.4 Methods

3.4.1 Data

We used the combined dataset described in Section 2.6 as a way to demonstrate automated sleep staging for a combined cohort of HCs and RBD participants. The dataset consisted of 50 RBD participants and 50 age-matched HCs. The gold-standard used for this chapter were the annotated sleep stages provided by clinical experts that recorded these PSG recordings.

3.4.2 Signal Pre-Processing

PSG recordings were first pre-processed to reduce noise and the effect of artefacts. To ensure consistency between the various recordings, all EEG, EOG, and EMG signals were resampled at 200Hz. The EEG and EOG signal were pre-processed with a 500^{th} order band pass finite impulse response (FIR) filter with a cut-off frequency of 0.3Hz and 40Hz. The EMG signal was filtered with a 500^{th} order notch filter at 50Hz and 60Hz (because data was sourced from European and Canadian institutions), in addition to a 500^{th} order high pass FIR filter with a cut-off frequency at 10Hz. Once pre-processed, features were then extracted from the signals so as to characterise every 30 second epoch. Often these techniques

rely on signal transformations, such as the Discrete Fourier Transform (DFT) and Discrete Wavelet Transform (DWT) for frequency domain features.

3.4.3 Feature Extraction

As detailed in Figure 3.1, feature extraction from PSG signals succeeds signal pre-processing. The following section details the extraction of features for each 30 second epoch for a single EEG, EOG, and EMG signal. The features extracted for this chapter were amalgamated from popular feature-based automated sleep stage classification studies (see Table 3.1). Once features are extracted for every annotated epoch they can be used to assess automated sleep staging.

Features to represent individual epochs were often extracted from mini-epochs and averaged across the entire 30 second epoch. The use of mini-epochs was particularly pertinent for frequency based features, which required signals to be stationary. For this thesis, mini-epochs were chosen to be 10 seconds, which was short enough to be considered stationary while also long enough to accurately describe a large proportion of the epoch.

Discrete Fourier Transform

The DFT is often used in signal processing and depicts a time series vector as a vector of its' frequency components. A formal definition of the DFT is presented in the appendix A.1.1. The DFT assumes that the time series signal is stationary, which means a constant mean and variance throughout the signal. However, all PSG signals are non-stationary, where the mean and variance change throughout the recording. The DFT is still widely used for PSG signals because the inherent information from frequency based components are invaluable and provide insightful information about the signal. To limit the impact of using non-stationary signals for the DFT, the features are extracted from short time windows (mini-epochs), each 10 seconds long, where signals can be considered stationary.

EEG Bandwidth	Frequency Range (Hz)
Delta (δ)	$0.5 - 4$
Theta (θ)	$4 - 8$
Alpha (α)	$8 - 12$
Lower Alpha (α_l)	$8 - 10$
Upper Alpha (α_u)	$10 - 12$
Lower Beta (β_l)	$12 - 15$
Upper Beta (β_u)	$15 - 30$
Gamma (γ)	$30 - 50$

Table 3.2: EEG bandwidth ranges for specific feature extraction.

Discrete Wavelet Transform

The DWT decomposes a signal into a series of Detail coefficients (Dl) and Approximation coefficients (Al), up to l levels of decomposition. In this study the DWT was applied to four levels, with $D1$-$D4$ coefficients and $D4$ used for feature extraction. This transform is defined in more detail in Appendix A.1.2.

Clinically Relevant Frequency Bands

As detailed in Chapter 1 Section 1.2.1 (see Table 1.1), there are clinically relevant frequency bands by which EEG signals are interpreted. Several features, described below, are calculated for each of these bands, where the α and β bands are divided into two groups as detailed in Table 3.2. For these features, the EEG signal was filtered by an 100^{th} order band-pass FIR filter with cut-off frequencies detailed in Table 3.2.

3.4.4 Electroencephalogram Feature Extraction

There are numerous studies that have explored automatic sleep staging based on EEG alone and were the motivation of EEG based features extracted in this work [89, 69, 57, 90, 91, 59, 92, 93]. The brain-wave activity recorded from the central and occipital areas of the scalp determine EEG features of the sleep/wake state and are often used to distinguish characteristic sleep stage attributes, including vertex sharp waves, K-complexes, sleep spindles, and delta waves [94].

Time Domain Features

From the amplitude of the EEG signal, simple features that describe visual characteristics were extracted, including maximum amplitude (Amax), mean amplitude (Amean), standard deviation (Astd), variance, percentiles (25^{th}, 75^{th}, and the percentile difference between the 75^{th} and 25^{th}). Additionally, the zero crossing rate was calculated by counting the number of zero-crossing instances within a given time window. These features were calculated for every mini-epoch (10 second), and averaged across each 30 second epoch.

Coastline

The coastline feature was defined by the summation of the absolute difference between successive amplitude values [79]. These features were calculated for every mini-epoch (10 second), and averaged across each 30 second epoch. This study also calculated the maximum and minimum value within a given 30 second epoch.

Change in Variance (DX_variance)

Change in variance is simply the differential of the mean variance calculated from adjacent 30 second epochs.

Kurtosis

Kurtosis is the measure detailing the tail end of a distribution. For a signal $\boldsymbol{x} \in \mathbb{R}^{1\times N}$, kurtosis, k, is defined as:

$$k_1 = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^4}{\left[\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2\right]^2} \tag{3.1}$$

$$k = \frac{N-1}{(N-2)(N-3)}((N+1)k_1 - 3(N-1)) + 3 \tag{3.2}$$

where, x_i is the i^{th} sample in $\boldsymbol{x}$ and $\bar{x}$ denotes the mean of the samples in $\boldsymbol{x}$.

A kurtosis value greater than three indicates the distribution has more outliers compared to a normal distribution, while a value less than three has fewer outliers.

Hjorth

These parameters are based on variance of the derivatives of the signal and have been an established feature used in EEG analysis and sleep classification [95, 59, 96]. The stationary Hjorth parameters were calculated for each mini-epoch to derive features indicative of activity, mobility, and complexity. For a signal $\boldsymbol{x} \in \mathbb{R}^{1\times N}$, the Hjorth parameters for activity, mobility, and complexity are defined as:

$$H_{activity} = \frac{1}{N}\sum_{i=1}^{N} x_i^2 \tag{3.3}$$

$$H_{mobility} = \sqrt{\frac{\frac{1}{N}\sum_{i=1}^{N} x_i'^2}{H_{activity}}} \tag{3.4}$$

$$H_{complexity} = \frac{\sqrt{\frac{\frac{1}{N}\sum_{i=1}^{N} x_i''^2}{\frac{1}{N}\sum_{i=1}^{N} x_i'^2}}}{H_{mobility}} \tag{3.5}$$

where, x_i is the i^{th} sample in $\boldsymbol{x}$, x' is the first derivative, and x'' is the second derivative.

The activity parameters provide a measure of the signal's variance or mean power. While the mobility parameter characterises the mean frequency of the signal [95]. The complexity parameter measures the change in frequency with reference to the mobility [95].

Autoregressive Coefficients (AR)

A popular technique for feature extraction from EEG signals is parametric modelling, where a signal is approximated by a mathematical model. In an autoregressive (AR) model of order p, the signal, $\boldsymbol{x} \in \mathbb{R}^{1\times N}$ (with N samples), can be defined as a linear combination of the past p values plus a white noise input.

$$\widetilde{\boldsymbol{x}}_{\boldsymbol{i}} = \sum_{n=1}^{p}(a_n x_{i-n}) + \boldsymbol{\xi_i} \tag{3.6}$$

where, $\widetilde{\boldsymbol{x}}_{\boldsymbol{i}}$, denotes each i^{th} sample in $\widetilde{x}$, which is the predicted AR model of $\boldsymbol{x}$ based on p AR coefficients denoted by a_n. $\boldsymbol{\xi}$ denotes white Gaussian noise.

The AR parameters, a_n, are determined using the Yule-Walker method and were calculated to the 10^{th} order. This order was inspired by Pardey *et al.* (1996) [97], that surmised for distinctly different stages of the sleep the optimal model order, p, ranged from 3 to 12 for various stages of sleep.

Time Domain Parameters (tdp)

The time domain parameters were inspired by and are a more generalised version of the Hjorth parameters described in Section 3.4.4 [98]. The time domain properties for each mini-epoch of the signal, $\boldsymbol{x} \in \mathbb{R}^{1\times N}$ (with N samples), are derived using the log-power (tdpf1) and log-amplitude (tdpf2) of each derivative up to the 10^{th} derivative as defined below:

$$P_j = \log\left(\frac{1}{N}\sum_{i=1}^{N}|\Delta^j x_i^2|\right) \tag{3.7}$$

$$A_j = \log\left(\frac{1}{N}\sum_{i=1}^{N}|\Delta^j x_i|\right) \tag{3.8}$$

where, $\Delta^j x_i$ is the i^{th} sample of the j^{th} derivative of $\boldsymbol{x}$. P_j and A_j denote the log-power and log-amplitude of the j^{th} derivative.

Square Energy Operator (SEO)

The square energy operator (SEO) features calculate the mean power and the standard deviation of a signal, $\boldsymbol{x} \in \mathbb{R}^{1\times N}$ (with N samples), as detailed below:

$$S_\mu = \frac{1}{N}\sum_{i=1}^{N} x_i^2 \tag{3.9}$$

$$S_\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}|x_i^2 - S_\mu|^2} \tag{3.10}$$

where, x_i is the i^{th} sample in $\boldsymbol{x}$, S_μ denotes the mean SEO, and S_σ denotes the standard deviation of the SEO.

Additionally, these features are calculated for each of the clinically relevant bands detailed in Table 3.2.

Teager-Kaiser Energy Operator (TKEO)

The Teager-Kaiser energy operator (TKEO) aims to measure the real physical energy of a system [99, 100]. The mean and standard deviation of the TKEO for a signal, $\boldsymbol{x} \in \mathbb{R}^{1\times N}$ (with N samples), is defined as:

$$T_\mu = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - x_{i-1}\cdot x_{i+1} \tag{3.11}$$

$$T_\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N} |x_i^2 - x_{i-1}\cdot x_{i+1} - T_\mu|^2} \tag{3.12}$$

where, x_i is the i^{th} sample in $\boldsymbol{x}$, T_μ denotes the mean Teager-Kaiser operator, and T_σ denotes the standard deviation of the Teager-Kaiser operator.

Once more, these features are calculated for each of the clinically relevant bands detailed in Table 3.2.

Ratios

From a signal the EEG ratio can be defined as the relative change of energy in a 20 second sliding window relative to the adjacent 640 second window. For a signal, $\boldsymbol{x} \in \mathbb{R}^{1\times N}$ (with N samples), the energy is calculated as the normalised squared amplitude:

$$s_i = \frac{x_i^2}{\max\{x_i^2 : i = 1 \ldots N\}} \tag{3.13}$$

where, x_i is the i^{th} sample in $\boldsymbol{x}$ and s_i is the i^{th} sample in the normalised squared amplitude $\boldsymbol{s} \in \mathbb{R}^{1\times N}$ (with N samples).

The ratio, $\boldsymbol{R} \in \mathbb{R}^{1\times N}$ (with N samples), is calculated as follows:

$$R_i = \frac{\frac{1}{20f_s}\sum_{j=i-10f_s}^{i+10f_s} s_j - \frac{1}{640f_s}\sum_{j=i-320f_s}^{i+320f_s} s_j}{\frac{1}{640f_s}\sum_{j=i-320f_s}^{i+320f_s} s_j} \tag{3.14}$$

where, R_i is the i^{th} sample of $\boldsymbol{R}$, s_j is the j^{th} sample in the normalised squared amplitude $\boldsymbol{s}$, and f_s is the sampling frequency. Once more, these features are calculated for each of the clinically relevant bands detailed in Table 3.2.

Permutation Entropy

Permutation entropy is a nonlinear measure that characterises the complexity of a time series signal [101]. For every mini-epoch the permutation entropy was calculated to the 10^{th} order. For a signal, $\boldsymbol{x} \in \mathbb{R}^{1\times N}$ (with N samples), where all n factorial ($n!$) permutations of their sequence in relative value, of length n (order), are represented by π. For each π the relative frequency was calculated as $p(\pi)$:

$$p(\pi) = \frac{\{\#t|t \leq N-n, (x_{t+1}, \ldots, x_{t+n}) \text{ has type } \pi\}}{N-n+1} \tag{3.15}$$

where, # denotes the counted number of combinations. In the case where there are equal values for the sequence, the order of appearance matters and the first occurrence was considered lower rank.

The permutation entropy for $n \geq 2$ is defined as:

$$H = -\sum_{\pi_1}^{\pi_{n!}} p(\pi) \log_2 p(\pi) \tag{3.16}$$

where the sum runs over all $n!$ permutations π of order n and H denotes the permutation entropy.

Frequency Domain Features

The frequency-domain features are derived using a one-sided periodogram (as defined in Appendix A.1.1), which characterises the power spectral density as a function of frequency. The DFT uses a windowing function of fixed width that separates the non-stationary signal into segments, which are considered close to stationary (10 seconds in this study). The Fourier transform was then applied to the segmented signal with a Hanning window. In automated sleep staging literature it is commonplace to see features such as peak frequency, mean frequency, peak frequency power, and mean Fourier magnitude for each clinically relevant band (detailed in Table 3.2). The following are some unconventional features from literature.

Median Frequency

The median frequency is defined as the frequency that equally divides the power spectral density function, $\boldsymbol{P} \in \mathbb{R}^{1\times N}$ (with N frequency bins), in two equal halves as detailed here:

$$\sum_{f=0}^{f_m} P_f = \frac{1}{2}\sum_{f=0}^{f_s/2} P_f \tag{3.17}$$

where, f_m denotes the median frequency and f_s the sampling frequency.

Spectral Edge Frequency

The spectral edge frequency (SEF) is defined as the frequency below which a certain fraction of the signal power is contained. The SEF at 95% is usually highest during REM stages [78]. For this study the spectral edge frequency (f_{edge}) was calculated at 95% of the signal power, $\boldsymbol{P} \in \mathbb{R}^{1\times N}$ (with N frequency bins), as shown below:

$$\sum_{f=0}^{f_{edge}} P_f = 0.95\sum_{f=0}^{f_s/2} P_f \tag{3.18}$$

where, f_{edge} denotes the spectral edge frequency at the 95^{th} percentile and f_s the sampling frequency.

Relative Spectral Power (RSP)

Relative power ratios were calculated for each clinically relevant frequency band as detailed in Table 3.2. Firstly for a power spectral density, $\boldsymbol{P} \in \mathbb{R}^{1\times N}$ (with N frequency bins), the relative spectral powers are calculated by taking the ratio between individual sub-band spectral powers and the total spectral power (the summation of each sub-band spectral power):

$$P_{total} = P_\delta + P_\theta + P_\alpha + P_\beta + P_\gamma \tag{3.19}$$

where, P_{total}, P_δ, P_θ, P_α, P_β, and P_γ denote the total, delta, theta, alpha, beta, and gamma spectral power, respectively.

$$RSP_\delta = \frac{P_\delta}{P_{total}} \tag{3.20}$$

$$RSP_\theta = \frac{P_\theta}{P_{total}} \tag{3.21}$$

$$RSP_\alpha = \frac{P_\alpha}{P_{total}} \tag{3.22}$$

$$RSP_\beta = \frac{P_\beta}{P_{total}} \tag{3.23}$$

$$RSP_\gamma = \frac{P_\gamma}{P_{total}} \tag{3.24}$$

where, RSP_δ, RSP_θ, RSP_α, RSP_β, and RSP_γ denotes the delta, theta, alpha, beta, and gamma relative spectral power, respectively.

Power Spectral Entropy

The power spectral entropy describes the spectral complexity of an uncertain system, also known as information entropy. The power spectral entropy was calculated from a power spectral density, $\boldsymbol{P} \in \mathbb{R}^{1\times N}$ (with N frequency bins) as follows:

$$p_i = \frac{P_i}{\sum_{i=1}^{N} P_i} \tag{3.25}$$

where, P_i denotes the power spectral density at the i^{th} frequency bin, where there are a total of N bins. This was calculated for frequencies between 0Hz and $\frac{f_s}{2}$Hz, where f_s is the sampling frequency.

The power spectral entropy was then defined by:

$$H = -\sum_{i=1}^{N} p_i \ln p_i \tag{3.26}$$

where, H denotes the power spectral entropy and was calculated for each mini-epoch.

3.4.5 Electrooculogram Feature Extraction

Additionally EOG signals were also processed and analysed to facilitate automated sleep staging, specifically to identify SEM and REM. In this work the EOG signal was derived from the difference between the right EOG electrode and the left EOG electrode. A number of these features included those extracted for EEG signals,

kurtosis, variance, percentiles (25^{th}, 75^{th}, and the percentile difference between the 75^{th} and 25^{th}), zero crossing rate, coastline (see 3.4.4 including the minimum and maximum value within each epoch), TKEO (see 3.4.4), relative spectral power (see 3.4.4), permutation entropy (see 3.4.4), and change in variance (see 3.4.4). Additional features were also inspired by Yetton *et al.* (2016), who demonstrated REM detection using only EOG features [79]. All EOG features were calculated for every 10 second mini-epoch, and averaged across each 30 second epochs.

Skewness

Skewness is a measure of the symmetry of a distribution when compared to a normal distribution. Positive values indicate the data is spread more towards values greater than the mean, while negative values mean they are spread to values less than the mean. For a signal, $\boldsymbol{x} \in \mathbb{R}^{1\times N}$ (with N samples), skewness is defined as:

$$s = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^3}{\left(\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}\right)^3} \tag{3.27}$$

where, s denotes the skewness, x_i is the i^{th} sample of $\mathbf{x}$, and $\bar{x}$ denotes the mean value of $\mathbf{x}$.

Max Rate of Change (MaxDx)

The maximum rate of change for the EOG signal was also calculated for each 10 second mini-epoch. For a signal, $\boldsymbol{x} \in \mathbb{R}^{1\times N}$ (with N samples), the max rate of change is defined as:

$$D = \max_{i=1\ldots N}\{|x_{i+1} - x_i|\} \tag{3.28}$$

where, D is the max rate of change and x_i is the i^{th} sample of $\boldsymbol{x}$. For this study we also calculated the maximum and minimum value within the 30 second epoch.

Maximum Peak, Prominence, Width, and Slope

For each mini-epoch the absolute maximum peak, peak prominence, width, and slope were found. Additionally, the second highest absolute maximum peak was also calculated. Peak prominence is defined as a measure of how much the peak extends from the baseline (not the zero line). The concept here was to describe an isolated peak with a greater prominence than a peak surrounded by other peaks. The prominence was calculated as follows:

1. Find maximum absolute peak.
2. Extend horizontal line from the peak to the left and right until line either:
   (a) Intersects signal.
   (b) Reaches the beginning or end of signal.
3. From the two intervals identified in step two (left and right from the peak), find the two minima for both intervals (either a valley or signal endpoint).
4. Find the maximum of the two minima found in step 3, this is the baseline to measure the height of the peak to calculate the prominence.

From the peak prominence, the peak width and slope can be calculated. The peak width is simply the width of the signal at half prominence from the baseline. The slope is represented by two values, the rise and the fall. This is calculated by measuring the rate of change from half prominence, before and after the peak for rise and fall, respectively. For further detail a graphical representation is given in Figure 3.2, detailing the calculation of prominence, width, and slope.

Autocorrelation (Autocorr)

Yetton *et al.* 2016 demonstrated the use of cross-correlation between vertical and horizontal eye movement for REM classification [79]. Inspired by these features, this study used autocorrelation on the single EOG signal. The autocorrelation gives an indication of a pattern within a given mini-epoch, enhancing REM spikes within the

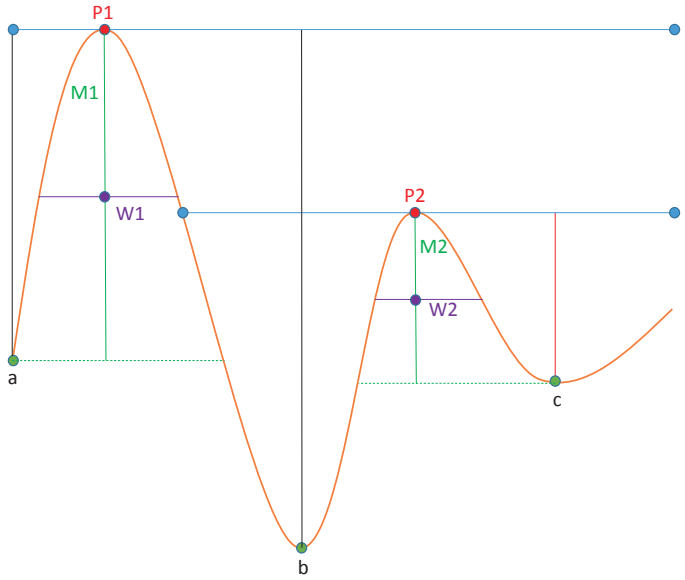


Figure 3.2: For a function shown in orange, there are two peaks ($P1$ and $P2$). The prominence of both points ($P1$ and $P2$) was measured by first identifying on the left and right of the peak, where the signal ends or intersects with the signal (this is depicted by the blue dots). Next the the minimum values between the peak value and each blue dot must be found. For $P1$ this corresponds to the green dots labelled a and b, while for $P2$ this refers to the green dots labelled b and c. From the green dots, the maximum values are used to calculate the prominence, this corresponds to points a for P1 and c for P2 (because $a > b$ and $c > b$). The prominence was then simply the height from this max value to the peak point (depicted by the green line $M1$ and $M2$ for $P1$ and $P2$, respectively). The peak width is given by measuring, from half prominence height (points $W1$ and $W2$), the width of the signal (given by the purple line $W1$ and $W2$ for $P1$ and $P2$, respectively).

EOG signal. The peak value of the autocorrelation sequence of every mini-epoch was calculated, for a signal, $\boldsymbol{x} \in \mathbb{R}^{1\times N}$ (with N samples), and is defined as:

$$R_{xx}(m) = \begin{cases} \sum_{n=0}^{N-m-1} x_{n+m} x_n^*, & m \geq 0, \\ R_{xx}(-m), & m < 0. \end{cases} \tag{3.29}$$

$$c_m = R_{xx}(m - N) \tag{3.30}$$

$$C = \max_{m=1,2,\ldots 2N-1} \{c_m\} \tag{3.31}$$

where, R_{xx} is autocorrelation for $\boldsymbol{x}$,$\boldsymbol{x}^*$ denotes the complex conjugate of $\boldsymbol{x}$, c_m is the m^{th} element of the autocorrelation sequence ($\boldsymbol{c} \in \mathbb{R}^{1\times 2N-1}$), and C the maximum value within the sequence. For this study, the maximum and minimum mini-epoch value for every 30 second epoch was calculated.

Haar Wavelet Transform

A DWT decomposes a discrete signal into a set of pre-specified basis functions. For automated sleep scoring Yetton *et al.* (2015) applied Haar wavelets as a basis function, because it resembles the shape of REM in EOG signals and is ideal for discontinuities [79]. For the wavelet the DWT was applied using four levels of decomposition followed by the four levels of the inverse DWT, where the maximum of the reconstructed signal was the extracted feature. The larger the amplitude of the inverse DWT signal the more the signal can be composed of wavelets (which resemble REM characteristics) and also more likely that the window contains REM [79].

3.4.6 Electromyogram Feature Extraction

EMG signals are used by clinicians to gauge levels of muscle atonia by observing phasic and tonic movement, which are also translated into features for sleep stage classification. Features calculated for EEG signals were also extracted for EMG signals, these included the 75^{th} percentile, spectral edge frequency (see 3.4.4), variance, and standard deviation. Features that quantify muscle atonia were also calculated and were inspired by work used to identify individuals with RBD, including the atonia index, quantified motor activity, and STREAM [102, 103, 104, 105]. All EMG features were calculated for every 30 second epoch.

Energy

EMG energy, as described by Liang *et al.* 2012, measures the mean absolute amplitude over each epoch in order to quantify body movement [65]. For a signal, $\boldsymbol{x} \in \mathbb{R}^{1\times N}$ (with N samples), it was defined as:

$$E = \frac{1}{N}\sum_{i=0}^{N}|x_i| \tag{3.32}$$

where, E was the energy and x_i the i^{th} element of $\boldsymbol{x}$. This feature was calculated for every 30 second epoch.

Entropy

The entropy of a signal measures the signal variability by deriving a distribution of its amplitude values. The algorithm used in this study was provided by Moddemeijer (1989) [106]. A histogram for signal $\boldsymbol{x} \in \mathbb{R}^{1\times N}$ (with N samples) is given by $\mathbf{k}$, where k_i is the number of samples in the i^{th} amplitude bin (total number of samples equals N).

$$H = -\sum_i (\frac{k_i}{N} \log \frac{k_i}{N}) + \log \Delta x + \beta \tag{3.33}$$

where, H denotes the calculated entropy, Δx the histogram bin width, and β the bias, which is defined as:

$$\beta = \frac{I-1}{2N} \tag{3.34}$$

where, I is the total number of bins from the histogram. This feature was calculated for every 30 second epoch.

Absolute Power (Abs_Power)

The average power between $10-50$Hz was calculated by integrating (rectangular method) the power spectral density function, $P(f)$. This feature was calculated for every 30 second epoch.

Relative Spectral Power (RSP)

Similar to the EEG spectral powers discussed in 3.4.4, the EMG relative spectral power was calculated, but only for frequencies greater than 10Hz, for alpha (10-12Hz), beta (12-30Hz), and gamma (30-40Hz).

Corrected Atonia Index

This metric has been associated with RBD identification since 2008 and further enhanced in 2010 [103, 104]. Firstly, the EMG signal, $\boldsymbol{x} \in \mathbb{R}^{1\times N}$ (with N samples), was rectified and averaged for each one second mini-epoch:

$$E_i = \frac{1}{f_s} \sum_{j=(i-1)\cdot f_s+1}^{(i-1)\cdot f_s+f_s} |x_j| \tag{3.35}$$

where, E_i denotes the i^{th} sample of the rectified one second averaged EMG signal ($\boldsymbol{E} \in \mathbb{R}^{1\times N/f_s}$) and f_s the sampling frequency.

The background noise was determined by calculating the minimum value over a moving 60 second sliding window. The background noise, $\boldsymbol{B} \in \mathbb{R}^{1\times N/f_s}$, was subtracted from the rectified and averaged EMG signal, $\boldsymbol{R} \in \mathbb{R}^{1\times N/f_s}$ to give the corrected, rectified, and averaged EMG signal ($\boldsymbol{A}$).

$$B_i = \min_{m=i\ldots(i+60)} \{E_m\} \tag{3.36}$$

$$A = E - B \tag{3.37}$$

where, B_i denotes the i^{th} sample of background noise ($\boldsymbol{B}$) and $\boldsymbol{A} \in \mathbb{R}^{1\times N/f_s}$ the corrected, rectified, and averaged one second amplitude.

For each 30 second epoch a normalised distribution histogram was used to calculate the percentage of amplitudes that fall into 20 classes (C_k, where k is the k^{th} class). C_0 is the percentage of values in A that are $\leq 1\mu V$. C_{20} is the percentage of values in A that are $\geq 19\mu V$. C_k is equal to the percentage of values $> k\mu V$ and $\leq (k+1)\mu V$. The corrected index for each mini-epoch was calculated from the ratio of values $\leq 1\mu V$ compared to all other values (excluding values $> 1\mu V$ and $\leq 2\mu V$), as shown in Equation 3.38. The corrected atonia index ranges from 0 (all amplitudes $> 1\mu V$, *i.e* absence of muscle atonia) to 1 (all amplitudes $\leq 1\mu V$, *i.e* complete muscle atonia):

$$I_i = \frac{C_0}{100 - C_1} \tag{3.38}$$

where, I_i denotes the corrected atonia index for the i^{th} epoch in the corrected, rectified, and averaged EMG signal. The corrected atonia index was calculated for each 30 second epoch.

Quantified Motor Activity

The quantified motor activity (QMA) was derived from an amplitude curve, defined by the difference between the maximum and minimum value using a sliding (51 sample) window across the EMG signal, $\boldsymbol{x} \in \mathbb{R}^{1\times N}$ (see Equation 3.39). The median value of the amplitude curve was then calculated for every epoch and was used as a feature (QMA amplitude), denoted by $\boldsymbol{m} \in \mathbb{R}^{1\times Q}$ (where Q is the number of epochs). To quantify the motor activity, a threshold signal was determined by the minimum value of $\boldsymbol{m}$, using a sliding window (121 epochs long), called the baseline (QMA baseline, $\boldsymbol{B} \in \mathbb{R}^{1\times Q}$, detailed in Equation 3.40) (this baseline value was also used as a feature). Whenever $\boldsymbol{m}$ exceeds four times the baseline ($\boldsymbol{B}$), it is considered movement activity. Movement activity segments less than 0.3 seconds are discarded. Similarly, segments separated by less that 0.5 seconds are merged into a single segment. The quantified motor activity provides two metrics, one derived accumulating the total duration of motor activity event segments as a percentage of the epoch (QMA duration). Second, the percentage of three second mini-epochs within each epoch with more than 50% of motor activity within their three second mini-epoch (QMA percentage). In the original study, the maximum value of these two metrics was used for RBD identification [105].

$$a_i = \max_{i-25\leq i\leq i+25}\{x_i\} - \min_{i-25\leq i\leq i+25}\{x_i\} \tag{3.39}$$

where, $\boldsymbol{a} \in \mathbb{R}^{1\times N}$ denotes the amplitude curve and i is the i^{th} sample point of $\boldsymbol{a}$. The QMA baseline was defined as:

$$B_j = \min_{j-60\leq j\leq j+60}\{m_j\} \tag{3.40}$$

where, B denotes the baseline, $\boldsymbol{m} \in \mathbb{R}^{1\times Q}$ denotes the amplitude curve and j is the j^{th} sample point of $\boldsymbol{m}$.

Electromyography Fractal Exponent

The fractal exponent feature was inspired by Krakovská *et al.* (2011), where it was used for automated sleep staging [68]. This feature aims to measure signal complexity by fitting a linear line (α) to a double logarithmic graph of spectral power density, $P(f)$, versus frequency, f, [68] (Equation 3.41). The fractal exponent was calculated for every 10 second epoch.

$$P(f) \sim f^{-\alpha} \tag{3.41}$$

where, α is the fractal exponent, for spectral power density $P(f)$ with frequencies f.

3.4.7 Miscellaneous Feature Extraction

In addition to extracting features from individual PSG signals, the following features combine signals and other miscellaneous attributes.

EEG & EOG Cross-correlation (CrossCorrelation)

At times EOG characteristics, like eye movement artefacts, can leak into EEG frontal electrodes [19]. Inspired by Yetton *et al.* 2016, this feature calculated the cross-correlation between the EEG and EOG signal. The peak of the cross-correlation sequence of every mini-epoch was calculated, for a EEG signal ($\boldsymbol{x} \in \mathbb{R}^{1\times N}$) and an EOG signal ($\boldsymbol{y} \in \mathbb{R}^{1\times N}$) (both with N samples) this is defined as:

$$R_{xy}(m) = \begin{cases} \sum_{n=0}^{N-m-1} x_{n+m} y_n^*, & m \geq 0, \\ R_{xy}(-m), & m < 0. \end{cases} \tag{3.42}$$

$$c_m = R_{xy}(m - N) \tag{3.43}$$

$$C = \max_{m=1,2,\ldots 2N-1} \{c_m\} \tag{3.44}$$

where, R_{xy} is the cross-correlation for $\boldsymbol{x}$ and $\boldsymbol{y}$, $\boldsymbol{y}^*$ denotes the complex conjugate of $\boldsymbol{y}$, c_m is the m^{th} element of the cross-correlation sequence ($\boldsymbol{c} \in \mathbb{R}^{1\times 2N-1}$), and C the maximum value within the sequence. For this study, the maximum absolute mini-epoch value for every 30 second epoch was calculated.

The Pearson cross-correlation coefficient was also calculated for each 10 second mini-epoch, which was defined as:

$$\rho = \frac{1}{N-1}\sum_{i=1}^{N}(\frac{x_i - \bar{x}}{\sigma_x})(\frac{y_i - \bar{y}}{\sigma_y}) \tag{3.45}$$

where, ρ denotes the Pearson correlation coefficient, x_i and y_i the i^{th} elements of $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. $\bar{x}$ denotes the mean value of $\boldsymbol{x}$, $\bar{y}$ the mean value of $\boldsymbol{y}$, σ_x the standard deviation of $\boldsymbol{x}$, and σ_y the standard deviation of $\boldsymbol{y}$.

EEG & EOG Coherence (MSCoherence)

Similar to the cross correlation features above, the coherence between the EEG and EOG signal was calculated for each epoch. The coherence aims to measure the similarity between the frequency spectrum of two signals ($\boldsymbol{x} \in \mathbb{R}^{1\times N}$ and $\boldsymbol{y} \in \mathbb{R}^{1\times N}$) using the magnitude squared coherence estimate. The magnitude squared coherence is a function of the power spectral density of $P_{xx}(f)$, $P_{yy}(f)$, and the cross power spectral density $P_{xy}(f)$ of the signals $\boldsymbol{x}$ and $\boldsymbol{y}$:

$$C_{xy}(f) = \frac{|P_{xy}(f)|^2}{P_{xx}(f)P_{yy}(f)} \tag{3.46}$$

$$C = \max_{m=1,2,\ldots f_s/2}\{C_{xy}\} \tag{3.47}$$

where, C_{xy} denotes the magnitude squared coherence estimate and was calculated for every 10 second mini-epoch using a hamming window of length $f_s/2$. The feature extracted was the maximum value of the magnitude squared coherence estimate, as shown in Equation 3.47.

Elapsed Time (HoursRec & HoursFromEnd)

Throughout the night, sleep stages are understood to change periodically, while REM is also documented to occur later in the night with more frequency [1]. As a result this study used elapsed time (in hours) and time until end of recording (in hours) as features. Elapsed time is the progression of time since the beginning of the first annotated epoch from the PSG recording.

3.4.8 Random Forest Classification

The random forest classifier was conceived by Breiman (2001) [107]. The concept involves many individual decision trees, where each tree casts a vote to determine the most popular class [108]. An individual decision tree can be represented as a list of if-then rules (nodes) that can partition a feature space into discrete classes (leaves). A random forest comprises of an ensemble of decision trees that use a random selection of features to split each node. The ensemble of trees provide a overall classification using majority voting, which helps dilute classification bias [108]. The random forest is a useful classifier to grasp the nature of a classification problem, because it requires only two parameters (where performance tends to be insensitive to parameter tuning), robust to outliers, relatively fast to train, resistant to over-fitting, and is capable of handling a large number of features. Furthermore, the algorithm is able to provide additional interpretive information regarding the importance of features.

Chief to building a random forest is the preparation of training data. A new set of training data is created by randomly sampling the original training data (sampling with replacement), which is known as the in-bag dataset. The in-bag set is equal to the original training dataset and will be used to "grow" trees. This is also known as a bootstrapping technique, which is often used to reduce variance and overcome over-fitting [107]. Each tree was constructed using a different bootstrap sample, a figure summarising this process is illustrated in Figure 3.3.

For each node of a tree, m_{try} random features are selected and used to construct nodes and leaves of the random tree classifier. The randomisation of features increases the accuracy of classification, decreases sensitivity to noise, and minimises the correlation between features [108].

Trees begin by splitting the first node into branches by identifying the best split within the group of randomly selected m_{try} features. For each node, the selected feature and splitting value is determined by calculating the split that achieves the greatest change in impurity. The impurity measures the extent of purity for a leaf node containing data points from possibly different classes. Thus the splitting

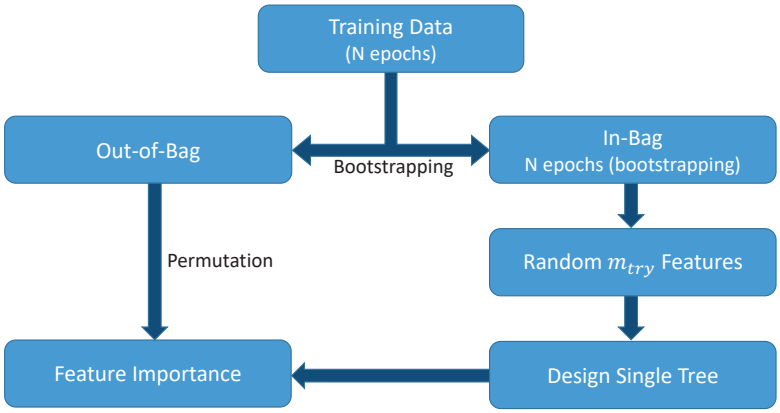


Figure 3.3: When constructing the ensemble of trees for the random forest, each tree begins by first bootstrapping the training dataset to form the in-bag set. A random m_{try} number of features are then used to design each node of the tree. The out-of-bag set can then be used to evaluate the tree and calculate the feature importance by randomly permutating the dataset and measuring the mean prediction error. This process is repeated for every tree of the random forest.

feature value is determined from every possible value from the randomly selected set of features that achieves the greatest decrease in impurity. This measure of impurity is called the Gini index, which is defined as:

$$I_g = \sum_{j=1}^{S} p_j(1 - p_j) = 1 - \sum_{j=1}^{S} p_j^2 \tag{3.48}$$

where S is the number of classification classes and p_j the percentage of data points belonging to class j. The node was split if impurity for the two descendent nodes was less than the parent node (Gini criterion), otherwise the node was not split and becomes a terminating node where the classified output is the most common class in that given subset.

This process is repeated recursively at each branch node using the subset that reached that branch with a new set of randomly selected m_{try} features. The selection of the next splitting feature is done based on the highest calculated decrease in the Gini impurity for the new set of randomly selected m_{try} features. The most common

occurring class in the training subset that reached the final node becomes the predicted classification output. For the entire ensemble (forest), each tree contributes a vote, where the majority becomes the final classification, as detailed in Figure 3.4.

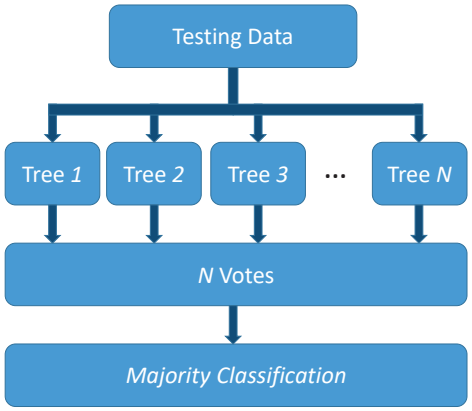


Figure 3.4: The structure of a random forest and the ensemble of trees, which each cast a vote to determine the final classification.

As detailed above, there are two parameters required for a random forest classifier, the number of trees and the number of randomly selected features for node branching. Depending on the application, either parameter can have an impact on classification performance. However, the results for classification are insensitive to the value of m_{try} [108], where a typical default value is $\sqrt{M}$ (rounded up), where M is the total number of features. Often a perturbation test is conducted to prove this is the case, where the value for m_{try} is halved and doubled to evaluate the impact on performance. For this work the perturbation test was satisfied and the impact was found to be minimal.

Breiman also concluded that increasing the number of trees simply increases the computational cost without impacting the classifier's accuracy, which converges to a limit [108]. Consequently this study did not focus on achieving computational efficiency but rather the performance of automated sleep staging, fixing the number of trees to 500 and m_{try} to the square root of the total number of features (rounded up). Once again a perturbation test was assessed for the number of trees and the value of 500 was found to be satisfactory.

Additionally, the random forest classifier is able to provide a measure of feature importance. For each feature, the importance measure is calculated by randomly permutating the variable across the "out-of-bag" samples and measuring the increase in prediction error (for classification this is accuracy). This is done for every tree and averaged over the entire ensemble and divided by the standard deviation over the entire ensemble.

3.4.9 Model training and evaluation

Evaluating machine learning models demands a systematic approach for splitting data into training and testing sets. In order to avoid over-fitting, data used to train models must be segregated from data used to test models in order to assess performance. Typically, K-fold cross-validation (CV) is adopted, where data is split into K equal (or near equal) folds/partitions.

Cross-Validation

For this chapter, a classifier was trained using $K-1$ folds and tested using the left-out fold. This is repeated K times, where each fold is cycled through as the testing set, meaning each fold is used for training while also being independently used for testing. The overall performance is derived by calculating the mean performance across all K folds [109]. Where datasets contain repeated samples from a single participant, for example sleep recordings with multiple sleep stages from a single or numerous nights, it is prudent to split folds by participant. As a result, data from a single participant does not end up in both the training and testing data, thereby avoiding bias and over fitting of the trained model.

Measures for Classification

In order to calculate the performance of the automated sleep staging, the predicted sleep stages were compared to the manually annotated sleep stages (the ground truth). For binary classification a true positive (TP) occurs when the ground truth is true and the predicted class is also true. A true negative (TN) occurs when the predicted class and the ground truth are both false. A false positive (FP) exists

when the predicted class is true but the ground truth is false. The opposite of this is a false negative (FN), where the predicted class is true while the ground truth is false. For multi-state sleep staging, each sleep stage was evaluated as a binary classification in order to generate results for each stage through macro-averaging (specifying a mean and standard deviation).

