Quantifying brain maturation in the preterm baby from EEG sleep analyses



Kirubin Pillay

St Catherine's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2018

Quantifying brain maturation in the preterm baby from EEG sleep analyses

Kirubin Pillay
St Catherine's College

Doctor of Philosophy
Michaelmas Term 2018

The preterm (premature) baby, born before term-age of 37 weeks Postmenstrual Age (PMA, the age since the last menstrual cycle of the mother) remains vulnerable and is admitted to the Neonatal Intensive Care Unit (NICU). Continual improvement of care in the NICU has limited their lifelong impact. However, clinical interventions and other environmental stresses can facilitate delays in brain maturation resulting in abnormal neurodevelopmental outcomes (mental and motor disabilities) by school-age. This is not easily detected by existing methods in the NICU, though Electroencephalography (EEG) is emerging as a promising modality to capture such deviations through the maturation of sleep. Due to a shortage of trained clinicians and the infeasibility of continually labelling EEG, developing fully automated methods is crucial.

Current methods are either not fully automated, do not consider the maturational effects across multiple sleep states, or have limited to no clinical validation on abnormal outcome data. Consequently, the aim of this research is to develop fully automated methods to quantify brain maturation from EEG sleep, through (i) robust automated sleep staging, (ii) tracking sleep-dependent developmental trajectories over PMA, (iii) assessing the quality of sleep-wake cycling with PMA, and (iv) validating these methods using both normal and abnormal outcome data.

This thesis first presents a robust algorithm for detecting the Quiet Sleep (QS) and non-QS sleep states from EEG over a wide PMA range, with median sensitivity and specificity of 0.82 and 0.92, respectively. With this method, sleep states are then estimated from abnormal outcome data and a data-driven feature extraction applied to identify maturational 'biomarkers' and derive sleep-dependent trajectories over PMA. Developmental deviations are detected that identify abnormal outcome patients early.

Probabilistic models are next developed to assess sleep-wake cycling by performing the first multi-state classification of sleep at term-age (with median sensitivity and specificity of 0.72 and 0.91, respectively). At this age, EEG differentiates into more complex states and the timing of this differentiation is also an important indicator of brain maturation. Finally, these models are expanded to novel, end-to-end solutions (employing Bayesian non-parametrics) that incorporate both age and sleep dependencies. These methods additionally classify severely delayed sleep-wake cycling behaviours, known as 'dysmaturity'.

Acknowledgements

First and foremost, I would like to thank my supervisor, Professor Maarten De Vos. Maarten, you've always been genuinely invested in my academic development, as well as my research, and provided continuous invaluable support. I will always be appreciative of your in-depth (and refreshingly blunt!) feedback.

To all the clinicians and engineers from KU Leuven who I've had the pleasure of collaborating with, thanks for all the advice and expertise that you provided throughout. Gunnar and Katrien, as senior neonatologists, I have greatly admired how your daily work stresses never affected your relaxed attitudes, propensity for making jokes, and ever-present optimism. Anneleen, I can honestly say that this research wouldn't have been possible without you. Thanks for painstakingly labelling the many EEG recordings for me, always proposing new and exciting clinical directions, and your general positivity. You maintained all of this while caring for two young children and it baffles me how you've managed to juggle it all!

To the CIBIM lab post-docs: Fernando Andreotti, Niclas Palmius, and Huy Phan, thanks for your continual guidance. Fernando, without your help, configuring Linux and Python would have remained the insurmountable hurdle that I had always thought it was. Nick, you brought out my 'inner nerd' and patiently showed me how to represent my research problems as Bayesian graphical models.

To the rest of the CIBIM lab. You are not just very intelligent colleagues, but also dear friends. Our daily 3pm coffee breaks provided a time for our (sometimes massively random) chats which kept me entertained and sane over the past four years.

I have gained many close friends here in Oxford (and London) through living in two fantastic shared houses, through the CDT program, college (to name but a few). I could write a long paragraph of thanks to each and every one of you, if space permitted. Thanks for all the great times and encouragement you have given me, through squash games, cricket matches, competing in pub quizzes, or grabbing a relaxed meal around Cowley road.

Finally to my loving parents and brother, Sudershan, thank you for your regular FaceTime calls, hilarious football chat, and many visits to see me in Oxford. Most of all, your advice and unwavering support throughout my life is the reason I am where I am right now.

Thank you all and God bless!

Contents

List of Figures **xi**

List of Tables **xiii**

List of Abbreviations **xv**

List of Notation **xix**

1 Introduction **1**
1.1 Motivation 1
1.2 Research aims 3
1.3 Thesis structure & key contributions 4
1.4 Research publications & presentations 6

2 Clinical Background & Existing Methods **9**
2.1 Introduction 9
2.2 The vulnerable preterm baby 9
2.2.1 Common diseases in the preterm baby 11
2.3 The NICU setting 12
2.3.1 Cerebral Ultrasonography (cUS) 14
2.3.2 Polysomnography (PSG) 14
2.3.3 Treating common diseases in the NICU 18
2.3.4 The vital link to brain maturation 18
2.3.5 Treating abnormal brain maturation in the NICU 19
2.4 Assessing neurodevelopmental outcome 20
2.4.1 Bayley Scales of Infant & Motor Development (BSID) 20
2.4.2 Treating abnormal outcome at follow-up 21
2.5 Preterm sleep & the EEG 21
2.5.1 EEG sleep ontogenesis 22
2.5.2 Automated sleep staging 26
2.6 Identifying EEG abnormalities 30
2.6.1 Dysmature patterns 31
2.6.2 Disorganized patterns 31

2.7 Quantifying brain maturation . . . 32
2.7.1 Maturational biomarkers . . . 32
2.7.2 Estimated Maturational Age (EMA) . . . 36
2.7.3 Quantifying sleep-wake cycling . . . 37
2.8 Summary & current limitations . . . 39

3 Datasets **41**
3.1 Introduction . . . 41
3.2 Patient recruitment & data collection . . . 41
3.3 Grading neurodevelopmental outcome . . . 42
3.4 EEG visual labelling . . . 43
3.5 Overview of EEG data & demographics . . . 43
3.5.1 Dataset 1 . . . 44
3.5.2 Dataset 2 . . . 44
3.5.3 Dataset 3 . . . 44
3.6 EEG pre-processing . . . 45

4 Essential Methodology & Technical Background **47**
4.1 Introduction . . . 47
4.2 EEG feature extraction . . . 48
4.2.1 Signal decompositions . . . 48
4.2.2 Time-domain features . . . 49
4.2.3 Frequency-domain features . . . 56
4.2.4 Summary of extracted features . . . 57
4.3 Regression methods . . . 59
4.3.1 Linear regression . . . 59
4.3.2 The random-intercepts model . . . 62
4.3.3 Regression trees & the Random Forest (RF) . . . 63
4.4 Feature selection methods . . . 66
4.4.1 Least Absolute Shrinkage & Selection Operator (LASSO) . . . 66
4.4.2 Minimum Redundancy Maximum Relevance (mRMR) . . . 67
4.5 Probabilistic models & classification . . . 68
4.5.1 Graphical form & Bayesian networks . . . 69
4.5.2 The Gaussian Mixture Model (GMM) . . . 72
4.5.3 The Hidden Markov Model (HMM) . . . 74
4.5.4 Dirichlet Process Mixture Models (DPMMs) . . . 79
4.6 Assessing model performance . . . 87
4.6.1 Cross-Validation (CV) . . . 87