For this work the performance of the classifier was evaluated using macro-averaged sleep stage accuracy, sensitivity, specificity, and F1 by using 10-participant-fold cross-validation with an even split between healthy and RBD participants. Results are also illustrated by a five-state Cohen Kappa score and a collection of confusion matrices. These methods and metrics are detailed below.

Accuracy, Sensitivity, Specificity, Precision, and F1

This collection of TPs, TNs, FPs, and FNs can be used to generate metrics to evaluate accuracy, sensitivity, specificity, precision, and F1.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{3.49}$$

Sensitivity (also known as recall) measures the proportion of true positives (correctly predicted positives) compared to the total positives from the ground truth:

$$Sensitivity = \frac{TP}{TP + FN} \tag{3.50}$$

Specificity describes the proportion of true negatives (correctly predicted negatives) compared to the total negatives from the ground truth:

$$Specificity = \frac{TN}{TN + FP} \tag{3.51}$$

Precision (also known as positive predictive value) describes the ratio of correctly predicted positives to all predicted positives:

$$Precision = \frac{TP}{TP + FP} \tag{3.52}$$

The F1 score measures the harmonic mean between sensitivity and precision:

$$F1 = 2.\frac{Sensitivity.Precision}{Sensitivity + Precision} \tag{3.53}$$

The focus on sensitivity and specificity allows the F1 score to compensate for imbalanced data, where there are more negative than positive classes.

Confusion Matrix

The aforementioned metrics detail the performance for binary classification but additional methods and scores are required to depict multi-stage classification. The confusion matrix represents the quantitative comparison between the ground truth (annotated multi-stage hypnogram) and the predicted multi-stage output through a two dimensional matrix. An example for a binary classification confusion matrix is given in Figure 3.5, which can then be extended for additional states.

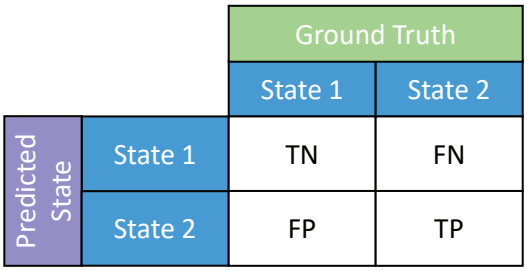


Figure 3.5: An example of a confusion matrix for a binary classification problem, where State 1 and State 2 represent the negative and positive state, respectively. This matrix is simply extended for additional states, where correctly classified states are found along the diagonal, and misclassified states elsewhere.

Cohen's Kappa

Multi-stage sleep classification was also assessed by Cohen's Kappa [110], which provides a measure of the agreement between two-raters (in this case between a human and an automated algorithm). The Cohen's Kappa is able to provide a score

for multi-state classification problems while also accounting for state imbalance, by calculating the proportion of agreement after chance agreement is removed from consideration [110]:

$$\kappa = \frac{p_0 - p_e}{1 - p_e} \tag{3.54}$$

where, p_0 denotes the relative observed agreement between two raters (defined by accuracy, shown in Equation 3.49) and p_e the expected agreement by chance (detailed below) [110]:

$$p_e = \frac{1}{N^2} \sum_{k=1}^{S} n_k^{(1)} n_k^{(2)} \tag{3.55}$$

where, N denotes the number of classified states, S is the total number of states, and $n_k^{(1)}$ and $n_k^{(2)}$ denotes the total number of predictions in each k^{th} state made by rater 1 and 2, respectively. For balanced binary classification $p_e \approx 0.5$. Detailed in Table 1.2 are arbitrary divisions to help identify the strength of agreement, ranging from poor (< 0.00), moderate ($0.41 - 0.60$), to almost perfect ($0.81 - 1.00$) [23].

Summary

To evaluate the ability of automated sleep staging on a combined cohort of age-matched HCs and RBD participants, a RF classifier (Section 3.4.8) was trained from the features detailed in the preceding sections (Sections 3.4.4, 3.4.5, 3.4.6, and 3.4.7), which are summarised in Table 3.3. With a total of 159 features, the number of m_{try} features was set to 13. The number of trees was set to 500 to achieve 5 stage automated sleep staging (W, N1, N2, N3, and REM). To assess the performance, the data described in Chapter 2 was divided into a training and testing set, and evaluated using K-participant-fold cross-validation (for this study K was set to 10).

3.5 Results

The following results detail the performance of automated sleep staging on a combined dataset of HCs and RBD participants, detailing the overall results in addition to results for individual cohorts.

Category	Feature	Feature Code	References
EEG - Time Domain	Maximum amplitude	Amax	
	Amplitude - Mean	Amean	
	Amplitude - Std	Astd	
	Amplitude - Variance	Variance	
	Percentile (25th and 75th)	Per25 & Per75	[63]
	Percentile Difference (75th - 25th)	PerDiff	[63]
	Zero crossing rate	ZeroCrossingRate	[111]
	Coastline (mean, max, and min)	Coastline	[79]
	Change in variance	DX_variance	
	Kurtosis	Kurtosis	
	Hjorth (activity, mobility, and complexity)	hjorth1, hjorth2 & hjorth3	[59, 95, 96]
	Autoregressive coefficients	AR2 to AR11	[97]
	Time domain parameters	tdpf1_1 to tdpf1_11	[98]
		tdpf2_1 to tdpf2_11	
	Square Energy Operator (mean and std)	mean_SEO	
		std_SEO	
	Band Pass (for each band in Table 3.2) Square Energy Operator	SEO_BPmean_	
		SEO_BPstd_	
	Teager-Kaiser Energy Operator (mean and std)	mean_TKEO	[99, 100]
		std_TKEO	
	Band Pass (for each band in Table 3.2) Square Energy Operator	TKEO_BPmean_	
		TKEO_BPstd_	
	Ratios (for each band in Table 3.2)	Ratios_	
	Permutation Entropy	Permutation_entropy	[101]
EEG - Frequency Domain	Peak Power Frequency	PeakPowerFreq	
	Mean Frequency	MeanFrequency	
	Median Frequency	MedianFrequency	
	Fourier Magnitude (for each band in Table 3.2)	Fourier_Magnitude_	
	Fourier Magnitude (max)	Fourier_Magnitude_max	
	Spectral Edge Frequency	SpectralEdgeFrequency	[78, 86]
	Relative Spectral Power (δ, θ, α, β, and γ)	RSP_delta	
		RSP_theta	
		RSP_alpha	
		RSP_beta	
		RSP_gamma	
	Power Spectral Entropy	SpectralEntropy	
EOG - Time Domain	Percentile (25th and 75th)	Per25	[63]
		Per75	
	Percentile Difference (75th - 25th)	PerDiff	[63]
	Zero crossing rate	ZeroCrossingRate	[111]
	Coastline (mean, max, and min)	Coastline	[79]
	Change in variance	DX_variance	
	Kurtosis	Kurtosis	
	Square Energy Operator (mean and std)	SEO	[112]
	Teager-Kaiser Energy Operator (mean and std)	TKEO	[99, 100]
	Permutation Entropy	Permutation_entropy	[101]
	Skewness	Skewness	
	Max rate of change	MaxDx	
	Maximum Peak	MaxPeak1	[79]
		MaxPeak2	
	Peak prominence	PeakProminence	[79]
	Peak width	PeakWidth	[79]
	Peak slope	PeakRise	[79]
		PeakFall	
	Autocorrelation (mean, max, and min)	Autocorr	[79]
EOG - Frequency Domain	Relative Spectral Power (δ, θ, α, β, and γ)	RSP_delta	[111]
		RSP_theta	
		RSP_alpha	
		RSP_beta	
		RSP_gamma	
	Haar wavelet transform	Wavelet	[79]
EMG - Time Domain	Energy	Energy	[90]
	Entropy	Entropy	[86]
	Corrected atonia index	AtoniaIndex	[103, 104]
	Quantified motor activity	Motor_Activity_Amp	[105]
		Motor_Activity_Baseline	
		Motor_Activity_Dur	
		Motor_Activity_Thresh	
	Percentile (75th)	Per75	[86]
	Variance	Variance	
	Amplitude - Standard Deviation	EMG_std	
EMG - Frequency Domain	Fractal exponent	fractal_exponent	[111]
	Absolute power	Abs_Power	
	Relative spectral power ($\alpha, \beta, and \gamma$)	RSP_alpha	
		RSP_beta	
		RSP_gamma	
	Spectral Edge Frequency	SpectralEdgeFrequency	[86]
EEG & EOG	Cross correlation	EEGEOG_CrossCorrelation	[79]
		EEGEOG_CrossCorrelationPVal	
	Magnitude squared coherence	MSCoherence	
Miscellaneous	Hours recorded	HoursRec	
	Hours from end	HoursFromEnd	

Table 3.3: EEG, EOG, and EMG feature summary

Overall, the automated sleep stage classifier performed well, achieving a multi-stage agreement score of 0.63 for the combined dataset, considered a substantial agreement [23]. When analysed individually, the HC cohort attained a substantial agreement score of 0.73, compared to the moderate score of 0.55 for the RBD cohort. These results are reflected in the confusion matrices in Figure 3.6, 3.7, and 3.8. An example of the predicted sleep stages compared to manual annotations are also detailed in Figure 3.9 and 3.10 for an HC and RBD participant, respectively. For comparison a pair of relatively poor performing results are given in Figures 3.11 and 3.12 for an HC and RBD participants, respectively.

Across all sleep stages there were greater rates of misclassification in participants with RBD compared to HCs. In total 21% of annotated N2 epochs for RBD individuals were misclassified, compared to 10.2% in HCs. Perhaps the greatest differentiator in performance for the RBD cohort was the misclassified REM epochs, which were misclassified as N2 and W by 31.2% and 16.1% (considerably higher than the HC at 8.88% and 1.94%, respectively). This was also reflected in N3, where 22.7% and 34% of annotated N3 states were misclassified for N2 for the HC and RBD cohorts, respectively.

	W	N1	N2	N3	REM
Accuracy	0.92 ± 0.079	0.87 ± 0.061	0.81 ± 0.083	0.94 ± 0.057	0.93 ± 0.047
Sensitivity	0.87 ± 0.16	0.27 ± 0.18	0.84 ± 0.13	0.61 ± 0.30	0.63 ± 0.31
Specificity	0.92 ± 0.097	0.96 ± 0.042	0.79 ± 0.11	0.98 ± 0.038	0.97 ± 0.033
Precision	0.79 ± 0.18	0.48 ± 0.23	0.72 ± 0.15	0.71 ± 0.29	0.77 ± 0.19
F1	0.81 ± 0.16	0.31 ± 0.18	0.76 ± 0.13	0.61 ± 0.28	0.64 ± 0.26

Table 3.4: The macro-averaged automated sleep staging performance (specifying mean and standard deviation), trained and evaluated on the entire dataset using 10-participant-fold cross-validation.

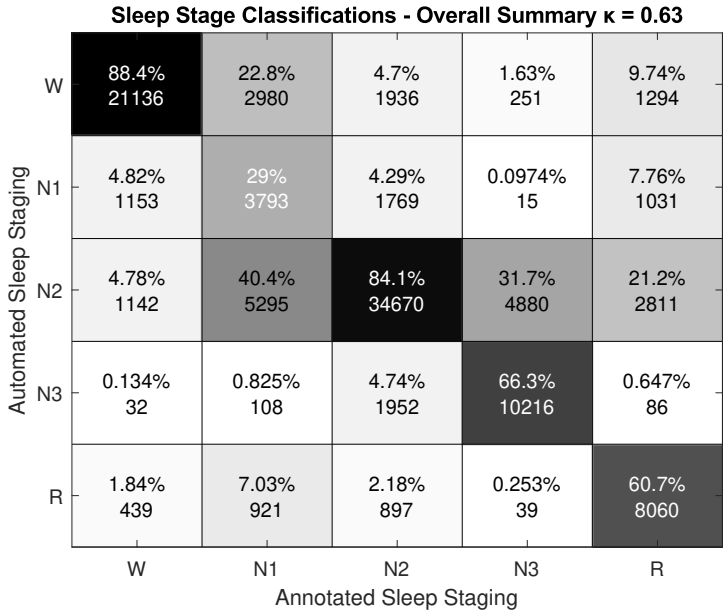


Figure 3.6: The confusion matrix for automated sleep staging, trained and evaluated on the entire dataset using 10-participant-fold cross validation. This technique achieved a five-state Cohen kappa score of 0.63. Each cell details the total number of epochs and the percentage of epochs from the manually annotated sleep stages (columns total 100%).

	W	N1	N2	N3	REM
Accuracy	0.95 ± 0.033	0.87 ± 0.046	0.87 ± 0.043	0.97 ± 0.027	0.96 ± 0.023
Sensitivity	0.90 ± 0.082	0.34 ± 0.16	0.89 ± 0.071	0.57 ± 0.32	0.83 ± 0.18
Specificity	0.96 ± 0.038	0.96 ± 0.042	0.84 ± 0.074	0.98 ± 0.022	0.98 ± 0.019
Precision	0.87 ± 0.11	0.60 ± 0.19	0.81 ± 0.079	0.57 ± 0.32	0.83 ± 0.15
F1	0.88 ± 0.085	0.41 ± 0.15	0.84 ± 0.053	0.53 ± 0.30	0.81 ± 0.16

Table 3.5: The macro-averaged automated sleep staging performance (specifying mean and standard deviation) isolated for HC participants only.

	W	N1	N2	N3	REM
Accuracy	0.88 ± 0.096	0.87 ± 0.073	0.76 ± 0.082	0.91 ± 0.063	0.90 ± 0.048
Sensitivity	0.83 ± 0.20	0.19 ± 0.16	0.79 ± 0.15	0.65 ± 0.27	0.42 ± 0.27
Specificity	0.89 ± 0.12	0.96 ± 0.042	0.75 ± 0.12	0.97 ± 0.048	0.97 ± 0.042
Precision	0.71 ± 0.19	0.36 ± 0.21	0.62 ± 0.15	0.86 ± 0.16	0.71 ± 0.22
F1	0.73 ± 0.18	0.22 ± 0.14	0.68 ± 0.12	0.70 ± 0.23	0.48 ± 0.24

Table 3.6: The macro-averaged automated sleep staging performance (specifying mean and standard deviation) isolated for RBD participants only. There is clear drop in performance when compared to HCs (see Table 3.5), with respect to REM, N2, N1, and W.

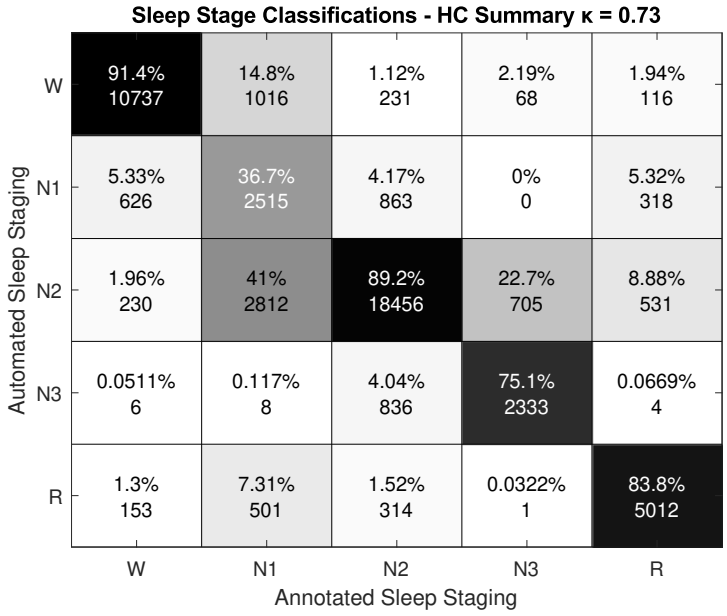


Figure 3.7: The confusion matrix for automated sleep staging isolated to observe the HC participants. For HC participants the five-state Cohen kappa score was 0.73. Each cell details the total number of epochs and the percentage of epochs from the manually annotated sleep stages (columns total 100%).

Robust detection of REM sleep stages is crucial for accurate diagnosis of RBD. Overall, the classifier achieved 0.93 ± 0.047 accuracy, 0.63 ± 0.31 sensitivity, and 0.97 ± 0.033 specificity for classification of REM sleep stages on the combined dataset. There was a substantial difference in the REM detection sensitivity between the healthy and RBD cohort, at 0.83 ± 0.18 and 0.42 ± 0.27, respectively. However the specificity for REM detection remains very high for both cohorts, which will prove pivotal in avoiding false positive RSWA results. A summary of multi-stage sleep classification performance for each cohort are provided in Table 3.4.

For N3 sleep it is interesting to observe the performance difference, where sensitivity within the RBD cohort outperforms HCs with a score of 0.65 ± 0.27 compared to 0.58 ± 0.32. This was also reflected in the F1 score where the RBD and HC cohorts attained a score of 0.70 ± 0.23 and 0.52 ± 0.30, respectively. The overall sensitivity, specificity, and F1 score for N3 classification performance was 0.61 ± 0.30, 0.98 ± 0.038, and 0.61 ± 0.28, respectively. This might be attributed to the greater prevalence of N3 sleep in the RBD cohort.

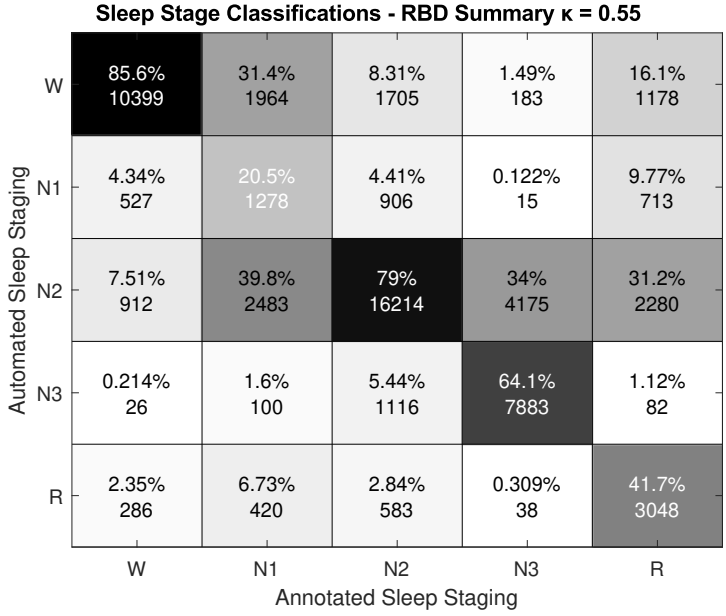


Figure 3.8: The confusion matrix for automated sleep staging isolated to observe the RBD participants. For RBD participants the five-state Cohen kappa score was 0.55. Compared to HCs (see Figure 3.7) the sensitivity of all sleep stages are lower in RBD participants (percentages along the diagonal). Each cell details the total number of epochs and the percentage of epochs from the manually annotated sleep stages (columns total 100%).

N2 is the most prevalent state in sleep and for this stage the HCs outperform the RBD cohort. The sensitivity of N2 classification achieved by the HC and RBD cohort was 0.89 ± 0.071 and 0.79 ± 0.15, respectively. This difference in performance was further demonstrated in the specificity, where the HC and RBD cohort scored 0.84 ± 0.074 and 0.75 ± 0.12. For the overall cohort the sensitivity, specificity, and F1 score was 0.84 ± 0.13, 0.79 ± 0.11, and 0.76 ± 0.13, respectively.

Typically N1 sleep provides the greatest challenge for automated sleep staging and was reflected in this dataset, with HCs outperforming the RBD cohort. The F1 score attained was 0.41 ± 0.15 and 0.22 ± 0.14 for the HC and RBD cohort, respectively. The overall sensitivity, specificity, and F1 score was 0.27 ± 0.18, 0.96 ± 0.042, and 0.31 ± 0.18, respectively.

Compared to all other sleep stages, W classification performs the best with an overall sensitivity, specificity, and F1 score of 0.87 ± 0.16, 0.92 ± 0.097, and 0.81 ± 0.16. For the individual cohorts, the HCs outperform the RBD cohort.

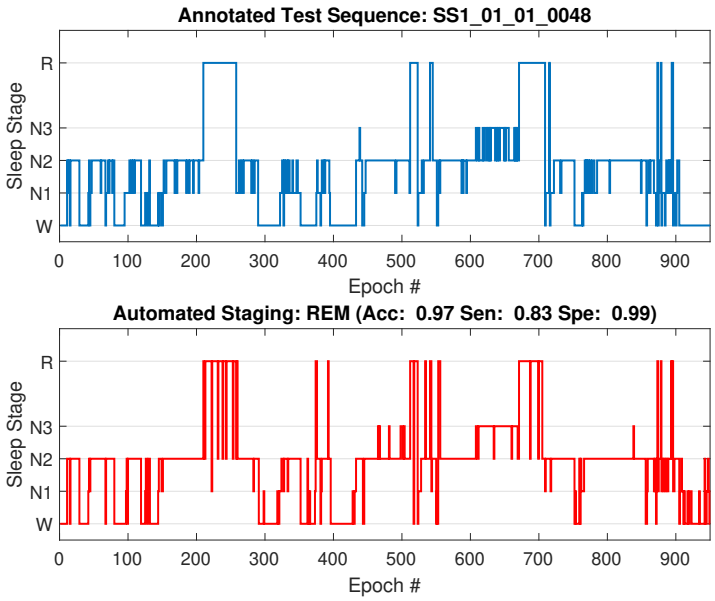


Figure 3.9: An example of a HC participant (SS48) depicting the results of automated sleep staging. The graph on top (blue) details the manually annotated sleep stages, while the graph below (red) details the predicted sleep stages. For this participant the REM sleep classification achieved an accuracy, sensitivity, and specificity of 0.97, 0.83, and 0.99, respectively. Note the erratic nature of REM sleep classification, which fails to maintain REM and often switches between N2 and REM.

As detailed in Section 3.4.8, the RF classifier was able to provide an importance measure for each feature, depicted in Figure 3.13. The top four features included EEG RSP (delta), EEG zero crossing rate, EEG median frequency, and EEG mean frequency. The first 14 features are all EEG based features, which was then followed by EMG entropy and EMG 75^{th} percentile. The first EOG based feature appeared in 19^{th} with EOG auto correlation. The time elapsed feature, hours recorded, appears in 34^{th} position.

It is also interesting to observe the features at the bottom of the list, detailed in Figure 3.14. It would appear the auto-regression coefficients (AR) and the log-power time domain parameters (tdp1) provide the least importance with respect to prediction error. A number of EMG based features also appear at the bottom, namely motor activity percent, atonia index, RSP (gamma), and RSP (beta). The only EOG features to appear unimportant include EOG rise, EOG fall, and EOG peak width.

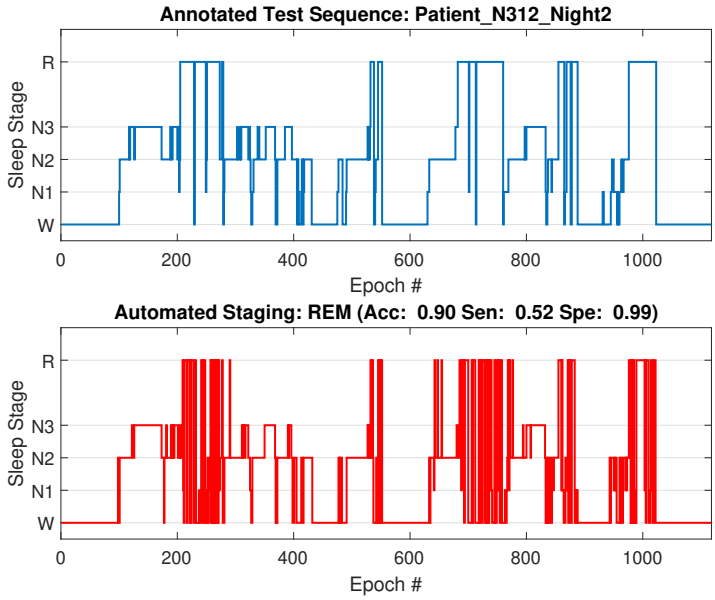


Figure 3.10: A example of automated sleep staging on an RBD participants, where the top graph (blue) details the manually annotated sleep stages and the predicted sleep stages depicted in the graph below (red). The REM classification accuracy, sensitivity, and specificity achieved was 0.97, 0.79, and 0.99, respectively.

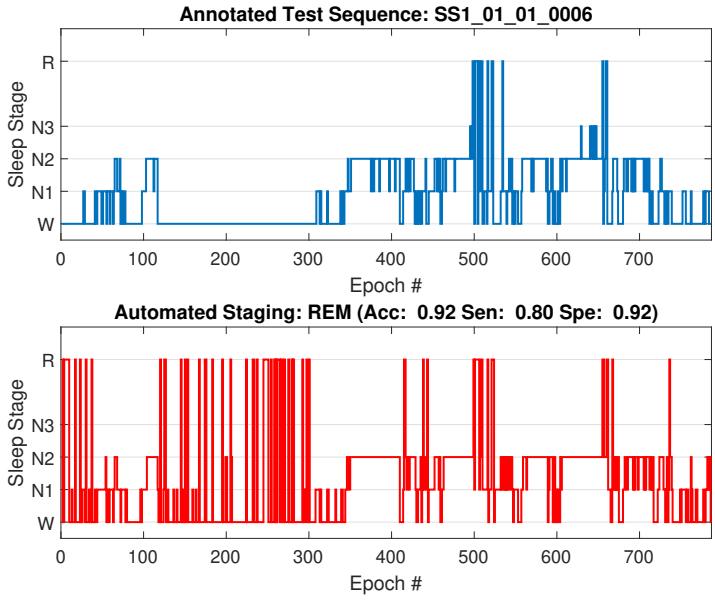


Figure 3.11: An example of a HC participant (SS06) that produced relatively poor automated sleep staging results. The graph on top (blue) details the manually annotated sleep stages, while the graph below (red) details the predicted sleep stages. For this participant the REM sleep classification achieved an accuracy, sensitivity, and specificity of 0.92, 0.80, and 0.92, respectively. Note the high prevalence of W stages and their tendency to be misclassified as REM epochs.

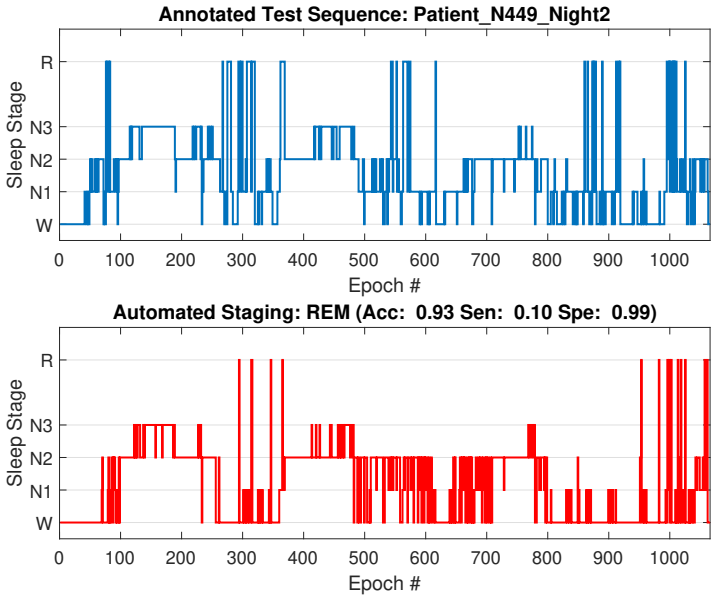


Figure 3.12: An example of poor automated sleep staging for an RBD participants, where the top graph (blue) details the manually annotated sleep stages and the predicted sleep stages depicted in the graph below (red). The REM classification accuracy, sensitivity, and specificity achieved was 0.93, 0.10, and 0.99, respectively. For this participant manually annotated REM epochs are sparse, erratic, and scarce.

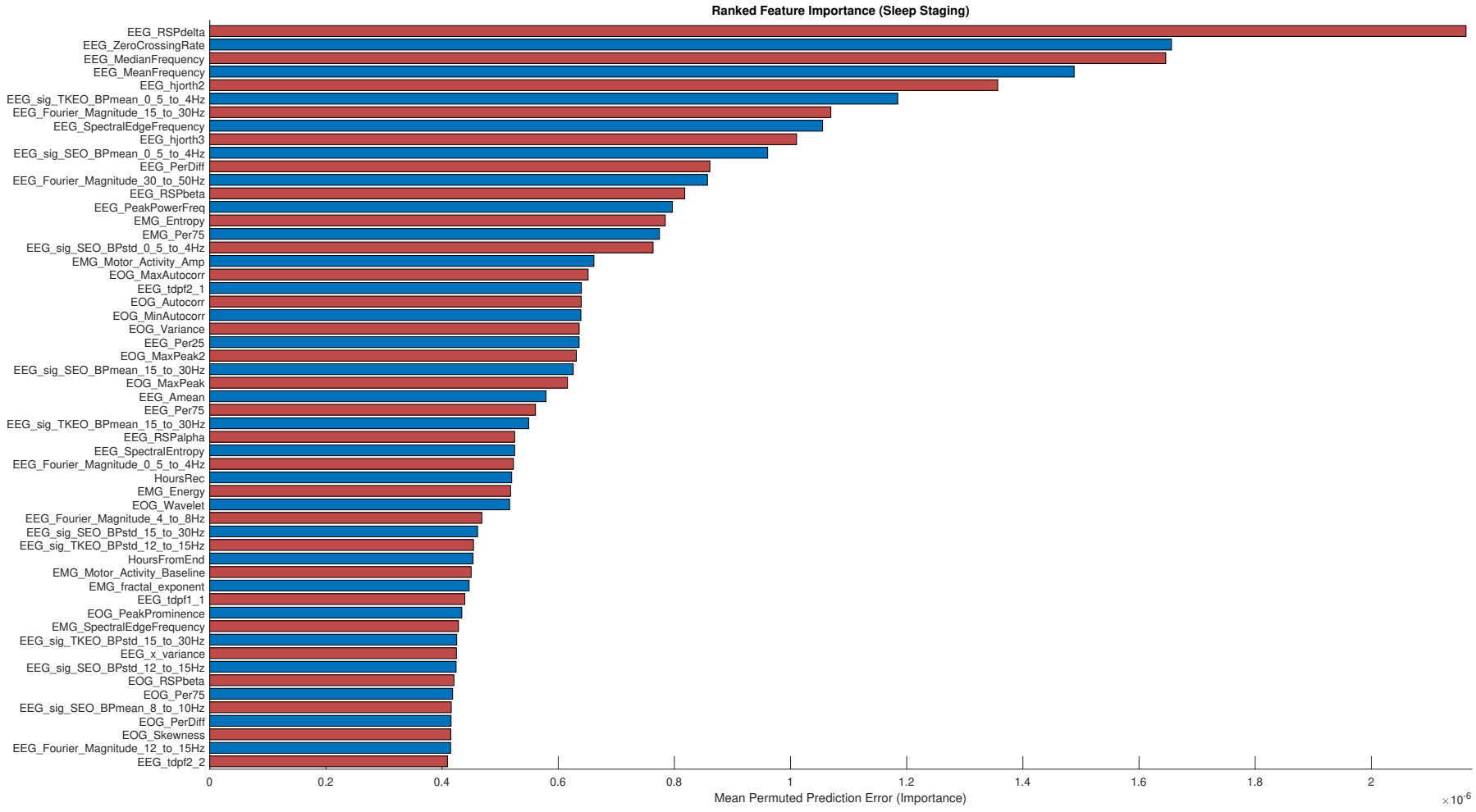


Figure 3.13: From 10-participant-fold cross-validation the mean permuted prediction error from each fold's RF classifier was used to understand feature importance. EEG features dominate the top features, particularly frequency based features. EMG and EOG features are also represented further down the list, along with time elapsed features (hours recorded).

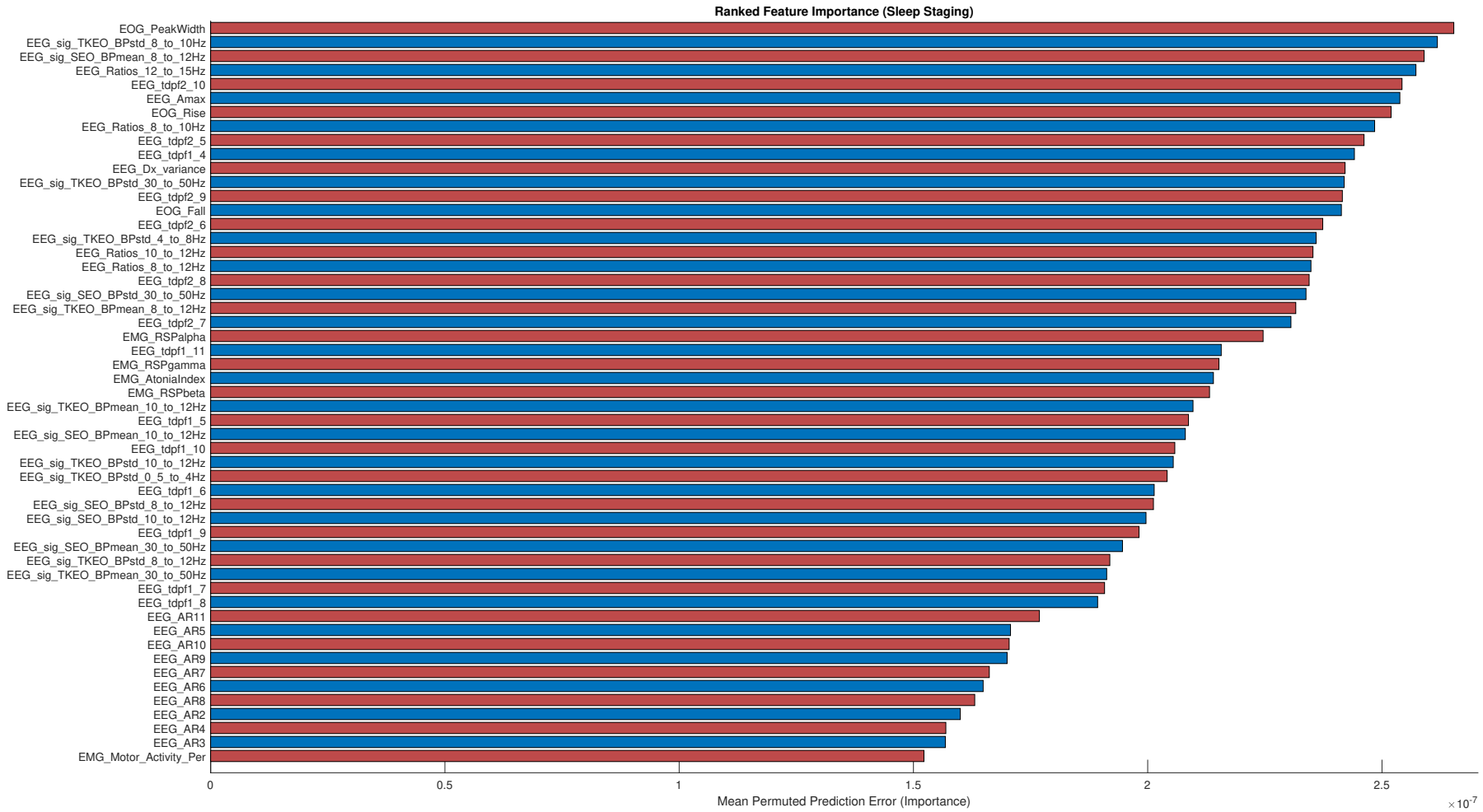


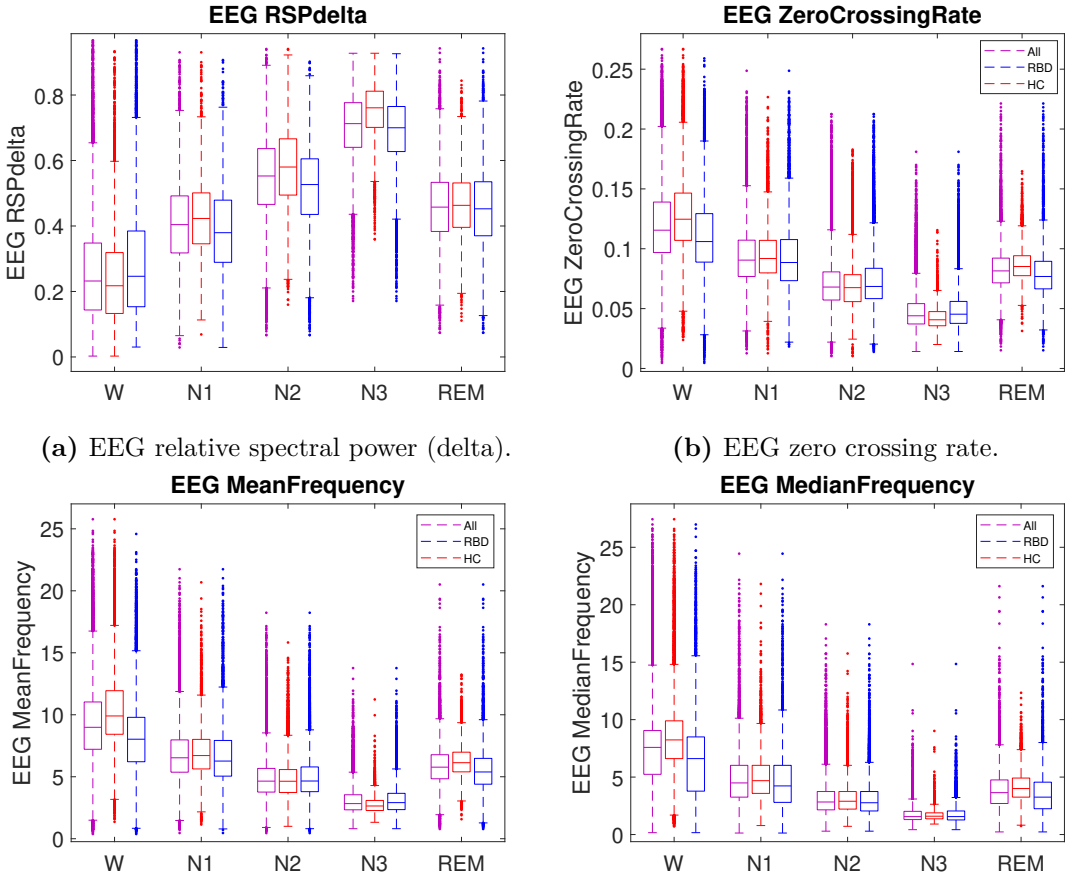
Figure 3.14: The least important features determined by the mean permuted prediction error. The autoregression coefficients and several EMG atonia level features (atonia index and motor activity) appear to produce low importance for automated sleep staging.

3.6 Discussion

The goal of this chapter was to validate an automated sleep stage classifier that would be functional on a combined cohort of HCs and RBD participants. This chapter demonstrated automated multi-stage sleep classification, providing a high REM sleep accuracy when compared to the gold standard of human expert annotation (Table 3.4). Despite a drop in performance for classification of REM stages in RBD individuals, specificity of REM stage detection remained high. This will prove beneficial when evaluating these REM epochs for RBD detection, explored in future chapters.

There is a significant degree of variability in manual sleep staging even with highly experienced scorers, with estimates of human inter-rater variability measured at a Cohen's Kappa of 0.68 to 0.76 [22, 21]. This variability can be further compounded by sleep disorders, where agreement scores can vary from 0.61 to 0.82 for individuals with PD and generalised anxiety disorder, respectively [21]. Our automated classifier achieved a similar benchmark, with a Cohen's Kappa for HCs of 0.73, and for the combined cohort of healthy and RBD participants of 0.63.

Compared to the HCs the drop in sleep staging performance in the RBD cohort was due to a greater rate of misclassification, especially with regards to N2 and REM sleep (depicted in Figures 3.7 and 3.8). Annotated N2 epochs in RBD participants were misclassified for W, N3, and REM by 8.31%, 5.44%, and 2.84%, respectively (compared to 1.12%, 4.04%, and 1.52% in HCs). This misclassification for N2 was due to greater variation in EEG characteristics compared to HCs. This coupled with the inability to exploit diminished levels of atonia contributes to a decrease in automated sleep staging performance. For HCs, 1.94% and 8.88% of manually annotated REM states were misclassified for W and N2, respectively (evident in Figure 3.9). For RBD participants, 16.1% and 31.2% of REM sleep were misclassified as W and N2, respectively (exemplified in Figures 3.10 and 3.12). Once again the stark difference in REM detection can be attributed to the loss of atonia, replicating W states. Additionally, the greater variation in RBD participants, contributes towards greater N2 misclassification (detailed in Figures 3.15, 3.16, and



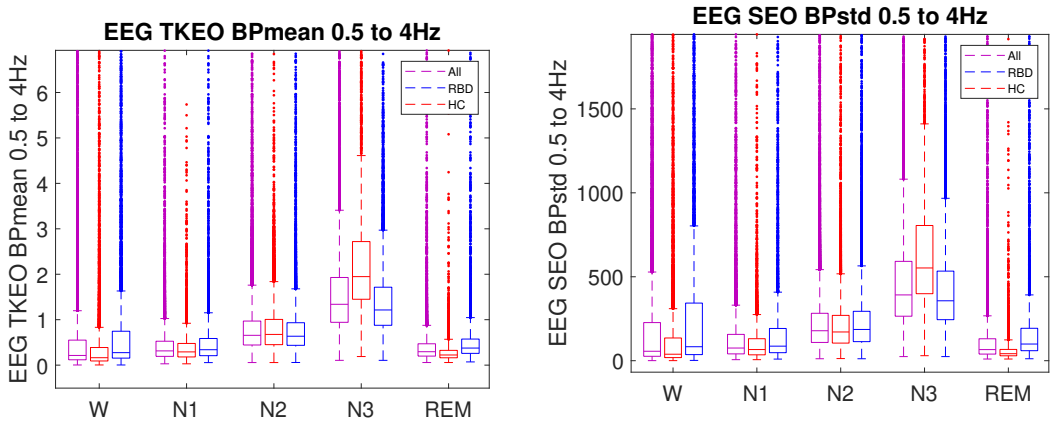
(a) EEG relative spectral power (delta).

(b) EEG zero crossing rate.

(c) EEG mean frequency.

(d) EEG median frequency.

Figure 3.15: These were the top four features as determined by the mean permutation error from each fold of the cross validation scheme. The RSP (delta) feature (3.15a) was able to clearly distinguish between W and N3, where delta waves are absent and strongly prevalent, respectively. The features shown in 3.15b, 3.15c, and 3.15d, capture the frequency properties of specific sleep stages and also display similar distributions. Frequencies during W are highest, followed by N1, N2, and REM which contain mixed frequencies. N3 sleep contains the lowest frequencies. For these top features it is also clear that HCs and RBD participants share the same distribution. However in all cases we can observe that RBD participants (blue) tend to have a wider standard deviation compared to HCs (red)



(a) EEG TKEO band pass mean 0.5 to 4Hz.

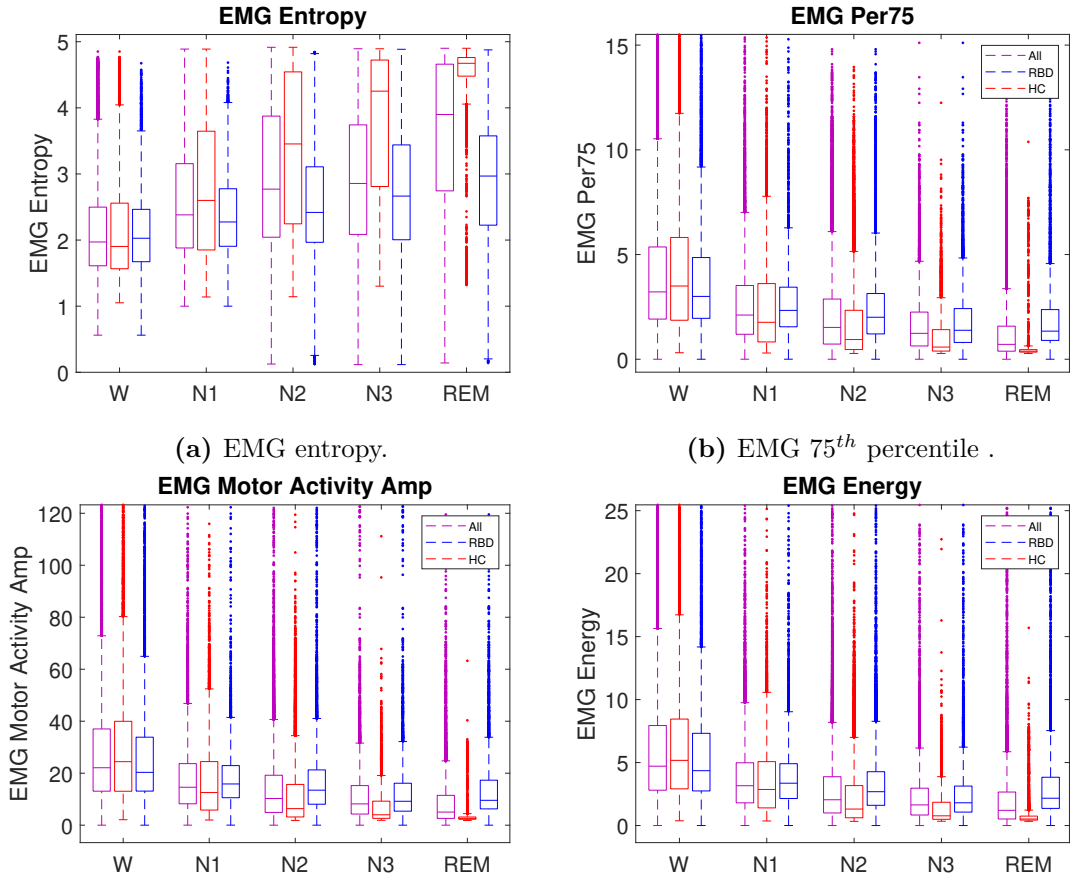
(b) EEG SEO band pass standard deviation 0.5 to 4Hz.

Figure 3.16: EEG Delta frequency features are quite prevalent in the top features based on the feature importance shown in Figure 3.13. The EEG TKEO delta (3.16a) and EEG SEO delta (3.16b) features are two examples. It is clear to see the distinguishing ability of these features in the high values within N3 sleep and the low values associated with W, N1, and REM. For these features we can also observe the distinct difference between HCs and RBD participants, especially during N3 and REM sleep. RBD participants (blue) exhibit higher values during REM, while having lower values during N3, when compared to HCs.

3.17). The increased misclassification of annotated W epochs in the RBD cohort, compared to HCs, can partly be linked to the greater prevalence of annotated W epochs. This is because individuals with RBD tend to have more interrupted and erratic sleep patterns. Furthermore the EMG signal helps distinguish REM from W in HCs, but this attribute is often not helpful in the context of RBD, where there can be an absence of atonia in REM.

Critical to RBD diagnosis is the identification of REM sleep, and while other studies in automated sleep staging produce better results in REM sleep detection, they benefit from primarily focusing on young HCs or a relatively smaller sample size [66, 57, 69, 54, 55, 65, 60, 58, 78, 81, 62, 63, 79]. Despite the lack of sensitivity, REM specificity remains high, which means that as long as REM is identified with a certain precision, the quantified absence of atonia will ideally remain indicative.

As a fairer direct comparison, previous literature that explored automated sleep staging in combined cohorts with RBD participants achieved REM detection sensitivity and specificity ranged from 0.88 to 0.94 and 0.96 to 0.97, respectively



(a) EMG entropy.

(b) EMG 75^{th} percentile .

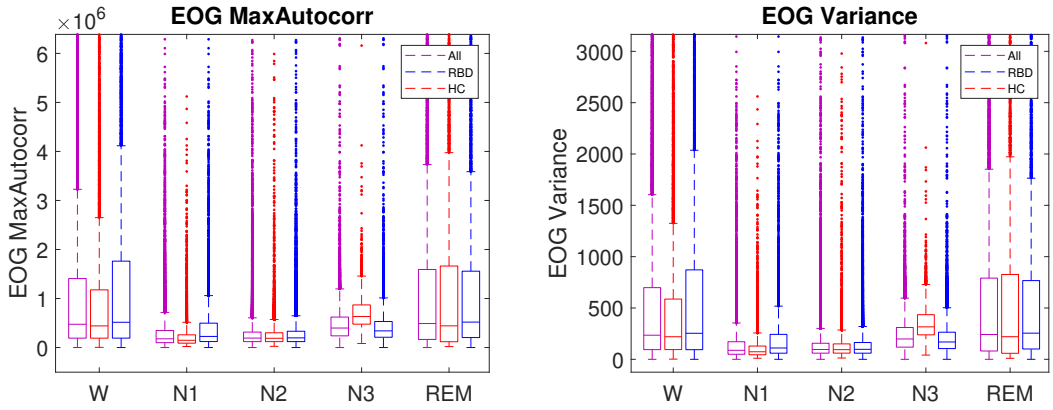
(c) EMG motor activity amplitude.

(d) EMG energy.

Figure 3.17: These were the top EMG features derived from the feature importance, depicted in Figure 3.13. For each of these features we can observe a clear difference between HCs (red) and RBD (blue) participants during REM sleep. For EMG entropy (3.17a), RBD participants provide lower values compared to HCs and this is apparent in all sleep stages (except W), but its most pronounced during REM. The distribution of EMG 75^{th} percentile 3.17b, EMG motor activity amplitude 3.17c, and EMG energy 3.17d are similar and all depict RBD participants with greater variation and higher values during N2, N3, and REM (compared to HCs).

[54, 55] (see Table 3.1). While this chapter did not outperform these values, the results of 0.62 ± 0.31 sensitivity and 0.97 ± 0.033 specificity is comparable. The noticeable difference in sensitivity might be explained by the greater number of participants included in this study which provided much greater feature variability contributing to REM misclassification.

From the feature importance it is clear to see the importance of EEG based features in automated sleep staging. The top four features focus on frequency based properties that are indicative of specific sleep stages. This includes the zero crossing



(a) EOG maximum autocorrelation. **(b)** EOG variance.

Figure 3.18: EOG maximum autocorrelation 3.18a and EOG variance 3.18b were the top EOG features to be represented in Figure 3.13. It is clear to see eye movement associated with W and REM provide high values for both these features, while SEM also produce slightly elevated values during N3. Whereas the complete absence of eye movement during N1 and N2 help differentiate these epochs from W, N3, and REM.

rate, which does not require a transformation into the frequency domain, but still evaluates the rate at which the signal crosses the zero line. As described in Chapter 1 Section 1.2.2, sleep stages are strongly associated with frequencies and were reflected in the feature importance of this dataset (see Figure 3.15). Low frequencies (delta) are associated with N3 sleep, while alpha frequencies are prevalent during wake, and mixed frequencies are found in REM and N2 sleep.

The delta frequency range appears numerous times in the list of top features, represented by EEG RSP delta, EEG TKEO mean (band pass 0.5 to 4Hz), EEG SEO mean (band pass $0.5-4$Hz), and EEG SEO standard deviation (band pass $0.5-4$Hz). These frequency ranges are most useful in distinguishing N3 sleep (3.16a).

EMG based features are not heavily represented in the top features, except for EMG entropy, EMG 75^{th} percentile, and EMG motor activity amplitude. EMG levels are used to distinguish between healthy REM sleep from others, but is also helpful in differentiating N2 and N3 sleep as depicted in Figure 3.17.

The elapsed time feature (hours recorded and hours from end) was also present in the top 50. These features provide insight into sleep staging because certain sleep stages are more likely to occur towards the beginning or end of sleep. For example, participants are expected to be awake at the beginning and end of a sleep

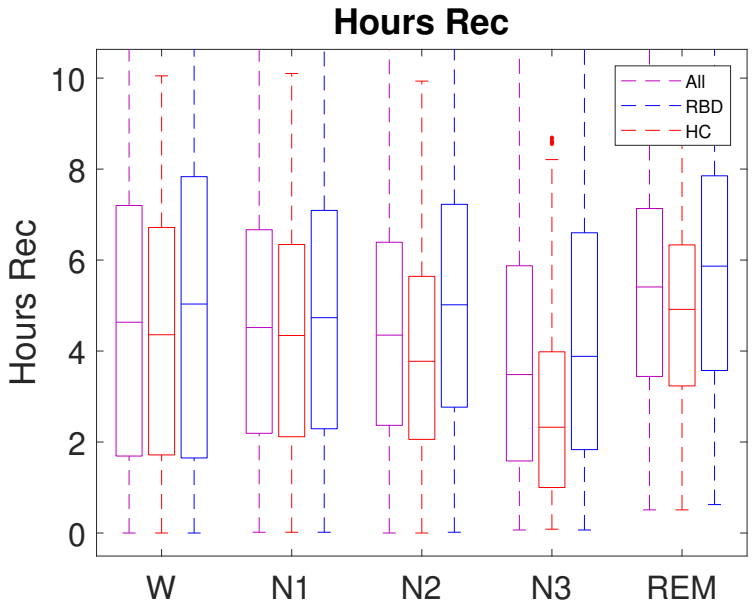


Figure 3.19: Hours recorded was also a high ranking feature and from this distribution we can see the discriminating ability to identify REM and N3 sleep. REM sleep appears to occur later in the night, while N3 sleep seems to occur earlier in the night. There also appears to be a distinct difference between HC (red) and RBD (blue) participants with respect to the occurrence of N3 sleep. It would seem HCs in this dataset tend to have instance of N3 sleep earlier in the night, when compared to RBD participants.

recording, while REM sleep is more prevalent in the latter half of sleep [1]. The start time of these PSG annotations are determined by a sleep technician that arbitrarily starts annotating sleep stages from a point they deem appropriate or roughly around the time of lights off. As a result the automated sleep staging may benefit from this feature, which was inherently derived from a clinical specialist and therefore might not be directly applicable to a fully automated take-home screening device. However, there is the near equivalent arbitrary starting point determined by the sleep participant at around the time of lights off. Participants would often provide a signature to ensure synchronisation between recording devices or simply to indicate lights off, which could provide an adequate substitute.

It is apparent by the prevalence of time domain parameters and autoregression coefficients that obtain low feature importance, that they are often made redundant or simply do not aid in automated sleep staging when compared to other features. Numerous EMG and EOG features also produce low importance values. A few of

the EMG features are those used to identify RBD participants (atonia index and motor activity percent). While they do not provide great assistance in automated sleep staging, they may prove useful in RBD detection, explored in later chapters.

This chapter focused on feature-based classifiers, but recent literature has demonstrated that deep learning techniques are producing state-of-the-art results [113, 114, 82, 115, 116, 46]. While these and other classifiers exist, this work chose to focus on a robust feature based classifier that provided options to interrogate the internal attributes so as to analyse the nature of the PSG dataset. Nonetheless, future work would look to employ deep learning techniques on RBD participants.

Although this study included a reasonably large number of RBD PSG recordings, further work is required in order to define the potential role of automated sleep staging in the clinical environment. This study was limited by simply focusing on RBD and HC participants, where actual clinical practise that would incorporate a mixed population of sleep disorders, which would introduce additional confounds and variations. This study is also limited in that sex ratios were not matched in the study cohorts and evidence suggests there are physiological differences in sleep between sexes [24]. Future extensions of this work would look to apply improved automated sleep staging algorithms, while including a range of common sleep disorders, and also incorporating other PSG and physiological signals that could better facilitate take-home sleep recording devices.

3.7 Conclusion

The work presented in this chapter explored automated sleep staging methods to be applied on a combined cohort of elderly HCs and RBD participants. The primary contributions of this chapter were:

- An exploration of features for automated sleep staging using EEG, EOG, and EMG signals.

- A rigorous evaluation of a random forest classifier for automated sleep staging on age-matched HCs and RBD participants (one of the largest published RBD datasets), achieving a level of accuracy comparable to manual annotation.
- Identifying the complexities and difficulties that RBD participants introduce to the problem of automated sleep staging.

While the performance of automated sleep staging presented in this chapter did not match those from studies that simply evaluated HCs, they were comparable. Once more this chapter demonstrated the feasibility of automated sleep staging on a combined cohort, while detailing the difficulties of greater variation introduced by RBD participants compared to HCs. Furthermore, by using a limited PSG montage, this study replicated the limitations associated with take-home PSG devices.

Golden slumbers
Fill your eyes
Smiles await you when you rise
Sleep pretty darling
Do not cry
And I will sing a lullaby

— Lennon & McCartney's *Golden Slumbers*

4

Automated Detection of REM Sleep Behaviour Disorder

Contents

4.1 Associated Publication 93
4.2 Introduction 94
4.3 Background 95
4.3.1 Quantification of Abnormal Muscle Movement 95
4.3.2 RBD Detection 97
4.4 Methods 99
4.4.1 Data 99
4.4.2 Pre-processing 99
4.4.3 Extract Abnormal Muscle Movement Metrics 99
4.4.4 RBD Classification 102
4.4.5 Evaluation 102
4.5 Results 103
4.6 Discussion 107
4.7 Conclusion 112

4.1 Associated Publication

- N. Cooray, F. Anderotti, C. Lo, M. Symmonds, M. T. Hu, and M. De Vos. "Automating the Detection of REM Sleep Behaviour Disorder". In: *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2018, pp. 1460–3

- N. Cooray, F. Andreotti, C. Lo, M. Symmonds, and M. T. M. Hu. "Detection of REM sleep behaviour disorder by automated polysomnography analysis". In: *Clinical Neurophysiology* 130.4 (2019), pp. 505–514

4.2 Introduction

The International Classification of Sleep Disorders (ICSD-3) [37] specifies that the diagnosis for RBD requires the lack of atonia during REM to be observed. Once REM has been identified, clinicians must visually evaluate the muscle movement captured by the EMG sensor to determine the absence of atonia. This presents a major challenge to clinicians, that an exact definition for healthy muscle atonia is lacking in literature, as little is known about what objectively constitutes muscle movement in an EMG signal. Many attempts have been made to define and quantify muscle activity during REM in order to differentiate healthy and abnormal movement [102, 105, 117, 55, 54, 118, 119, 84, 120, 104, 103, 121, 122]. An objective metric would enable clinical experts to develop a standardised pipeline that could be employed globally for the evaluation of REM sleep without atonia (RSWA) in EMG signals, in a similar way the metric of heart rate was used for ECG signals in terms of cardiac activity. Furthermore, an objective quantification of muscle atonia during REM would provide a convenient metric to facilitate an automated decision support tool to identify RBD. Designing a unified decision support-tool provides a novel and useful scenario to integrate signal processing and machine learning techniques to provide state-of-the-art performance in RBD detection. This chapter details the final step in devising a fully automated RBD detection support tool. Using the collated data, described in Chapter 2, muscle atonia during REM was evaluated using established techniques in literature. New metrics are presented and compared to established techniques in order to assess their benefits and limitations. These metrics include sleep architecture ratios and the change in atonia index and fractal exponent values between sleep stages. Finally, the feasibility of classification using these metrics are explored with the intention to enable an RBD screening support-tool.

The principle contributions of this chapter to RBD detection are:

- The development of new metrics compared to established techniques to identify RBD. Particularly the use of sleep architecture and fractal exponents to distinguish RBD.
- New classification strategies are evaluated using established and new metrics (described in this chapter) to assess RBD detection performance.
- Evaluating the efficacy of metrics to classify RBD and their feasibility in the context of a take-home screening tool.

4.3 Background

Standard EMG electrodes are placed on the chin submentalis to measure muscle activity. These electrodes detect electrical activity from muscle neurons, which are signalling muscles to contract. Clinical experts evaluate the observed signals as qualitative estimates of overall muscle activity [1]. EMG provides crucial insight into sleep stage transitions as detailed in the AASM, and is critical in defining several sleep disorders including RBD, as per the ICSD [1, 37].

4.3.1 Quantification of Abnormal Muscle Movement

Since the formal definition of RBD in 1986, the first study to specify a visual technique to quantify motor events during REM was called the RBD PSG scoring method, developed by Lapierre and Monplaisir [123]. Since inception this method has further been developed to include specific details on thresholds and their performance on RBD identification [124]. This technique aims to distinguish EMG signals into tonic and phasic movement. Tonic muscle activity is defined as long-lasting activity accompanied by complex movement, resulting in epochs with abnormally elevated background activity, while phasic activity describes motor activity associated with twitching depicted by EMG bursts. Monplaisir *et al.* (2010) specified tonic epochs as those $20s$ periods where the amplitude of the chin EMG is at least twice that of the background or greater than $10\mu V$ [124]. Scoring epochs

as tonic or atonic enabled the calculation of EMG tonic density. While phasic chin EMG density was defined as the percentage of $2s$ mini-epochs containing EMG events lasting 0.1 to 10 seconds, with an amplitude exceeding four times the amplitude of background EMG activity [124].

Monplaisir *et al.* (2010) studied 80 RBD participants and 80 age and gender matched controls to determine the cut-offs that achieved the best performance with respect to RBD identification [124]. Achieving RBD detection accuracy of 85.6% (sensitivity and specificity of 88.9% and 82.5%, respectively), with tonic REM EMG density $\geq 30\%$ or when phasic REM EMG density $\geq 15\%$. The RBD PSG scoring method has been used in several studies and has been shown to perform well in discriminating RBD from healthy participants [124, 123, 125]. However, this technique still demands visual analysis to identify EMG events and differentiate tonic and phasic movement. As a result this methodology is susceptible to subjectivity (from the rater), potentially yielding varying classification performances due to inter-rater variability. This was overcome with the digitisation and computerisation of PSG, enabling the development of numerous metrics, including the supra-threshold REM EMG activity metric (STREAM), REM atonia index, and the Quantified Motor Activity (QMA).

The Supra-Threshold REM EMG Activity Metric (STREAM) uses EMG variance to distinguish RBD participants from HCs [102]. STREAM is defined as the percentage of REM epochs, which exceed a threshold of four times the 5^{th} percentile of variance during NREM. This technique identified potential RBD with 100% sensitivity and 71% specificity in 23 participants (17 with neurological disorders and six controls) [102].

The atonia index provides a metric to measure the level of muscle atonia during any stage of sleep [103]. The index is calculated using the average rectified EMG amplitude, comparing the ratio of values $\leq 1\mu V$ to all values (excluding values $> 1\mu V and \leq 2\mu V$). To overcome background noise, a correction value is subtracted from the averaged rectified EMG amplitude, using a moving minimum value over a 60 second sliding window.

Ferri *et al.* (2010) further improved the atonia index by accounting for background noise [104]. This involved using a moving 60 second window to determine the minimum rectified EMG amplitude, which was interpreted as the background noise. This background noise was subtracted from the rectified EMG signal and then, as described above, the atonia index was calculated. This score is known as the corrected atonia index. A corrected atonia index greater than 0.9 achieved RBD detection sensitivity and specificity of 74.3% and 91.4%, respectively [104].

One factor that was preventing a perfect performance was the variability of RBD episodes, which does not occur in every instance of REM sleep. Another study by Ferri *et al.* (2013) documented the night-to-night variability of the atonia index in RBD participants to be under 20% [120]. Furthermore, this metric has been used to investigate the effects of Clonazepam as a treatment for RBD [121, 122].

The QMA metric, developed by Frandsen *et al.* (2015), used a sliding window and a threshold to derive features that captured motor activity events [105]. These events measured motor activity events in terms of total duration and their percentage of REM mini-epochs (relative to total REM sleep). This study achieved RBD detection with an area-under-the-curve (AUC) performance of 0.81. This collection of EMG based metrics provided an opportunity for machine learning based techniques to be explored further.

4.3.2 RBD Detection

Only a small number of studies have investigated machine learning techniques to achieve automated RBD detection [119, 84, 117]. The first study to describe RBD classification was a conference paper by Kempfner *et al.* 2010 [119]. Using manually annotated sleep stages (AASM) and a single chin EMG channel, four features were derived (kurtosis, percentile ratio, log-power, and median frequency) for RBD classification. Through Principle Component Analysis (PCA), features were reduced to two components, which in-turn were used to train a classifier based on quadratic discriminant analysis. This study achieved 100% sensitivity and 100%

specificity with an area-under-the-curve (AUC) of 1.0, although it only had a small cohort of six HCs and six patients diagnosed with RBD.

Kempfner *et al.* (2013) described another algorithm that measured and detected abnormal muscle activity during REM sleep [84]. This technique required three EMG channels and from each channel a single feature was calculated by comparing the mean envelope of a mini-epoch to the minimum envelope of the entire epoch. A one-class support vector machine (OC-SVM) was then used to classify the epochs as an inlier or outlier. This detection algorithm was able to distinguish all PD (12/12) participants, Periodic Limb Movement (PLM) participants (12/12) and HCs (12/12) from iRBD participants (12/12) obtaining an AUC of 1.0.

The follow-up study from Kempfner *et al.* (2014) combined RBD detection technique with automated REM classification [117]. With a collection of 48 PSG recordings (12 HC, RBD, and PLM participants each), a binary-class support vector machine (BC-SVM) was trained using a single EOG channel (eight features) and three EMG channels (five feature each). The classifier simply detected REM mini-epochs (three seconds long) versus everything else (NREM and W), and was post-processed with a normalised Blackman window. Subsequently, an outlier classifier was trained on REM sleep from HC participants using three EMG channels (submentalis, left tibialis, and right tibialis). The percentage of outlier mini-epochs compared to all mini-epochs are used as a muscle score to identify RBD participants. The algorithm was trained using a training set (36 participants) and achieved a REM detection sensitivity of 0.96 and specificity of 0.93 on the validation set (12 participants). Using annotated and automated REM mini-epochs, it produced a correlated muscle score of 0.99. RBD classification achieved a sensitivity of 0.92 and specificity of 0.89, with an area-under-the-curve of 0.98.

These aforementioned techniques are the current established quantitative metrics to identify RBD participants. The results presented above are used as a benchmark to compare the introduction of new metrics for RBD classification using a random forest (RF) classifier.

4.4 Methods

4.4.1 Data

The combined dataset described in Section 2.6 was used to train and test RBD detection techniques. The dataset consisted of 50 RBD participants and 50 age-matched HCs. The gold-standard used for this chapter was the clinical diagnosis of RBD, confirmed by clinical experts that provided these PSG recordings.

4.4.2 Pre-processing

Conventional filtering techniques are required in order to prepare the EMG signal for feature extraction and suppress noise. As detailed previously (see Chapter 2), to ensure consistency between various recordings and datasets, all signals were resampled at 200Hz. EMG signals are understood to contain information between $10-100$Hz [1], therefore satisfying the Nyquist-Shannon sampling criterion [126]. As detailed in 3 Section 3.4.2, EMG signals undergo notch filtering at 50Hz and 60Hz (500^{th} order notch filter) and high pass filtering with a 500^{th} order FIR filter at a 10Hz cut-off.

4.4.3 Extract Abnormal Muscle Movement Metrics

The aim of quantifying abnormal muscle movement is to capture discriminating metrics to identify RBD. These metrics will focus on EMG features and incorporate sleep architecture that are inherent in the definition of RBD provided by the AASM. The metrics are derived using time and frequency based features. This work utilised metrics described in Section 4.3.1 to take advantage of their documented merit [104, 102, 105] and compare to newly devised metrics.

Established Metrics

The aforementioned established metrics include the corrected atonia index (Section 3.4.6), quantified motor activity (3.4.6), and the super-threshold REM EMG activity metrics (STREAM, defined below).

Super-Threshold REM EMG Activity Metric (STREAM)

STREAM requires the variance of the EMG signal to be calculated during REM and NREM sleep for every epoch, which was described and defined in Chapter 3 Section 3.4.6. STREAM is then calculated as the percentage of REM sleep epochs with EMG variance greater than four times the 5^{th} percentile of variance during all NREM epochs.

Sleep Architecture Metrics

Individuals diagnosed with RBD have been known to have disrupted sleep architecture and thus features that reflect these attributes were calculated [127, 128]. This includes percentage of N2, N3, and REM sleep relative to all non-wake states (see Equations 4.1-4.3). Sleep efficiency, E_{ff} was also calculated, and is the percentage of all non-wake time compared to all recorded sleep as in Equation 4.4.

$$A_{N2} = \frac{N2_{epochs}}{N1_{epochs} + N2_{epochs} + N3_{epochs} + REM_{epochs}} \tag{4.1}$$

$$A_{N3} = \frac{N3_{epochs}}{N1_{epochs} + N2_{epochs} + N3_{epochs} + REM_{epochs}} \tag{4.2}$$

$$A_{REM} = \frac{REM_{epochs}}{N1_{epochs} + N2_{epochs} + N3_{epochs} + REM_{epochs}} \tag{4.3}$$

$$E_{ff} = \frac{N1_{epochs} + N2_{epochs} + N3_{epochs} + REM_{epochs}}{W_{epochs} + N1_{epochs} + N2_{epochs} + N3_{epochs} + REM_{epochs}} \tag{4.4}$$

where $N1_{epochs}$, $N2_{epochs}$, $N3_{epochs}$, REM_{epochs}, and W_{epochs} are the number of epochs for N1, N2, N3, REM and wake, respectively. A_{N2}, A_{N3}, and A_{REM} denote the N2, N3, and REM ratio.

Atonia Index Ratios

The corrected atonia index has been established as an effective metric to distinguish RBD individuals. However, the possibility of false positives and false negatives might be averted with the use of atonia index ratios that compare the indexes between specific sleep stages. As defined by the AASM, the levels of muscle atonia changes between sleep stages [19]. Therefore, the ratio of the atonia index between REM and N2 as well as REM and N3 can be explored as a feature to distinguish RBD individuals.

$$R_{N2REM} = \frac{C_{N2} + 1}{C_{REM} + 1} \tag{4.5}$$

$$R_{N3REM} = \frac{C_{N3} + 1}{C_{REM} + 1} \tag{4.6}$$

where C_{N2}, C_{N3}, and C_{REM} is the average atonia index during N2, N3, and REM, respectively. R_{N2REM} and R_{N3REM} is the atonia index ratio between N2 and N3 compared to REM. The additional constant found in the denominator avoids the possibility of dividing by zero.

Electromyography Fractal Exponent

The fractal exponent feature was inspired by Krakovská *et al.* (2011), where it was used for automated sleep staging. This feature aims to measure signal complexity by fitting a linear line (α) to a double logarithmic graph of spectral power density, $P(f)$, versus frequency, f [68] (Equation 4.7). The fractal exponent was calculated for every 30 second epoch. For every N2, N3, and REM sleep stage the fractal exponent was averaged. The fractal exponent ratios were then calculated by dividing the average values during N2 and N3 by the average REM value (see Equation 4.8 and 4.9).

$$P(f) \sim f^{-\alpha} \tag{4.7}$$

$$\alpha_{N2REM} = \frac{\alpha_{N2} + 1}{\alpha_{REM} + 1} \tag{4.8}$$

$$\alpha_{N3REM} = \frac{\alpha_{N3} + 1}{\alpha_{REM} + 1} \tag{4.9}$$

where α is the fractal exponent for spectral power density, $P(f)$, with frequencies, f. α_{N2} and α_{REM} is the average fractal exponent during N2 and REM, respectively. α_{N2REM} and α_{N3REM} denote the fractal exponent ratios for N2 and N3 compared to REM. The additional constant found in the denominator avoids the possibility of dividing by zero.

4.4.4 RBD Classification

As described in Section 4.3, existing literature describes RBD classification using simple thresholds and machine learning techniques. This chapter analyses the use of established measures and new metrics that incorporate sleep architecture to identify RBD. Specifically, we explore the use of a random forest (RF) classifier for RBD detection. Results are then compared to established metrics and existing classifiers. The RF algorithm was an ideal candidate because it is essentially parameter-free, robust to outliers, fast to train, and resistant to over-fitting [129]. Furthermore, RF classifiers provide incisive information on feature importance, which detail the effectiveness of specific features to distinguish RBD participants.

4.4.5 Evaluation

The RBD detection classifier was trained and evaluated using 10-participant-fold cross-validation with even split between HC and RBD participants. For the purpose of this study, two RF classifiers were evaluated, one using established EMG metrics (four features), while the other combined features detailing sleep architecture, atonia index ratios, and fractal exponents (11 features in total). These two classifiers will illustrate the impact of using additional EMG features based on sleep architecture to detect RBD. Additionally, feature importance from the trained classifiers (the averaged permuted delta prediction error from each fold) will aim to rank each feature by their ability to distinguish RBD participants.

Classifier Performance

Evaluation of the classifier was done by calculating the RBD detection accuracy, sensitivity, specificity, precision, and F1-score (detailed in Equations 3.49 to 3.53 in Chapter 3 Section 3.4.9).

The Wilcoxon Rank Sum Test

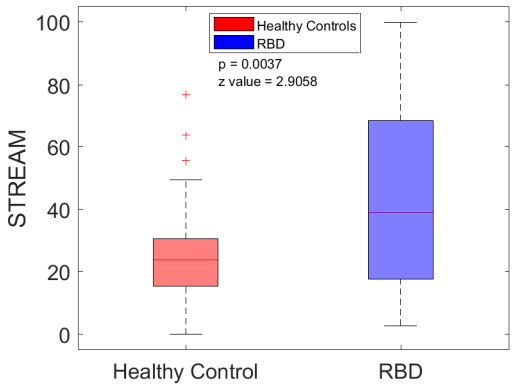
The Wilcoxon rank sum test is a non-parametric method to test for two populations (where the samples are independent), which in this case are RBD participants and HCs [130]. This measure was used to test the significance of metrics to distinguish RBD participants from HCs. The null hypothesis being tested is that the data for RBD participants and HCs are samples from continuous distributions with equal medians. The null hypothesis was rejected when the p-value is less than 0.05, which is at the 5% significance level. The Wilcoxon rank sum test approximates the p-value by calculating the z-statistic. This chapter presents the distribution of each feature depicting the p-value and z-statistic.

4.5 Results

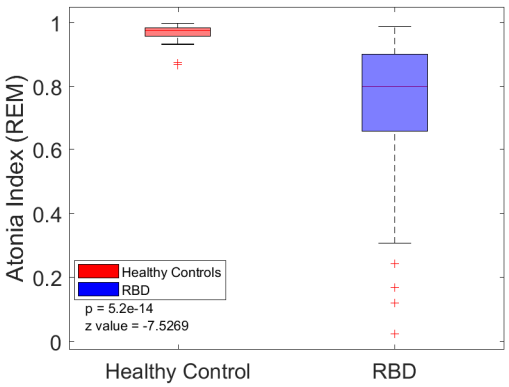
RBD Metrics

For each of the features described in Section 4.4, the feature distribution derived from the collated dataset is detailed for each cohort through a box plot. Figure 4.1 illustrates the distribution of established RBD detection metrics, which reasonably differentiate RBD participants from HCs. STREAM, atonia index (REM), and motor activity detection all prove that RBD participants provide greater variation and a differing mean to HCs.

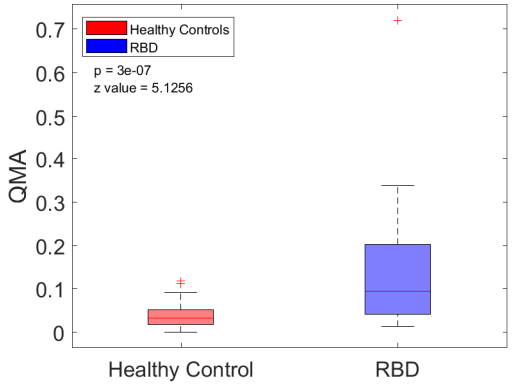
It is also interesting to observe the difference between the RBD and HC participants with respect to sleep architecture, where the distribution for these features are detailed in Figures 4.2 and 4.3. The N2 and N3 ratio metrics proved to be significant, while REM ratio and sleep efficiency did not. RBD participants appear to have greater ratio of N3 sleep and less N2 sleep compared to HCs.



(a) STREAM.



(b) Atonia index (REM).

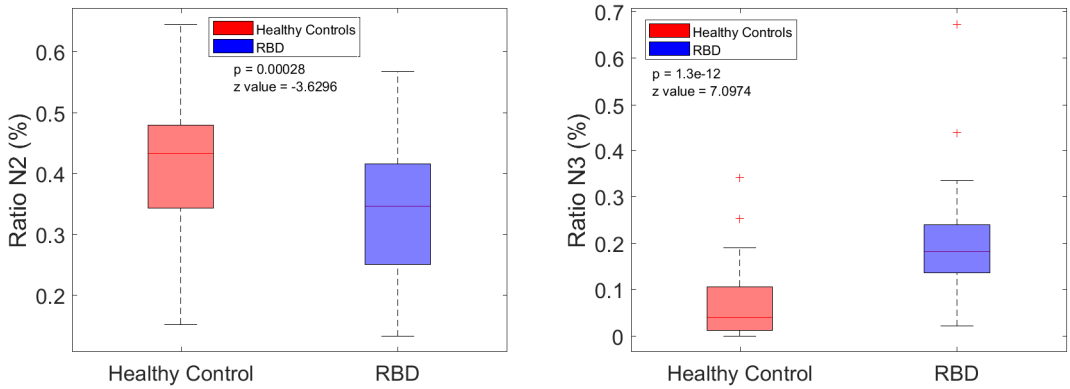


(c) Quantified motor activity.

Figure 4.1: These figures depict the distribution of (a) STREAM, (b) atonia index (REM), and (c) QMA. These three established techniques clearly demonstrate their ability to significantly distinguish RBD participants from HCs. In all cases the variation of RBD participants are greater when compared to HCs.

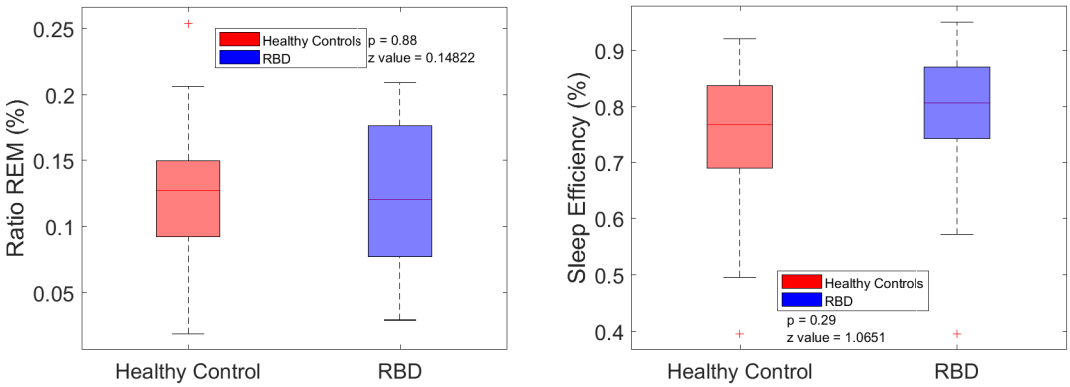
Furthermore, we explored the use of incorporating sleep architecture with the atonia index, where the distributions are shown in Figure 4.4. These metrics also proved to significantly differentiate RBD participants, which have a lower atonia index during REM than both N3 and N2, when compared to HCs.

Additionally the fractal exponent ratio and sleep architecture are combined to distinguish RBD participants. These metrics also proved to significantly differentiate RBD participants from HCs as shown in Figure 4.3. Notably, the fractal exponent during REM is lower when compared to the exponent calculated during N3 and N2.



(a) Ratio of N2 sleep (%). **(b)** Ratio of N3 sleep (%).

Figure 4.2: These figures detail the sleep architecture distribution of the datasets, specifically (a) N2 ratio (%) and (b) N3 ratio (%). These distributions prove to be significantly different, as detailed by their p-value. RBD participants appear to have a greater ratio of N3 sleep at the expense of N2 sleep when compared to HCs.



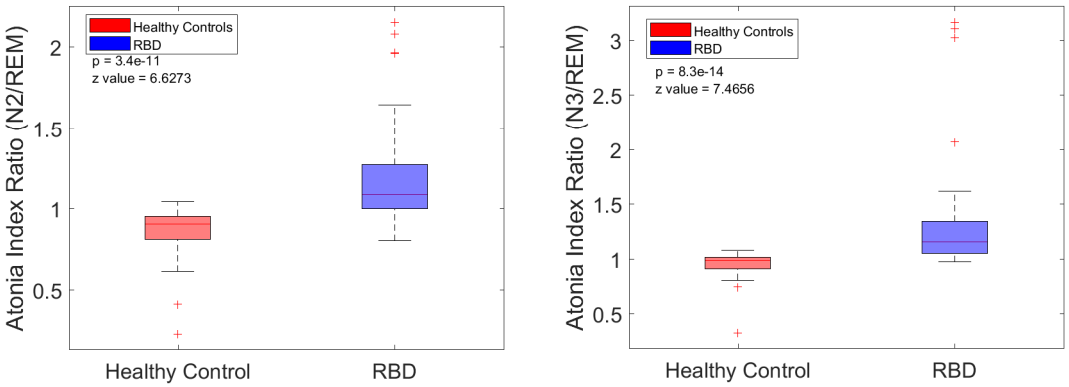
(a) Ratio of REM sleep (%). **(b)** Sleep efficiency (%).

Figure 4.3: These are additional sleep architecture features, (a) REM ratio (%) and (b) sleep efficiency (%). These two metrics don't prove to be significantly different (see p-value).