4.6.2 Measures for classification . . . 89
4.6.3 Measures for regression . . . 94
4.6.4 The Wilcoxon signed-rank test . . . 95

5 Unsupervised Quiet Sleep Detection Across Preterm Ages 97
5.1 Introduction . . . 97
5.2 Data . . . 98
5.3 The CLASS algorithm . . . 99
5.3.1 Artefact Subspace Reconstruction (ASR) . . . 99
5.3.2 Adaptive Segmentation (ASG) . . . 102
5.3.3 Feature extraction & cluster-time profiles . . . 104
5.3.4 QS classification . . . 105
5.4 CLASS parameter optimization . . . 106
5.5 Assessing CLASS performance . . . 107
5.6 Results . . . 109
5.6.1 Parameter selection by RO . . . 109
5.6.2 CLASS performance with PMA . . . 110
5.6.3 Comparing CLASS algorithm stages . . . 112
5.6.4 Comparing CLASS with SAT% and CNN . . . 114
5.7 Discussion . . . 114
5.8 Conclusion . . . 117

6 Deriving Developmental Trajectories Across Preterm Ages 119
6.1 Introduction . . . 119
6.2 Data . . . 120
6.3 Feature pre-processing . . . 120
6.4 Identifying maturational biomarkers . . . 121
6.5 The RF model for brain-age prediction . . . 122
6.6 Model training & assessing performance . . . 123
6.7 Results . . . 124
6.7.1 Significant biomarkers for PMA . . . 124
6.7.2 Significant biomarkers for PMA & outcome . . . 124
6.7.3 Comparing RF and LASSO+OLS . . . 126
6.7.4 Assessing outcome across *mtry* . . . 128
6.7.5 Developmental trajectories and *mtry* . . . 129
6.8 Discussion . . . 131
6.8.1 Linking biomarkers and EEG characteristics . . . 131
6.8.2 Linking developmental trajectories and outcome . . . 133
6.9 Conclusion . . . 135

7 Multi-state Sleep Classification at Term-age 137
7.1 Introduction . . . 137
7.2 Data . . . 138
7.3 Feature pre-processing . . . 138
7.3.1 Patient-wise standardization . . . 139
7.3.2 Feature selection . . . 139
7.4 Probabilistic sleep classification . . . 140
7.4.1 Classification using the state-dependent GMM . . . 140
7.4.2 Classification using the HMM-GMM . . . 141
7.5 Model training & assessing performance . . . 142
7.6 Results . . . 143
7.6.1 Binary classification performance . . . 143
7.6.2 Multi-state classification performance . . . 145
7.7 Discussion . . . 149
7.8 Conclusion . . . 152

8 End-to-end Models for Assessing Sleep-Wake Cycling 155
8.1 Introduction . . . 155
8.2 Data . . . 156
8.3 Feature pre-processing . . . 157
8.4 The aGMM . . . 159
8.4.1 Inference . . . 159
8.5 The aHMM-GMM . . . 161
8.5.1 Inference . . . 162
8.6 The aDP-HMM-GMM . . . 162
8.6.1 Inference . . . 164
8.7 Model training & assessing performance . . . 164
8.8 Results . . . 166
8.8.1 Feature performance . . . 166
8.8.2 Varying δ in the aDP-HMM-GMM . . . 167
8.8.3 Fitting GMMs by EM & Neal's Algorithm 8 . . . 169
8.8.4 Comparing overall model performances . . . 169
8.8.5 Comparing brain-age classification performance . . . 172
8.8.6 Comparing sleep state classification performance . . . 174
8.8.7 Assessing sleep-wake cycling . . . 175
8.9 Discussion . . . 177
8.10 Conclusion . . . 179

9 Conclusions & Future Work **181**
9.1 Summary of results & contributions . . . 182
9.1.1 Automated sleep staging . . . 182
9.1.2 EEG biomarkers & trajectory estimation . . . 182
9.1.3 Assessing sleep-wake cycling . . . 183
9.2 Future work . . . 184
9.2.1 Data quality . . . 184
9.2.2 Labelling sleep state & outcome . . . 185
9.2.3 Incorporating additional modalities . . . 187
9.2.4 Extending the probabilistic models . . . 187
9.2.5 Further avenues for machine learning . . . 188

Appendices

A Additional Distributions & Algorithms **193**
A.1 Statistical distributions . . . 193
A.1.1 The z-distribution . . . 193
A.1.2 The student's t-distribution . . . 193
A.1.3 The F-distribution . . . 194
A.2 Signal decompositions . . . 195
A.2.1 The Discrete Fourier Transform (DFT) . . . 195
A.2.2 The Discrete Wavelet Transform (DWT) . . . 196
A.2.3 Empirical Mode Decomposition (EMD) . . . 198
A.2.4 The Hilbert-Huang Transform . . . 199
A.3 Principal Component Analysis (PCA) . . . 199
A.4 K-means clustering . . . 200
A.5 The SAT% algorithm . . . 201

B Supplementary Tables **205**
B.1 Chapter 6: Significant feature correlations . . . 205
B.2 Chapter 7: Recording-specific performance . . . 210
B.3 Chapter 8: aDP-HMM-GMM δ performance . . . 211

References **213**

List of Figures

2.1 Schematic defining the different terminology used to describe preterm and full-term ages . . . 10
2.2 The effect of IVH on the preterm brain ventricles . . . 12
2.3 Photographs of preterm monitoring in a typical NICU setting . . . 13
2.4 A screenshot of a typical PSG output from an NICU monitor . . . 15
2.5 The standard 10-20 electrode placement system for EEG electrodes 16
2.6 Examples of the different EEG sleep states that develop with PMA. 23
2.7 Summary of the changing nature of the EEG sleep states with PMA 26
2.8 A typical disorganized multichannel EEG pattern . . . 32

4.1 Illustration of how estimates of total signal length change with length scale . . . 53
4.2 Illustration of a single regression tree which selects four different features . . . 64
4.3 Chosen notations for representing probabilistic graphical models . . 70
4.4 Typical plate notation for the GMM . . . 73
4.5 The typical graphical model for an HMM . . . 76
4.6 Plate notation for an HMM-GMM . . . 80
4.7 Plate notation for a DPMM. . . . 81
4.8 Illustration of an ROC curve. . . . 91
4.9 The general form of the confusion matrix in a binary classification problem. . . . 92

5.1 Age distribution of the recordings in dataset 1 . . . 98
5.2 Flowchart and illustrations of the processing stages of CLASS . . . 100
5.3 Illustration of the ASR method on 30 minutes of EEG. . . . 103
5.4 Illustration of the changes in mean F1-error and parameter values during RO . . . 111
5.5 Assessing the performance of CLASS on the test set. . . . 112

6.1 Pre-processing procedure for extracting features from each EEG recording . . . 121
6.2 The top statistically significant features with PMA . . . 125

6.3 The top statistically significant features with PMA and outcome . . 127
6.4 Comparison of the RF and LASSO+OLS regression models 128
6.5 RF performance on dataset 2 when varying *mtry*. 129
6.6 RF-derived trajectories of brain-age against PMA in dataset 2 . . . 130

7.1 Illustration of patient-wise standardization on a single feature . . . 140
7.2 Median ranking coefficients of the top 30 features for binary classification 143
7.3 Median ranking coefficients of the top 30 features for multi-state classification 146
7.4 κ performance for the GMM and HMM-GMM against feature number 147
7.5 Results from multi-state sleep classification 148

8.1 Histogram of the total number of EEG recordings by age group . . 157
8.2 Graphical models for the aGMM, aHMM-GMM, and aDP-HMM-GMM 160
8.3 Median ranking coefficients for the top 15 features 166
8.4 I performance for the GMM and HMM-GMM against feature number 168
8.5 The change in $\alpha^{(a)}$, $\alpha^{(s)}$ and component number with δ 169
8.6 Density distributions for the aDP-HMM-GMM and aHMM-GMM . 170
8.7 Confusion matrices for brain-age classification of recordings across PMA groups. 173
8.8 Sleep state classification performance across PMA groups 174
8.9 Estimated hypnograms by the aDP-HMM-GMM for QS and non-QS 176

A.1 The standard z-distribution. 194
A.2 The t-distribution for various values of d. 194
A.3 The F-distribution for various values of $d1$ and $d2$. 195
A.4 The cascading DWT filter bank for decomposing a signal into frequency bands 197
A.5 Results of grid search over key parameters in the SAT% algorithm . 203

List of Tables

3.1 Summary of patient demographics and EEG information for dataset 1. 45
3.2 Summary of patient demographics and EEG information for dataset 2. 46

4.1 Frequencies extracted from/defined after performing the DFT, DWT, and EMD. 49
4.2 The full list of the 113 extracted EEG features, separated by time- and frequency- domain. 58

5.1 Summary of the CLASS parameters before and after RO. 109
5.2 Comparing CLASS performance at different stages of the algorithm 113
5.3 Comparing classification performance between CLASS and the CNN 114

6.1 The top ten outcome-sensitive features chosen by RF for $mtry = 200$ 131

7.1 Comparing performance for binary sleep classification 145
7.2 Comparing performance for multi-state sleep classification 149

8.1 Comparing overall brain-age and sleep state performance between models 171

B.1 Significant QS features with PMA 205
B.2 Significant non-QS features with PMA 207
B.3 Significant QS and non-QS features with PMA and outcome 209
B.4 Individual patient κ scores for the HMM-GMM and GMM models . 210
B.5 $\alpha^{(a)}$ and $\alpha^{(s)}$ for different values of δ in the aDP-HMM-GMM 211

List of Abbreviations

AASM . . . . . American Academy of Sleep Medicine
aEEG . . . . . Amplitude-integrated EEG
ApEn . . . . . Approximate Entropy
AS . . . . . Active Sleep
ASG . . . . . Adaptive Segmentation
ASR . . . . . Artefact Subspace Reconstruction
AUC . . . . . Area Under the Curve
BAC . . . . . Brain Activity Cycling
BSID-II . . . . . Bayley Scales of Infant & Motor Development II
CLASS . . . . . Cluster-based Adaptive Sleep Staging
CNN . . . . . Convolutional Neural Network
CRP . . . . . Chinese Restaurant Process
cUS . . . . . Cerebral Ultrasonography
CV . . . . . Cross Validation
DAG . . . . . Directed Acyclic Graphs
DFT . . . . . Discrete Fourier Transform
DP . . . . . Dirichlet Process
DPMM . . . . . Dirichlet Process Mixture Model
DWT . . . . . Discrete Wavelet Transform
ECG . . . . . Electrocardiogram
EEG . . . . . Electroencephalogram
EM . . . . . Expectation Maximization
EMA . . . . . Estimated Maturational Age
EMD . . . . . Empirical Mode Decomposition
EOG . . . . . Electrooculogram
ETDF . . . . . Energy-weighted Temporal Dependency Function
FIR . . . . . Finite Impulse Response
GA . . . . . Gestational Age
GMM . . . . . Gaussian Mixture Model
GP . . . . . Gaussian Process

HIE . . . . . Hypoxic Ischaemic Encephalopathy
HM . . . . . Higher Margin
HMM . . . . . Hidden Markov Model
HVS . . . . . High Voltage Slow-wave
IBI . . . . . Inter-Burst Interval
IBP . . . . . Indian Buffet Process
IBR . . . . . IBI-to-Bursts Ratio
IID . . . . . Independent and Identically Distributed
IMF . . . . . Intrinsic Mode Function
IQR . . . . . Inter-Quartile Range
IS . . . . . Indeterminate Sleep
IVH . . . . . Intraventricular Haemorrage
LASSO . . . . . Least Absolute Shrinkage & Selection Operator
LL . . . . . Line Length
LM . . . . . Lower Margin
LOSO . . . . . Leave-One-Subject-Out
LVI . . . . . Low-Voltage Irregular
MA . . . . . Moving Average
MDI . . . . . Mental Development Index
ML . . . . . Maximum Likelihood
MRI . . . . . Magnetic Resonance Imaging
mRMR . . . . . Minimum Redundancy Maximum Relevance
MSE . . . . . Multiscale Entropy
$\sqrt{MSE}$ . . . . . Root Mean Squared Error
NICU . . . . . Neonatal Intensive Care Unit
OLS . . . . . Ordinary Least Squares
PCA . . . . . Principal Component Analysis
PDI . . . . . Psychomotor Development Index
PI . . . . . Prediction Interval
PMA . . . . . Postmenstrual Age
PSG . . . . . Polysomnography
PVL . . . . . Periventricular Leukomalacia
QS . . . . . Quiet Sleep
rEEG . . . . . range-EEG
REM . . . . . Rapid Eye Movement
RF . . . . . Random Forest

RO . . . . . Random Optimization
ROC . . . . . Receiver Operating Characteristic
SampEn . . . . Sample Entropy
SATs . . . . . Spontaneous Activity Transients
SAT% . . . . . % of SATs
SD . . . . . Standard Deviation
SSE . . . . . Sum-of-Squares Error
TA . . . . . Tracé Alternant
TD . . . . . Tracé Discontinu