RBD Detection

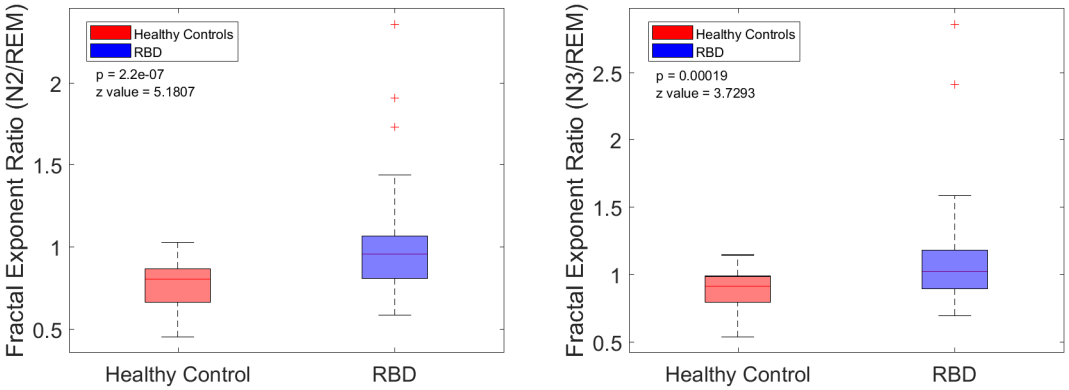
	Accuracy	Sensitivity	Specificity	Precision	F1
STREAM	0.60 ± 0.14	0.64 ± 0.20	0.56 ± 0.20	0.60 ± 0.12	0.61 ± 0.074
Atonia Index	0.85 ± 0.10	0.74 ± 0.20	$\mathbf{0.96 \pm 0.08}$	$\mathbf{0.96 \pm 0.085}$	0.82 ± 0.13
QMA	0.77 ± 0.078	0.66 ± 0.092	0.88 ± 0.13	0.87 ± 0.14	0.74 ± 0.078
Established Metrics (RF)	0.88 ± 0.087	0.86 ± 0.16	0.90 ± 0.13	0.92 ± 0.11	0.87 ± 0.096
Combined Metrics (RF)	$\mathbf{0.94 \pm 0.049}$	$\mathbf{0.94 \pm 0.092}$	0.94 ± 0.092	0.95 ± 0.076	$\mathbf{0.94 \pm 0.05}$

Table 4.1: A summary of RBD classification performance using established metrics and a simple threshold compared with RF classifiers. Performance values in bold detail the best performance, where both trained RF classifiers outperformed established individual metrics with respect to accuracy, sensitivity, and F1 scores.



(a) Atonia index ratio (N2/REM). **(b)** Atonia index ratio (N3/REM).

Figure 4.4: These additional features, (a) atonia index ratio (N2/REM) and (b) atonia index ratio (N3/REM), proved to significantly differentiate RBD participants from HCs.



(a) Fractal exponent ratio (N2/REM). **(b)** Fractal exponent ratio (N3/REM).

Figure 4.5: The feature distribution of (a) fractal exponent ratio (N2/REM) and (b) fractal exponent ratio (N3/REM). These new features proved to distinguish RBD participants from HCs with a significance level of 5%.

In order to evaluate the performance of RBD classification we calculated the accuracy, sensitivity, specificity, precision, and F1 score for each fold of the cross validation. Table 4.1 details the classification performance of established metrics (STREAM, atonia index, and QMA), compared with two RF classifiers trained on established metrics and a combination of new and established features. The best performance was achieved by the RF classifier trained in the combination of established and new metrics with a accuracy, sensitivity, specificity, and F1 score of 0.94 ± 0.049, 0.94 ± 0.092, 0.94 ± 0.092, and 0.94 ± 0.05 respectively.

RBD Feature Importance

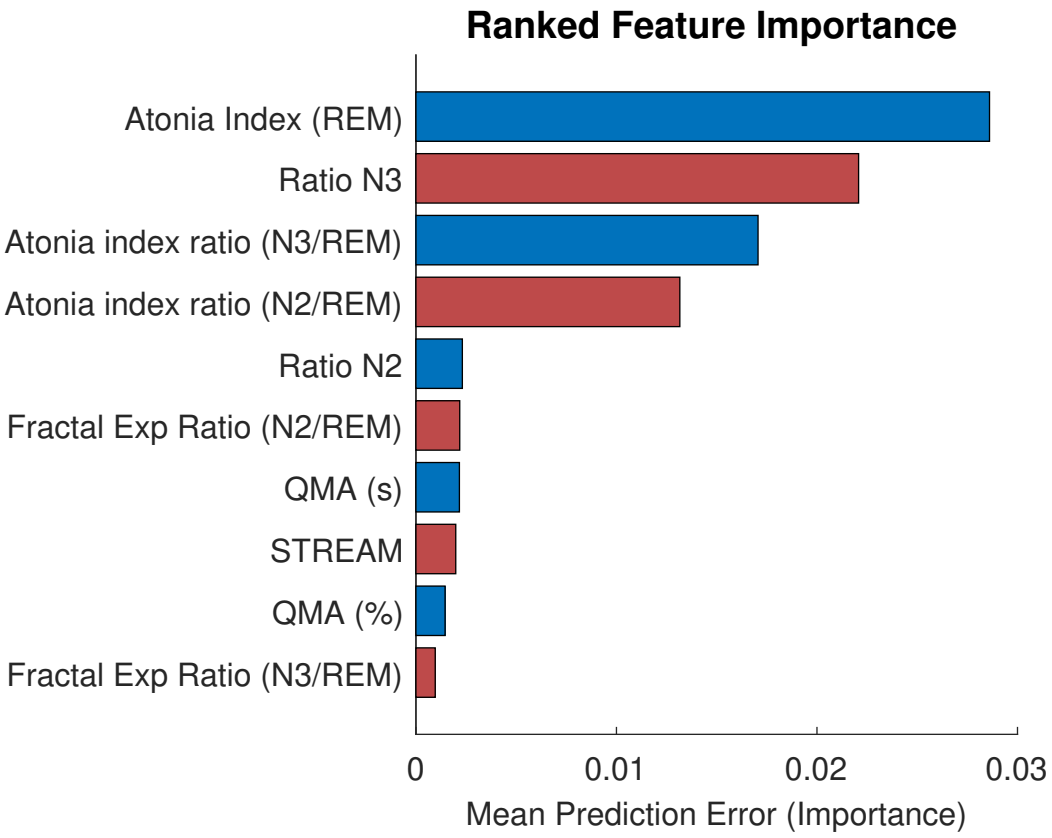


Figure 4.6: The ranked feature importance (mean predicted error) from the RF classifier trained on the combination of established and new metrics.

A ranked list of feature importance was generated using the RF classifier trained on the combination of established and new metrics, illustrated in Figure 4.6. The established atonia index (REM) proved most important, followed by new features that include sleep architecture and relative changes in atonia index and fractal exponent between sleep stages.

4.6 Discussion

The goal of this chapter was to evaluate established techniques on this newly collated set of data of PSG recordings, while subsequently providing a platform to test and validate new metrics to improve upon RBD identification. Lastly the

evaluation of these metrics for RBD classification were validated and compared in order to rank their effectiveness.

Established RBD Metrics

Our results (see Figure 4.1) reaffirms the ability of established metrics to distinguish RBD participants from HCs using the collated dataset for this work [102, 105, 104]. The STREAM, atonia index (REM), and QMA have provided a significant p-value to a level of 5% (as calculated by the Wilcoxon rank sum). Of these established metrics the atonia index (REM) produces the best p-value in terms of discriminating cohorts. For these metrics the variation in RBD participants was greater than HCs, as is expected by the loss to REM muscle atonia. These metrics were able to effectively quantify levels of both tonic and phasic EMG activity during REM, to replicate the visual diagnostic inspection of sleep clinicians. It is also apparent there was a clear overlap between both cohorts, where RBD participants produce scores similar to HC participants and *vice versa*. This was an expected result, because RBD participants display symptoms intermittently and do not exhibit loss of muscle atonia for every instance of REM sleep or for every night of sleep. Conversely, there are instances of HCs that produce scores indicative of RBD participants, this was due to movement occurring during REM coupled with artefacts within the signal.

New Sleep Architecture Metrics

One objective of this chapter was to explore the utility of sleep architecture to distinguish RBD participants. Figures 4.2 to 4.3 detail the feature distribution of sleep efficiency, N2, N3, and REM ratios. The REM ratio and sleep efficiency provided no effective differentiation, however N2 and N3 ratios showed significant disparity with a level of 5%. RBD participants appear to have a greater percentage of N3 sleep compared to HCs, while N2 sleep was more prevalent in HCs. In literature Massicotte-Marquez *et al.* (2005) provided similar evidence that iRBD participants spent more time in slow-wave sleep compared to HC, in addition to increased all-night delta power in NREM sleep from EEG recordings [127]. Therefore validating these metrics as potentially important information to aid in distinguishing

RBD participants from HCs. While the p-value was significant for the ratio of N2, N3, and sleep efficiency there was still a large overlap in scores, as demonstrated in the box plots (see Figure 1.8, 1.9, and 4.2). This suggests that individually these metrics might not be effective in identifying RBD participants but would provide support when used in conjunction with other metrics.

New RBD Metrics

As described in Chapter 1 at Section 1.2.2, typical sleep stages are, in-part, described by the gradual loss of atonia through the various stages. Capturing the relative changes between these stages was the aim of developing the atonia index and fractal exponent ratios. From Figure 4.4 to 4.5, we observed that the atonia index and fractal exponent ratio between N2 and REM was significantly different between RBD participants and HCs, to a level of 5%. These ratio values were higher in RBD participants compared to HCs because the denominator (REM) in Equations 4.5, 4.6, 4.8, and 4.9 was smaller than the numerator (N2 and N3). This can be attributed to symptoms of RBD, which include the loss of atonia during REM (providing smaller values compared to N2 or N3 for both the atonia index and fractal exponent).

The atonia index ratios simply take advantage of the virtues of the REM atonia index and uses it as a denominator to scale the atonia index during N2 and N3 (see Equations 4.5 and 4.6). This has the benefits of averaging the general level of muscle atonia across stages and using it comparatively to get a relative and average change across the entire night of sleep. In HCs the atonia index is closer to 1.0 during REM, the maximum value that can be achieved to represent perfect muscle atonia. While for N2 and N3 these values are less than 1.0 because participants don't exhibit muscle atonia during these stages. As a result for HCs the atonia index ratio for N2 and N3 compared to REM tends to be less than 1.0 However, for RBD participants, the atonia index during REM tend to drop to be much less than 1.0 (a median of 0.80 as illustrated in Figure 4.4). Because the atonia index during REM is lower than the index scored during N2 and N3, the atonia index ratio tends to be greater than 1.0 for RBD participants (see Figure 4.4).

The fractal exponent ratios are similar to the atonia index ratios in that they use the values during REM as a comparative denominator to the values calculated during N2 and N3. During REM, the slope of the power spectrum (log-log graph) tends to be steeper and in-turn larger than during N2 or N3 for HCs, which provide ratio values less than 1.0 (see Figure 4.5). This is a result of muscle atonia during REM, where muscle relaxation reduces the power spectrum, specifically more in the lower frequencies (creating a larger gradient). However, for RBD participants the absence of muscle atonia means the power spectrum slope becomes flatter and in some cases has a smaller gradient than those calculated during N2 or N3, providing ratio values closer to and greater than 1.0 (see Figure 4.5). Fractal exponent ratios are similar to the atonia index ratios, in that they utilise the changes in muscle atonia between sleep stages. However, fractal exponents take advantage of the frequency spectrum instead of being amplitude based.

RBD Classification

This chapter evaluated the classification of RBD participants using a RF classifier trained using a combination of metrics that quantify the loss of atonia and incorporate sleep architecture. These performances were compared to established individual metrics as a reference. STREAM, atonia index, and QMA achieved results that were reflective of the literature [104, 102, 105], where the atonia index provided the best results with a F1 score of 0.82 ± 0.13.

While the atonia index outperformed STREAM and QMA, it was surpassed by both trained RF classifiers, as detailed in Table 4.1. The RF classifier trained only on established metrics was able to exploit the information provided by the atonia index and incorporate the added detail within STREAM and QMA to boost sensitivity. Supplementing these established metrics with additional features (including the atonia index ratios, EMG fractal exponent, and sleep architecture) to train a RF classifier provided the best performance in RBD detection, with respect to the accuracy, sensitivity, and F1 score (0.94 ± 0.049, 0.94 ± 0.092, and 0.94 ± 0.05, respectively). These results align with the state of the art, described in

literature, which achieved 0.96 sensitivity and 0.93 specificity [117]. These results also appear to outperform the corrected atonia index, which in literature achieved 0.734 sensitivity and 0.914 specificity [104].

The atonia index slightly outperforms the combined metrics RF classifier in regards to specificity and precision (by 2% and 1%, respectively). This small reduction in specificity and precision was comparatively small compared to the increase of 20% in sensitivity and 12% in the F1 score. For a classification algorithm designed for a potential screening tool, one could argue for a performance that optimised both sensitivity and precision, enabling the tool to identify a maximised number of people with likely RBD while maintaining a high degree of correct classification. Once more this relatively lower specificity may not be a disadvantage because a highly sensitive classifier might enable the discovery of participants on the cusp of developing a neurological or sleeping disorder.

Therefore, based on these F1 scores, the RF classifier trained on combined metrics would make an ideal candidate for a screening tool. The boost in performance can be attributed to the additional features that incorporate the transitional changes in atonia between sleep stages [1] and the altered sleep architecture that are inherently different between HCs and those with RBD [127]. RBD participants can have a very good night's sleep, with very low muscle movement during REM, indicative of healthy sleep. But by incorporating sleep features describing the percentage of N3 or N2 sleep, the RF has the opportunity to correctly classify these participants based on their unusual sleep architecture.

The importance of these metrics was further substantiated by Figure 4.6, where the ratio of N3 sleep proved to be the second most important feature. We can also observe that ratios of the atonia index, fractal exponent, and N2 sleep proved more important than QMA and STREAM. While the ratio of N3 and N2 sleep was able to help distinguish between HCs and those with RBD, it may prove too general when applied to individuals with various other neurological and confounding sleep disorders (such as insomnia and narcolepsy, which are defined by their fractured and erratic sleep patterns). Nonetheless, this collection of features that provided

the best RBD classification performance, simply requires a single EMG sensor and annotated sleep stages (specifically N2, N3, and REM).

Although this chapter validated these additional metrics, future studies may further validate to what extent these metrics are generalisable through other independent datasets. A further extension would consider incorporating additional recordings of HCs and cohorts with numerous other sleep disorders, which may demand additional differentiating features.

4.7 Conclusion

This chapter outlines the automated classification of RBD using a single EMG sensor and manually annotated sleep staging. The main contributions and findings of this chapter were:

- Based on the collated PSG recordings used in this study (one of the largest collected on RBD participants), established metrics for RBD detection were reaffirmed as significant in identifying RBD participants.
- Additional metrics that incorporated sleep architecture into the atonia index and fractal exponent proved significant (to a level of 5%) in distinguishing RBD participants from HCs.
- The use of a RF classifier trained on established metrics proved to outperform individual metrics with an accuracy, sensitivity, specificity, and F1 score of 0.88 ± 0.087, 0.86 ± 0.16, 0.90 ± 0.13, and 0.87 ± 0.096 respectively.
- The best RBD detection performance was produced by an RF classifier trained on the combination of established and new metrics (incorporating sleep architecture) with an accuracy, sensitivity, specificity, and F1 score of 0.94 ± 0.049, 0.94 ± 0.092, 0.94 ± 0.092, and 0.94 ± 0.05 respectively.
- Ranking feature importance by calculating the mean permutated prediction error, indicated that REM atonia index was the most important, followed by N3 sleep ratio, and atonia index ratios.

The proposed RF classifier that used a combination of established objective RBD metrics and new features based on sleep architecture represents the final step in the detection of RBD participants. However, there is the dependency on manually annotated sleep staging, a laborious and time-consuming task. In the aim of achieving complete automated RBD identification for a potential screening-tool, automated sleep staging from a limited number of sensors would be necessary and is described in the chapters to follow.

Goodnight, my angel
Now it's time to dream
And dream how wonderful your life will be
Someday your child may cry
And if you sing this lullaby
Then in your heart
There will always be a part of me

— Billy Joel's *Lullaby (Goodnight, My Angel)*

5

Automated End-to-End RBD Detection

Contents

5.1 Associated Publication 115
5.2 Introduction 116
5.3 Background 117
5.3.1 Automated Sleep Stage Classification 117
5.3.2 Automated RBD Detection 118
5.4 Methods 118
5.4.1 Data 118
5.4.2 Signal Pre-Processing 119
5.4.3 Feature Extraction 119
5.4.4 Automated Sleep Staging 119
5.4.5 RBD Detection 119
5.4.6 Evaluation 120
5.5 Results 121
5.5.1 RBD Detection 121
5.5.2 Correlation of Automated RBD Metrics 121
5.5.3 Bland-Altman Analysis of Automated RBD Metrics . . 122
5.5.4 Automatic RBD Classification 124
5.6 Discussion 126
5.7 Conclusion 130

5.1 Associated Publication

- N. Cooray, F. Andreotti, C. Lo, M. Symmonds, and M. T. M. Hu. "Detection of REM sleep behaviour disorder by automated polysomnography analysis".

In: *Clinical Neurophysiology* 130.4 (2019), pp. 505–514

5.2 Introduction

This chapter explores a fully automated pipeline for RBD detection that emulates the diagnostic process, as described in Figure 1.12. The framework combines sleep stage classification and abnormal EMG quantification in a cohort of age-matched healthy and RBD-diagnosed participants. The automated sleep classification technique presented in Chapter 3 was used to evaluate PSG signals and provide sleep stages, which in-turn contributed towards calculating RBD metrics that aim to quantify abnormal EMG activity. These metrics were then used in an RBD detection classifier, described in Chapter 4 to distinguish HCs from RBD participants. Only one study in literature exists that aims to achieve a fully automated pipeline on RBD detection in this manner [117].

The motivation of this work was to establish the feasibility of a fully automated RBD screening tool, given PSG recordings from a clinical setting. This work reaffirms existing work in the field ([117]), by demonstrating the effectiveness of machine learning techniques by incorporating more participants and less sensors, thereby lending more credence towards a potential screening tool. This chapter evaluates the impact automated sleep staging has on abnormal EMG quantification and the resulting detection of RBD participants.

The principal contributions of this chapter to the field of automated RBD detection are:

- Demonstrated a rigorous evaluation of fully-automated RBD detection using automated sleep staging.

- RBD metrics, used to quantify abnormal EMG activity, are highly correlated and within the limits of agreement when using manually and automatically annotated sleep stages (using the RF classifier described in Chapter 3).

- Sufficient precision and sensitivity of the automated sleep stage classifier coupled with the RF RBD detection technique (presented in Chapter 4), successfully distinguishes RBD participants from HCs.

5.3 Background

In previous chapters we were able to demonstrate that automated sleep staging and RBD detection was possible from a full PSG, including EEG, EOG, and EMG sensors (see Chapters 3 and 4, respectively). For RBD diagnosis, clinicians complete these tasks sequentially, to evaluate RSWA, and encapsulates the motivation for this chapter. This end-to-end detection of RBD participants from raw PSG recordings has been detailed in one previous study by Kempfner *et al.* (2014) and achieved excellent results [117]. Kempfner *et al.* (2014) proposed a full-automatic method for RSWA, that begins with REM sleep identification followed by abnormal EMG activity detection, from a dataset of 48 participants (16 HCs, 16 RBD, and 16 PLM participants) [117].

5.3.1 Automated Sleep Stage Classification

Kempfner *et al.* (2014) demonstrated with three EEG ($F_3 - A_2$, $C_3 - A_2$, and $O_1 - A_2$) and two EOG channels automated REM sleep classification as the first step towards RBD identification. All 30 second epochs were segmented into 3 second mini-epochs, where EEG and EOG features were extracted using a sliding window of 11 mini-epochs with 10 mini-epoch overlap. As a result features were extracted for each window representing a 3 second mini-epoch. Features extracted included EOG covariance (between the left and right channel) and the median rectified EEG amplitude within clinically relevant frequency bands (detailed in Chapter 1 Table 1.1). Features were then scaled and merged together, to be used in a binary class SVM to identify REM sleep. Post-processing on the SVM classification was then used to emulate the cyclical nature of REM and NREM throughout the night using a normalised Blackman window. Evaluated on the testing dataset (12 participants), this technique achieved REM sensitivity and specificity of 0.96

and 0.93, respectively. These are impressive results and were achieved across a cohort containing 12 participants and included 2 sleep disorders (PLM and RBD). These results might slightly benefit from an up-sampling of the annotated sleep stages, which were evaluated in 3 second mini-epochs based on their 30 second annotation (a factor increase of 10). Nonetheless this technique achieves high sensitivity and specificity for a diverse cohort of participants, which inevitably contributed towards the success of RBD detection.

5.3.2 Automated RBD Detection

For the detection of abnormal EMG activity, Kempfner *et al.* (2014) used EMG channels from the submentalis (chin) and the left and right anterior tibialis (lower leg). The EMG signals were expressed as an envelope signal through smoothing and full-wave rectification. The activity feature was derived from the mean of the EMG envelope and comparing it to the minimum value within a reference window. The features were then used in a one-class SVM to identify abnormal muscle activity during REM sleep. The classifier was trained using only HCs, inspired by Mourão-Miranda *et al.* (2011) [131]. For each participants the percentage of outliers with respect to inliers was used to provide a metric called the "muscle score". Using manually and automatically annotated REM sleep, muscles scores were calculated that achieved a correlation of 0.99, due to the high sensitivity and specificity of automatic REM detection. Using their testing dataset, this study was able to separate all RBD participants (4 in total) from the HC and PLM participants (4 participants each). These impressive results were the first to establish the feasibility of fully automated RBD detection technique. While this study was relatively small, it does set the foundation for this thesis and for future work.

5.4 Methods

5.4.1 Data

Once again, the combined dataset detailed in Chapter 2 was used and consists of 50 RBD participants and 50 age-matched HCs. The gold standard for sleep

staging and the clinical diagnosis was given by clinical specialists that provided the PSG recordings.

5.4.2 Signal Pre-Processing

As described in Chapter 3 PSG recordings underwent preprocessing. The EEG, EOG, and EMG signals were resampled at 200Hz. The EEG and EOG signal were pre-processed with a 500^{th} order band pass FIR filter with a cut-off frequency of 0.3Hz and 40Hz. The EMG signal was filtered with a 500^{th} order notch filter at 50Hz and 60Hz (because data was sourced from either European or Canadian institutions), in addition to a 500^{th} order high pass FIR filter with a cut-off frequency at 10Hz. Once pre-processed features were then extracted from the signals so as to characterise every 30 second epoch. Features are extracted for the purposes of both automated sleep staging and RBD detection.

5.4.3 Feature Extraction

Features used for automated sleep staging are detailed in Chapter 3 Section 3.4.3. Metrics used to identify RBD are described in Chapter 4 Section 4.4.3. Automated sleep staging and RBD detection were both achieved using an RF classifier.

5.4.4 Automated Sleep Staging

The results from Chapter 3 Section 3.5 provided automated sleep staging results (through an RF classifier) that achieved REM detection with a sufficient sensitivity and specificity to ideally facilitate successful RBD identification.

5.4.5 RBD Detection

Based on the metrics evaluated in Chapter 4 Section 4.4.3, this chapter uses automatically annotated sleep stages to generate RBD metrics to evaluate their performance in RBD detection. Results are then compared to metrics generated from manually annotated sleep stages to assess the impact of automated sleep staging on RBD detection. These results provide insight into the viability of a fully automated RBD detection through a limited PSG recording.

5.4.6 Evaluation

The RBD detection classifier was trained and evaluated using a 10 participant fold cross-validation scheme with even split between HC and RBD participants. For this chapter the training fold was used to train a classifier for automated sleep staging and another classifier for RBD detection classifier. Automated sleep staging results from the testing fold, were then used to generate RBD metrics, which were evaluated in the RBD detection. For this chapter, two RF classifiers for RBD detection, were evaluated, one using established EMG metrics (four features), while the other included additional features detailing sleep architecture, atonia index ratios, and fractal exponents (11 features in total). These two classifiers illustrate the impact of using additional EMG features based on sleep architecture to detect RBD when using automated sleep staging. The performance of RBD detection from automated sleep staging was evaluated by calculating the macro accuracy, sensitivity, specificity, and F1 score (as described in Chapter 3 Section 3.4.9). The impact of automated sleep staging on the calculation of these metrics was also carefully analysed.

Firstly, the correlation of RBD metrics using manually and automatically annotated sleep staging was analysed to assess the impact of automated sleep staging on RBD metrics. Secondly, Bland and Altman (B&A) plots were introduced to describe the agreement between two quantitative measures as an improvement to simple correlation factors [132, 133, 134]. This chapter used B&A plots to evaluate the agreement between RBD metrics derived from manual and automatic sleep staging. The B&A plot measures agreement by plotting the mean against the difference of both quantitative methods, while specifying the limits of agreement (where Bland and Altman recommend that 95% of all data points fall within two standard deviations of the mean difference) [132]. Furthermore, B&A plots detail the bias between the mean differences of both quantitative methods and their significance. For this study B&A plots were used to establish a level of agreement between RBD metrics, derived using manually and automatically annotated sleep stages within a certain agreement interval. The decision on whether these intervals are acceptable

must be decided by clinical evaluation and in our case through the performance of automated RBD detection with metrics derived from automated sleep staging.

Lastly, the performance of RBD detection are presented for each fold using automated sleep staging, detailing the macro-averaged accuracy, sensitivity, specificity, and F1 (as detailed in Chapter 4 Section 4.4.5).

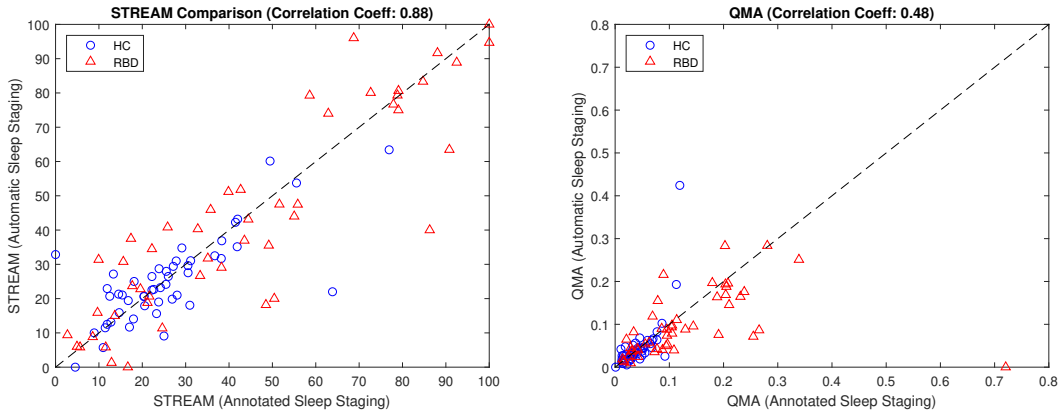
5.5 Results

5.5.1 RBD Detection

The performance of RBD detection using metrics described in Chapter 4 Section 4.4.3, are depicted again in Table 5.1a (derived from manually annotated sleep staging). The results from automated sleep staging are detailed in Chapter 3 Section 3.5 and their impact on calculating RBD metrics are detailed by the Pearson correlation with metrics derived from manually annotated sleep stages.

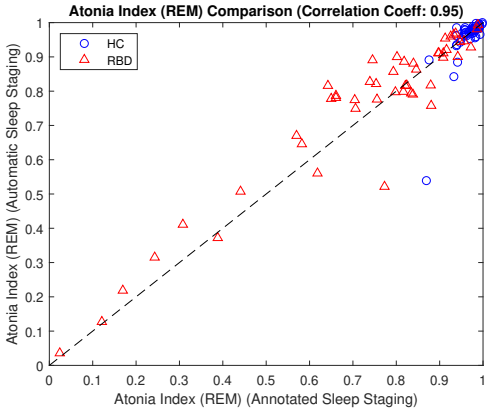
5.5.2 Correlation of Automated RBD Metrics

For the established RBD metrics, the correlation between metrics derived from manually and automatically annotated sleep stages are depicted in Figure 5.1. The Pearson correlation was calculated to assess the effect of automated sleep staging on the calculation of these metrics. The correlation for STREAM, QMA, and the atonia index (REM) were 0.88, 0.48, and 0.95, respectively. STREAM and the atonia index (REM) produce high correlation scores, while the QMA metric appears relatively low. For the new proposed RBD metrics, the correlation between metrics derived from manually and automatically annotated sleep staging are detailed in Figure 5.2. The atonia index ratios between REM compared to N2 and N3 were 0.86 and 0.93, respectively. While the correlation of fractal exponent ratios between REM compared to N2 and N3 were 0.60 and 0.54, respectively. The ratio of total N2 sleep and N3 sleep achieved a correlation of 0.58 and 0.77, respectively.



(a) STREAM.

(b) QMA.

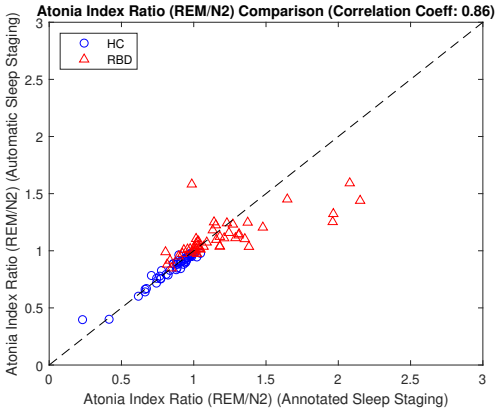


(c) Atonia Index (REM).

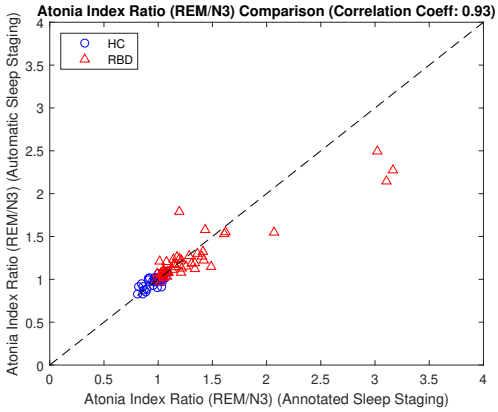
Figure 5.1: The correlation of established RBD metrics using automatically and manually annotated sleep staging are detailed in these graphs for 5.1a STREAM, 5.1b QMA, and 5.1c atonia index (REM). In each of these cases the RBD participants (red triangles) appear to have greater variation from the manually annotated scores, particularly for the QMA metric.

5.5.3 Bland-Altman Analysis of Automated RBD Metrics

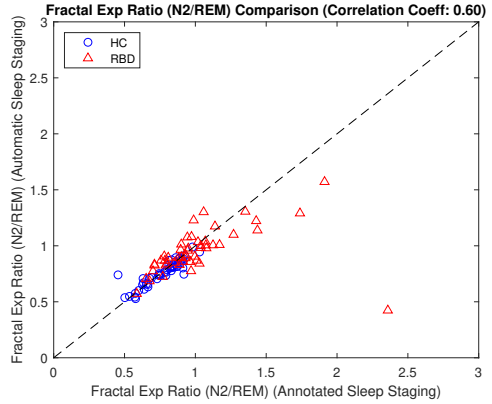
The impact of established RBD metrics calculated from automated sleep staging were evaluated by the B&A plots given in Figure 5.3. STREAM, QMA, and the atonia index (REM) appear to provide agreeable results where the limits of agreement are represented by horizontal dotted lines. The mean difference bias, defined by the solid black line in Figure 5.3, all appear to have little to no bias when using automated sleep staging (shaded area around mean difference includes the zero line). The B&A plots for the atonia index ratios, fractal exponent ratios, and the sleep stage ratios are illustrated in Figure 5.4. Each of these metrics appear to provide



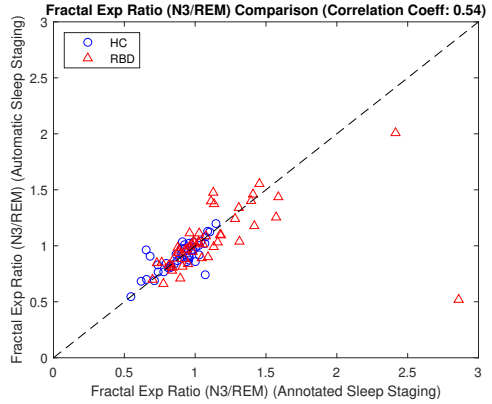
(a) Atonia index ratio (REM/N2).



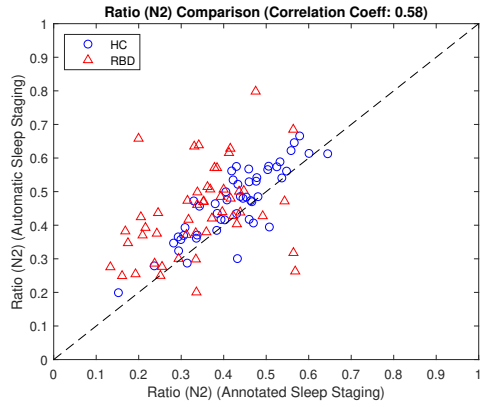
(b) Atonia index ratio (REM/N3).



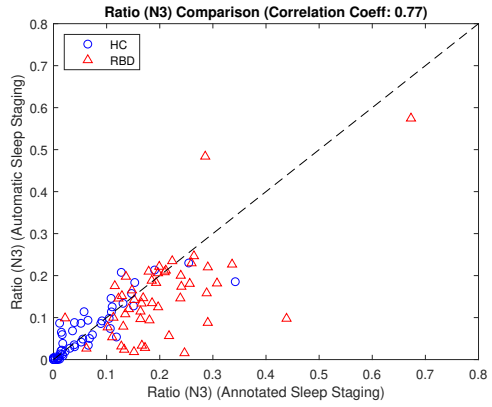
(c) Fractal exponent ratio (N2/REM).



(d) Fractal exponent ratio (N3/REM).

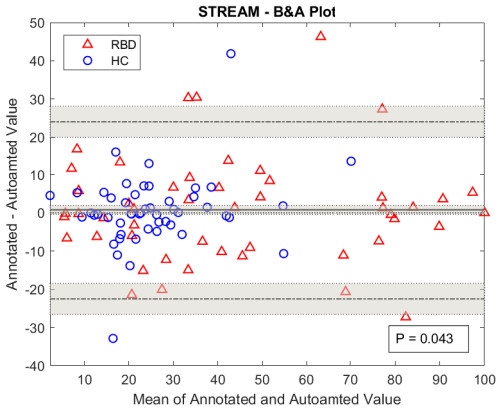


(e) Ratio N2 (%).

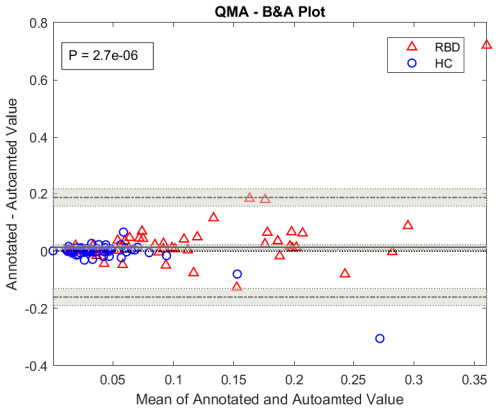


(f) Ratio N3 (%).

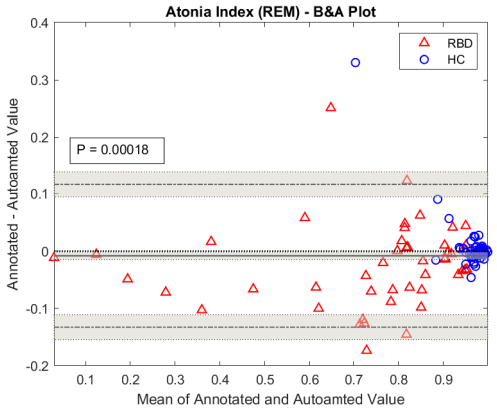
Figure 5.2: The correlation for additional RBD metrics derived from manually and automatically annotated sleep stages. The atonia index ratios between 5.2a REM/N2 and 5.2b REM/N3 provide high correlation scores. The fractal exponent ratios between 5.2c N2/REM and 5.2d N3/REM achieve lower correlation scores, with a greater variation in metrics for RBD participants. The correlation for the ratio of total sleep in 5.2e N2 was lower than the ratio in 5.2f N3 sleep.



(a) STREAM.



(b) QMA.



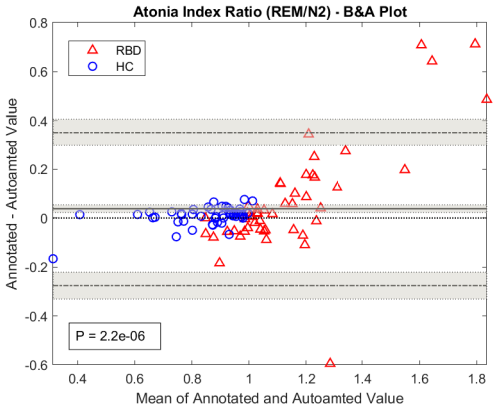
(c) Atonia Index (REM).

Figure 5.3: The BA plots for the established RBD metrics using manually and automatically annotated sleep staging for 5.3a STREAM, 5.3b QMA, and 5.3c atonia index (REM). All metrics appear to be within the limits of agreement. However both 5.3b QMA and 5.3c atonia index (REM) incur a bias when using automatically annotated sleep stages, as demonstrated by the dotted line (mean difference) and the zero line (solid black line and the surrounding grey shaded area). The STREAM metrics appears to have no significant bias when using automatically annotated sleep stages. Once again these plots confirm that the RBD values appear to generate the greatest variation between annotated and automatic metrics.

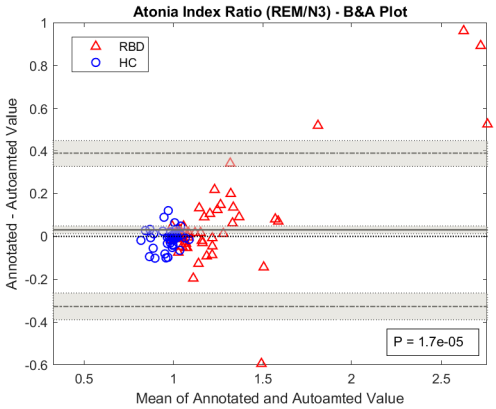
agreeable results, where calculations are within the limits of agreement. However, all these metrics appear to have a slight bias when using automated sleep staging (where the shaded area around the mean difference does not include the zero line).

5.5.4 Automatic RBD Classification

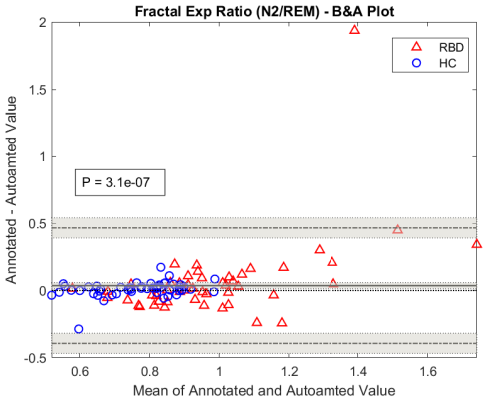
The ability to discriminate individuals with RBD from HCs using individual metrics and two RF classifiers (trained on established metrics and established metrics



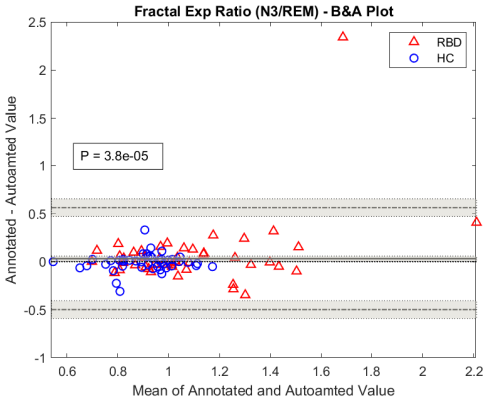
(a) Atonia index ratio (REM/N2).



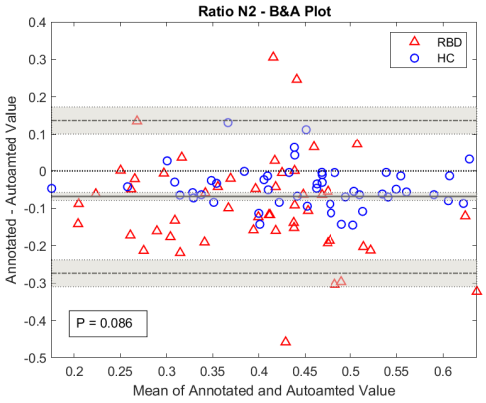
(b) Atonia index ratio (REM/N3).



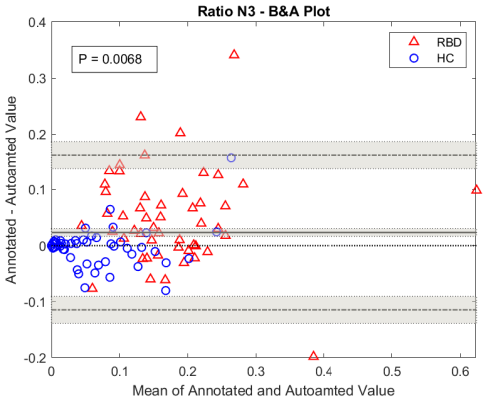
(c) Fractal exponent ratio (N2/REM).



(d) Fractal exponent ratio (N3/REM).



(e) Ratio N2 (%).



(f) Ratio N3 (%).

Figure 5.4: The BA plots for the new RBD metrics using automatically and manually annotated sleep stages. The atonia index ratios between 5.4a REM/N2 and 5.4b REM/N3 appear to be within the limits of agreement. The fractal exponent ratios between 5.4c N2/REM and 5.4d REM/N3 also produce scores within the limits of agreement. The ratio of total sleep in 5.4e N2 and 5.4f N3 are also agreeable between manually and automatically annotated sleep stages. All of these metrics appear to incur a bias to scores when using automatically annotated sleep stages and detailed by the mean difference (dotted line) found outside the zero line and the surrounding shaded grey area. From these plots we can observe the greater variation of values for RBD participants (red triangles).

supplemented with additional features) is depicted in Table 5.1b. Of the established metrics, the atonia index (REM) achieved the best results, with an accuracy, sensitivity, and specificity of 0.81 ± 0.094, 0.70 ± 0.16, and 0.92 ± 0.098, respectively. While the best overall results were obtained by the RF classifier that combined all metrics with an accuracy, sensitivity, and specificity of 0.90 ± 0.063, 0.86 ± 0.13, and 0.94 ± 0.092, respectively.

	Accuracy	Sensitivity	Specificity	Precision	F1
STREAM	0.60 ± 0.19	0.64 ± 0.20	0.56 ± 0.25	0.62 ± 0.20	0.62 ± 0.17
Atonia Index	0.85 ± 0.092	0.74 ± 0.16	$\mathbf{0.96 \pm 0.08}$	$\mathbf{0.96 \pm 0.091}$	0.82 ± 0.11
QMA	0.77 ± 0.13	0.66 ± 0.13	0.88 ± 0.16	0.86 ± 0.17	0.74 ± 0.14
Established Metrics (RF)	0.88 ± 0.11	0.86 ± 0.16	0.90 ± 0.10	0.90 ± 0.10	0.87 ± 0.12
Combined Metrics (RF)	$\mathbf{0.92 \pm 0.075}$	$\mathbf{0.94 \pm 0.092}$	0.90 ± 0.13	0.92 ± 0.10	$\mathbf{0.92 \pm 0.07}$

(a) RBD classification performance using manually annotated sleep stages.

	Accuracy	Sensitivity	Specificity	Precision	F1
STREAM	0.62 ± 0.12	0.66 ± 0.22	0.58 ± 0.23	0.63 ± 0.22	0.62 ± 0.16
Atonia Index	0.81 ± 0.094	0.70 ± 0.16	0.92 ± 0.098	0.91 ± 0.11	0.78 ± 0.12
QMA	0.77 ± 0.13	0.66 ± 0.13	0.88 ± 0.16	0.86 ± 0.17	0.74 ± 0.14
Established Metrics (RF)	0.87 ± 0.10	$\mathbf{0.88 \pm 0.13}$	0.86 ± 0.13	0.87 ± 0.11	0.87 ± 0.10
Combined Metrics (RF)	$\mathbf{0.90 \pm 0.063}$	0.86 ± 0.13	$\mathbf{0.94 \pm 0.092}$	$\mathbf{0.95 \pm 0.082}$	$\mathbf{0.89 \pm 0.073}$

(b) A summary of end-to-end RBD classification performance using automatic sleep staging.

Table 5.1: These tables summarise the impact of automated sleep staging on automated RBD detection, which only incurred a small drop in performance compared to manual sleep staging. From Table 5.1b we can observe the success of the atonia index metric was diminished considerably, compared to Table 5.1a. As detailed in Chapter 4, the performance from the RF classifiers achieved the best performance, with respect to accuracy, specificity, precision, and F1 (mean and standard deviation are presented). The numbers in bold detail the highest score achieved in that column.

5.6 Discussion

The goal of this study was to validate a fully-automated pipeline for identifying individuals with RBD. The first step began with automated multi-stage sleep classification, which provided a high REM sleep accuracy when compared to the gold standard of human expert annotation (see details in Chapter 3 Section 3.5). Despite a drop in performance for classification of REM stages in RBD individuals, specificity of REM stage detection remained high and EMG quantification metrics

based on automated sleep staging were shown to be highly correlated to those from manual annotation (Figure 5.1 and 5.2).

Despite the lack of sensitivity, REM specificity remained high, which means that as long as REM was identified with a certain precision, the quantified absence of atonia will remain indicative. The high correlation coefficients depicted in Figures 5.1 and 5.2 illustrate the success of automated sleep staging in preserving EMG metrics. From the established metrics, the QMA metric provided the lowest correlation (0.48), which might be attributed to its dependence on several adjustable parameters such as the QMA detection threshold, minimum QMA event duration and the inter-event interval. Once more this technique was susceptible to sleep stage misclassification because it relies on calculating variance during REM, which was especially variable in individuals with RBD. The STREAM algorithm proved more robust and immune to high variability, because it compares variance between REM and NREM. The atonia index (REM) not only proves robust but was also parameter free (excluding window lengths) and provided a correlation scores of 0.88. From the established metrics, the atonia index (REM) provided the best score of 0.95. Additional features that incorporate sleep architecture also remained unaffected from the use of automated sleep staging. The atonia index ratios (N2/REM and N3/REM) have correlation scores (0.86 and 0.93, respectively) less than the traditional atonia index (REM) but remained relatively high due to the sleep stage classification performance of N2 and N3. The success of the multi-stage classification also contributes to the high correlation scores for N3 sleep ratio (0.77). However, it would appear for the N2 ratio there was a strong bias in overestimating caused by N2 misclassification (correlation of 0.58). The EMG fractal exponent ratios (N2/REM and N3/REM) prove to provide average correlation scores, with 0.60 and 0.54, respectively. These metrics did not seem to benefit from the robustness of information derived from the frequency spectrum (specifically during REM, N2 and N3).

Bland-Altman analysis was also used to evaluate the agreement between metric quantification using manually and automatically annotated sleep staging (as shown in Figures 5.3 and 5.4). From these plots it is clear that automated sleep staging

(derived from EEG, EOG, and EMG features) sufficiently produced RBD metrics that were within the limits of agreement with metrics derived from manually annotated sleep staging. Because automated sleep staging was poorest on RBD participants (red triangles), we observed a greater variation in the y-axis (difference), especially as the x-axis values (mean value) moved towards values associated with RBD (red triangles). However we can observe a bias introduced to RBD metrics derived from automatically annotated sleep staging, where the mean difference (solid black line in Figures 5.3 and 5.4) and its confidence interval (grey shaded area) does not include the line of equality (zero axis, dotted line through the origin). This is the direct result from automated sleep stage misclassification. When REM was misclassified for N2 and W, the impact on the calculation of the RBD metrics meant a bias towards values associated with N2 and W sleep. For each metric detailed Figures 5.3 and 5.4 there was a significant bias introduced from automated sleep staging, as detailed by the p-value detailed in the graphs, except for the N2 sleep ratio. The bias from automated sleep staging was expected and is the natural result of REM sleep misclassification for N2 and W sleep, as metrics start to reflect values associated with N2 and W. Note that the limits of agreement are calculated by assuming that the difference between the metrics calculated from manually and automatically annotated sleep stages are normally distributed, however this is not necessarily required as stipulated by Bland and Altman (1999) [134]. From these plots we can observe the resilience and fidelity of these metrics when using automated sleep staging.

Finally the effect of automated sleep staging on RBD detection was summarised in Table 5.1b, where the success of accurate automated sleep staging ensured the detection performances remained high. Metrics used for RBD detection are largely derived from REM sleep, the impact of misclassifying actual REM for other sleep stages means less epochs are available to analyse and detect RSWA (hence a bias towards negative RBD detection and false negatives). Inversely, other sleep stages misclassified as REM, might have the impact of biasing positive RBD detection (false positives), however due to high specificity of REM classification, this effect was

nullified. Subsequently, RBD detection performance was only slightly diminished when using automatically classified sleep stages (see Table 5.1).

From the manually annotated sleep stages, the best performance in RBD detection was attained by the RF classifier trained on a combination of established metrics supplemented by features that incorporate sleep architecture with an accuracy, sensitivity, specificity, and F1 of 0.92 ± 0.075, 0.94 ± 0.092, 0.90 ± 0.13, and 0.92 ± 0.07, respectively. The performance of RBD detection using automated sleep staging was only marginally lower compared with manual sleep annotation (accuracy, sensitivity, specificity, and F1 of 0.90 ± 0.063, 0.86 ± 0.13, 0.94 ± 0.092, and 0.89 ± 0.073 respectively). The only direct comparison in literature comes form Kempfner *et al.* (2014) that achieved RBD detection with a sensitivity of 0.92 and specificity of 0.89 from a cohort of 48 participants with automated REM classification [117]. These results are fairly similar and both provide high sensitivity and specificity, which would be ideal from the perspective of a clinical screening tool. Potentially providing an accurate tool that identifies individuals with RBD with a low false positive and false negative rate.

The performance of the atonia index noticeably drops when using automated sleep staging, when compared to annotated sleep staging, and can be attributed to the dependency on REM sleep classification, especially where there is RSWA. We can also observe from the B&A plot that there is a slight bias towards HC values (higher atonia index scores), this was caused by not including REM epochs that capture RSWA, which were misclassified for other stages.

Therefore, these results could be further improved upon with better automated sleep stage classification. Although perfect automated sleep staging may never be attainable, given human inter/intra variability [21, 22], this study has demonstrated that this is not necessarily required for successful RBD detection. Given sufficiently precise automated sleep staging, enough information was retained to identify RBD participants using an RF classifier and a combination of metrics that measure the absence of atonia, the atonia changes between different sleep stages, and sleep architecture.

Although this study included a reasonably large number of RBD PSG recordings, further work is required in order to define the potential role of automated RBD detection in the clinical environment. This study was limited by simply focusing on distinguishing RBD individuals from HCs, and there maybe additional confounds in clinical practise when applied to a mixed population of sleep disorders. To help further delineate RBD, a promising signature might be found with the inclusion of other non-motor features, such as altered heart rate variability during sleep [135], which is explored in the next chapter. Furthermore there is evidence to suggest EMG signals provided from the flexor digitorum superficialis and exterior digitorum brevis best capture phasic EMG activity and might help increase the sensitivity of RSWA detection, especially in the case of mild RBD episodes [136]. This study was also limited in that sex ratios were not matched in the study cohorts and evidence suggests there are physiological differences in sleep between sexes [24]. Future directions of this work would look to apply improved automated sleep staging algorithms across a range of common sleep disorders, whilst also incorporating other PSG and physiological signals to improve RBD detection.

5.7 Conclusion

This study proposed a fully-automated pipeline for RBD detection based on a combination of established EMG metrics (atonia index, QMA, and STREAM) and new metrics that incorporate sleep architecture. In the previous chapters we have detailed the process of RBD diagnosis (based on PSG evaluation) by clinical experts and emulated these elements through machine learning techniques. This was all achieved using a limited montage, consisting of a single EEG, EOG, and EMG channel, which paves the way to design and implement an automated system that could easily be incorporated into lightweight wearable technology, conducive for take-home purposes.

The main contributions and findings of this chapter were:

- Established RBD metrics calculated from automated sleep staging, including STREAM and the atonia index (REM), were highly correlated to their manually annotated derivations (with a Pearson correlation of 0.88 and 0.95, respectively). Metrics that incorporated sleep architecture also demonstrated similar correlation scores, measuring 0.86, 0.93, 0.60, and 0.77 for the atonia index ratio (REM/N2), atonia index ratio (REM/N3), fractal exponent ratio (N2/REM), and the N3 sleep ratio, respectively.

- These RBD metrics also demonstrated that they were within the limits of agreement, as defined by the B&A plots, when using manually and automatically annotated sleep stages. However, misclassification from automated sleep staging did impart a bias to RBD metrics, although they did not compromise the success of automated RBD detection.

- Sufficient precision and sensitivity of the automated sleep stage classifier coupled with the RF RBD detection technique presented in Chapter 4, enabled RBD metrics to retain their discriminative properties (as demonstrated by the correlation and B&A analysis) to successfully distinguish RBD participants from HCs. Yielding RBD detection accuracy, sensitivity, and specificity of 0.90 ± 0.063, 0.86 ± 0.13, and 0.94 ± 0.092, respectively (from an age-matched cohort of 50 HCs and 50 RBD participants).

Weary with toil, I haste me to my bed,
The dear repose for limbs with travel tired;
But then begins a journey in my head
To work my mind, when body's work's expired:

— William Shakespeare's *Sonnet* 27

6

Automated RBD Detection with Minimal Sensors

Contents

6.1 Associated Publication 134
6.2 Introduction 134
6.3 Background 136
6.3.1 The Autonomic Nervous System 136
6.3.2 R-Peak Detection 136
6.3.3 Automated Sleep Staging with Heart Rate 138
6.3.4 Summary 142
6.3.5 Autonomic Dysfunction in RBD 142
6.3.6 Minimal Automated Sleep Staging 143
6.4 Methods 143
6.4.1 Data 143
6.4.2 Pre-processing 144
6.4.3 Electrocardiogram Feature Extraction 144
6.4.4 Minimal Sleep Stage Classification 157
6.4.5 Extract Autonomic Dysfunction Metrics 158
6.4.6 Minimal RBD Detection 159
6.4.7 Evaluation 160
6.5 Results 161
6.5.1 Minimal Automated Sleep Staging 161
6.5.2 Minimal RBD Detection 162
6.5.3 Bland-Altman Analysis on RBD Metrics 165
6.6 Discussion 165
6.7 Conclusion 174

6.1 Associated Publication

- N. Cooray, F. Andreotti, C. Lo, M. Symmonds, M. T. M. Hu, and M. De Vos. "Screening for REM Sleep Behaviour Disorder with Minimal Sensors (In Review)". In: *Clinical Neurophysiology* (2020), pp. 1–22

6.2 Introduction

To further explore the mechanisms behind RBD conversion to PD and to develop neuroprotective therapies, larger cohorts of individuals with RBD must be identified. As discussed earlier, the greatest hurdle to designing such studies is the slow and arduous RBD diagnostic process involving PSG recordings within sleep clinics, which are often under-staffed and oversubscribed (with severely limited availability). PSG recordings are expensive, time consuming to setup, potentially uncomfortable and rely on labour intensive visual inspection of PSG data. Previous studies have demonstrated that automated sleep staging and RBD detection is possible from a limited PSG consisting of EEG, EOG, and EMG sensors ([44, 47, 117]) and was also demonstrated in Chapters 3, 4, and 5. This work sought to further minimise this limited framework by excluding EEG sensors and including an ECG sensor to determine an effective yet minimal set of sensors necessary for automated RBD detection. ECG sensors are ubiquitous and user-friendly and therefore provide an opportunity to further economise sleep staging by excluding the need for cumbersome and expensive EEG sensors. Once more these sensors may also aid in the detection of RBD.

Figure 6.1 details the conventional process for automated sleep staging using ECG signals, including the final step for RBD detection derived from sleep stages and ECG features. The layout of this chapter begins with a literature review of ECG based automated sleep staging and autonomic dysfunction in RBD participants. Secondly, this chapter details the features extracted from ECG based signals for automated sleep staging, including the description of data used and the necessary

pre-processing. Lastly, this chapter presents the evaluation of various combinations of PSG signals for automated sleep staging with a focus to achieve automated RBD detection with minimal sensors in order to reach best-in-class performance.

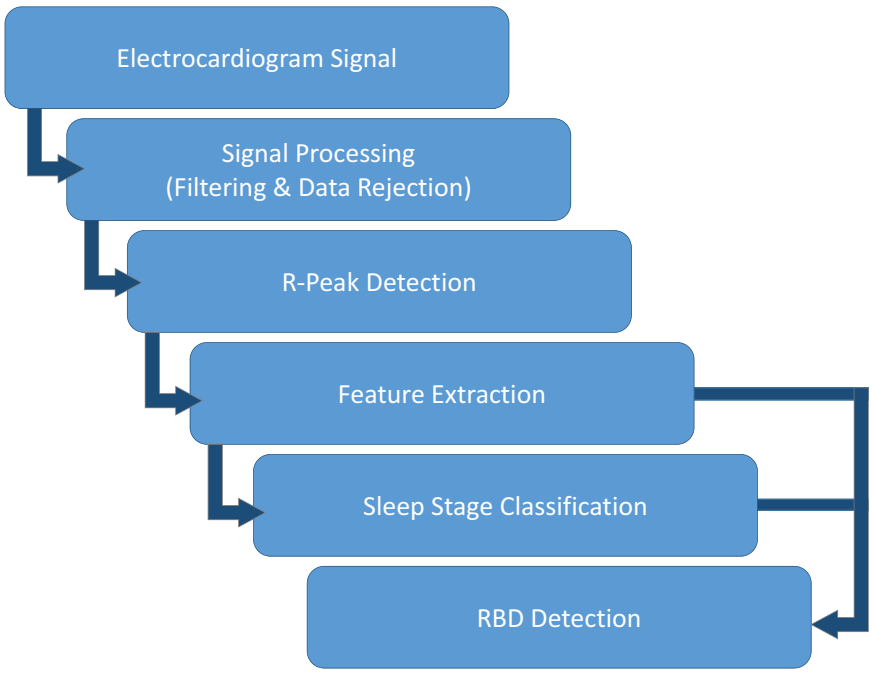


Figure 6.1: Common automatic sleep stage classification methodology using ECG sensors. The final step includes RBD detection using automated sleep staging and features extracted from the ECG signal.

The principal contributions of this chapter to the field of automated RBD detection are:

- A demonstration validating the feasibility of a fully automated RBD screening tool using minimal sensors.
- Accurate and reliable REM sleep stage classification in individuals with RBD, without the need for EEG sensors.
- A demonstration of ECG features with limited success in REM classification and RBD detection.

6.3 Background

In previous chapters we have demonstrated automated sleep staging and RBD detection from a limited set of PSG channels (Chapters 3, 4, and 5), consisting of EEG, EOG, and EMG signals. Recent sleep research is uncovering an interplay between the central nervous system (CNS) and the autonomic nervous system (ANS) that is regulated throughout sleep and its various stages [137].

6.3.1 The Autonomic Nervous System

The ANS is responsible for regulating the majority of the body's internal systems, such as blood pressure, breathing, body temperature, digestion, urine, and myocardial function [138]. Of particular interest are the changes and regulation of heart rate variability (HRV) during sleep and how it might be interpreted for automated sleep staging [139, 140, 141, 142, 143, 144, 145, 146, 147], but traditionally limited to three stages (W/NREM/REM). These ECG based automated sleep stage classifiers show promise but their application thus far has been limited to only relatively young HCs and unexplored in individuals with sleep disorders.

6.3.2 R-Peak Detection

Automated R-peak detection is the first step in deriving ECG based features and one of the most cited methods in literature is the Pan-Tomkins algorithm [148]. This real time algorithm is able to detect QRS complexes (an example wave shown in Figure 6.2) using characteristics from an ECG signal and was applied in this work [148]. The algorithm adaptively adjusts thresholds and parameters to account for variations in the ECG signal that occur due to changes in QRS morphology and heart rate [148]. This section briefly details the algorithm described by Pan-Tomkins (1985) [148].

R-Peak Detection Algorithm Overview

This QRS detection method consists of three processing steps that begins with linear digital filtering, followed by a non-linear transformation, and concludes with

[2]Anthony Atkielski [public domain] via Wikimedia Commons.

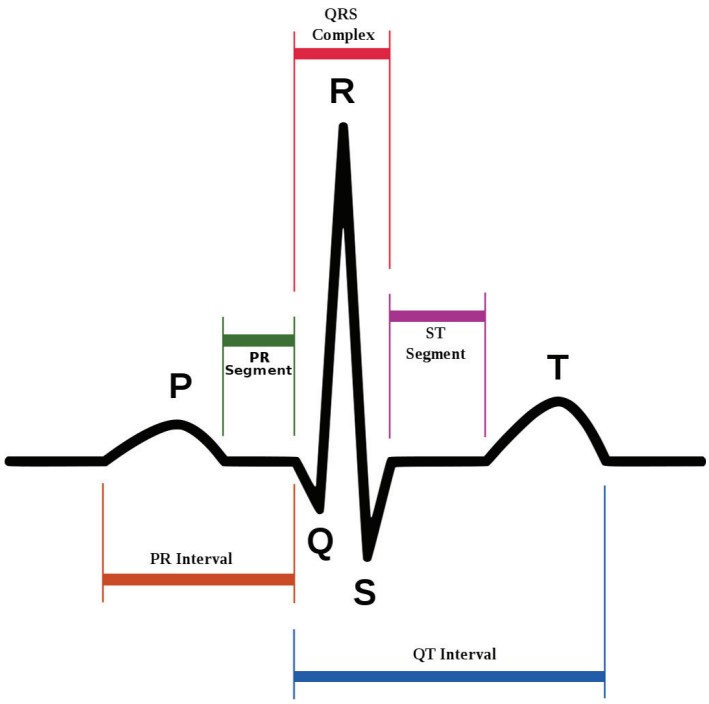


Figure 6.2: A typical PQRST wave from an ECG signal, detailing the characteristics of the QRS complex used to identify R peaks.[2]

a decision rule algorithm. Throughout the process the algorithm adjusts thresholds and parameters periodically in order to adapt to changes in the ECG signal, with respect to QRS morphology and heart rate. First the ECG signal is band-pass filtered to reduce the impact of noise and interference found in the signal.

The band-pass filter cut-off frequencies were chosen to maximise the QRS energy, found between 5-15Hz [148]. After filtering, differentiation was achieved to describe the nature of the QRS-complex slope. After differentiation the signal was squared, rectifying the signal and exaggerating its derivative. Next, a moving integration window using N samples was used to create the final feature waveform and detail the slope of the R-wave. For this study, N was set to the number of samples in 100 milliseconds, as a default value, roughly the expected length of a QRS complex. This waveform was then smoothed by a third order median filter. Lastly, detection thresholds are used to find QRS complexes and missed beats, where these parameters are continually adjusted to adapt to changes in the ECG characteristics.

The threshold value was based on the amplitude distribution from a segment of the smoothed waveform (from between the first to the 90^{th} second). The threshold

was set to 60% of the 98^{th} percentile of the amplitude distribution. From all possible peaks identified from this threshold, a search back was then used to identify missed beats. A potential missed beat was found when RR intervals are greater than 150 percent of the median RR interval (where intervals are greater than 0.01 seconds, given that beats cannot occur this close apart). A potential missed beat was then found between these RR intervals where the smoothed waveform exceeds half of the initial threshold, described above. This methodology provides the RR intervals from which features can be extracted for automated sleep staging and RBD detection.

The parameters described above were chosen based on default values that achieved effective beat detection in healthy control participants [149]. Potential future work might explore the impact of fine-tuning these parameters to achieve optimal automated sleep staging for ECG signals.

6.3.3 Automated Sleep Staging with Heart Rate

While most automated sleep staging literature includes numerous sensors and channels, this section reviews studies that focus on ECG sensors exclusively. Once again, the evaluation of these studies focus on the performance of REM classification, due to its' importance in RBD identification.

Threshold Based Algorithm

Kurihara *et al.* (2012) designed a study using a bed to non-invasively extract signals to measure the heart beat and body movement [142]. The bed provided a body movement signal known as ballistocardiography (BCG). The study consisted of 10 male participants with an average age of 22.2 years. Using each signal, two functions were created to provide an indication of REM and depth of sleep [142]. These functions were used to classify sleep stages and achieved a 3-stage kappa score of 0.48 ± 0.07. This study established the feasibility of ECG based signals to aid in sleep stage classification but was limited to 10 young male participants and also benefited greatly from an additional BCG signal.

Linear Discriminant Analysis

Two studies that explored automated sleep staging using linear discriminant analysis (LDA) on ECG based signals included Redmond *et al.* (2007) and Fonseca *et al.* (2015) [139, 145]. Redmond *et al.* (2007) used a Hilbert transform based R-peak detector to generate RR-based features [139]. Normalisation was applied on the RR intervals for each participant by dividing values by the mean RR interval. As a form of error correction, a median filter was also applied to compensate for missed beats and erroneous beat detection. The RR interval was also detrended to account for variation in the ECG caused by the circadian rhythm [139]. This study also measured respiratory effort (directly from the ECG signal) and included an inductance plethysmography, which estimates rib-cage effort. In total 30 features were extracted, including RR-interval series, ECG derived respiratory (EDR), RR-EDR cross-spectral, and rib-cage respiratory effort. Redmond *et al.* (2007) demonstrated with a cohort of 31 male participants (average age of 42 ± 7.4) automated sleep staging with a 3-stage kappa score of 0.45.

Similarly, Fonseca *et al.* (2015) also used ECG and a inductance plethysmography to extract 142 features for 3-stage classification on 48 HC participants [145]. A total of 86 features were computed from an ECG sensor and 44 features derived from the respiratory signal. Features included time domain, frequency domain, non-linear features, and visibility graph based features. To overcome participant variability, Z-score normalisation was applied for each feature. Interestingly, this study also used cubic spline fitting to smoothen features in an attempt to emulate the cyclical nature of sleep structure. Feature selection was achieved through sequential forward selection and classification through Bayesian linear discriminant analysis. This study achieved a 3-stage kappa score of 0.56 ± 0.15 and a REM classification accuracy of 0.80 ± 0.08. Both these studies demonstrated the advantage of using ECG in combination with a respiratory sensor for sleep stage classification.

Support Vector Machines

Ebrahimi *et al.* (2015) explored automated sleep staging using heart rate and respiratory rate in combination with an SVM classifier [144]. By extracting time domain, time-frequency, fractal and entropy based features from 5 minute long epochs this technique achieved REM classification accuracy, sensitivity, and specificity of 0.692, 0.864, and 0.519, respectively. This study confirmed the importance of using long epochs for ECG based features and also demonstrated the difficulty of achieving high performance for participant-independent sleep stage classification. Additionally this study demonstrated that respiratory rate features from a respiratory sensor are superior to respiratory derived featured from an ECG based sensor, as determined by the recursive feature elimination algorithm.

Hidden Markov Models

EGG based features used in combination with Hidden Markov Models (HMMs) have been explored in literature [141, 140]. HMMs aim to take advantage of sleep's inherent cyclical nature and include temporal information for sleep stage classification. With 18 HC participants that provided heart beat intervals along with movement activity (all derived from a BCG), Kortelainen *et al.* (2010) achieved 3-stage classification with a kappa score of 0.44 ± 0.19 [141]. This study benefited from a bed sensor to quantify movement, which aided specifically with wake classification. However, using only HRV for sleep stage classification produced a 3-stage kappa score of 0.36 ± 0.19, where differentiating between NREM and REM became more difficult. Heart rate intervals derived from a BCG sensor were not ideal and was reflected in the classification performance.

Mendez *et al.* (2010) also detailed the difficulty of 3-stage classification and focused instead on NREM and REM classification [140] from a single ECG sensor. Using a dataset of 24 HC participants this study achieved a REM classification accuracy, sensitivity, and specificity of 0.80, 0.63, and 0.85, respectively. This study had the added benefit of excluding wake periods but demonstrated the utility of HMMs for sleep stage classification from HRV. Both these studies were able to take

advantage of sleep's macro-structure and cyclical nature through a HMM. These models require very few features, which provided computational efficiency and was one major benefit of a HMM. However these studies benefited greatly from the addition of respiratory based sensors, which appeared necessary.

Random Forest

The RF classifier has been shown to be quite effective at sleep stage classification using ECG signals, exemplified by Xiao *et al.* (2013) and Yucelbas *et al.* (2018) [143, 147]. Xiao *et al.* (2013) demonstrated with 45 PSG recordings, rough and convenient 3-stage sleep stage classification with 41 features [143], producing a kappa score of 0.463. This study benefited from RR interval extraction from automatic rhythmus analysis with manual review, where annotated periods labelled non-stationary and artefacts were excluded. For this study, features extracted included time domain, frequency domain, and non-linear analysis using a 5 minute windows centred around each 30 second epoch. Each feature was then normalised by fractile based values. The fractile value at 5% and 95% are first calculated, and all values between these fractiles were transformed to be between 0 and 1. This study demonstrated a slight improvement to performance using this fractal based normalisation. This study was the first to establish the ability of RF classifier and ECG features extracted from 5 minute windows to achieve satisfactory automated sleep staging for HC participants.

Yucelbas *et al.* (2018) was one of the first to demonstrate automated sleep staging from ECG signals on a patient cohort of OSA participants [147]. This study used two datasets with 10 and 18 participants, containing HCs and individuals with OSA, to achieve successful 3-stage classification with an RF classifier and morphological features derived from a single ECG signal [147]. This study explored several feature extraction methods, including singular value decomposition, variational mode decomposition, Hilbert Huang transformation, and morphological, where morphological features produced the best performance. Morphological features were extracted from R peaks, which were detected using the Pan-Tompkins QRS detection algorithm [148]. The R-peaks were then used to derive P, Q, R, S, and T

waves, whose x and y components were used to calculate 15 morphological features [147]. This study also found the RF classifier outperformed artificial neural networks, decision trees, Naive Bayes, and linear regression classifiers. For the first dataset (10 participants), this study achieved a 3-stage kappa score of 0.737 ± 0.025 and 0.572 ± 0.019 for HC and OSA individuals, respectively [147]. While in the second dataset (18 participants) attained a kappa score of 0.431 ± 0.019 and 0.523 ± 0.012 for HC and OSA individuals, respectively [147]. This study confirmed patient participants (OSA individuals) perform lower compared to HCs, because it was evident that OSA participants have structurally disordered ECG. It was also clear from this study, using only ECG features meant is was difficult to differentiate wake from NREM and also REM and NREM. As a result REM classification accuracy was 0.410 and 0.521 for OSA and HC participants, respectively (for the first dataset).

6.3.4 Summary

This section has detailed numerous studies that have demonstrated automated sleep staging using ECG based features. It is clear from previous work that achieving five stage automated sleep staging was difficult for HCs, let alone the prospect of applying them on RBD participants. Therefore this chapter focused on automated three stage sleep staging (W, NREM, and REM). While the aforementioned classifiers have each demonstrated their worth (with similar outcomes), the focus of this chapter was not to identify the best classifier, but rather to evaluate the usefulness of a single classifier on an RBD cohort that has already proved its viability in literature. For automated sleep staging, this chapter applied an RF classifier, which is a powerful classifier that has been demonstrated in the previous studies ([143, 147]) and in Chapter 3.

6.3.5 Autonomic Dysfunction in RBD

Recent literature has indicated cardiac autonomic dysfunction as a potential characteristic of RBD [150, 135, 151]. These studies indicate that HRV properties might be able to distinguish RBD participants from HCs, providing an additional metric beyond abnormal EMG behaviour.

Postuma *et al.* (2010) demonstrated through RR-standard deviation and low frequency components that there is a clear autonomic dysfunction in RBD participants compared to HCs. Similarly, Sorenson *et al.* (2013) illustrated attenuated sympathetic activity in heart rate variability in participants with PD and/or RBD compared to HCs [135]. Bugalho *et al.* (2018) attributed to RBD participants a reduction in parasympathetic modulation of heart rate variability in relation to sleep stage transitions [151]. It is common for HRV to be used to evaluate components of ANS functionality [152], conventionally measured by ECG sensors. These studies have demonstrated the ability of HRV attributes to differentiate RBD participants from HCs. This chapter quantified these attributes to evaluate their performance to detect RBD participants. However, while RBD participants exhibit cardiac autonomic dysfunction this might compromise the ability to achieve automated sleep staging, which was also explored in this chapter.

6.3.6 Minimal Automated Sleep Staging

Single sensor studies in sleep staging have been proposed for EOG and EMG sensors, which are equally economical and convenient substitutes for EEG sensors [85, 66, 79]. With respect to RBD detection, our previous chapters explored established objective techniques to evaluate RWSA and demonstrated how they could be improved using a combination of EMG metrics (see Chapters 4 and 5). This chapter investigated the inclusion of ECG based metrics that exploit changes in ANS, as a substitute or enhancement to RBD detection. Furthermore, using an economical combination of ECG, EOG, and EMG sensors, this chapter explored the feasibility of a simpler RBD screening tool, being minimalist while proving as effective at identifying RBD compared to conventional PSG techniques.

6.4 Methods

6.4.1 Data

The combined dataset described in Chapter 2, consisting of 100 age-matched HC (50) and RBD (50) participants, was used once again for this chapter. The

gold standard for sleep staging and the clinical diagnosis was given by clinical specialists that provided the PSG recordings. The consistency of the datasets used in Chapters 3, 4, and 5 provides the ability for this chapter to compare results with previously described techniques.

6.4.2 Pre-processing

To reduce the impact of noise and artefacts, PSG signals were pre-processed. Firstly, all ECG, EOG and EMG signals were resampled at 200Hz. The EOG signal was pre-processed with a 500^{th} order band pass FIR filter with a cut-off frequency of 0.3Hz and 40Hz. The EMG signal was filtered with a 500^{th} order notch filter at 50Hz and 60Hz (because the recordings were sourced from either Europe or Canada), in addition to a 500^{th} order band pass FIR filter between 10Hz and 100Hz. The ECG signal was filtered by a 10^{th} order Butterworth band pass filter between 5Hz and 45Hz.

6.4.3 Electrocardiogram Feature Extraction

Feature extraction for EOG and EMG signals were as described in Chapter 3. Literature that described ECG features for automated sleep staging provided motivation for this study [139, 140, 143, 144, 145, 153, 146, 147] and are summarised in this section. For this study the pre-processed ECG signal was segmented into five-minute segments, centred on each 30 second epoch (used in some HRV analysis [144, 147]). Features were then derived after Pan-Tompkins QRS detection ([148]) from each segment, providing a sequence of R-peak locations, $\boldsymbol{x} \in \mathbb{R}^{1\times P}$ (with P peaks). The interval between each adjacent R-peak then provides another sequence, $\boldsymbol{r} \in \mathbb{R}^{1\times N}$ (with N adjacent intervals). This sequence of intervals was then used to extract features for each segment. Features were then averaged across a moving 150-second sliding window, demonstrated in numerous other studies [139, 143, 145].

Amplitude Around R Peaks

For these amplitude based features, the ECG signal was first normalised by the median amplitude value at the detected R peaks. From this normalised signal the variance, standard deviation, and the mean difference was calculated for every epoch. Given a sequence of ECG amplitude values at detected R peaks, $\boldsymbol{x} \in \mathbb{R}^{1\times N}$ (with N number of peaks), the mean difference was given by:

$$a = \frac{1}{N-1}\sum_{i=1}^{N-1} x_{i+1} - x_i \tag{6.1}$$

where, a denotes the mean difference and x_i the i^{th} sample of the sequence of amplitude values derived from detected R peaks.

RR Intervals - Time Domain

From RR intervals the following features are calculated for each epoch:

- **meanRR**: Mean RR Interval (ms).
- **medianRR**: Median RR Interval (ms).
- **SDNN**: Standard deviation of RR intervals.
- **RMSSD**: Square root of mean of squares of differences between adjacent RR intervals.
- **SDSD**: Standard deviation of differences between adjacent RR intervals.
- **NN50**: Number of pairs of adjacent RR intervals differing by more than 50ms.
- **pNN50**: Percentage of NN50 (NN50/total number of RR intervals).

Additionally, the meanRR, medianRR, SDNN are recalculated with a normalised RR interval, where the RR interval series was divided by the mean RR interval.

HRV Bandwidth	Frequency Range (Hz)
Low Frequency (LF)	$0 - 0.15$
High Frequency (HF)	$0.15 - 0.40$

Table 6.1: RR frequency bands.

RR Intervals - Frequency Domain

It is thought that frequency domain features of HRV can be interpreted as sympathetic and parasympathetic activity associated with ANS [143]. The frequency bands used to describe HRV are detailed in Table 6.1, and are separated into low frequency (LF) and high frequency (HF) bands.

The total power between both bands was calculated as total power (total_power). For each band described in Table 6.1 the following features were calculated:

- **HFpeak & LFpeak**: Frequency of maximum peak in frequency band.
- **HFpower_perc & LFpower_perc**: Percentage of total power in frequency band.
- **HFpower_norm & LFpower_norm**: Percentage of low and high frequency power within band.
- **HFnorm & LFnorm**: Normalised power within band.

The ratio of the low to high frequency power was also calculated, termed the LFHF ratio.

Additionally, the total_power, peak, power_norm, norm, and LFHF ratio were recalculated with a normalised RR interval, where the RR interval series was divided by the mean RR interval.

Poincare Plot

A Poincare plot is simply a scatter plot of successive R-peak intervals, describing $RR(i)$ versus $RR(i+1)$. For this work the standard deviation of instantaneous beat-to-beat interval variability (standard deviation along the $y = -x$ direction) and the the continuous long-term RR interval variability (standard deviation along

the $y = x$ direction) was calculated [154]. The standard deviation of instantaneous beat-to-beat interval variability for a RR interval sequence, $\boldsymbol{r} \in \mathbb{R}^{1\times N}$ (with N intervals), was defined as:

$$Sd_1 = \frac{\sqrt{2}}{2}\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\Delta r_i - \overline{\boldsymbol{\Delta r}})^2} \tag{6.2}$$

$$Sd_2 = \sqrt{\left|\frac{2}{N-1}\sum_{i}^{N}(r_i - \overline{\boldsymbol{r}})^2 - Sd_1^2\right|} \tag{6.3}$$

where, r_i denotes the i^{th} sample of the heart beat interval sequence ($\boldsymbol{r}$), Δr_i denotes the i^{th} difference between adjacent heart beat intervals r_i and r_{i+1}, $\boldsymbol{\Delta r}$ denotes the mean value of Δr, and $\overline{\boldsymbol{r}}$ denotes the mean value of the heart beat interval sequence ($\boldsymbol{r}$).

Additionally, $SD1$ and $SD2$ are recalculated with a normalised RR interval, where the RR interval series was divided by the mean RR interval.

Approximate Entropy (ApEn)

Approximate entropy is an algorithm to determine the regularity of a data series.

Given a sequence, $\boldsymbol{r} \in \mathbb{R}^{1\times N}$ (with N samples), with $i = 1...N - m + 1$, two segments can be given by:

$$x_i = \{r_i, r_{i+1}, r_{i+2}...r_{i+m-1}\} \tag{6.4}$$

$$x_j = \{r_j, r_{j+1}, r_{j+2}...r_{j+m-1}\} \tag{6.5}$$

where, m denotes the length of each segment.

A distance vector is then created by:

$$d[x_i, x_j] = max_{k=1,2,...m}[|r(i+k-1) - r(j+k-1|] \tag{6.6}$$

where, $i = 1..., N - m + 1$, $j = 1...N - m + 1$, and $r[n]$ denotes the n^{th} index of the segment $\boldsymbol{r}$.

The numerator of C_m^i counts, within the resolution r, the number of blocks of consecutive values of length m which are similar to a given block. This can then be used to calculate the approximate entropy:

$$C_i^m[a] = \frac{\{number\ of\ j, j \leq N - m + 1, such\ that\ d[x_i, x_j] \leq r\}}{N - m + 1} \tag{6.7}$$

$$\Phi^m[a] = \frac{1}{N - m + 1} \sum_{i=1}^{N-m+1} \log C_i^m[a] \tag{6.8}$$

$$ApEn(m, r, N) = \Phi^m[r] - \Phi^{m+1}[r] \tag{6.9}$$

where, $m \geq 1$.

Approximate entropy measures the regularity, where low values indicate a consistent and repetitive series. High values mean irregular and complex patterns, indicative of randomness. This feature was also recalculated with a normalised RR interval, where the RR interval series was divided by the mean RR interval.

Sample Entropy (SampEn)

Sample entropy aims to estimate the randomness of a signal and is a variation on the approximate entropy [155]. The sample entropy aims to overcome the approximate entropy's dependency on the length of the signal and the bias present within the signal, where it can produce inconsistent values between datasets [155].

The sample entropy measures the chances that two segments, within a signal for m samples are similar for $m + 1$ samples. Where similarity is calculated as the absolute difference between the matched pair of segments, below a specified tolerance, r.

Given a sequence, $\boldsymbol{r} \in \mathbb{R}^{1\times N}$ (with N samples), with $i = 1...N - m + 1$, a segment can be given by:

$$u_i = \{r_i, r_{i+\tau}, r_{i+2\tau}...r_{i+(m-1)\tau}\} \tag{6.10}$$

where, m denotes the length of each segment and τ the constant time lag, which was set to 1.

A distance vector is then created by:

$$d[u_i, u_j] = max_{k=1,2,...m}[|u_i[k] - u_j[k]|] \tag{6.11}$$

where, $j = 1...N - m + 1$, $u_i[k]$ denotes the k^{th} index of the segment $\boldsymbol{u}$.