List of Notation

a, A	Scalar value, a or A (variable, parameter or constant)
$\boldsymbol{x}$	Column or row vector, $\boldsymbol{x}$
$\boldsymbol{X}$	Matrix, $\boldsymbol{X}$
x_i	The i^{th} element of $\boldsymbol{x}$ or i^{th} instance of x
x_{ij}	The $(i \times j)^{th}$ element of $\boldsymbol{X}$ or $(i, j)^{th}$ instance of x
$\boldsymbol{x}'$	First derivative of $\boldsymbol{x}$
$\boldsymbol{x}''$	Second derivative of $\boldsymbol{x}$
$\boldsymbol{x}^\intercal$	Transpose of $\boldsymbol{x}$
$\bar{x}$	Mean value of elements in $\boldsymbol{x}$ or across instances of x
$\boldsymbol{X}^{-1}$	Matrix inverse of $\boldsymbol{X}$
$\boldsymbol{X}^{+}$	Moore-Penrose pseudo-inverse of $\boldsymbol{X}$
$\mathcal{S}$	A set of values, $\mathcal{S}$
$\varnothing$	Empty set
$\in$	belongs to the set
$\notin$	does not belong to the set
$\forall$	for all
:	where
$\mathbb{R}$	The set of real numbers
$\mathbb{N}$	The set of natural numbers (positive integers)
$\mathbb{C}$	The set of complex numbers
$\mathbb{R}^N$	Column vector of real numbers, of length N
$\mathbb{R}^{1\times N}$	Row vector of real numbers, of length N
$\mathbb{R}^{P\times N}$	Matrix of real numbers, of dimensions $(P \times N)$
$p(\cdot)$	Probability density distribution
$p(\,\cdot\,\vert\,\cdot\,)$	Conditional density distribution
$p(\,\cdot\,,\cdot\,)$	Joint density distribution

$\sim$	is distributed by
$\equiv$	is equivalent to
$\mathcal{N}(\cdot)$	Normal (Gaussian) distribution
$\mathcal{DP}(\cdot)$	Dirichlet Process distribution
$\mathcal{NIW}(\cdot)$	Normal-inverse Wishart distribution
$\mathcal{W}^{-1}(\cdot)$	Inverse Wishart distribution
$\text{Cat}(\cdot)$	Categorical distribution
$\text{unif}\{\cdot\}$	Discrete uniform distribution
$\Gamma(\cdot)$	Gamma function
$\mathcal{FT}\{\cdot\}$	Discrete Fourier Transform
$\log(\cdot)$	Natural logarithm
$\lvert\cdot\rvert$	Absolute value
$\lvert\lvert\cdot\rvert\rvert$	Euclidian distance
Δt	A difference/change in t

1 Introduction

1.1 Motivation

Preterm (premature) births account for around 11% of births worldwide on an annual basis, which corresponds to approximately 15 million babies born before the expected term-age of 37 weeks Postmenstrual Age (PMA, the age since the last menstrual cycle of the mother at the onset of pregnancy) [1]. The majority of these babies exhibit a wide range of developmental complications but the continuing improvement of perinatal care (immediately after birth) in the Neonatal Intensive Care Units (NICUs)[1] has limited their lifelong impact. This said, many preterm babies will experience disturbances in brain development (maturation) that may lead to more severe mental (cognitive) and motor disabilities materializing as late as school-age, with extreme cases resulting in neonatal (newborn) death. One clear cause of this is a restricted oxygen supply to the brain tissues (hypoxia) that can result in severe brain lesions and injuries to the white and grey matter [2, 3].

Most of these complications are discovered at a late stage in the NICU due to the shortage of appropriate diagnostic tools necessary for real-time monitoring of the baby. This highlights a need to introduce automated methods to facilitate the earlier identification of developmental deviations in brain maturation that may lead to later abnormal 'neurodevelopmental' outcomes. Furthermore, fully automated methods are crucial to move towards a standardized, high-throughput

[1] In many European centres, the NICU refers to the overall neonatal care unit which includes the intensive care unit as well as the low and high dependency units. This convention is used here.

data analysis within the NICU, as it is infeasible for trained clinicians to continuously manually label these outputs.

In the clinical setting, normal or abnormal neurodevelopmental outcomes in the preterm baby are currently only determined later through follow-up tests at 9-24 months of age. Cerebral Ultrasonography (cUS) remains the main modality [4] for brain monitoring during the patient's stay in the NICU, and is utilized to directly identify structural damage (or symptoms thereof) associated with severe brain injuries. More complex methods have been proposed in the research setting, using alternative modalities such as Magnetic Resonance Imaging (MRI) [5–11], Electrocardiogram (ECG), oxygen saturation [12–14], as well as high-resolution video-recording [15]. However, these approaches do not tell the full story.

As early as the 1980s, clinical research has identified cases where preterm babies develop abnormal outcomes without any evidence of clear structural damage [16]. This creates a limitation with certain modalities, such as cUS, and facilitates an interest in better understanding the underlying processes of brain maturation. To this end, 'sleep ontogenesis' (the changing nature of sleep with age) has been identified as a means to quantify functional brain maturation and can potentially reflect early deviations linked to abnormal outcome [17, 18].

Below 28 weeks PMA, the baby exhibits very crude sleep behaviours but beyond this age, the evolution of deeper neuronal connections results in the clear separation between periods of wakefulness, Active Sleep (AS, or Rapid Eye Movement (REM) sleep), and Quiet Sleep (QS, or non-REM sleep) [19]. Electroencephalogram (EEG) measurements of brain activity provide means to distinctly quantify this sleep ontogeny, clearly reflecting the maturation of the sleep states as the baby matures. From a practical standpoint, such a modality is also affordable and non-invasive, with such a utilization improving clinical applicability and feasibility for real-time monitoring in the NICU.

During normal brain maturation, it is assumed that the EEG will exhibit patterns (or characteristics) that reflect the actual PMA of the baby. However, when there are disturbances to brain maturation, these characteristics may begin

to exhibit behaviours reflecting a delayed (or accelerated) rate of sleep development [20], as well as more extreme, deviant behaviours [21]. By term-age, deviating EEG is further characterized by a poor quality of differentiation into more complex sleep states that begin to resemble infant and adult sleep [20].

Previous clinical and engineering research have identified two broad approaches for utilizing EEG sleep maturation to identify developmental deviations. The first approach proposes features as candidate 'biomarkers' to track specific EEG characteristics across sleep states and PMA [16, 22–30]. The second approach considers the quality of the sleep-wake cycling (the transitions between sleep states and wakefulness), through measures of overall cyclicity [31–34] or through sleep state-specific changes in durations, proportions etc. [23, 24, 35, 36].

However, as will be discussed further in chapter 2, the majority of these methods are not fully automated (using visually labelled sleep periods), and/or do not consider the maturational effects across multiple sleep states. More significantly, focus remains largely on EEG from preterm babies with normal neurodevelopmental outcome, such that there remains limited clinical knowledge of the potential of such methods for the early identification of long-term abnormalities.