The distance vector describes the maximum absolute value found when subtracting the individual elements from two of the $\boldsymbol{u}$ vectors. The probability of then finding a distance vector within a threshold value r from the template epoch u_i is calculated as:

$$C_i^m[r] = \frac{\{number\ of\ j, j \neq i, j \leq N - m + 1, such\ that\ d[u_i, u_j] \leq r\}}{N - m + 1} \tag{6.12}$$

This count represents the number of distance vectors that are below the threshold r, divided by the total number of distance vectors, providing a probability of finding a distance vector below the threshold. This probability score is then averaged across the template segment:

$$\Phi^m[r] = \frac{1}{N - m + 1} \sum_{i=1}^{N-m+1} C_i^m[r] \tag{6.13}$$

The sample entropy is then given by:

$$SampEn(m, r, N) = -\ln[\frac{\Phi^m[r]}{\Phi^{m+1}[r]}] \tag{6.14}$$

where, $SampEn$ denotes the sample entropy.

Sample entropy depicts the logarithmic probability that vectors, which are close in m points, are also close in $m+1$ points. Where for m points, a high value suggests $m + 1$ will also be similar, indicative of a regular and consistent signal. A low value for m points means a low probability $m+1$ points will be similar, suggesting a more complex signal. This feature was also recalculated with a normalised RR interval, where the RR interval series was divided by the mean RR interval.

Recurrence Plot

The recurrence plot aims to describe the properties of a deterministic dynamic system in order to illustrate the characteristics of time dependency [156]. The first step is to construct the following $N \times N$ matrix, from a sequence, $\boldsymbol{r} \in \mathbb{R}^{1\times N}$ (with N samples):

$$R_{i,j} = \Theta\left(\varepsilon - \|r_i - r_j\|\right), i, j = 1, 2..., N \tag{6.15}$$

where, r_i denotes the i^{th} samples from $\boldsymbol{r}$, ε the predefined cutoff distance, and $\Theta(r)$ the Heaviside function.

The cutoff distance ε describes an orbit around a point, r_i, where those elements within the orbit are given a binary value of $R_{i,j} = 1$ and when outside given a value $R_{i,j} = 0$. From this matrix the following features are calculated: recurrence rate (RR), determinism (DET), entropy (ENTR), and average line length (L). The recurrence rate (RR) was calculated by:

$$RR = \frac{1}{N^2} \sum_{i,j=1}^{N} R_{i,j} \tag{6.16}$$

where, $R_{i,j}$ is the i^{th} and j^{th} element of matrix R.

The recurrence rate simply measures the density of points within the orbit. The determinism was calculated as the ratio of recurrence points on the diagonal structures compared to all recurrence points, defined as:

$$DET = \frac{\sum_{l=l_{min}}^{N} lP(l)}{\sum_{i,j=1}^{N} R_{i,j}} \tag{6.17}$$

where, $P(l)$ denotes the frequency distribution of the lengths of the diagonal structures in the recurrence plot, l_{min} the threshold excluding lines formed by tangential motion of the phase trajectory [157]. Determinism measures the predictability of a system.

Entropy uses the Shannon entropy of the frequency distribution of the diagonal line lengths and was calculated as:

$$ENTR = -\sum_{l-l_{min}}^{N} p(l) \ln p(l) \tag{6.18}$$

$$p(l) = \frac{P(l)}{\sum_{l=l_{min}}^{N} P(l)} \tag{6.19}$$

where ENTR denotes entropy.

Entropy measures the randomness of a deterministic structure in a dynamical system. Larger values indicate a more complex system. The average length of the diagonal lines is given by:

$$L = \frac{\sum_{l=l_{min}}^{N} lP(l)}{\sum_{l=l_{min}}^{N} P(l)} \tag{6.20}$$

where, the parameters are those denoted in Equation 6.17.

Teager-Kaiser Energy Operator (TKEO)

The mean of the TKEO, for a sequence of R-peak intervals, $\boldsymbol{r} \in \mathbb{R}^{1\times N}$ (with N intervals), was calculated for this work as was defined in Chapter 3 Section 3.4.4. This feature was also recalculated with a normalised RR interval, where the RR interval series was divided by the mean RR interval.

Detrended Fluctuation Analysis Exponent (DAF)

Detrended fluctuation analysis was used to determine the statistical self-similarity of a signal [158]. Given a signal, $\boldsymbol{r} \in \mathbb{R}^{1\times N}$ (with N intervals), we first calculate:

$$R_i = \sum_{i=1}^{N} (r_i - \overline{\boldsymbol{r}}) \tag{6.21}$$

where, i denotes the i^{th} sample in $\boldsymbol{r}$, $\overline{\boldsymbol{r}}$ the mean value of $\boldsymbol{r}$, and R_i is the cumulative sum.

R_i is then subdivided into windows of length t and for each window a line of best fit is found using the least squares method. The resulting piecewise sequence for a given line of fit for a window i is given by Y_i, where the fluctuation can then be defined as:

$$F(t) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(R_i - Y_i)^2} \tag{6.22}$$

where, F denotes the fluctuation and R_i the i^{th} sample of the cumulative sum for window n.

The fluctuation measurement was then repeated over a range of different window sizes, t, and a log-log graph of $F(t)$ against t was constructed. The slope of this graph, α, is termed the scaling exponent and can be derived using the least squares method. The exponent value will range between 0 and 1, where the series is considered anti-correlated, uncorrelated, and correlated for values $< \frac{1}{2}$, $\backsimeq \frac{1}{2}$, and $> \frac{1}{2}$, respectively [158].

Lempel Ziv Complexity (LZ)

The Lempel-Ziv (LZ) complexity is a measure of binary sequence complexity ([159]) and was used to show a correlation between HRV and bipolar disorder during sleep [160]. First the given sequence, $\boldsymbol{r} \in \mathbb{R}^{1\times N}$ (with N intervals) must be categorised in the following manner:

$$z(i) = \begin{cases} 0, & if\, r_i \leq T_h \\ 1, & if\, r_i > T_h \end{cases} \tag{6.23}$$

where, n denotes the length of $\mathbf{r}$, and T_h the threshold to convert the sequence to a binary sequence. For this work, the threshold (T_h) was chosen as the median RR value.

The LZ complexity, $c(n)$ can then be calculated for a sequence $\boldsymbol{z}$. The binary sequence $\boldsymbol{z}$ is scanned from left to right, where the counter, $c(n)$, is increased by one when a new sub-sequence (pattern) of consecutive characters are encountered when scanning. The counting adheres to the following rules:

1. Let S and Q denote two sequences, SQ be the concatenation of S and Q, sequence SQ_π is derived from SQ after its last character is deleted (π means the operation to delete the last character in the sequence). Let $v(SQ_\pi)$ denote the dictionary of all different sub-sequences of SQ_π. Initially, $c(n) = 1$, $S = s_1$, and $Q = s_2$, and so $SQ_\pi = s_1$.

2. Given $S = s_1 s_2, ..., s_r$, $Q = s_{r+1}$, so $SQ_\pi = s_1 s_2, ..., s$; if Q belongs to $v(SQ_\pi)$, then s_{r+1}, (equivalent to Q) is a sub-sequence of SQ_π, and so S does not change. Update Q to be $s_{r+1}s_{r+2}$, and then judge whether Q belongs to $v(SQ_\pi)$. Repeat this process until Q does not belong to $v(SQ_\pi)$.

3. Now, $Q = s_{r+1}s_{r+2}, ..., s_{r+i}$, which is not a sub-string of $SQ_\pi = s_1 s_2, ..., s_r s_{r+1}, ..., s_{r+i-1}$, thus $c(n)$ increases by one.

4. Thereafter, S is updated to be $S = s_1 s_2, ..., s_{r+i}$, and $Q = s_{r+i+1}$.

This procedure is repeated until Q is the last character and then $c(n)$ is the number of different sub-strings contained in $\boldsymbol{r}$, depicting the unique patterns within the sequence. $c(n)$ might vary with sequence length and is therefore normalised as follows:

$$C(n) = \frac{c(n)}{b(n)} \tag{6.24}$$

$$b(n) = \frac{n}{n \log_2(n)} \tag{6.25}$$

where, n denotes the length of sequence $\boldsymbol{r}$

Mutual Information (BD & PD)

Mutual information is a statistical measure of mutual dependence of two variables and was explored as a feature for ECG based automated sleep stating by Xiao *et al.* (2013) [143]. The mutual information between variable $\boldsymbol{x}$ and $\boldsymbol{y}$ is given by:

$$m(x, y) = H(x) + H(y) - H(x, y) \tag{6.26}$$

where, $H(x)$ and $H(y)$ denotes marginal entropies, and $H(x, y)$ the joint entropy of $\boldsymbol{x}$ and $\boldsymbol{y}$.

Xiao *et al.* (2013) demonstrated using the mutual information from a sequence of RR intervals and its time-delayed sequence could reflect the complexity of the RR sequence by quantifying the information flow with dependence on time delay

[143]. Given a sequence of RR intervals, $\boldsymbol{r} \in \mathbb{R}^{1\times N}$ (with N intervals), the beat and peak delay can be defined as:

$$B = m(r, r) - m(r, r + 1) \tag{6.27}$$

$$P = m(r, r) - m_{peak} \tag{6.28}$$

where, $m(r, r + k)$ denotes the mutual information of $\boldsymbol{r}$ with a delay of k samples and m_{peak} the first maximum value of $m(r, r + k)$ (where k ranges from two to seven) and represents the modulation of respiratory to HR.

Both these features depict the complexity of the RR sequence by quantifying the information and its dependence on time delay [143]. The beat decay describes the difference between the mutual information of an RR sequence with no delay and the mutual information of an RR sequence with a delay of one heart beat. The peak delay quantifies the difference of the mutual difference of an RR sequence with no delay and the mutual information of an RR sequence delayed by the peak delay that provided the maximum mutual information (between two and seven delayed heart beats).

Auto-correlation (BDa & PDa)

In a similar sense to the mutual information, Xiao *e al.* (2013) aimed to use the inter-beat auto-correlation coefficient to measure the difference between RR intervals during REM and NREM sleep [143]. For a sequence of RR intervals, $\boldsymbol{r} \in \mathbb{R}^{1\times N}$ (with N intervals), the auto-correlation beat and peak delay can be defined as:

$$a(k) = \frac{Cov[r(n), r(n + k)]}{\sqrt{D[x(n)]} \cdot \sqrt{D[x(n + k)]}} \tag{6.29}$$

where, $Cov(\cdot)$ denotes the covariance, $D(\cdot)$ the variance, and $r(n + k)$ the time-delayed series with a delay of k.

The auto-correlation beat and peak delay were then calculated as follows for each epoch:

$$B_a = a(0) - a(1) \tag{6.30}$$

$$P_a = a(0) - a_{peak} \tag{6.31}$$

where, $a(k)$ denotes the auto-correlation of $\boldsymbol{r}$ with a delay of k samples and a_{peak} the first maximum value of $a(k)$ (where k ranges from two to seven).

Zero Crossing Interval (ZCI)

The zero crossing interval (ZCI) was calculated as the number of times the sequence crosses the mean value for a given time window. The mean zero crossing interval (mZCI) was the average distance between zero crossing intervals. Lastly the normalised zero crossing interval (nZCI) was calculated as the number of zero crossing intervals divided by the mean zero crossing interval.

Wavelet Transform

As detailed previously, the DWT decomposes a discrete signal into a set of pre-specified basis functions and for this study was applied to the ECG signal. This work applied the Daubechies 5 wavelet as the basis function. For the wavelet, the DWT was applied using 10 levels of decomposition and used the maximum amplitude of the inverse DWT (for both the approximation (Clvl) and detail coefficients (Dlvl)) as the calculated feature for each epoch.

Tachycardia (tacy) and Bradycardia (brady)

In a clinical setting, heart rate is often described in beats per minutes (BPM), which is useful in evaluation and assessing conditions like tachycardia and bradycardia. Typically, tachycardia occurs in adults who have a BPM greater than 100 and bradycardia when less than 60 BPM. To provide a probability of an individual having tachycardia, the median BPM was calculated and evaluated on a normal cumulative distribution function with a mean of 120 and a standard deviation of 20. While the probability of an individual having bradycardia was calculated by evaluating median BPM on an inverted normal cumulative distribution function with a mean of 60 and a standard deviation of 20.

Irregular Index

Literature has suggested that PD is significantly co-morbid with atrial fibrillation (AF) [161], consequently a measure termed irregular evidence [162, 163] metric was used to quantify irregularly-irregular heartbeats during NREM sleep. Irregular index is a metric calculated using RR intervals and is a measure of irregular heart beat intervals. It might aid in the identification of AF, but does not prove participants are likely to have AF. However, it does illustrate the general randomness of the RR intervals.

The irregular index metric was calculated by first plotting the differential of $RR(i-2) - RR(i-1)$ against $RR(i) - RR(i-1)$ for all RR intervals from automated QRS detection, this is known as a Lorenz plots. A participant with regular heartbeats will have a very centrally located distribution of points. While a participant with regularly-irregular heartbeats will have points skewed in a certain direction of the Cartesian quadrant (e.g. a sequence associated with arrhythmia). A participant suffering from irregularly-irregular heartbeats (e.g. atrial fibrillation symptoms) would have a very scattered plot in all directions. As a result the irregular evidence metric measures the sparseness of the distribution of points.

From the Lorenz plot, a bi-variate histogram was calculated, where all points around the origin are counted and termed origin count (a point count), these represent the number of regular beat intervals (OriginCount). The irregular evidence (IrrEv) represented the number of bins that have points outside the origin space, which depicts the sparsity of irregular beat intervals (bin count, not a point count). The equation for irregular index (IrrIndex) was equal to irregular bin count minus origin bin count minus compensatory pause evidence. The compensatory pause evidence (PACEv) aims to measure beat compensatory changes [162].

Elapsed Time (HoursRec & HoursFromEnd)

As demonstrated previously, the use of elapsed time proved beneficial in automated sleep staging in Chapter 3 and was applied once again for this chapter. Elapsed time

Category	Feature	Feature Code	References
ECG - Time Domain	R-peak amplitude (mean, variance, and standard deviation)	Mean - amp_mean	
		Variance - amp_var	
		Std - amp_std	
	RR Intervals	Mean - meanRR/meanRR_norm	[143]
		Median - medianRR/medianRR_norm	
		Standard deviation - SDNN/SDNN_norm	
		Root of mean of squares of differences - RMSSD	
		Standard deviation of differences - SDSD	
		No. of adjacent pairs difference >50ms - NN50	
		Percentage of NN50 - pNN50	
	Poincare Plot	PoincareSD1/PoincareSD1_norm	[154]
		PoincareSD2/PoincareSD2_norm	
	Approximate Entropy	ApEn/ApEn_norm	
	Sample Entropy	SampEn/SampEn_norm	[155]
	Recurrence Plot	Recurrence rate (RR)	[156]
		Determinism (DET)	
		Entropy (ENTR)	
		Line length (L)	
	Teager-Kaiser Energy Operator	TKEO/TKEO_norm	
	Detrended Fluctuation Analysis Exponent	DAFa2	[158]
	Lempel Ziv Complexity	LZ	[159]
	Mutual Information	Peak delay (PD)	[143]
		Beat delay (BD)	
	Auto-correlation	Peak delay (PDa)	[143]
		Beat delay (BDa)	
	Zero crossing interval	ZCI	
		Mean - mZCI	
		Normalised - nZCI	
	Wavelet Transform	Approximation (Clvl1 to Clvl10)	
		Detail coefficient (Dlvl1 to Dlvl10)	
	Tachycardia and Bradycardia	tacy	
		brady	
	Irregular Index	OriginCount	[162, 163]
		Irregular Evidence (IrrEv)	
		Irregular Index (IrrIndex)	
		Compensatory pause evidence (PACEv)	
ECG - Frequency Domain	Total Power	total_power/total_power_norm	[143]
	High Frequency Band Power (peak and normalised)	HFpeak/HFpeak_norm	[143]
		nHF	
	Low Frequency Band Power (peak and normalised)	LFpeak/LFpeak_norm	[143]
		nLF	
	Percentage of total power	HFpower_perc	[143]
		LFpower_perc	
	Percentage of total power normalised	HFpower_norm	[143]
		LFpower_norm	
Miscellaneous	Elapsed time	Hours recorded (HoursRec)	
		Hours from end (HoursFromEnd)	

Table 6.2: ECG feature summary

(in hours) and time until end of recording (in hours) were used as features and are referenced from the first and last annotated epoch from the PSG recording, respectively.

begin

Summary of ECG Features

end

6.4.4 Minimal Sleep Stage Classification

The RF algorithm has proven effective for automated sleep staging and RBD detection (as demonstrated in Chapters 3, 4, and 5) and was once again used for this chapter. The classifier was trained to classify 30 second epochs into one of

three sleep stages (REM, NREM, and W) using a combination of 25 EOG features, 17 EMG features and 75 ECG features (as described in Section 6.3, where the number of trees was set to 500, $m_{try} = \sqrt{M}$ (rounded down) randomly selected features, where M is the total number of features). For every combination of sensors described in Table 6.3 an RF classifier was trained and tested and the results compared to evaluate REM sleep classification.

Each combination of the sleep stage classifiers was evaluated using macro-averaged sleep stage accuracy, sensitivity, specificity, and Cohen-Kappa score (three stage) by using 10-participant-fold cross-validation with an even split between healthy and RBD participants. Additionally, multi-stage classification was assessed by Cohen's Kappa [110]. These performances were compared to results obtained using EEG, EOG and EMG features described in Chapter 3 (combination Z3).

Chapter 5 demonstrated that accurate automated sleep staging (especially REM detection) yielded a high performance in RBD detection. The top performing PSG combinations were then used to evaluate automatic RBD detection using previously defined RBD metrics and ECG based metrics.

Sleep Staging Combination	# Features	# Sensors	ID
ECG	75	1	A1
EOG	25	1	B1
EMG	17	1	C1
ECG+EOG	98	2	A2
ECG+EMG	90	2	B2
EOG+EMG	40	2	C2
ECG+EOG+EMG	113	3	A3
EEG+EOG+EMG	159	3	Z3

Table 6.3: Combination of minimal PSG sensors to be used and analysed for automated sleep staging (three state). The Z3 combination is based on results from Chapter 3.

6.4.5 Extract Autonomic Dysfunction Metrics

This study aimed to use ECG based metrics that incorporate sleep architecture and HRV in order to identify RBD participants. These metrics included the variability and mean sample entropy for each REM epoch, calculated for every participant

(SampEn). Similarly the standard deviation of the RR intervals for every REM epoch was averaged for each participant (RR Std). Often low frequency and high frequency components of RR intervals are used to describe sympathetic and parasympathetic activity in the autonomic nervous system and as a result the mean peak low frequency and high frequency was calculated for all participants (LFpeak and HFpeak). To capture the changes in heart rate variation between sleep stages, the ratio between REM and NREM (RR_Index) values was calculated for the mean RR interval (see Equation 6.32). The low frequency and high frequency ratio is a popular feature used in heart rate variability, and the ratio between NREM and REM (LFHD_Index) epochs was calculated for each participant as an RBD metric (see Equation 6.33). Literature has also suggested that PD is significantly co-morbid with AF [161], consequently the metrics termed irregular index and origin count ([162, 163]) were used to quantify irregularly-irregular heartbeats during NREM (IrrIndex) and REM (OriginCount) sleep. These ECG based metrics were used to automatically identify RBD participants and were compared to metrics from Chapter 4.

$$RR_{index} = \frac{RR_{NREM} + 1}{RR_{REM} + 1} \tag{6.32}$$

where RR_{NREM} and RR_{REM} denotes the average RR interval during NREM and REM, respectively. The additional constant found in the denominator avoids the possibility of dividing by zero.

$$LFHF_{index} = \frac{LFHF_{NREM} + 1}{LFHF_{REM} + 1} \tag{6.33}$$

where $LFHF_{NREM}$ and $LFHF_{REM}$ denotes the average LFHF ratio during NREM and REM, respectively. The additional constant found in the denominator avoids the possibility of dividing by zero.

6.4.6 Minimal RBD Detection

In RBD diagnosis, literature stipulates that after the identification of REM sleep, RSWA must be confirmed visually [37]. Chapter 4 demonstrated that established features, which quantify RSWA in combination with sleep architecture, provided

RBD detection performances approaching that of the gold standard (manual clinical diagnosis) by using RF classifiers. This chapter replicated previous results and compared them to RBD detection using additional ECG based metrics in isolation and combination with EMG based metrics through RF classification.

RF classifiers were used to achieve RBD detection, as per the combinations detailed in Table 6.4. Classifiers were trained using 500 trees; $m_{try} = \sqrt{M}$ (where M is the number of features detailed in Table 6.4), and evaluated using a 10-participant-fold cross-validation scheme. In addition to the metrics detailed in our previous study, we proposed new ECG metrics that capture changes in heart rate variability between sleep stages (see Section 6.4.5).

RBD Detection Combination	# Features	# Sensors	ID
ECG	75	1	D1
EMG	25	1	E1
ECG+EMG	17	1	D2

Table 6.4: These are the minimal PSG sensors to be analysed for RBD detection in coordination with automated sleep staging.

6.4.7 Evaluation

The evaluation of automated sleep staging was measured by the 3-stage kappa score and confusion matrices. Furthermore the feature importance for the ECG based sleep stage classifier (combination A1) was derived using the mean permuted error, to determine the ECG based features necessary for sleep stage discrimination. These important features were then compared to the most effective minimal sleep stage classifier to ascertain the value of ECG based features (combination A3). The performance of RBD detection was evaluated measuring the accuracy, sensitivity, specificity, and F1 score using both manually and automatically annotated sleep stages. For each of the RF classifiers trained for RBD detection, feature importance was calculated to rank the importance of RBD metrics. Once again, for this chapter B&A plots were used to establish a level of agreement between RBD metrics, derived using manually and automatically annotated sleep stages (further details are found in Chapter 5 Section 5.4.6).

6.5 Results

The following results are presented in three sections, detailing 1) minimal automated sleep stage classification, 2) minimal RBD detection using manually/automatically annotated sleep stages, and 3) Bland and Altman plots for RBD metrics using automatically/manually annotated sleep stages.

6.5.1 Minimal Automated Sleep Staging

The best performance of automated sleep staging was provided by all three signals (EOG, ECG, and EMG) attaining an agreement score of 0.58 (0.70 and 0.48 for the individual HC and RBD cohort, respectively). These results are also comparable to the previous chapter that included EEG features (Chapter 3), shown in Figure 6.3 as combination Z3 (EEG, EOG, and EMG). Once again it is clear that automated sleep staging performs considerably better on HC participants than RBD participants, echoing the results from Chapter 3. The best performing combination of two sensors proved to be the EOG and EMG sensors (C2) followed by the EOG and ECG combination (A2), achieving a 0.57 and 0.51 kappa score, respectively. The combination of ECG and EMG features attained an agreement score of 0.41. The best single sensor performance was given by EOG features (see Figure 6.3 B1), achieving a 3-stage kappa score of 0.50, considered moderate agreement [23]. The single ECG and EMG sensor (A1 and C1, respectively) had limited success, never exceeding fair agreement (0.30 and 0.28, respectively) [23].

For ECG features, the ranked feature importance for sleep stage classification are detailed in Figure 6.4, while Figure 6.5 details importance for combined EOG, ECG, and EMG features. From Figure 6.4 it appears elapsed time is the most important, followed by ZCI and RR interval based features. In Figure 6.5 EOG features dominate the list of important features, while EMG and ECG features make an appearance in the top 30.

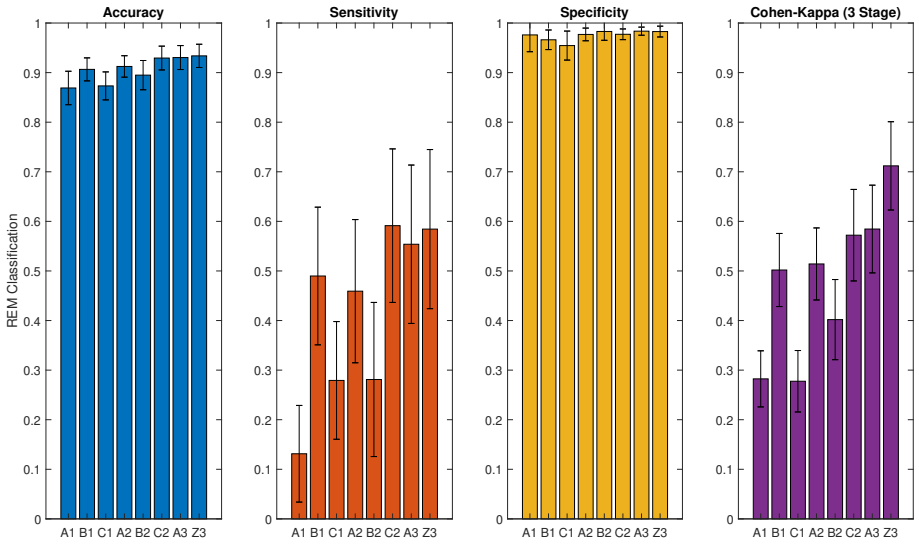


Figure 6.3: The depiction of REM detection performance for various PSG signal combinations. For a single sensor (A1, B1, and C1) the best sensitivity and F1 score was given by the EOG signal (B1). For two sensors, (A2, B2, C2), the best performance was given by EOG and EMG (C2). This exceeds the performance of a single EOG sensor (B1), which provided an increase in sensitivity, while maintaining equal specificity.

6.5.2 Minimal RBD Detection

The performances of RBD detection by ECG and/or EMG metrics are depicted in Figure 6.6. ECG metrics demonstrated they are able to distinguish RBD participants with an accuracy, sensitivity, and specificity of 0.62±0.20, 0.56±0.23, and 0.68±0.24, respectively. The combination of EMG metrics and sleep architecture (detailed in Chapter 4) achieved an accuracy, sensitivity, and specificity of 0.93 ± 0.09, 0.92 ± 0.13, and 0.94 ± 0.092, respectively. Combining EMG and ECG metrics for RBD detection provided a similar accuracy, sensitivity, and specificity of 0.93 ± 0.10, 0.94 ± 0.092, and 0.92 ± 0.13, respectively.

The ranking of ECG and/or EMG metrics for RBD detection are depicted in Figure 6.7. Simply using ECG metrics (Figure 6.7a) indicated that NREM origin count (a measure or regular heart beats), NREM irregular index (IrrIndex, a measure of irregularly-irregular heartbeats), and the ratio or NREM sleep were the most valuable. Figure 6.7b ranks EMG metrics, where the atonia index ratio (between NREM and REM), atonia index (REM), and the fractal exponent ratio

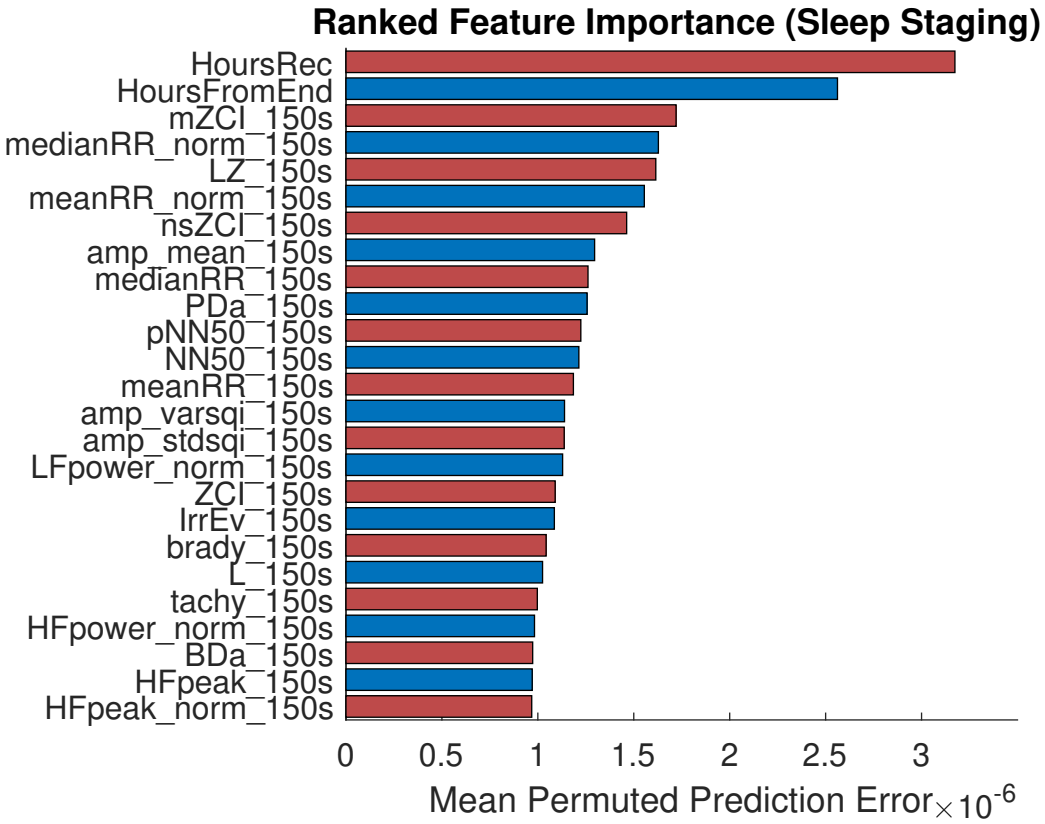


Figure 6.4: The order of feature importance for automated sleep staging (REM detection) using ECG features. The elapsed recording time represents the top two features, followed by ZCI and RR interval based features. Also highly represented were the ECG amplitude based features such as amplitude mean, variance and stdsqi.

(between NREM and REM) proved to be the most important. ECG and EMG metrics used in combination (Figure 6.7c), demonstrated that EMG metrics remain the best identifiers in a similar order to Figure 6.7b.

Figure 6.8 details the best RBD detection performance from automated sleep staging classifiers using a combination of EOG, ECG, and EMG features. The best performance (F1 score) was given by C2D2 (three unique sensors - EOG and EMG for sleep staging followed by EMG and ECG for RBD detection), followed by C2E1 combination (two sensors - EOG and EMG for sleep staging followed by EMG for RBD detection) with an accuracy, sensitivity, specificity and F1 score of 0.90 ± 0.11, 0.88 ± 0.13, 0.92 ± 0.098, and 0.90 ± 0.12, respectively. These results are similar and only marginally less than when using a cumbersome EEG combination (Chapter 5), depicted by Z3 (Figure 6.8).

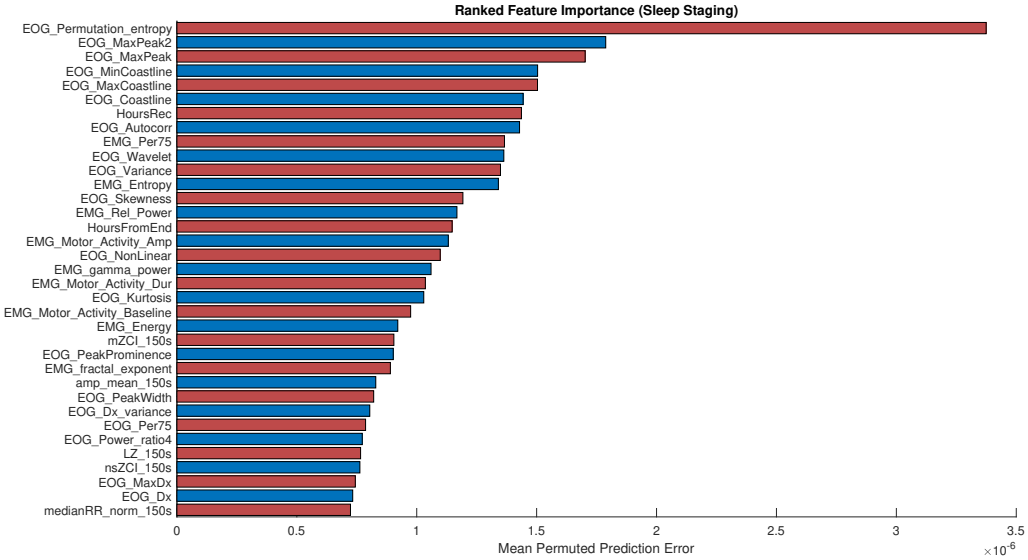


Figure 6.5: The order of feature importance for automated sleep staging (REM detection) using ECG, EOG, and EMG features. EOG features appeared most important for REM detection, specifically permutation entropy, max peak, and coastline features. This was followed by elapsed recording time and EMG features, such as the 75^{th} percentile, entropy, relative power and motor activity. These additional features from another signal provide the boost in REM detection performance shown in Figure 6.3

	W	NREM	REM
Accuracy	0.79 ± 0.089	0.72 ± 0.08	0.87 ± 0.067
Sensitivity	0.42 ± 0.17	0.90 ± 0.10	0.13 ± 0.19
Specificity	0.91 ± 0.08	0.38 ± 0.17	0.98 ± 0.068
Precision	0.60 ± 0.19	0.73 ± 0.096	0.35 ± 0.35
F1	0.45 ± 0.12	0.80 ± 0.082	0.15 ± 0.19

(a) A summary of sleep stage classification performance using ECG features.

	W	NREM	REM
Accuracy	0.85 ± 0.11	0.81 ± 0.097	0.93 ± 0.048
Sensitivity	0.64 ± 0.21	0.89 ± 0.13	0.59 ± 0.31
Specificity	0.91 ± 0.12	0.67 ± 0.18	0.98 ± 0.022
Precision	0.72 ± 0.21	0.83 ± 0.095	0.76 ± 0.21
F1	0.64 ± 0.17	0.85 ± 0.097	0.62 ± 0.27

(b) A summary of sleep stage classification performance using EOG and EMG features.

Table 6.5: Performance of automatic sleep stage classification using (a) ECG features and (b) combined EOG and EMG features. For (a) performance was relatively low compared to literature on ECG sleep staging, but interestingly this was overwhelmingly due to performance on RBD participants, where sensitivity was many times smaller than HCs. The performance of (b) was substantially better than using an ECG feature and a single EOG feature.

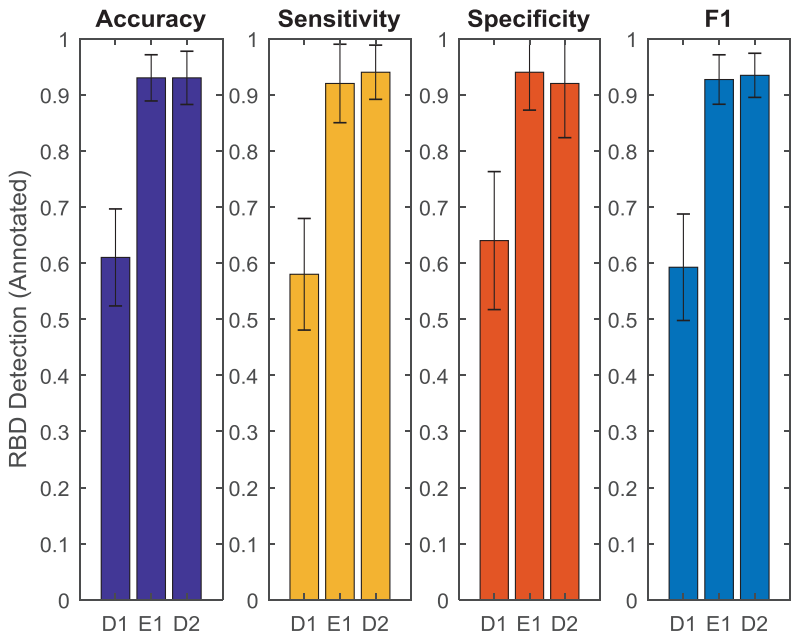


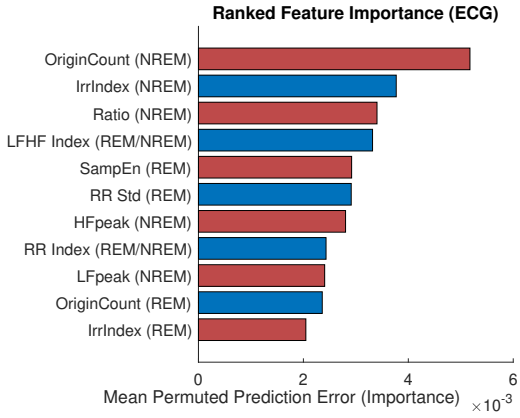
Figure 6.6: Performance of RBD detection using ECG (D1), EMG (E1), and both ECG and EMG signals (D2) with manually annotated sleep stages. With annotations ECG features (D1) prove effective at RBD identification, but are clearly outperformed by an EMG sensors (E1). Combining these sensors (D2) only improves performance marginally with respect to the F1 score.

6.5.3 Bland-Altman Analysis on RBD Metrics

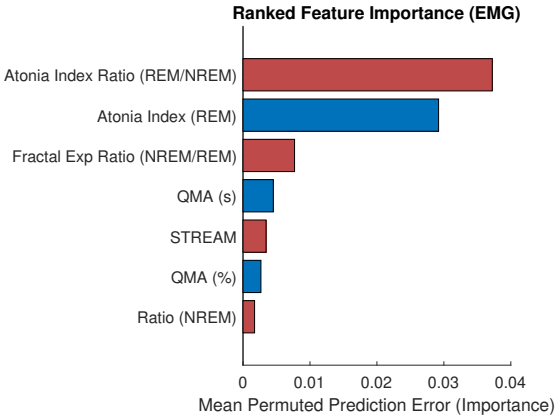
The impact of RBD metrics calculated from automated sleep staging from a single EOG and EMG was evaluated by the Bland-Altman plots given in Figures 6.9 and 6.10. A selection of important RBD metrics (provided by Figure 6.7) appear to provide agreeable results where the limits of agreement are represented by horizontal dotted lines. The mean difference bias was defined by the solid black line in Figures 6.9 and 6.10, where of the top three EMG and ECG based metrics all appear to have a slight bias when using automated sleep staging (shaded area around mean difference that includes the zero line).

6.6 Discussion

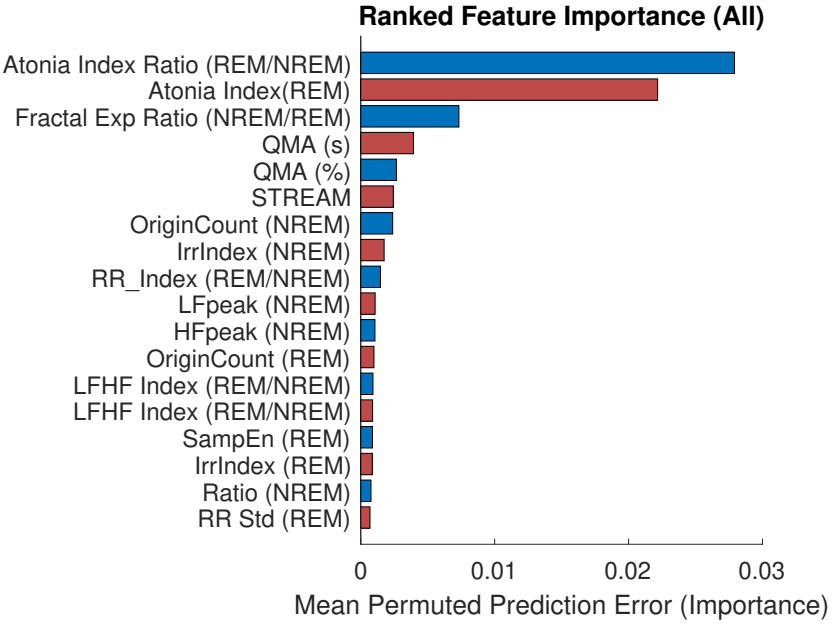
The aim of this study was to investigate a minimalist approach for identifying RBD through a fully automated pipeline. This study excluded the use of cumbersome-to-



(a) Feature Importance on ECG metrics.



(b) Feature Importance on EMG metrics.



(c) Feature Importance on ECG and EMG metrics.

Figure 6.7: Features importance for RBD detection using (a) ECG, (b) EMG and (c) both EMG and ECG metrics. From the ECG features (a), irregular evidence during NREM proves the most effective for RBD detection, followed by the LFHF index ratio between REM and NREM.

apply EEG sensors and instead focused on a combination of ECG, EOG, and EMG sensors to achieve automated three-stage sleep classification and RBD identification. The use of EOG and EMG sensors for automated sleep staging was the most effective, achieving REM detection that was comparable with literature and near human expert annotation (Figure 6.3). This was expected, given visual staging (especially REM) relies heavily on EOG and EMG signals. ECG based features did not appear to be helpful in automated sleep staging, especially for RBD participants

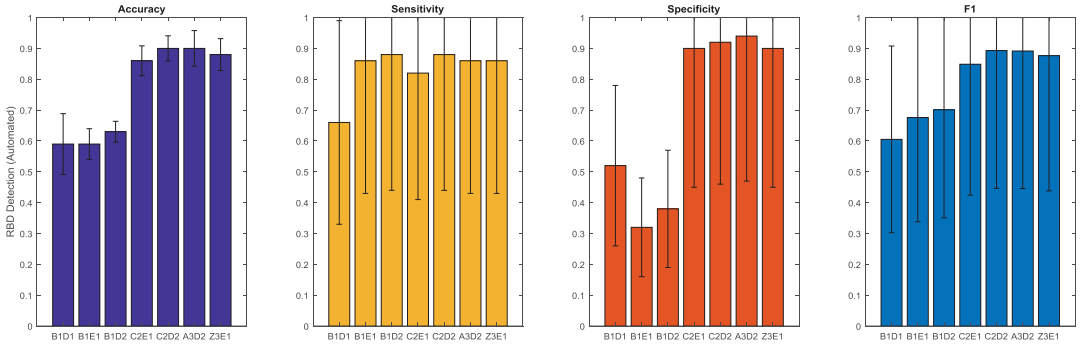
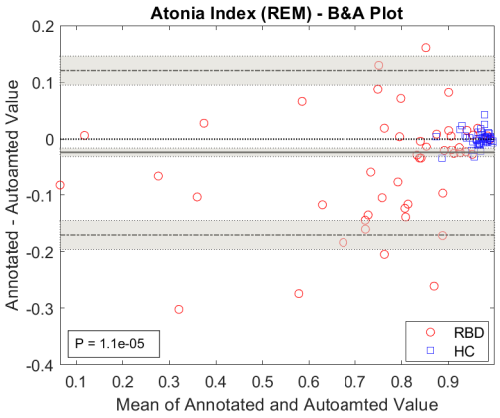


Figure 6.8: RBD detection performance when using automated sleep staging. For this figure B1, C2, and A3 represent automated sleep staging using only EOG, both EOG and EMG, or ECG, EOG and EMG features, respectively. For comparison the Z3 combination was also included, which used EEG, EOG, and EMG features. D1, E1 and D2 represent metrics derived from EMG, ECG, and both EMG and ECG sensors for RBD detection, used in combination with automated sleep staging.

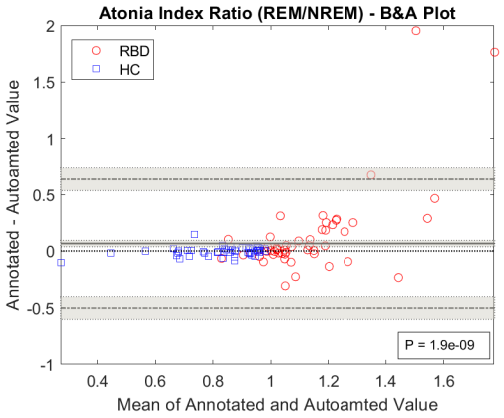
(see Table 6.5a). However, this study demonstrated the use of ECG based RBD metrics to identify RBD participants using an RF classifier with some success (Figure 6.6). Once more the performance of RBD detection remained high when using automated sleep staging from EOG and EMG features (shown in Figure 6.8 and Table 6.6). This can be attributed to the high positive prediction value and specificity provided by automated sleep staging that facilitated congruent RBD metrics (Figures 6.9 and 6.10).

As has been established in literature, the inter-rater variability of manual sleep staging is high and often exacerbated by sleep disorders that provide additional variation in sleep characteristics [22, 21]. This was also demonstrated in Chapter 3, where automated sleep staging performed better on HCs than RBD participants and was again observed in this chapter (for all combination of sensors). For a single ECG sensor, sleep staging was particularly difficult for RBD participants.

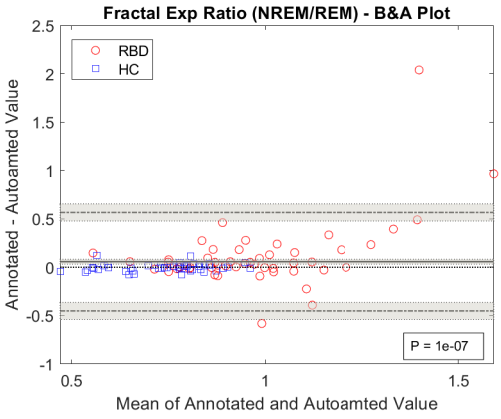
The automated three-stage classification performance of ECG features (Table 6.5a) on a combined dataset (kappa of 0.30), reflected the lower range of results achieved by similar studies (kappa scores between 0.35 and 0.73) [139, 143, 145, 147]. However these studies did benefit from focusing on relatively younger HCs



(a) Atonia Index (REM).



(b) Atonia Index Ratio (REM/NREM).



(c) Fractal Exp Ratio (NREM/REM).

Figure 6.9: Bland and Altman plots of important RBD metrics comparing scores derived from manually and automatically annotated sleep stages. These include the (a) atonia index during REM, (b) atonia index ratio between REM and NREM, and (c) fractal exponent ratio between REM and NREM. The p-value details the Kolmogorov-Smirnov test to determine if the difference between metrics calculated from manual and automatic sleep staging is from a normal distribution (rejected when < 0.05). For metrics shown above, the line of equality (the zero dotted line) does not fall within the limits of confidence around the mean difference. Therefore when using automated sleep staging the calculated metrics for RBD detection slightly shifts towards values associated with RBD or HC participants. This was unsurprising given misclassification from automated sleep staging would have an impact on the calculation of these metrics.

	Accuracy	Sensitivity	Specificity	Precision	F1
STREAM	0.62 ± 0.12	0.66 ± 0.22	0.58 ± 0.23	0.63 ± 0.22	0.62 ± 0.16
Atonia Index	0.81 ± 0.094	0.70 ± 0.16	0.92 ± 0.098	0.91 ± 0.11	0.78 ± 0.12
QMA	0.77 ± 0.13	0.66 ± 0.13	0.88 ± 0.16	0.86 ± 0.17	0.74 ± 0.14
Established Metrics (RF)	0.87 ± 0.10	$\mathbf{0.88 \pm 0.13}$	0.86 ± 0.13	0.87 ± 0.11	0.87 ± 0.10
Combined Metrics (RF)	$\mathbf{0.90 \pm 0.063}$	0.86 ± 0.13	$\mathbf{0.94 \pm 0.092}$	$\mathbf{0.95 \pm 0.082}$	$\mathbf{0.89 \pm 0.073}$

(a) A summary of end-to-end RBD detection performance using automatic sleep staging from Chapter 5.

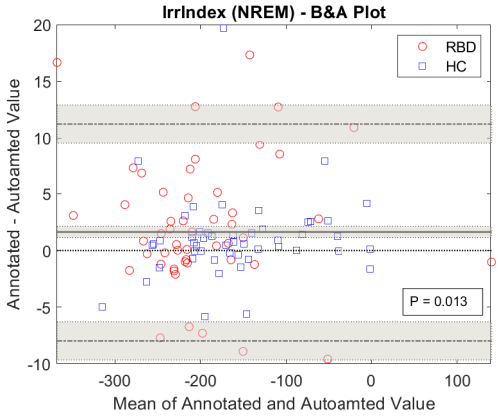
	Accuracy	Sensitivity	Specificity	Precision	F1
STREAM	0.57 ± 0.14	0.62 ± 0.21	0.52 ± 0.22	0.57 ± 0.16	0.58 ± 0.15
Atonia Index	0.83 ± 0.11	0.70 ± 0.22	$\mathbf{0.96 \pm 0.08}$	$\mathbf{0.96 \pm 0.08}$	0.78 ± 0.16
QMA	0.77 ± 0.17	0.66 ± 0.25	0.88 ± 0.16	0.88 ± 0.18	0.72 ± 0.22
Established Metrics (RF)	$\mathbf{0.89 \pm 0.094}$	$\mathbf{0.88 \pm 0.13}$	0.90 ± 0.16	0.92 ± 0.13	$\mathbf{0.89 \pm 0.093}$
Combined Metrics (RF)	0.86 ± 0.092	0.82 ± 0.17	0.90 ± 0.13	0.91 ± 0.11	0.85 ± 0.10

(b) A summary of end-to-end RBD detection performance using automatic sleep staging from the minimal combination of PSG sensors.

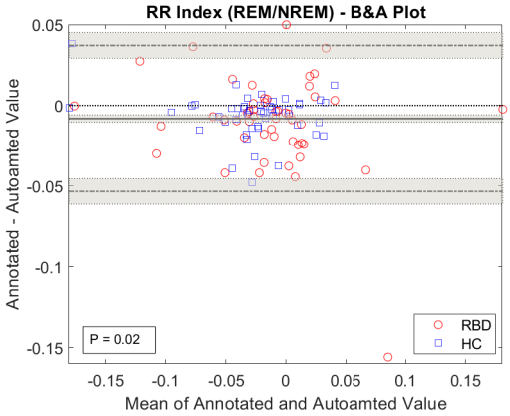
Table 6.6: RBD detection results using automatically annotated sleep stages from (a) Chapter 3 and through (b) EOG and EMG features only. Despite the reduced automated sleep staging performance using only two PSG sensors (EOG and EMG), the RBD detection performance is still comparable to RBD detection from Chapter 5, which included three PSG sensors (EEG, EOG, and EMG). However, the combined established RF classifier now marginally provides the best performance (with respect to the F1 score).

(participants of 28 to 48 people with a mean age that varies from 41 to 43). Yucelbas *et al.* (2018) achieved the highest score of 0.74, but even this technique applied to another dataset achieved a kappa of 0.43 and participants with OSA scored between 0.52 and 0.57, indicating a high degree of variability [147]. Xiao *et al.* (2013) had the added benefit of using manually annotated ECG segments, termed "stationary", achieving a score of 0.47 [143].

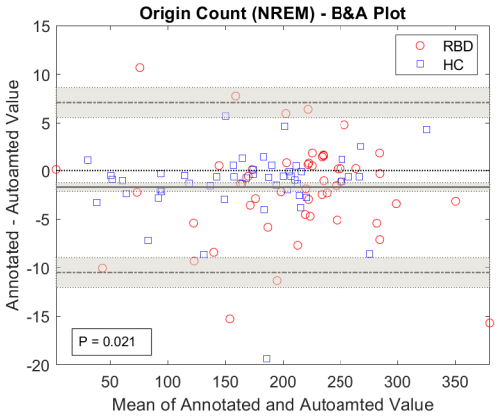
Interestingly, while this chapter achieved a kappa score of 0.30 (using only ECG features) on a mixed cohort of 100 participants, the RBD cohort severely underperformed with respect to REM detection. This was due to the misclassification of REM states for NREM and W, potentially caused by changes to HRV regulation and from movement artefacts, which could be attributed to symptoms of RBD. The feature importance for REM detection, detailed in Figure 6.4, ranks elapsed-time as more important than ECG features. Elapsed time is the progression of time since the beginning of the first annotated epoch from the PSG recording. This feature might benefit from the sleep technician that arbitrarily starts annotating sleep



(a) IrrIndex (NREM).



(b) RR Index (REM/NREM).



(c) Origin Count (NREM).

Figure 6.10: Bland and Altman plots for ECG based RBD metrics, detailing (a) irregular index during NREM, (b) RR interval index between REM and NREM, and (c) origin count during NREM. In all cases we can observe that metrics are within the limits of agreement. The p-value details the Kolmogorov-Smirnov test to determine if the difference between metrics calculated from manual and automatic sleep staging is from a normal distribution (rejected when $p < 0.05$). For metrics shown above, the line of equality (the zero dotted line) does not fall within the limits of confidence around the mean difference. For these metrics it is also apparent that automated sleep staging causes a shift in values, as was expected from sleep stage misclassification.

stages from a point they deem appropriate or around the time of lights off. While this attribute might not be directly applicable to a fully automated take-home screening device, there is the near equivalent arbitrary starting point determined by the sleep participant at around the time of lights off. Participants often provide a signature to ensure synchronisation between recording devices or simply to indicate lights off. The highest ranked ECG features included median RR intervals and ZCI, which are illustrative of HRV but ultimately these features do not contribute

towards high REM detection sensitivity.

Alternatively, EOG features provided a boost to automated sleep staging performance (kappa score of 0.50 and mean REM F1 score of 0.52), as shown in Figure 6.3. These results are comparable to EOG sleep staging literature that achieved four-state kappa scores of 0.59 ([66]) and REM F1 scores of $0.64 - 0.80$ ([164, 66, 79]). It was clearly evident that for automated sleep staging, the ability of ECG features was superseded by EOG features. Combining EMG and EOG features, further improved automated sleep staging with a kappa of 0.57 and mean REM F1 score of 0.62. Using all three sensors (EOG, EMG, and ECG) provided an almost negligible improvement (kappa score of 0.58 and mean REM F1 score of 0.60). Essentially providing no additional merit for an additional ECG sensor, which would only add cost and hassle to a potential screening support tool. The inclusion of EEG features (Z3 combination, see Figure 6.3), as per our previous chapter (Chapter 3), provided only a marginal improvement to REM detection, but markedly improved three-stage kappa, due to improved NREM and W classification (kappa score of 0.72 and mean REM F1 score of 0.62).

The feature ranking for REM classification, when using all signals, are detailed in Figure 6.5. The prevalence of EOG features only confirms their correlation with REM sleep and specifically detecting rapid eye movement through permutation entropy, max peak amplitude, and coastline features. In the ranking order, these features are closely followed by elapsed time and EMG features that measure muscle atonia (namely, amplitude percentile, entropy, and relative power). The addition of ECG features did not improve EOG, EMG, or EOG and EMG combinations, indicating that these features are often redundant and perhaps only help to reduce variation (see Figure 6.3). While ECG features did not appear fruitful in automated sleep staging, they had some success in RBD detection.

RBD classification through ECG metrics proved effective but ultimately did not outperform EMG metrics (see Figure 6.6). The ECG metric ranking for RBD detection (Figure 6.7) indicates that NREM origin count and NREM irregular index as the most informative. Literature described a loss in HRV regulation during

REM ([150, 135, 151]) and differences in atrial fibrillation ([161]) in individuals diagnosed with RBD. Combining EMG and ECG metrics provided a marginal improvement to RBD detection, with regards to the F1 score (Figure 6.6). This suggests that ECG and EMG metrics are correlated, in that a reduction of HRV regulation during REM is synonymous with a loss of muscle atonia. The impact of automated sleep staging on RBD detection was also evaluated using the most successful combination of sensors (detailed in Figure 6.3).

The performance of RBD detection using the best sensor combinations for both automated sleep staging and RBD classification are detailed in Figure 6.8. The boost in performance from using more than one sensor for sleep staging can be observed in C2E1, C2D2, and A3D2 (multiple sensors) compared to B1D1, B1E1, and B1D2 (single EOG sensor). Previously our results confirmed that C2 and A3 provide better REM classification sensitivity compared to B1 (Figure 6.3), thereby enabling better RBD detection through better REM classification. Interestingly, when using automated sleep staging (C2 – EOG and EMG), RBD detection using ECG metrics (C2E1) are equivalent to EMG metrics (C2D2), see Figure 6.8. Once again proving that there was considerable overlap in the discriminating abilities of both these metrics, where performance can be differentiated with better REM detection. In terms of achieving economy and efficiency, C2E1 used two sensors (EOG and EMG) compared to three for C2D2 and A3D2 (EOG, EMG and ECG), a potential advantage for a simple and cost-effective screening tool. Furthermore these results are comparable to results derived using EEG features illustrated by the Z3E1 combination, thereby bypassing the complications of applying EEG electrodes.

From Table 6.6 we can observe the RBD detection results using manually and automatically annotated sleep staging (using the minimalist combination of a EMG and EOG sensor). While RBD detection performance dropped slightly when using automated sleep staging, it still remained high (mean accuracy, sensitivity and specificity of 0.90, 0.88, and 0.92, respectively).

Bland-Altman analysis was used to evaluate the agreement between metric quantification using manually and automatically annotated sleep staging from the

C2 combination (EOG and EMG). Figures 6.9 and 6.10 provides the Bland-Altman plots for the best metrics for RBD detection using the EMG sensor (atonia index, fractal exponent, and atonia index ratio) and the ECG sensor (irregular index, RR interval index ratio, and NREM origin count) as detailed in Figures 6.9 and 6.10. From these plots it was clear that automated sleep staging (derived from EOG and EMG features) sufficiently produced RBD metrics that were correlated to metrics derived from manually annotated sleep staging. However we can observe a bias introduced to RBD metrics derived from automatically annotated sleep staging, where the mean difference (solid black line in Figures 6.9 and 6.10) and its' confidence interval (grey shaded area) does not include the line of equality (zero axis, dotted line through the origin). This was the direct result of automated sleep stage misclassification. When REM was misclassified for NREM and W, the impact on the calculation of RBD metrics provided a bias towards values associated with NREM and W sleep. Because automated sleep staging was poorest on RBD participants (red circles), we observed a greater variation in the y-axis (difference), especially as the x-axis values (mean value) moved towards values associated with RBD (red circles). Note that the limits of agreement are calculated by assuming that the difference between the metrics calculated from manually and automatically annotated sleep stages are normally distributed, however this does not necessarily have to be the case [134]. From these plots we can observe the resilience and fidelity of these metrics using automated sleep staging from the most successful combination of minimal PSG sensors (EOG and EMG).

These results verify the use of only an EOG and EMG sensor for automated RBD detection, enabling the potential for a minimal, cost-effective, and readily accessible take-home sleep device for RBD screening purposes. However, often epidemiological studies incur much greater costs for logistical reasons and include additional sensors to optimise their collection of data and therefore include ECG and EEG sensors. A common weakness in existing PSG are the inconsistent and intermittent signals that potentially plague recordings. By exploring various sensor combinations, this

study also demonstrated the redundancy and ability of PSG sensors to function interchangeably for both automated sleep staging and RBD detection.

While this study included a large number of participants, its application to a clinical setting are limited because it simply focused on HC and RBD cohorts. A clinical application would demand greater resilience to a myriad of sleeping disorders and other population variations that could easily share confounding RBD attributes. These include periodic limb movement of sleep, OSA, parasomnias, and severe insomnia (very fragmented sleep), which should also be included and evaluated. While ECG metrics did not prove as effective in RBD detection as EMG metrics, there was still no clear indication of which characteristic manifests earlier in RBD participants, loss of REM atonia or loss of HRV regulation during sleep. The loss of HRV regulation might even provide an early bio-marker for RBD, and could potentially offer insight into the PD conversion of RBD participants, although literature exists that suggests there is currently no correlation for the latter [150]. This study also confirmed that successful REM detection was paramount to effective RBD detection and future work to improve automated REM detection would only prove beneficial. Advances in deep learning techniques are achieving promising results and may prove useful in this context [77, 46]. Through the use of limited PSG, this study verified the feasibility of a simple and practical take-home device for sleep evaluation, of which only a limited number of clinical grade devices supported by literature exist [165, 166]. Applying such wearable tools would be the next logical step forward towards validating a fully automated RBD detection pipeline.

6.7 Conclusion

This study proved the feasibility of a fully automated pipeline for RBD detection using an EOG and EMG sensor. The primary contributions of this chapter were:

- Demonstrated automated sleep staging comparable to manual annotation, using minimal PSG sensors (EOG and EMG).

- Demonstrated that RBD detection through ECG based metrics was effective but did not out-perform EMG based metrics.
- Validated the feasibility of a fully automated RBD screening tool without an EEG sensor.

This chapter was able to bypass time-consuming manual sleep staging and cumbersome EEG sensors to achieve best-in-class RBD detection using fully-automated analysis of minimal sensors. This chapter proposes a cost-effective, practical, and simple RBD identification support tool using only two sensors (EMG and EOG).

O soft embalmer of the still midnight,
Shutting, with careful fingers and benign,
Our gloom-pleas'd eyes, embower'd from the light,
Enshaded in forgetfulness divine:

— John Keats' *Sonnet to Sleep*

7 Conclusion and Future Work

Contents

7.1 Introduction 177
7.2 Summary 178
7.2.1 Automated Sleep Stage Classification 178
7.2.2 Automated RBD Detection 179
7.2.3 End-to-End RBD Detection 180
7.2.4 Minimal End-to-End RBD Detection 181
7.3 Limitations and future work 182
7.3.1 Data 182
7.3.2 Automated Sleep Staging 182
7.3.3 RBD Detection 183
7.3.4 Take-home Settings 184
7.4 Conclusion 185

7.1 Introduction

This final chapter summarises the main contributions of the work presented in this thesis and how these have contributed to the field of automated RBD detection. A discussion on the limitations of this work is then presented, alongside potential future work in this field. Lastly the chapter details the implementation issues and hurdles needed to overcome developing an actual RBD screening tool to be used within a take-home setting.

7.2 Summary

This thesis has demonstrated the viability of machine learning techniques, fusing a range of PSG signals, and supporting the automated detection of individuals with RBD. Furthermore this research was designed to facilitate an automated take-home RBD screening device, by utilising minimal number of signals and sensors to provide an effective and efficient tool for epidemiological studies. In summary this thesis was able to demonstrate the following:

- An RF algorithm was able to sufficiently automate the sleep staging of a combined cohort of HC and RBD participants
- The performance of automated sleep staging was more difficult in RBD participants compared to HCs.
- For automated RBD detection, an RF classifier using established metrics along with additional metrics that incorporate sleep architecture delivered best-in-class performance.
- Successfully demonstrated a viable end-to-end automated RBD detection methodology, using automated sleep staging to identify RBD participants.
- The ECG signal did not aid in automated sleep staging for RBD participants. Furthermore the addition of ECG based metrics for RBD detection did not improve performance compared to established techniques.
- This thesis demonstrated that a minimum of two PSG signals (EOG and EMG) can provide sufficient end-to-end RBD detection.

7.2.1 Automated Sleep Stage Classification

While clinicians are given guidelines and rules for evaluating sleep through PSG signals, they are often not explicitly defined, which causes inter- and intra-rater variability among experts. Given this challenging problem, existing literature on automated sleep staging is vast and able to perform at near-human levels,

but these studies are often limited to young HCs. Therefore, this study explored popular feature-based automated sleep staging techniques and amalgamated popular features to evaluate their performance on a mixed cohort of older HCs and RBD participants (detailed in Chapter 3).

Automated REM classification (binary, NREM/REM) on RBD participants has only been demonstrated in two previous studies [54, 117], which achieved high performance when tested on 16 RBD participants (mixed cohort of 48 participants). These papers provided a foundation, upon which this thesis builds by performing automated multi-stage sleep classification to achieve a five-stage kappa score of 0.63 on a mixed cohort of 100 participants (50 RBD participants). With respect to REM classification, this work achieved an accuracy, sensitivity, and specificity of 0.93 ± 0.047, 0.63 ± 0.31, and 0.97 ± 0.033, respectively. As determined by the RF classifier, the most important features in automated sleep staging proved to be EEG relative spectral power (delta), EEG zero crossing rate, EEG median frequency, and EEG mean frequency. The most important EMG and EOG features came in 15^{th} (EMG entropy) and 19^{th} position (EOG maximum auto-correlation), respectively. Intuitively, these features capture visual properties perceived by clinicians when scoring PSG recordings.

7.2.2 Automated RBD Detection

After identifying REM sleep, the next major step in diagnosing RBD, is to evaluate EMG activity during these periods to observe the absence of muscle atonia. Numerous studies have aimed to quantify and objectively measure abnormal muscle activity in order to identify RBD participants and are often quite successful [102, 104, 103, 120, 125, 124, 105, 119, 84, 117]. Instances of false negatives occur when RBD participants do not exhibit violent or obvious episodes of muscle movement during REM. In Chapter 4, our research explored the use of relative changes between muscle atonia during different sleep stages, while also incorporating sleep architecture metrics to identify RBD participants.

Using perfect sleep staging from the data described in Chapter 2, Chapter 4 was able to validate metrics that had been previously tested in other independent studies for RBD detection. [104, 102, 105]. Furthermore, Chapter 4 detailed an RF classifier trained on established metrics that measure muscle activity in addition to metrics that incorporate sleep architecture, which outperformed individual metrics with an accuracy, sensitivity, and specificity of 0.94 ± 0.049, 0.94 ± 0.092, and 0.94 ± 0.092. The atonia index (REM) proved to be the most important feature from the RF classifier, followed by the N3 ratio of sleep and the atonia index ratios (N3/REM and N2/REM). The combination of numerous metrics and the inclusion of sleep architecture improved sensitivity and the F1 score compared to all tested metrics. The next natural progression was to evaluate the performance of RBD detection using automated sleep staging.

7.2.3 End-to-End RBD Detection

In Chapter 5 a fully automated end-to-end RBD detection tool was evaluated. By calculating the EMG metrics discussed in the Chapter 4 from automated sleep staging (derived from the techniques presented in Chapter 3), detected RBD with an accuracy, sensitivity, and specificity of 0.90 ± 0.063, 0.86 ± 0.13, 0.94 ± 0.092. Once again a RF classifier trained on a combination of RBD metrics outperformed individual established metrics and was able to demonstrate an imperfect automated multi-stage sleep classifier using a limited number of PSG channels was able to successfully identify RBD participants with a high degree of precision (0.95 ± 0.082).

The performance of the automated sleep stage classifier ensured a high correlation between RBD metrics calculated from manually and automatically annotated sleep staging. This was also evident in the B&A plots that detailed agreement between RBD metrics derived from manually and automatically annotated sleep stages. However, there was a clear bias introduced to RBD metrics when using automated sleep staging, with a tendency for metrics to reflect values associated with HCs. As a result RBD detection from manually annotated sleep stages was better, with a slight decrease in accuracy, sensitivity, and F1. One of the greatest challenges

of take-home sleep devices are the application of EEG sensors. This thesis then explored automated end-to-end RBD detection without EEG sensors and instead incorporated ECG features for automated sleep staging and RBD detection.

7.2.4 Minimal End-to-End RBD Detection

In Chapter 6 this thesis explored the use of an ECG sensor and the minimal set of PSG signals to achieve effective fully automated RBD detection. ECG sensors are far more prevalent and accessible when compared to EEG sensors and provided a potentially more economical and practical alternative for a take-home screening device. Once more, literature had advocated the potential use of ECG based features [150, 135, 151] for RBD identification, while automated sleep staging based on ECG features had also been well established in HCs [139, 140, 142, 143, 144, 141, 145, 147]. This thesis demonstrated the ability of ECG sensors to be used for automated sleep staging with limited success. The ability for ECG metrics to identify RBD was also explored, but did not outperform EMG based metrics.

ECG based features for automated sleep staging achieved a 3-stage kappa score of 0.28 and a REM classification accuracy, sensitivity, and specificity of 0.87 ± 0.06, 0.13 ± 0.19, and 0.98 ± 0.06, respectively. While ECG based features did not provide noticeable benefits over other sensors, the optimal minimal combination of PSG signals for end-to-end RBD detection included EOG and EMG features. These two sensors achieved a 3-stage kappa score of 0.58 and the RBD detection derived from this combination attained an accuracy, sensitivity, and specificity of 0.90 ± 0.11, 0.88 ± 0.13, and 0.92 ± 0.098, respectively. These results are comparable to those derived in Chapter 5, which had the benefit of an EEG channel. Furthermore the RBD metrics were within levels of agreement when derived between manually and automatically annotated sleep staging using this minimal combination (as detailed by B&A plot analysis).

While the complete removal of EEG sensors from a take-home screening might not be wise given the cost and logistical investment taken to complete such studies, this chapter confirmed the redundancy of such sensors. Often take-home screening

devices are plagued with intermittent sensors that can fall off or simply fail for numerous reasons. This work demonstrated that PSG sensors can be used interchangeably to achieve end-to-end RBD detection.

7.3 Limitations and future work

In this thesis the methods presented for automated sleep staging and RBD detection work well for the datasets that were analysed. However there are several limitations which are discussed below, alongside future directions of research and how to implement a potential RBD screening tool.

7.3.1 Data

While the collection of RBD PSG recordings used in this thesis, was one of the largest in literature, it presented some limitations. By using participants that simply contained RBD and HC individuals, this limited the potential application in a clinical setting. Once more the datasets were also limited in that sex ratios were not matched and evidence suggests there are physiological differences in sleep between sexes [24]. Therefore this work and future work would benefit from including PSG recordings from individuals with confounding sleep disorders and from a better distribution of male and female participants.

7.3.2 Automated Sleep Staging

This thesis merged commonly used features for PSG recordings in a powerful RF classifier to achieve automated sleep stage classification on a mixed cohort of HC and RBD participants. The variability in RBD participants meant these individuals performed poorer compared to HCs. A further extension of this work would be to apply this classifier on a broader range of individuals with various other sleep disorders. This would emulate clinical practise, where additional confounds from a mixed population (that includes other sleep disorders) would increase the difficulty of the problem.

Recent literature on automated sleep staging is dominated by deep neural networks, which are producing state-of-the-art performances on HCs [77, 113, 167, 168, 114, 115, 116, 46]. Furthermore these techniques are also proving equally successful on participants with sleep disorders such as narcolepsy, hypersomnia, and PD [83, 82], implying these strategies are able to adjust for variability amongst participants. These techniques utilise multiple layers of linear and non-linear processing *'neurons'* to learn hierarchical representations from input data [77]. Some of these techniques take advantage of convolutional neural networks that learn multiple filters that are used to convolve with input data to extract time-invariant features [77, 168, 115]. A few studies demonstrated the use of recurrent neural networks, which condition outputs based on all previous inputs, as they retain memory and utilise feedback connections to learn temporal information from sequences of input data (emulating human visual techniques in sleep staging) [113]. These techniques are data hungry and demand a large training set to produce effective automated sleep staging results. The studies mentioned above included datasets that range from 62 to $10,000$ PSG recordings, indicating the expected size for satisfactory deep neural network. Therefore the PSG recordings used in this thesis (100 in total) would prove sufficient for exploring deep neural networks, however the inclusion of participants diagnosed with RBD might complicate the effectiveness of such models and would prove interesting for future work. In most cases the ability to interrogate the neural network for details on how classification was determined is not possible, which is unattractive from a clinical perspective that demands transparency. Nonetheless, the results from neural networks are unmatched, but as this work has demonstrated, perfect automated sleep staging might not be necessary for effective RBD detection.

7.3.3 RBD Detection

This thesis presented an effective automated technique to detect RBD participants. However, these techniques would benefit from being applied to data comprising

of numerous other sleep disorders, especially those that might mimic RBD characteristics. Potentially confounding disorders include PLM, OSA, parasomnias, and severe insomnia (very fragmented sleep). Short comings to RBD detection performance can be attributed to imperfect automated sleep staging and also failure to classify instances of RBD participants having no episodes of abnormal muscle movement during REM sleep. This was to be expected from a disorder that displays symptoms intermittently (20% in night-to-night variability [120]) and provides an argument towards developing a take-home PSG device that could be applied over multiple nights.

There is evidence to suggest EMG signals provided from the flexor digitorum superficialis and exterior digitorum brevis best capture phasic EMG activity and might help increase the sensitivity of RSWA detection, especially in the case of mild RBD episodes [136]. Video monitoring of sleep might also prove advantageous, by capturing signature momentary movements and vocalisations, localised to periods of REM sleep [169]. Zhang *et al.* 2008 manually generated video based metrics to measure complex movements (punching, sitting up, kicking, and jumping), vocalisations (muttering, talking, laughing, and shouting), and simple motor movements (twitching) [169]. This study concluded that video based metrics in combination with PSG improved the RBD diagnostic ability [169].

7.3.4 Take-home Settings

In collaboration with Aintree University Hospital (Liverpool), Royal Hallamshire Hospital (Sheffield), Royal Papworth Hospital (Cambridge), and John Radcliffe Hospital a current study is being undertaken to use sleep-monitoring-kits to record sleep from 100 RBD participants and 50 HCs. As an extension of the work used in this thesis, the research aims to further explore automated sleep staging and the ability to diagnose RBD, all within a take-home setting. These sleep-monitoring-kits are given to participants every 6 months to provide several nights and include EEG, EOG, EMG, pulse oximeter, ECG, chest and tummy bands (measure breathing

effort), and video monitoring (to visually confirm movements). An example of the kit being worn is given in Figure 7.1.

Multiple night recordings will also provide an opportunity to explore and detect changes in participants, enabling an opportunity to predict which people with RBD are at higher risk of developing PD. This work would aim to extend the work from this thesis to a take-home environment, which will provide its own complications and difficulties, but would further establish the feasibility of a possible self-administered take-home RBD screening tool.



Figure 7.1: An example of the new sleep study that utilises a take-home sleep recording device to analyse HC and RBD participants across the UK. Photographed by Claire El Mouden in 2018.

7.4 Conclusion

This thesis has investigated the design of an automated RBD diagnostic support-tool, in the aim to facilitate the development of a take-home screening tool to identify individuals with RBD. Without take-home recordings, this was explored using minimal PSG sensors from a cohort of 50 healthy individuals and 50 RBD participants, one of the largest collections of data in literature of its' kind. This work focused on the diagnostic procedure of clinicians by applying signal processing

and machine learning techniques to a minimal combination of PSG signals. This procedure begins with automated sleep staging to identify clinically relevant sleep stages on a mixed cohort of HC and RBD participants. Followed by automated RBD detection, achieved by evaluating muscle movement in combination with sleep architecture. Lastly, this work evaluated minimal PSG signal combinations that would produce effective RBD detection to enable a simple, cost-effective, and practical screening tool for epidemiological studies.

Appendices

Sweet dreams are made of this
Who am I to disagree?
I travel the world
And the seven seas,
Everybody's looking for something.

Lennox and Stewart's *Sweat Dreams (Are Made of This)*

A Additional Algorithms

Contents

A.1 Signal Decomposition 189
A.1.1 Discrete Fourier Transform (DFT) 189
A.1.2 Discrete Wavelet Transform (DWT) 190

A.1 Signal Decomposition

A.1.1 Discrete Fourier Transform (DFT)

The DFT decomposes a time series signal into its' frequency components by approximating the signal through, potentially, an infinite sum of sinusoidal waves of varying frequency, amplitude, and phase. For a signal, $\boldsymbol{x} \in \mathbb{R}^{1\times N}$ (where N is the length of the signal), the DFT, $\mathcal{FT}\{\boldsymbol{x}\}$, is given by:

$$\mathcal{FT}\{\boldsymbol{x}\} \rightarrow X_k = \sum_{n=0}^{N-1} x_n e^{-j\frac{2\pi}{N}kn} \tag{A.1}$$

where x_n is the n^{th} sample from $\boldsymbol{x}$, $j = \sqrt{-1}$, and $k \in \mathbb{N} : 0 \leq k \leq N-1$.

This is efficiently implemented using the Fast Fourier Transform algorithm [170]. From the DFT, the periodogram estimate of the power spectral density can be defined for a signal $\boldsymbol{x}$:

$$P_k = \frac{1}{N.f_s}|X_k|^2 \tag{A.2}$$

where f_s is the sampling frequency, X_k is the value of the DFT of $\boldsymbol{x}$ at k, and $k \in \mathbb{N} : 0 \leq k \leq N-1$.

To compensate for leakage caused by the calculation of the periodogram for a finite length, the time series signal is multiplied by a window function to help taper the edges of the signal. For this study the Hamming window was used.

A.1.2 Discrete Wavelet Transform (DWT)

While the DFT decomposes a signal into a set of sinusoidal signals, the Discrete Wavelet Transform (DWT) similarly decomposes a signal into a set of basis functions (wavelets). When calculating the wavelet transform, a windowing function is used to analyse short segments of the signal. These windowing functions, called wavelets, can be scaled and translated in order to provide a variable time-frequency resolution.

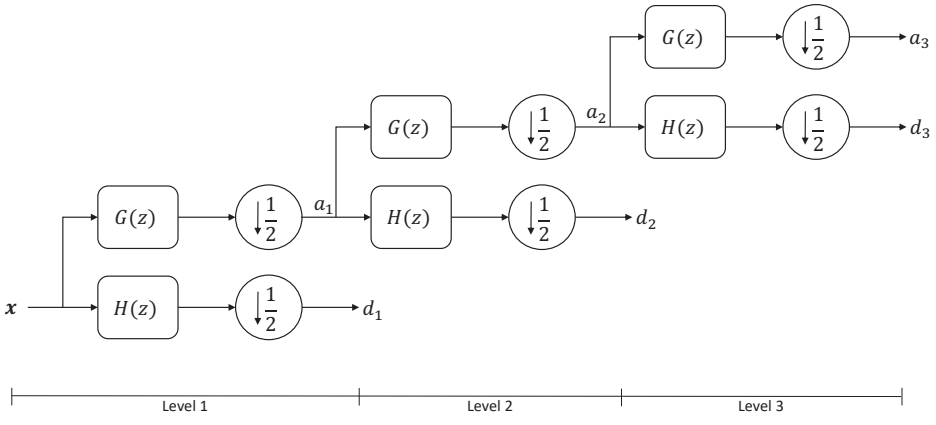


Figure A.1: The DWT and the cascading filter bank to achieve three levels ($l = 3$) of decomposition. The low-pass and high-pass filters are described by $G(z)$ and $H(z)$, respectively, where the outputs are sub-sampled by two. The detail (d_l) and approximation (a_l) coefficients are illustrated after each level.

To achieve computational efficiency, the DWT was calculated using a filter bank, as detailed in Figure A.1. The signal is split into two channels by a low-pass and high-pass filter, also known as the scaling ($H(z)$) and wavelet filter ($G(z)$), respectively. The outputs of these filters are then sub-sampled by a factor of two. The output from the low-pass filter are termed the detail coefficients, while the output from the high-pass filter are the approximation coefficients. The approximation coefficients can be further processed with the same set of filters, this can be achieved multiple times through cascading filters to provide multiple levels, l, of approximation and detail coefficients. At each level the frequency resolution is doubled, while the time resolution is halved [171]. The parameters of the low-pass and high-pass filters are determined by the mother-wavelet, where a dictionary of wavelets exists for numerous specialised applications.

He that is thy friend indeed,
He will help thee in thy need:
If thou sorrow, he will weep;
If thou wake, he cannot sleep;
Thus of every grief in heart
He with thee doth bear a part.

— Richard Barnfield's *An Ode*

References

[1] M. H. Kryger, T. Roth, and W. C. Dement. *Principles and Practice of Sleep Medicine*. 5th. Elsevier, 2011.

[2] B. Y. Mileykovskiy, L. I. Kiyashchenko, T. Kodama, Y. Y. Lai, and J. M. Siegel. "Activation of pontine and medullary motor inhibitory regions reduces discharge in neurons located in the locus coeruleus and the anatomical equivalent of the midbrain locomotor region." In: *The Journal of neuroscience* 20.22 (2000), pp. 8551–8.

[3] B. Y. Mileykovskiy, L. I. Kiyashchenko, and J. M. Siegel. "Cessation of activity in red nucleus neurons during stimulation of the medial medulla in decerebrate rats." In: *The Journal of physiology* 545.Pt 3 (2002), pp. 997–1006.

[4] C. H. Schenck, J. Y. Montplaisir, B. Frauscher, B. Hogl, J. Gagnon, R. Postuma, K. Sonka, P. Jennum, M. Partinen, I. Arnulf, V. C. D. Cock, Y. Dauvilliers, P. Luppi, A. Heidbreder, G. Mayer, F. Sixel-döring, C. Trenkwalder, M. Unger, P. Young, Y. K. Wing, L. Ferini-strambi, R. Ferri, Y. Y. Lai, M. Pavlova, C. Saper, P. Schmidt, J. M. Siegel, C. Singer, E. S. Louis, A. Videnovic, and W. Oertel. "Rapid eye movement sleep behavior disorder : devising controlled active treatment studies for symptomatic and neuroprotective therapy — a consensus statement from the International Rapid Eye Movement Sleep Behavior Disorder Study Group". In: *Sleep Medicine* 14.8 (2013), pp. 795–806.

[5] A. Iranzo, J. Santamaria, and E. Tolosa. "The clinical and pathophysiological relevance of REM sleep behavior disorder in neurodegenerative diseases". In: *Sleep Medicine Reviews* 13.6 (2009), pp. 385–401.

[6] D. Claassen, K. Josephs, J. Ahlskog, M. Silber, M. Tippmann-Peikert, and B. Boeve. "REM sleep behavior disorder preceding other aspects of synucleinopathies by up to half a century(e–Pub ahead of print)(CME)". In: *Neurology* 75.6 (2010), pp. 494–499.

[7] R. B. Postuma, J. F. Gagnon, M. Vendette, M. L. Fantini, J. Massicotte-Marquez, and J. Montplaisir. "Quantifying the risk of neurodegenerative disease in idiopathic REM sleep behavior disorder". In: *Neurology* (2009).

[8] B. F. Boeve, M. H. Silber, T. J. Ferman, J. A. Lucas, and J. E. Parisi. "Association of REM Sleep Behavior Disorder and Neurodegenerative Disease May Reflect an Underlying Synucleinopathy". In: *Movement Disorders* 16.4 (2001), pp. 622–630.

[9] B. F. Boeve, M. H. Silber, T. J. Ferman, S. C. Lin, E. E. Benarroch, A. M. Schmeichel, J. E. Ahlskog, R. J. Caselli, S. Jacobson, M. Sabbagh, C. Adler, B. Woodruff, T. G. Beach, A. Iranzo, E. Gelpi, J. Santamaria, E. Tolosa, C. Singer, D. C. Mash, C. Luca, I. Arnulf, C. Duyckaerts, C. H. Schenck, M. W. Mahowald, Y. Dauvilliers, N. R. Graff-Radford, Z. K. Wszolek, J. E. Parisi, B. Dugger, M. E. Murray, and D. W. Dickson. "Clinicopathologic correlations in 172 cases of rapid eye movement sleep behavior disorder with or without a coexisting neurologic disorder". In: *Sleep Medicine* 14.8 (2013), pp. 754–762.