1.2 Research aims

The general aim of this research is thus:

To develop fully automated methods to quantify brain maturation in the preterm baby from EEG sleep, aiding the early prediction of neurodevelopmental outcome.

The more specific goals are:

1. To develop robust methods for automated EEG sleep staging, essential to achieve fully automated approaches for quantifying brain maturation.

2. To identify relevant EEG maturational biomarkers to estimate sleep-dependent developmental trajectories and identify early deviations linked to long-term abnormal neurodevelopmental outcomes.

3. To develop automated solutions to assess EEG sleep-wake cycling, as an alternative means to quantify brain maturation and identify developmental deviations.

4. To develop such methods using both normal and abnormal outcome data, allowing a more objective and practical assessment of their clinical potential as diagnostic tools in the NICU.

1.3 Thesis structure & key contributions

This thesis presents a series of novel machine learning applications and models, to achieve the research goals outlined in section 1.2. Below is an overview of the thesis structure and the key contributions. Motivations for the chosen methods are provided in the corresponding chapters.

- **Chapter 1** has provided the overall scope, motivations, and key goals of this research. **Chapter 2** builds on this, delving deeper to provide a detailed clinical background on preterm abnormalities, maturation, and EEG sleep. Existing methodology (and current limitations) for automated sleep staging and quantifying brain maturation are also discussed.

- **Chapter 3** details the EEG datasets available for this research, including the procedure for patient recruitment and data collection.

- **Chapter 4** provides a technical background of the existing signal processing and machine learning methodology used throughout this research. This includes descriptions of EEG features extracted, feature selection methods, as well as regression and probabilistic models.

- **Chapter 5** presents a robust, automated algorithm for the detection of QS from EEG. An existing method is extended, resulting in the first unsupervised and clinically robust algorithm that overcomes the challenges of inter-patient variability, while preventing the need to train over large datasets.

- **Chapter 6** proposes a regression-based framework for estimating patient developmental trajectories from the EEG recordings, across PMA. The QS detection algorithm of chapter 5 is utilized to first extract sleep states, before novel sleep-dependent maturational biomarkers are identified through a large 'data-driven' extraction of features. By varying the biomarkers used, this method achieves the first prediction of patients with abnormal outcome, by classifying deviating trajectories.

- **Chapter 7** presents a method that utilizes the same data-driven feature set to perform automated sleep staging of term-age recordings. Focus now shifts to classifying the differentiated EEG sleep states at this age, as a means to assess the timing of brain maturation during this period. Probabilistic methods are investigated that model the sleep state transitions. This results in the first automated method for multi-state sleep classification at term-age.

- **Chapter 8** extends the methods applied in chapter 7 to novel, age-dependent probabilistic models, capable of classifying sleep states across a wide PMA range. This provides the first end-to-end diagnostic tool for quantifying EEG sleep-wake cycling and estimating extreme cases of delayed brain development.

- **Chapter 9** provides a final summary of the research, contributions and limitations. This leads to a discussion on a number of potential avenues for future work.

1.4 Research publications & presentations

Peer-reviewed journal articles & proceedings:

K. Pillay*, A. Dereymaeker*, K. Jansen, G. Naulaers & M. De Vos. "Deriving developmental trajectories from EEG to predict normal and abnormal brain maturation across preterm ages". Submitted to: *Sci. Rep.* (2018) **Joint first authors.*

K. Pillay, A. Dereymaeker, K. Jansen, G. Naulaers & M. De Vos. "A Bayesian parametric model for quantifying brain maturation from sleep-EEG in the vulnerable newborn baby". In: *2018 IEEE Eng. Med. Biol. 40th Annu. Conf.*, IEEE, 2018, pp. 1-4. [37]

K. Pillay, A. Dereymaeker, K. Jansen, G. Naulaers, S. Van Huffel & M. De Vos. "Automated EEG sleep staging in the term-age baby using a Generative Modelling approach". In: *J. Neural Eng.* 15.3 (2018), p. 036004. [38]

K. Pillay & M. De Vos. "A unifying Bayesian approach for preterm brain-age prediction that models EEG sleep transitions over age". In: *2017 NIPS workshop on Machine Learning for Health (ML4H)*, NIPS, 2017. arXiv:1809.07102. [39]

A. Dereymaeker*, **K. Pillay***, J. Vervisch, S. Van Huffel, G. Naulaers, K. Jansen & M. De Vos. "An Automated Quiet Sleep Detection Approach in Preterm Infants as a Gateway to Assess Brain Maturation". In: *Int. J. Neural Syst.* 27.6 (2017), pp. 1750023(1)-1750023(18). **Joint first authors.* [40]

A. Dereymaeker, **K. Pillay**, J. Vervisch, M. De Vos, S. Van Huffel, K. Jansen & G. Naulaers. "Review of sleep-EEG in preterm and term neonates". In: *Early Hum. Dev.* 113 (2017), pp. 87-103. [41]

Oral & poster presentations:

K. Pillay, A. Dereymaeker, K. Jansen, G. Naulaers & M. De Vos. "An end-to-end Bayesian approach for modelling preterm EEG sleep". In: *2019 Intl. Brain Monitoring and Neuroprotection in the Newborn Conf.*, 2019. Poster presentation.

M. Buckle, **K. Pillay**, C. Hartley, G. Schmidt, A. Hoskin, M. De Vos, C. K. Patel & R. Slater. "Characterising changes in infant brain activity during retinopathy of prematurity screening". In: *2019 Intl. Brain Monitoring and Neuroprotection in the Newborn Conf.*, 2019. Oral presentation.

A. H. Ansari, **K. Pillay**, O. De Wel, A. Dereymaeker, K. Jansen, A. Caicedo, J. Vervisch, S. Van Huffel, G. Naulaers, & M. De Vos. "An improved deep convolutional neural network for sleep stage classification in preterm and term neonates" In: *2019 Intl. Brain Monitoring and Neuroprotection in the Newborn Conf.*, 2019. Poster presentation.

K. Pillay, A. Dereymaeker, K. Jansen, S. Van Huffel, G. Naulaers & M. De Vos. "An automated brain-age prediction model by modelling EEG sleep stages across age". In: *2017 Intl. Brain Monitoring and Neuroprotection in the Newborn Conf.*, 2017. Poster presentation.

A. Dereymaeker*, **K. Pillay***, M. De Vos & G. Naulaers. "Tracking EEG-brain maturation in preterm infants with the aid of an automated Quiet Sleep detection algorithm". In: *2017 Intl. Brain Monitoring and Neuroprotection in the Newborn Conf.*, 2017. **Joint first authors.* Oral presentation.

A. Dereymaeker, **K. Pillay**, M. De Vos, S. Van Huffel, K. Jansen & G. Naulaers. "Brain maturation related to early neurodevelopmental outcome can be separated by automated EEG biomarkers in preterm infants". In: *2017 Intl. Brain Monitoring and Neuroprotection in the Newborn Conf.*, 2017. Poster presentation.

K. Pillay, A. Dereymaeker, J. Vervisch, S. Van Huffel, K. Jansen, G. Naulaers & M. De Vos. “Detecting the Quiet Sleep stage from EEG in the preterm neonate”. In: *MEIBioeng2016 Conf.*, 2016. Poster presentation.

2

Clinical Background & Existing Methods

Associated publication(s):

A. Dereymaeker, K. Pillay et al. "Review of sleep-EEG in preterm and term neonates". In: *Early Hum. Dev.* 113 (2017), pp. 87-103.

2.1 Introduction

Chapter 1 provided an overview of the clinical motivation behind the research presented in this thesis. Specifically, the aim of developing fully automated approaches for quantifying brain maturation in the preterm baby from EEG sleep, was defined.

This chapter now delves deeper into the clinical background, providing more in-depth descriptions of the preterm baby, the NICU setting, associated diseases, treatments, and the importance of EEG as an emerging modality for tracking brain maturation. Existing methods for automated sleep staging and quantifying brain maturation from EEG are also discussed, highlighting the current limitations.

2.2 The vulnerable preterm baby

At the onset of pregnancy, which is defined as the first day after the last menstrual cycle of the mother, the baby is expected to be born at least 37 weeks after this event although the healthy, expected delivery date is defined at 40 weeks. This

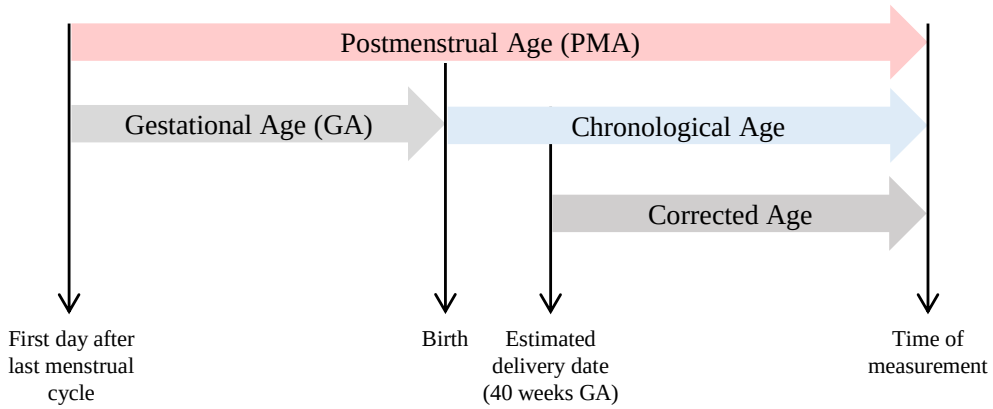


Figure 2.1: Schematic defining the different terminology used to describe preterm and full-term ages.

period between conception and birth is known as the gestation, and the age at birth is consequently called the Gestational Age (GA) [42]. When a baby has a GA between 37-40 weeks, they are deemed full-term, as they are born at term-age. In cases when GA<37 weeks, the baby is deemed premature (or preterm) with some babies born as young as 24 weeks GA.

As GA is fixed at the time of birth, the subsequent age of the baby measured at any point after birth is known as the PMA. The PMA is simply the GA plus the age between birth and the point of assessment (the chronological age) [42].

These are currently the most widely-accepted measures of preterm and full-term ages, and are defined by the American Academy of Pediatrics (AAP) [42]. A summary of these ages and how they relate is shown in figure 2.1. The corrected age is also included and defined as the chronological age minus the number of weeks born before 40 weeks GA [42], often used in studies that compare cohorts of preterm and full-term babies.

A preterm baby will continue to develop *ex utero*, both physically and mentally, and has a heightened susceptibility to various diseases due to an increased exposure to external stresses from the environment at a much earlier age. Such diseases are of greater likelihood and severity with decreasing GA [43].

2.2.1 Common diseases in the preterm baby

During early development *ex-utero*, a number of diseases may present itself and affect various regions of the body. *Retinopathy of prematurity*, for example, results in impaired vision, due to an underdevelopment of the blood vessels surrounding the retina of the eyes when the baby is born. The blood vessels developing outside the womb are more fragile and can leak, causing scarring to the surrounding tissue [1, 44].

General respiratory diseases are also common, due to the immaturity of the preterm lung which has a limited number of air sacs (alveoli) preventing adequate functioning. *Necrotizing enterocolitis* is a condition when a portion of the bowel dies, with the underlying cause thought to be a combination of poor blood flow and intestinal infection [45–47]. *Single* (or *spontaneous*) *intestinal perforation* is also a related (but distinct) disease. As the name suggests, it consists of a single perforation to the gastrointestinal tract and the underlying cause remains unclear [46].

Patent ductus arteriosus occurs when the *ductus arteriosus* fails to close at birth. This is a blood vessel which allows the underdeveloped lungs to be initially bypassed in the circulatory system. When the lungs are sufficiently developed, this vessel closes. If not closed, the blood flow is adversely effected and can result in heart failure [46].

In addition, underdevelopment of the immune system can result in an enhanced risk of bacterial and viral infections to tissues, which is known as sepsis. This is a major cause of death and, if spreading to brain tissue, can lead to cerebral palsy, seizures, and hearing loss [48]. Seizures are often also indicative of epilepsy. Traumatic brain injuries leading to epilepsy and cerebral palsy also includes *Periventricular Leukomalacia* (PVL), damage to white matter tissue (lesions) as a result of reduced blood flow and oxygen to the brain. Holes resulting from damaged tissue can fill up with fluid resulting in the formation of cysts [49]. PVL is a specific form of brain injury due to oxygen deprivation, known as *Hypoxic Ischaemic Encephalopathy* (HIE), and is most prevalent in term-age babies [50].

Another common traumatic brain injury (leading to similar physical disabilities) is an *Intraventricular Haemorrage* (IVH), a bleeding into the brain's ventricular

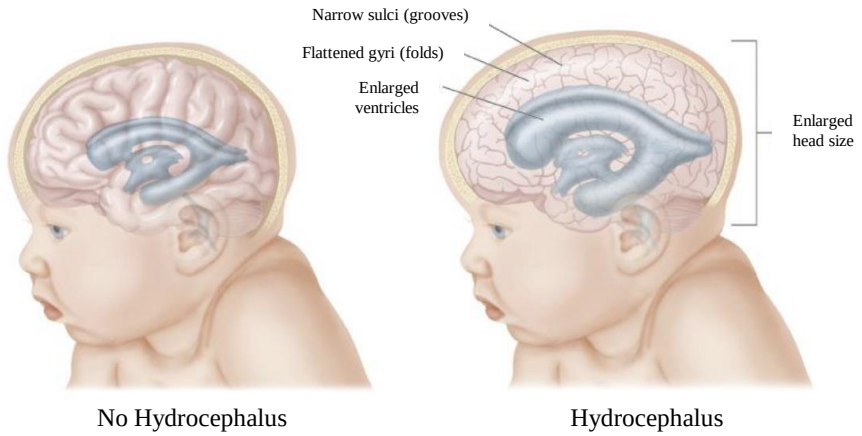


Figure 2.2: Illustration of the enlargement in brain ventricles due to *hydrocephalus*, as a result of IVH[1].

system that can produce enlarged ventricles. This causes an increased pressure within the skull (*hydrocephalus*) [51] and is illustrated in figure 2.2.