[10] M. C. Chen, H. Yu, Z.-L. Huang, and J. Lu. "Rapid eye movement sleep behavior disorder." In: *Current Opinion in Neurobiology* 23.5 (2013), pp. 793–798.

[11] A. Iranzo, J. Santamaria, and E. Tolosa. "Idiopathic rapid eye movement sleep behaviour disorder : diagnosis , management , and the need for neuroprotective interventions". In: *The Lancet Neurology* 15.4 (2016), pp. 405–419.

[12] J.-x. Yang and L. Chen. "Economic Burden Analysis of Parkinson ' s Disease Patients in China". In: *Parkinson's Disease* 2017 (2017).

[13] M. Goedert, R. Jakes, and M. G. Spillantini. "The Synucleinopathies: Twenty Years on". In: *Journal of Parkinson's Disease* 7.s1 (2017), S53–S71.

[14] M. Martí, E. Tolosa, and J. Campdelacreu. "Clinical Overview of Thesynucleopathies". In: *Movement disorders : official journal of the Movement Disorder Society* 18 Suppl 6 (2003), S28–33.

[15] I. G. McKeith, D. W. Dickson, J. Lowe, M. Emre, J. T. O'Brien, H. Feldman, J. Cummings, J. E. Duda, C. Lippa, E. K. Perry, D. Aarsland, H. Arai, C. G. Ballard, B. Boeve, D. J. Burn, D. Costa, T. Del Ser, B. Dubois, D. Galasko, S. Gauthier, C. G. Goetz, E. Gomez-Tortosa, G. Halliday, L. A. Hansen, J. Hardy, T. Iwatsubo, R. N. Kalaria, D. Kaufer, R. A. Kenny, A. Korczyn, K. Kosaka, V. Lee, A. Lees, I. Litvan, E. Londos, O. L. Lopez, S. Minoshima, Y. Mizuno, J. A. Molina, E. B. Mukaetova-Ladinska, F. Pasquier, R. H. Perry, J. B. Schulz, J. Q. Trojanowski, and M. Yamada. "Diagnosis and management of dementia with Lewy bodies". In: *Neurology* 66.9 (2006), p. 1455.

[16] P. L. Lantos and M. I. Papp. "Cellular pathology of multiple system atrophy: A review". In: *Journal of Neurology Neurosurgery and Psychiatry* 57.2 (1994), pp. 129–133.

[17] A. B. Neikrug and S. Ancoli-Israel. "Diagnostic tools for REM sleep behavior disorder". In: *Sleep Medicine Reviews* 16.5 (2012), pp. 415–429.

[18] P. Jennum, J. S. Cano, C. Bassetti, P. Clarenbach, J. Mathis, R. Poirrier, K. Sonka, E. Svanborg, L. D. Groselj, D. Kaynak, M. Kruger, a. Papavasiliou, and Z. Zahariev. "Sleep disorders in neurodegenerative disorders and stroke". In: *European Handbook of Neurological Management: Volume 1* 1 (2011), pp. 529–543.

[19] R. Berry, R. Brooks, C. Gamaldo, S. Harding, R. Lloyd, C. Marcus, and B. Vaughn. *The AASM Manual for Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications*. 2.0. American Academy of Sleep Medicine, 2012.

[20] Rechtschaffen and Kales. *Rechtschaffen Kales - A Manual of Standardized Terminology Techniques and Scoring System for Sleep Stages of Human Subjects*. 1968.

[21] H. Danker-Hopfe, D. Kunz, G. Gruber, G. Klösch, J. L. Lorenzo, S. L. Himanen, B. Kemp, T. Penzel, J. Röschke, H. Dorn, a. Schlögl, E. Trenker, and G. Dorffner. "Interrater reliability between scorers from eight European sleep laboratories in subjects with different sleep disorders." In: *Journal of sleep research* 13.1 (2004), pp. 63–69.

[22] H. Danker-Hopfe, P. Anderer, J. Zeitlhofer, M. Boeck, H. Dorn, G. Gruber, E. Heller, E. Loretz, D. Moser, S. Parapatics, B. Saletu, A. Schmidt, and G. Dorffner. "Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard". In: *Journal of Sleep Research* 18.1 (2009), pp. 74–84.

[23] J. R. Landis and G. G. Koch. "The Measurement of Observer Agreement for Categorical Data". In: *Biometrics* 33.1 (1977), p. 159.

[24] G. Luca, J. Haba Rubio, D. Andries, N. Tobback, P. Vollenweider, G. Waeber, P. Marques Vidal, M. Preisig, R. Heinzer, and M. Tafti. "Age and gender variations of sleep in subjects without sleep disorders." In: *Annals of Medicine* 47.6 (2015), pp. 482–91.

[25] S. Redline, S. Redline, H. L. Kirchner, H. L. Kirchner, S. F. Quan, S. F. Quan, D. J. Gottlieb, D. J. Gottlieb, V. Kapur, V. Kapur, A. Newman, and A. Newman. "The effects of age, sex, ethnicity, and sleep-disordered breathing on sleep architecture." In: *Arch Intern Med* 164 (2004), pp. 406–418.

[26] J. Carrier, S. Land, D. J. Buysse, D. J. Kupfer, and T. H. Monk. "The effects of age and gender on sleep EEG power spectral density in the middle years of life (ages 20-60 years old)." In: *Psychophysiology* 38.2 (2001), pp. 232–242.

[27] N. Darchia, I. G. Campbell, and I. Feinberg. "Rapid eye movement density is reduced in the normal elderly." In: *Sleep* 26 (2003), pp. 973–977.

[28] D. L. Bliwise, L. He, F. P. Ansari, and D. B. Rye. "Quantification of electromyographic activity during sleep: a phasic electromyographic metric." In: *Journal of clinical neurophysiology* 23.1 (2006), pp. 59–67.

[29] M. M. Ohayon, M. Caulet, and R. G. Priest. "Violent behavior during sleep." In: *The Journal of clinical psychiatry* 58.8 (Aug. 1997), 369–76, quiz 377.

[30] H. F. Chiu and Y. K. Wing. "REM sleep behaviour disorder: an overview." In: *International journal of clinical practice* 51.7 (Oct. 1997), pp. 451–4.

[31] C. H. Schenck and M. W. Mahowald. "REM sleep behavior disorder: clinical, developmental, and neuroscience perspectives 16 years after its formal identification in SLEEP." In: *Sleep* 25.2 (2002), pp. 120–138.

[32] B. Högl, A. Stefani, and A. Videnovic. "Idiopathic REM sleep behaviour disorder and neurodegeneration - An update". In: *Nature Reviews Neurology* 14.1 (2018), pp. 40–56.

[33] T. M. Bovolenta, S. Maria, C. D. Azevedo, R. A. Saba, V. Borges, H. B. Ferraz, and A. C. Felicio. "Systematic Review and Critical Analysis of Cost Studies Associated with Parkinson ' s Disease". In: *Parkinson's Disease* 2017.3410946 (2017), p. 11.

[34] L. Ballesteros and P. Mir. “The social impact of Parkinson’s disease in Spain: Report by the Spanish Foundation for the Brain”. In: *Neurologia* 31.6 (2016), pp. 357–430.

[35] Y. Winter, M. Balzer-geldsetzer, A. Spottke, J. P. Reese, E. Baum, J. Klotsche, and J. Rieke. “Longitudinal study of the socioeconomic burden of Parkinson’ s disease in Germany”. In: *European Journal of Neurology* 17.9 (2010), pp. 1156–1163.

[36] C. H. Schenck, S. R. Bundlie, M. G. Ettinger, and M. W. Mahowald. “Chronic behavioral disorders of human REM sleep: a new category of parasomnia.” In: *Sleep* 9.2 (June 1986), pp. 293–308.

[37] M. J. Sateia. “International classification of sleep disorders-third edition highlights and modifications”. In: *Chest* 146.5 (2014), pp. 1387–1394.

[38] R. Eye, M. Sleep, and B. Disorder. “Rapid Eye Movement Sleep Behavior Disorder”. In: 4 (2007), pp. 1833–1841.

[39] R. B. Postuma and J. Montplaisir. “Update on the pharmacology of REM sleep behavior disorder”. In: *Neurology* 67.5 (2006), pp. 742–747.

[40] I. Arnulf. “REM Sleep Behavior Disorder : Motor Manifestations and Pathophysiology”. In: *Movement Disorders* 27.6 (2012), pp. 677–689.

[41] D. Kunz and F. Bes. “Melatonin as a Therapy in REM Sleep Behavior Disorder Patients : An Open-Labeled Pilot Study on the Possible Influence of Melatonin on REM-Sleep Regulation”. In: 14.3 (1999), pp. 507–511.

[42] D. Kunz and R. Mahlberg. “Sleep and melatonin A two-part , double-blind , placebo-controlled trial of exogenous melatonin in REM sleep behaviour disorder”. In: *Journal of Sleep Research* 19.4 (2010), pp. 591–596.

[43] E. J. Olson, B. F. Boeve, M. H. Silber, and M. H. Silber. “Rapid eye movement sleep behaviour disorder : demographic , clinical and laboratory findings in 93 cases”. In: *Brain* 123.2 (2000), pp. 331–339.

[44] N. Cooray, F. Anderotti, C. Lo, M. Symmonds, M. T. Hu, and M. De Vos. “Automating the Detection of REM Sleep Behaviour Disorder”. In: *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2018, pp. 1460–3.

[45] H. Phan, F. Andreotti, N. Cooray, O. Y. Chén, and M. De Vos. “Joint Classification and Prediction CNN Framework for Automatic Sleep Stage Classification”. In: *IEEE Transactions on Biomedical Engineering* (2018), pp. 1–11.

[46] H. Phan, F. Andreotti, N. Cooray, O. Y. Ch, and D. Vos. “SeqSleepNet : End-to-End Hierarchical Recurrent Neural Network for Sequence-to-Sequence Automatic Sleep Staging”. In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 27.3 (2019), pp. 400–410.

[47] N. Cooray, F. Andreotti, C. Lo, M. Symmonds, and M. T. M. Hu. “Detection of REM sleep behaviour disorder by automated polysomnography analysis”. In: *Clinical Neurophysiology* 130.4 (2019), pp. 505–514.

[48] N. Cooray, F. Andreotti, C. Lo, M. Symmonds, M. T. M. Hu, and M. De Vos. "Screening for REM Sleep Behaviour Disorder with Minimal Sensors (In Review)". In: *Clinical Neurophysiology* (2020), pp. 1–22.

[49] B. Frauscher, D. Gabelia, M. Biermayr, A. Stefani, H. Hackner, T. Mitterling, W. Poewe, and B. Högl. "Validation of an integrated software for the detection of rapid eye movement sleep behavior disorder." In: *Sleep* 37.10 (2014), pp. 1663–71.

[50] M. Cesari, J. A. Christensen, L. Kempfner, A. N. Olesen, G. Mayer, K. Kesper, W. H. Oertel, F. Sixel-Döring, C. Trenkwalder, H. B. Sorensen, and P. Jennum. "Comparison of computerized methods for rapid eye movement sleep without atonia detection". In: *Sleep* 41.10 (2018), pp. 1–11.

[51] C. O'Reilly, N. Gosselin, J. Carrier, and T. Nielsen. "Montreal Archive of Sleep Studies : an open-access resource for instrument benchmarking and exploratory research". In: *Journal of Sleep Research* (2014), pp. 628–635.

[52] M. G. Terzano, L. Parrino, A. Sherieri, R. Chervin, S. Chokroverty, C. Guilleminault, M. Hirshkowitz, M. Mahowald, H. Moldofsky, A. Rosa, R. Thomas, and A. Walters. "Atlas, rules, and recording techniques for the scoring of cyclic alternating pattern (CAP) in human sleep." In: *Sleep medicine* 2.6 (Nov. 2001), pp. 537–53.

[53] C. H. Schenck, T. D. Hurwitz, and M. W. Mahowald. "REM sleep behaviour disorder: an update on a series of 96 patients and a review of the world literature". In: *Journal of Sleep Research* 2.4 (1993), pp. 224–231.

[54] J. Kempfner, P. Jennum, M. Nikolic, J. A. E. Christensen, and H. B. D. Sorensen. "Automatic detection of REM sleep in subjects without atonia." In: *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. Vol. 2012. 2012, pp. 4242–5.

[55] J. Kempfner, P. Jennum, H. B. D. Sorensen, J. A. E. Christensen, and M. Nikolic. "Automatic Sleep Staging: From young adults to elderly patients using multi-class support vector machine". In: *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2013, pp. 5777–5780.

[56] R. Agarwal and J. Gotman. "Computer-assisted sleep staging". In: *IEEE Transactions on Biomedical Engineering* 48.12 (2001), pp. 1412–1423.

[57] S. Güneş, K. Polat, and S. Şebnem Yosunkaya. "Efficient sleep stage recognition system based on EEG signal using k-means clustering based feature weighting". In: *Expert Systems with Applications* 37.12 (2010), pp. 7922–8.

[58] S. Khalighi, T. Sousa, G. Pires, and U. Nunes. "Automatic sleep staging: A computer assisted approach for optimal combination of features and polysomnographic channels". In: *Expert Systems with Applications* 40.17 (2013), pp. 7046–7059.

[59] B. Koley and D. Dey. "An ensemble system for automatic sleep stage classification using single channel EEG signal". In: *Computers in Biology and Medicine* 42.12 (2012), pp. 1186–95.

[60] V. Bajaj and R. B. Pachori. "Automatic classification of sleep stages based on the time-frequency image of EEG signals". In: *Computer Methods and Programs in Biomedicine* 112.3 (2013), pp. 320–8.

[61] G. Zhu, Y. Li, and P. P. Wen. "Analysis and classification of sleep stages based on difference visibility graphs from a single-channel EEG signal". In: *IEEE Journal of Biomedical and Health Informatics* 18.6 (2014), pp. 1813–1821.

[62] T. Sousa, A. Cruz, S. Khalighi, G. Pires, and U. Nunes. "A two-step automatic sleep stage classification method with dubious range detection". In: *Computers in Biology and Medicine* 59 (2015), pp. 42–53.

[63] T. Lajnef, S. Chaibi, P. Ruby, P.-E. Aguera, J.-B. Eichenlaub, M. Samet, A. Kachouri, and K. Jerbi. "Learning machines and sleeping brains: Automatic sleep stage classification using decision-tree multi-class support vector machines". In: *Journal of Neuroscience Methods* 250 (2015), pp. 94–105.

[64] C. Stepnowsky, D. Levendowski, D. Popovic, I. Ayappa, and D. M. Rapoport. "Scoring accuracy of automated sleep staging from a bipolar electroocular recording compared to manual scoring by multiple raters". In: *Sleep Medicine* 14.11 (2013), pp. 1199–1207.

[65] S. F. Liang, C. E. Kuo, Y. H. Hu, and Y. S. Cheng. "A rule-based automatic sleep staging method". In: *Journal of Neuroscience Methods* 205.1 (2012), pp. 169–76.

[66] J. Virkkala, R. Velin, S.-L. Himanen, A. Värri, K. Müller, and J. Hasan. "Automatic sleep stage classification using two facial electrodes." In: *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. Vol. 2008. 2008, pp. 1643–6.

[67] D. Álvarez-Estévez, J. M. Fernández-Pastoriza, E. Hernández-Pereira, and V. Moret-Bonillo. "A method for the automatic analysis of the sleep macrostructure in continuum". In: *Expert Systems with Applications* 40.5 (2013), pp. 1796–1803.

[68] A. Krakovská and K. Mezeiová. "Automatic sleep scoring: A search for an optimal combination of measures". In: *Artificial Intelligence in Medicine* 53.1 (2011), pp. 25–33.

[69] L. Fraiwan, K. Lweesy, N. Khasawneh, H. Wenz, and H. Dickhaus. "Automated sleep stage identification system based on time-frequency analysis of a single EEG channel and random forest classifier". In: *Computer Methods and Programs in Biomedicine* 108.1 (2012), pp. 10–9.

[70] P. Memar and F. Faradji. "A Novel Multi-Class EEG-Based Sleep Stage Classification System". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 26.1 (2018), pp. 84–95.

[71] B. Şen, M. Peker, A. Çavuşoğlu, and F. V. Çelebi. "A comparative study on classification of sleep stage based on EEG signals using feature selection and classification algorithms". In: *Journal of Medical Systems* 38.3 (2014).

[72] M. M. Rahman, M. I. H. Bhuiyan, and A. R. Hassan. "Sleep stage classification using single-channel EOG". In: *Computers in Biology and Medicine* 102.June (2018), pp. 211–220.

[73] T. L. da Silveira, A. J. Kozakevicius, and C. R. Rodrigues. "Single-channel EEG sleep stage classification based on a streamlined set of statistical features in wavelet domain". In: *Medical and Biological Engineering and Computing* 55.2 (2017), pp. 343–352.

[74] A. R. Hassan and M. I. H. Bhuiyan. "Computer-aided sleep staging using Complete Ensemble Empirical Mode Decomposition with Adaptive Noise and bootstrap aggregating". In: *Biomedical Signal Processing and Control* 24 (2016), pp. 1–10.

[75] A. R. Hassan and M. I. H. Bhuiyan. "Automated identification of sleep states from EEG signals by means of ensemble empirical mode decomposition and random under sampling boosting". In: *Computer Methods and Programs in Biomedicine* 140 (2017), pp. 201–210.

[76] S. F. Liang, C. E. Kuo, Y. C. Lee, W. C. Lin, Y. C. Liu, P. Y. Chen, F. Y. Cherng, and F. Z. Shaw. "Development of an EOG-Based Automatic Sleep-Monitoring Eye Mask". In: *IEEE Transactions on Instrumentation and Measurement* 64.11 (2015), pp. 2977–2985.

[77] A. Supratak, H. Dong, C. Wu, and Y. Guo. "DeepSleepNet : A Model for Automatic Sleep Stage Scoring Based on Raw". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 25.11 (2017), pp. 1998–2008.

[78] S. A. Imtiaz and E. Rodriguez-Villegas. "A Low Computational Cost Algorithm for REM Sleep Detection Using Single Channel EEG". In: *Annals of Biomedical Engineering* 42.11 (2014), pp. 2344–2359.

[79] B. D. Yetton, M. Niknazar, K. A. Duggan, E. A. McDevitt, L. N. Whitehurst, N. Sattari, and S. C. Mednick. "Automatic detection of rapid eye movements (REMs): A machine learning approach". In: *Journal of Neuroscience Methods* 259 (2016), pp. 72–82.

[80] S. Herscovici, A. Pe'er, S. Papyan, and P. Lavie. "Detecting REM sleep from the finger: an automatic REM sleep algorithm based on peripheral arterial tone (PAT) and actigraphy". In: *Physiological Measurement* 28.2 (2007), pp. 129–140.

[81] D. E. McCarty, P. Y. Kim, C. Frilot, A. L. Chesson, and A. a. Marino. "The Fingerprint of Rapid Eye Movement: Its Algorithmic Detection in the Sleep Electroencephalogram Using a Single Derivation." In: *Clinical EEG and Neuroscience* 47.4 (2014), pp. 298–304.

[82] A. Patanaik, J. L. Ong, J. J. Gooley, S. Ancoli-Israel, and M. W. Chee. "An end-to-end framework for real-time automatic sleep stage classification". In: *Sleep* 41.5 (2018), pp. 1–11.

[83] A. Malafeev, D. Laptev, S. Bauer, X. Omlin, A. Wierzbicka, A. Wichniak, W. Jernajczyk, R. Riener, J. Buhmann, and P. Achermann. "Automatic human sleep stage scoring using deep neural networks". In: *Frontiers in Neuroscience* 12.November (2018), pp. 1–15.

[84] J. Kempfner, P. Jennum, M. Nikolic, J. A. E. Christensen, and H. B. D. Sorensen. "SLEEP phenomena as an early biomarker for Parkinsonism". In: *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2013, pp. 5773–6.

[85] J. Virkkala, J. Hasan, V. Alpo, S.-l. Himanen, and M. Kiti. "Automatic detection of slow wave sleep using two channel electro-oculography". In: *Journal of Neuroscience Methods* 160.1 (2007), pp. 171–177.

[86] S. Charbonnier, L. Zoubek, S. Lesecq, and F. Chapotot. “Self-evaluated automatic classifier as a decision-support tool for sleep/wake staging”. In: *Computers in Biology and Medicine* 41.6 (2011), pp. 380–389.

[87] M. Ronzhina, O. Janoušek, J. Kolářová, M. Nováková, P. Honzík, and I. Provazník. “Sleep scoring using artificial neural networks”. In: *Sleep Medicine Reviews* 16.3 (2012), pp. 251–263.

[88] J. Y. Tian and J. Q. Liu. “Automated Sleep Staging by a Hybrid System Comprising Neural Network and Fuzzy Rule-based Reasoning.” In: *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference* 4 (2005), pp. 4115–8.

[89] U. R. Acharya, E. C.-P. Chua, K. C. Chua, L. C. Min, and T. Tamura. “Analysis and Automatic Identification of Sleep Stages Using Higher Order Spectra”. In: *International Journal of Neural Systems* 20.6 (2010), pp. 509–521.

[90] Y. L. Hsu, Y. T. Yang, J. S. Wang, and C. Y. Hsu. “Automatic sleep stage recurrent neural classifier using energy features of EEG signals”. In: *Neurocomputing* 104 (2013), pp. 105–14.

[91] H. G. Jo, J. Y. Park, C. K. Lee, S. K. An, and S. K. Yoo. “Genetic fuzzy classifier for sleep stage identification”. In: *Computers in Biology and Medicine* (2010).

[92] R. K. Sinha. “Artificial neural network and wavelet based automated detection of sleep spindles, REM sleep and wake states”. In: *Journal of Medical Systems* 32.4 (2008), pp. 291–299.

[93] Y. Wang, K. a. Loparo, M. R. Kelly, and R. F. Kaplan. “Evaluation of an automated single-channel sleep staging algorithm”. In: (2015), pp. 101–111.

[94] B. Frauscher, A. Iranzo, C. Gaig, V. Gschliesser, M. Guaita, V. Raffelseder, L. Ehrmann, N. Sola, M. Salamero, E. Tolosa, W. Poewe, J. Santamaria, and B. Höogl. “Normative EMG Values during REM Sleep for the Diagnosis of REM Sleep Behavior Disorder.” In: *Sleep* 35.6 (2012), pp. 835–847.

[95] B. Hjorth. “EEG analysis based on time domain properties”. In: *Electroencephalography and Clinical Neurophysiology* 29.3 (1970), p. 4.

[96] S. T.-B. Hamida and B. Ahmed. “Computer based sleep staging: Challenges for the future”. In: *2013 7th IEEE GCC Conference and Exhibition (GCC)*. IEEE, Nov. 2013, pp. 280–285.

[97] J. Pardey, S. Roberts, and L. Tarassenko. “A review of parametric modelling techniques for EEG analysis”. In: *Medical Engineering and Physics* 18.1 (1996), pp. 2–11.

[98] C. Vidaurre, N. Krämer, B. Blankertz, and A. Schlögl. “Time Domain Parameters as a feature for EEG-based Brain-Computer Interfaces”. In: *Neural Networks* 22.9 (2009), pp. 1313–1319.

[99] J. F. Kaiser. “International Conference on Acoustics, Speech, and Signal Processing”. In: *International Conference on Acoustics, Speech, and Signal Processing*. Vol. 2. 10. Albuquerque: IEEE, 1990, pp. 381–384.

[100] J. F. Kaiser. “Some useful properties of Teager’s energy operators”. In: *International Conference on Acoustics, Speech and Signal Processing*. Vol. 3. Minneapolis: IEEE, 1993, pp. 149–152.

[101] C. Bandt and B. Pompe. “Permutation Entropy: A Natural Complexity Measure for Time Series”. In: *Physical Review Letters* 88.17 (Apr. 2002), p. 174102.

[102] J. W. Burns, F. B. Consens, R. J. Little, K. J. Angell, S. Gilman, and R. D. Chervin. “EMG variance during polysomnography as an assessment for REM sleep behavior disorder.” In: *Sleep* 30.12 (2007), pp. 1771–8.

[103] R. Ferri, C. Franceschini, M. Zucconi, S. Vandi, F. Poli, O. Bruni, C. Cipolli, P. Montagna, and G. Plazzi. “Searching for a marker of REM sleep behavior disorder: submentalis muscle EMG amplitude analysis during sleep in patients with narcolepsy/cataplexy.” In: *Sleep* 31.10 (2008), pp. 1409–1417.

[104] R. Ferri, F. Rundo, M. Manconi, G. Plazzi, O. Bruni, A. Oldani, L. Ferini-Strambi, and M. Zucconi. “Improved computation of the atonia index in normal controls and patients with REM sleep behavior disorder”. In: *Sleep Medicine* 11.9 (2010), pp. 947–949.

[105] R. Frandsen, M. Nikolic, and M. Zoetmulder. “Motor activity in REM sleep Analysis of automated quanti fi cation of motor activity in REM sleep behaviour disorder”. In: *Journal of Sleep Research* 24.5 (2015), pp. 583–590.

[106] R. Moddemeijer. “On estimation of entropy and mutual information of continuous distributions”. In: *Signal Processing* 16.3 (1989), pp. 233–248.

[107] L. Breiman. “Bagging predictions”. In: *Machine Learning* 24.2 (1996), pp. 123–140.

[108] L. Breiman. “Random Forests”. In: *Machine learning* 45.1 (2001), pp. 5–32.

[109] S. Arlot and A. Celisse. “A survey of cross-validation procedures for model selection”. In: *Statistics Surveys* 4 (2010), pp. 40–79.

[110] J. Cohen. “A Coefficient of Agreement for Nominal Scales”. In: *Educational and Psychological Measurement* XX.1 (1960), pp. 37–46.

[111] K. Susmáková and A. Krakovská. “Discrimination ability of individual measures used in sleep stages classification.” In: *Artificial intelligence in medicine* 44.3 (Nov. 2008), pp. 261–77.

[112] A. Tsanas, M. A. Little, P. E. Mcsharry, and L. O. Ramig. “Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson ’ s disease symptom severity”. In: *Journal of the Royal Society Interface* (2010).

[113] S. Biswal, H. Sun, B. Goparaju, M. Brandon Westover, J. Sun, and M. T. Bianchi. “Expert-level sleep scoring with deep neural networks”. In: *Journal of the American Medical Informatics Association* 25.12 (2018), pp. 1643–1650.

[114] A. N. Olesen, P. Jennum, P. Peppard, E. Mignot, and H. B. Sorensen. “Deep residual networks for automatic sleep stage classification of raw polysomnographic waveforms”. In: *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS* 2018-July.1 (2018), pp. 3713–3716.

[115] A. Sors, S. Bonnet, S. Mirek, L. Vercueil, and J. F. Payen. "A convolutional neural network for sleep stage scoring from raw single-channel EEG". In: *Biomedical Signal Processing and Control* 42 (2018), pp. 107–114.

[116] J. B. Stephansen, A. N. Olesen, M. Olsen, A. Ambati, E. B. Leary, H. E. Moore, O. Carrillo, L. Lin, F. Han, H. Yan, Y. L. Sun, Y. Dauvilliers, S. Scholz, L. Barateau, B. Hogl, A. Stefani, S. C. Hong, T. W. Kim, F. Pizza, G. Plazzi, S. Vandi, E. Antelmi, D. Perrin, S. T. Kuna, P. K. Schweitzer, C. Kushida, P. E. Peppard, H. B. Sorensen, P. Jennum, and E. Mignot. "Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy". In: *Nature Communications* 9.1 (2018), pp. 1–15.

[117] J. Kempfner, H. B. D. Sorensen, M. Nikolic, and P. Jennum. "Early Automatic Detection of Parkinson's Disease Based on Sleep Recordings". In: *Journal of clinical neurophysiology* 31.5 (2014), pp. 409–415.

[118] J. Kempfner, G. L. Sorensen, H. B. D. Sorensen, and P. Jennum. "Automatic REM sleep detection associated with idiopathic rem sleep Behavior Disorder." In: *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference* 2011 (2011), pp. 6063–6.

[119] J. Kempfner, G. Sorensen, M. Zoetmulder, P. Jennum, and H. B. D. Sorensen. "REM behaviour disorder detection associated with neurodegenerative diseases." In: *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference* 2010 (2010), pp. 5093–6.

[120] R. Ferri, S. Marelli, F. I. I. Cosentino, F. Rundo, L. Ferini-Strambi, and M. Zucconi. "Night-to-night variability of automatic quantitative parameters of the chin EMG amplitude (atonia index) in REM sleep behavior disorder". In: *Journal of Clinical Sleep Medicine* 9.3 (2013), pp. 253–258.

[121] R. Ferri, S. Marelli, L. Ferini-Strambi, A. Oldani, F. Colli, C. H. Schenck, and M. Zucconi. "An observational clinical and video-polysomnographic study of the effects of clonazepam in REM sleep behavior disorder". In: *Sleep Medicine* (2013).

[122] R. Ferri, M. Zucconi, S. Marelli, G. Plazzi, C. H. Schenck, and L. Ferini-Strambi. "Effects of long-term use of clonazepam on nonrapid eye movement sleep patterns in rapid eye movement sleep behavior disorder". In: *Sleep Medicine* (2013).

[123] O. Lapierre and J. Montplaisir. "Polysomnographic features of REM sleep behavior disorder: development of a scoring method." In: *Neurology* 42.7 (July 1992), pp. 1371–4.

[124] J. Montplaisir, J. F. Gagnon, M. L. Fantini, R. B. Postuma, Y. Dauvilliers, A. Desautels, S. Rompre, and J. Paquet. "Polysomnographic diagnosis of idiopathic REM sleep behavior disorder". In: *Movement Disorders* 25.13 (2010), pp. 2044–2051.

[125] Y. Dauvilliers, S. Rompré, J. F. Gagnon, M. Vendette, D. Petit, J. Montplaisir, and S. Rompre. "REM sleep characteristics in narcolepsy and REM sleep behavior disorder". In: *Sleep* 30.7 (2007), pp. 844–849.

[126] C. E. Shannon. "Communication in the Presence of Noise". In: *Proceedings of the IRE* 37.1 (1949), pp. 10–21.

[127] J. Massicotte-marquez, J. Carrier, A. De, A. Mathieu, D. Petit, and J. Montplaisir. "Slow-wave Sleep and Delta Power in Rapid Eye Movement Sleep Behavior Disorder". In: *Annals of Neurology* 57.2 (2005), pp. 277–282.

[128] X. Zhang, Z. Song, J. Ye, Y. Fu, J. Wang, L. Su, X. Zhu, M. Zhang, Y. Cheng, W. Wu, and R. Xue. "Polysomnographic and neuropsychological characteristics of rapid eye movement sleep behavior disorder patients". In: *Brain and Behavior* 9.3 (2019), pp. 1–9.

[129] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.

[130] R. Sokal and F. Rohlf. *Biometry: the principles and practice of statistics in biological research. 2nd ed.* Vol. 133. 2012.

[131] J. Mourão-Miranda, D. R. Hardoon, T. Hahn, A. F. Marquand, S. C. Williams, J. Shawe-Taylor, and M. Brammer. "Patient classification as an outlier detection problem: An application of the One-Class Support Vector Machine". In: *NeuroImage* 58.3 (2011), pp. 793–804.

[132] J. M. Bland and D. G. Altman. "Statistical Methods For Assessing Agreement Between Two Methods of Clinical Measurement". In: *The Lancet* 327.8476 (1986), pp. 307–310.

[133] J. M. Bland and D. G. Altman. "Comparing methods of measurement : why plotting difference against standard method is misleading". In: *The Lancet* 346.8982 (1995), pp. 1085–1087.

[134] J. M. Bland and D. G. Altman. "Measuring agreement in method comparison studies with heteroscedastic measurements". In: *Statistical Methods in Medical Research* 8.2 (1999), pp. 135–160.

[135] G. L. Sorensen, J. Mehlsen, and P. Jennum. "Reduced sympathetic activity in idiopathic rapid-eye-movement sleep behavior disorder and Parkinson's disease". In: *Autonomic Neuroscience: Basic and Clinical* 179.1-2 (2013), pp. 138–141.

[136] B. Frauscher, A. Iranzo, B. Högl, J. Casanova-Molla, M. Salamero, V. Gschliesser, E. Tolosa, W. Poewe, and J. Santamaria. "Quantification of electromyographic activity during REM sleep in multiple muscles in REM sleep behavior disorder". In: *Sleep* 31.5 (2008), pp. 724–731.

[137] M. de Zambotti, J. Trinder, A. Silvani, I. M. Colrain, and F. C. Baker. "Dynamic coupling between the central and autonomic nervous systems during sleep: A review". In: *Neuroscience and Biobehavioral Reviews* 90.February (2018), pp. 84–103.

[138] R. Gordan, J. K. Gwathmey, and L.-H. Xie. "Autonomic and endocrine control of cardiovascular function". In: *World Journal of Cardiology* 7.4 (2015), p. 204.

[139] S. J. Redmond and W. T. Mcnicholas. "Sleep staging using cardiorespiratory signals". In: 256 (2007), pp. 245–256.

[140] M. O. Mendez, M. Matteucci, V. Castronovo, L. Ferini-Strambi, S. Cerutti, and A. M. Bianchi. "Sleep staging from Heart Rate Variability : time-varying spectral features and Hidden Markov Models Matteo Matteucci Vincenza Castronovo and Luigi Ferini-Strambi Sergio Cerutti and Anna Maria Bianchi". In: 3 (2010), pp. 246–263.

[141] J. M. Kortelainen, M. O. Mendez, A. M. Bianchi, M. Matteucci, and S. Cerutti. "Sleep Staging Based On Signals Acquired Through Bed Sensor". In: *IEEE Transactions on Information Technology in Biomedicine* 14.3 (2010), pp. 776–785.

[142] Y. Kurihara and K. Watanabe. "Sleep-stage decision algorithm by using heartbeat and body-movement signals". In: *IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans* 42.6 (2012), pp. 1450–1459.

[143] M. Xiao, H. Yan, J. Song, Y. Yang, and X. Yang. "Biomedical Signal Processing and Control Sleep stages classification based on heart rate variability and random forest". In: *Biomedical Signal Processing and Control* 8.6 (2013), pp. 624–633.

[144] F. Ebrahimi, S.-k. Setarehdan, and H. Nazeran. "Biomedical Signal Processing and Control Automatic sleep staging by simultaneous analysis of ECG and respiratory signals in long epochs". In: *Biomedical Signal Processing and Control* 18 (2015), pp. 69–79.

[145] P. Fonseca, X. Long, M. Radha, R. Haakma, R. M. Aarts, and J. Rolink. "Sleep stage classification with ECG and respiratory effort". In: *Physiological Measurement* 36.10 (2015), pp. 2027–2040.

[146] J. Malik, Y.-L. Lo, and H.-t. Wu. "Sleep-wake classification via quantifying heart rate variability by convolutional neural network". In: (2018).

[147] Ş. Yücelbaş, C. Yücelbaş, G. Tezel, S. Özşen, and Ş. Yosunkaya. "Automatic sleep staging based on SVD , VMD , HHT and morphological features of single-lead ECG signal". In: *Expert Systems with Applications* 102 (2018), pp. 193–206.

[148] J. Pan and W. J. Tompkins. "Real-Time Qrs Detection Algorithm." In: *IEEE Transactions on Biomedical Engineering* BME-32.3 (1985), pp. 230–236.

[149] F. Andreotti, O. Carr, M. A. F. Pimentel, A. Mahdi, and M. De Vos. "Comparing Feature Based Classifiers and Convolutional Neural Networks to Detect Arrhythmia from Short Segments of ECG". In: *2017 Computing in Cardiology Conference (CinC)* 44 (2018), pp. 1–4.

[150] R. B. Postuma, P. A. Lanfranchi, H. Blais, J.-F. Gagnon, and J. Y. Montplaisir. "Cardiac autonomic dysfunction in idiopathic REM sleep behavior disorder". In: *Movement Disorders* 25.14 (2010), pp. 2304–2310.

[151] P. Bugalho, M. Mendonça, T. Lampreia, R. Miguel, R. Barbosa, and M. Salavisa. "Heart rate variability in Parkinson disease and idiopathic REM sleep behavior disorder". In: *Clinical Autonomic Research* 28.6 (2018), pp. 557–564.

[152] F. Andreotti, H. Phan, N. Cooray, C. Lo, M. T. M. Hu, and M. D. Vos. "Multichannel Sleep Stage Classification and Transfer Learning using Convolutional Neural Networks". In: (2018), pp. 171–174.

[153] H. Yoon, S. H. Hwang, J.-w. Choi, Y. J. Lee, D.-u. Jeong, and K. S. Park. "REM sleep estimation based on autonomic dynamics using R – R intervals REM sleep estimation based on autonomic dynamics using R – R intervals". In: (2017).

[154] B. Roy and S. Ghatak. "Métodos não-lineares para avaliar mudanças na variabilidade da frequência cardíaca em pacientes com diabetes tipo 2". In: *Arquivos Brasileiros de Cardiologia* 101.4 (2013), pp. 317–326.

[155] J. S. Richman and J. R. Moorman. "Physiological time-series analysis using approximate entropy and sample entropy maturity in premature infants Physiological time-series analysis using approximate entropy and sample entropy". In: *Americal Journal of Physiology Heart and Circulatory Physiology* 278 (2000), H2039–H2049.

[156] J. P. Eckmann, O. Oliffson Kamphorst, and D. Ruelle. "Recurrence plots of dynamical systems". In: *Epl* 4.9 (1987), pp. 973–977.

[157] X. Li, G. Ouyang, X. Yao, and X. Guan. "Dynamical characteristics of pre-epileptic seizures in rats with recurrence quantification analysis". In: *Physics Letters, Section A: General, Atomic and Solid State Physics* 333.1-2 (2004), pp. 164–171.

[158] R. M. Bryce and K. B. Sprague. "Revisiting detrended fluctuation analysis". In: *Scientific Reports* 2 (2012).

[159] J. Ziv and A. Lempel. "On the Complexity of Finite Sequences". In: *IEEE Transactions on Information Theory* IT-22.1 (1976).

[160] M. Migliorini, M. O. Mendez, and A. M. Bianchi. "Study of heart rate variability in bipolar disorder: Linear and non-linear parameters during sleep". In: *Frontiers in Neuroengineering* 4.January (2012), pp. 1–7.

[161] C. T. Hong, L. Chan, D. Wu, W. T. Chen, and L. N. Chien. "Association between Parkinson's disease and atrial fibrillation: A population-based study". In: *Frontiers in Neurology* 10.FEB (2019).

[162] S. Sarkar, D. Ritscher, and R. Mehra. "Communications A Detector for a Chronic Implantable Atrial". In: *IEEE Transactions on Biomedical Engineering* 55.3 (2008), pp. 1219–1224.

[163] J. Oster and G. D. Clifford. "Impact of the presence of noise on RR interval-based atrial fibrillation detection". In: *Journal of Electrocardiology* 48.6 (2015), pp. 947–951.

[164] R. Agarwal, T. Takeuchi, S. Laroche, and J. Gotman. "Detection of rapid-eye movements in sleep studies". In: *IEEE Transactions on Biomedical Engineering* 52.8 (2005), pp. 1390–1396.

[165] K. B. Mikkelsen, D. B. Villadsen, M. Otto, and P. Kidmose. "Automatic sleep staging using ear-EEG". In: *BioMedical Engineering OnLine* (2017), pp. 1–15.

[166] A. Sterr, J. K. Ebajemito, K. B. Mikkelsen, M. A. Bonmati-Carrion, N. Santhi, G. Atzori, V. Revell, S. Debener, D.-J. Dijk, and M. De Vos. "Reliable detection of sleep stages derived from behind-the-ear electrodes (cEEGrid): A comparison to standard polysomnography". In: *Frontiers in Human Neuroscience* (2018).

[167] S. Chambon, M. N. Galtier, P. J. Arnal, G. Wainrib, and A. Gramfort. "A Deep Learning Architecture for Temporal Sleep Stage Classification Using Multivariate and Multimodal Time Series". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 26.4 (2018), pp. 758–769.

[168] Z. Cui, X. Zheng, X. Shao, and L. Cui. "Automatic sleep stage classification based on convolutional neural network and fine-grained segments". In: *Complexity* 2018 (2018).

[169] J. Zhang, S. P. Lam, C. Kwok, W. Ho, A. M. Li, F. Paed, J. Tsoh, and F. Psych. "Diagnosis of REM Sleep Behavior Disorder by Video-Polysomnographic Study: Is One Night Enough?" In: *Sleep* (2008).

[170] The MathWorks. *Signal Processing Toolbox User's Guide R2017a*. Natick, MA, 2017.

[171] A. Graps. "An Introduction to Wavelets". In: *IEEE Computational Science and Engineering* 2.2 (1995), pp. 50–61.