In order to mitigate, diagnose and treat such a large variety of diseases, all preterm babies are admitted to the NICU.

2.3 The NICU setting

Figure 2.3 shows some images of preterm care within a NICU setting (at the UZ Leuven Hospitals, Leuven, Belgium). Newborn babies are typically placed inside incubators, which simulate a constant, womb-like environment by which to minimize external stresses to the baby. In tandem with the incubator, a variety of treatments, real-time measurements, and interventions are applied, with the specific modalities and practises varying from NICU to NICU [43].

To alleviate diseases that arise due to oxygen deprivation, most preterm babies are provided with external oxygen through a tracheal tube (mechanical ventilation) or through a tight-fitting mask (non-invasive ventilation). This also works the lungs and encourages further development. However, such ventilation can also lead to a chronic lung disease known as *Bronchopulmonary Dysplasia* [52].

Retinopathy of prematurity is diagnosed through routine screening by an ophthalmologist (particularly if the baby has a GA <32 weeks) [44]. *Spontaneous*

[1]CDC [Public domain], via Wikimedia Commons.

intestinal perforation and *patent ductus arteriosus* can be identified through X-rays [53]. The latter is also diagnosed through recordings of heart beat, using an ECG, as part of real-time vital sign monitoring [53]. To diagnose *necrotizing enterocolitis*, Bell's scores are routinely used which categorize the disease in three stages, ranging from a 'suspected' to 'advanced' level. This is based on blood tests as well as also through vital-sign monitoring [45, 47].

Real-time vital sign monitoring is performed using a combination of modalities which is collectively known as Polysomnography (PSG). This is carried out continuously throughout the preterm baby's stay in the NICU. In addition, varying grades of IVH and PVL can be identified through cUS scans which are also measured on a regular basis. In cases of severe lesions, brain MRI is also performed to confirm the findings of cUS (provided that the baby remains physiologically stable [43]). Both cUS and PSG are discussed in further detail below.



Figure 2.3: Photographs of preterm monitoring in a typical NICU setting (images taken of the NICU in the UZ Leuven Hospitals, Leuven, Belgium).

2.3.1 Cerebral Ultrasonography (cUS)

cUS is particularly useful for identifying and assessing the severity of IVH and PVL in the baby. Regular monitoring and observations of the spread of the condition over time results in diagnoses using defined clinical grades.

In the case of IVH, severity is graded from I to IV. Grades I and II denote cases where bleeding is localized, or there is mild bleeding into the ventricles without any enlargement. Grade III specifies enlargement of the ventricles due to accumulated blood, while grade IV denotes bleeding that has extended outside the ventricles and into the brain tissue [54].

For PVL, there are also four grades that identify lesions through visualizing the size and formation of the cysts across the brain. Grade I denotes mild PVL without any cyst formation occurring beyond seven days. Grades II-IV categorize increasing formation from small cysts in the surrounding ventricular area, to extensive cysts that spread into the deep white matter regions [54].

Finally, a cUS phenomenon known as 'flaring' is sometimes observed. These are regions of slightly higher-than-normal echo responses (brighter spots) that may appear around the ventricular regions. If sustained for long periods of time, these can also be correlated with brain injury [55].

2.3.2 Polysomnography (PSG)

The PSG comprises a series of modalities used to capture real-time data of the baby's vital signs, particularly during sleep. A screenshot of the output from a typical NICU PSG monitor is shown in figure 2.4 and the main components are as follows [19, 43]:

- EEG: A multichannel recording of brain function. This typically utilizes non-invasive scalp electrodes positioned using the international standard 10-20 electrode system, as shown in figure 2.5. Monitored EEG channels include four lobe regions: Temporal (T), Central(C)[2], Parietal (P), and Occipital (O). There are also specific scalp sites that provide additional channels, including the

[2]There is in fact no existing central lobe in the brain, with C merely providing another separable 'sub-lobe' region.

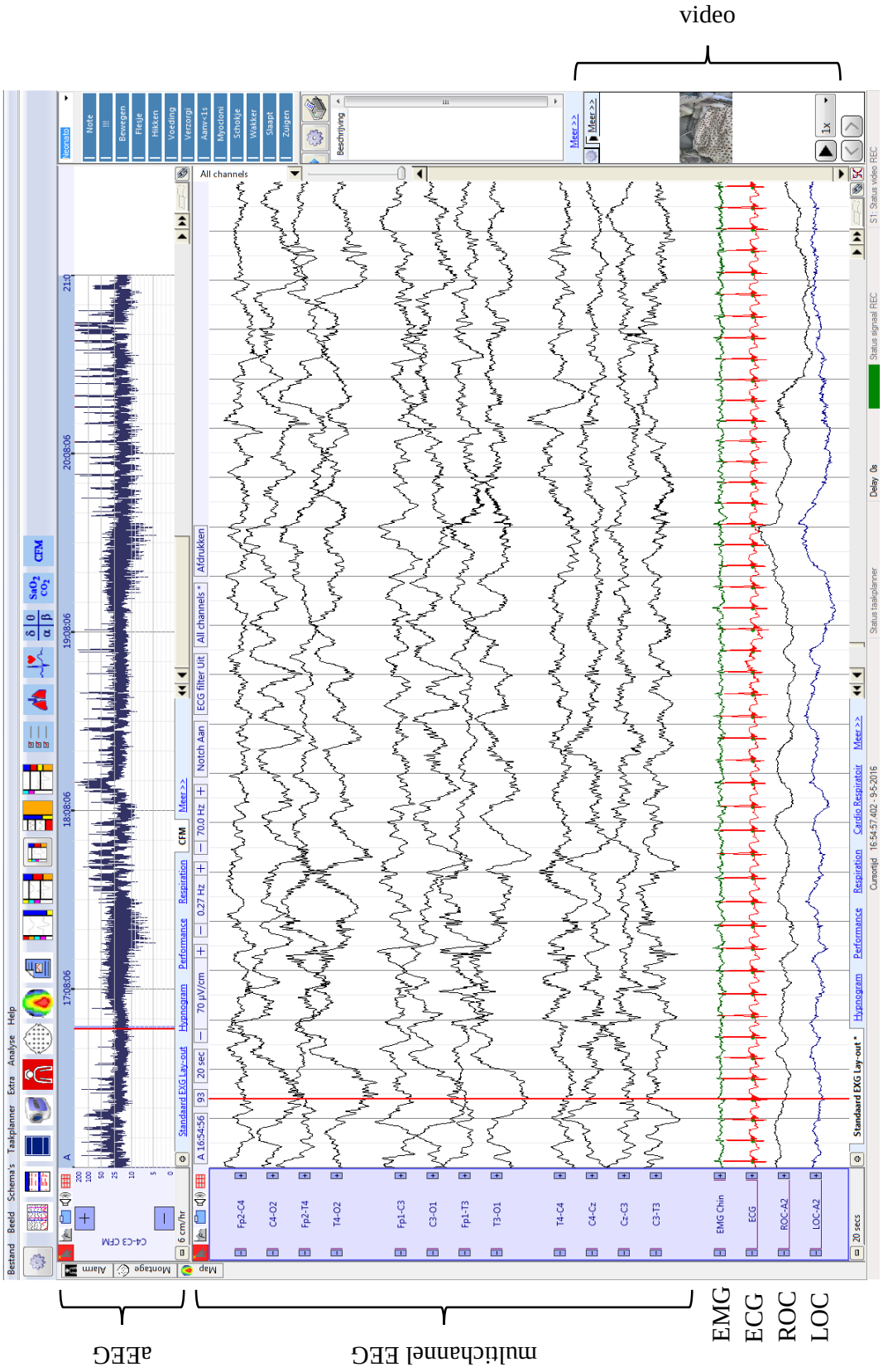


Figure 2.4: A screenshot of a typical PSG output from a NICU monitor at the UZ Leuven Hospitals, Leuven, Belgium. This is measured from a term-age preterm baby of 39 weeks PMA. Multichannel EEG, aEEG, EMG, ECG, EOG (ROC and LOC), and continuous video monitoring are included. The bipolar EEG montage used for visual assessment in Leuven is shown here.

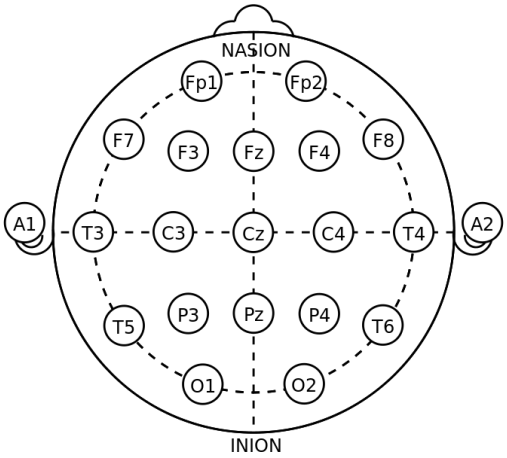


Figure 2.5: The standard 10-20 electrode placement system for EEG electrodes[3]. In this schematic, the T, C, P, O, A, Fp and F sites are included.

Earlobes (A), Frontal-Polar (Fp), and Frontal (F) sites [56]. The EEG output is presented with respect to a reference. Different referencing schemes include a bipolar montage (taking differences between specific channel pairs), a single referential montage (with respect to a single channel/electrode on the scalp), or an average referential montage (using the average activity across the scalp). For visual assessment in the NICU setting, a bipolar montage is usually chosen (as also shown in figure 2.4). EEG is currently mostly used in the NICU setting to diagnose brain structural damage, seizures and to visually assess sleep quality. However, its use remains limited and is not standardized across NICUs.

- Amplitude-integrated EEG (aEEG): The aEEG (also known as cerebral function monitoring) is a reduced form of the EEG, and displays the peak-to-peak amplitude of specific EEG channels in a semi-logarithmic plot. It proves useful for summarizing EEG over very long periods and captures sufficient information to identify seizure events or unusually suppressed brain activity. Even if the full multichannel EEG is not used, aEEG is often included in the monitoring procedure.

- Electrooculogram (EOG): A recording of eye motion using two electrodes, placed above (location ROC) and below (location LOC) of the eye. Electrodes are the

[3][Public domain], from Wikimedia Commons.

same as for EEG, however, the amplitudes are orders of magnitude higher. It proves useful in detecting REM behaviours that are associated with specific sleep states, as well as seizures.

- Electromyogram (EMG): This is typically four electrodes that measure the degree of muscle movement (tone). Leads are usually placed on the chin and above and below the jaw line. The degree of muscle tone provides an indicator of sleep state and certain seizures, particularly in cases of unusually high levels of atonia (the lack of muscle tone).

- ECG: Measures the heart beat of the baby in a reduced form (typically consisting of two to three leads), as opposed to the typical twelve leads used for adult ECG measurement. Through the quality of the 'QRS complex' in the ECG signal (which represents a heart beat), it is possible to diagnose a range of heart defects, including *patent ductus arteriosus*. Heart rate can also be inferred from this signal and is indicative of particular sleep states.

- Pressure transducers: Typically a nasal thermistor, fitted near or inside the nostrils (if possible), which allows the monitoring of nasal airflow and respiratory rate. Deviations can indicate a variety of the aforementioned diseases, and also helps in identifying specific sleep states. In some cases, pneumographs are introduced, consisting of piezo-electric belts placed across the chest (thoracic and/or abdominal) to measure breathing rate. These are either accelerometer- or impedance-based, where in the latter the changing resistance of current flow due to lung inflation (and deflation) is measured.

- Continuous video monitoring: A continuous live video feed of the baby while in the incubator. It is used to record seizures, physical responses to pain, and transitions from sleep into wakefulness. It is not strictly part of conventional PSG monitoring but is usually coupled with it.

2.3.3 Treating common diseases in the NICU

Diagnoses of diseases through PSG and imaging modalities have become easier and more accurate, leading to prompt interventions within the NICU. Depending on the nature and severity of the disease, these interventions focus on alleviating chronic symptoms all the way to completely curing it.

When *retinopathy of prematurity* is diagnosed, the baby will most commonly undergo laser therapy where the fragile blood vessels are burned away to preserve the main region of the visual field. This intervention stops disease progression and ensures a sharp central vision at the cost of permanently reduced side vision [44].

Severe cases of *necrotizing enterocolitis*, *spontaneous intestinal perforation* and *patent ductus arteriosis* usually require surgical interventions, while a variety of medications are also administered in less extreme cases [47, 53]. Antibiotics are also commonly used to treat certain cases of sepsis, while anti-epileptic medications are provided if epilepsy is diagnosed [43]. These treatments are usually successful if the baby is not severely ill, completely curing the disease [47, 53]. However, in cases of extreme severity or a generally poor state of the baby, there may be post-treatment complications that require subsequent operations. Surgery to treat *patent ductus arteriosis* may require follow-up care by cardiologists throughout life, facilitating specific changes to lifestyle (exercise, diet etc.).

For high grades of PVL (>I), there remains no focussed treatments in the NICU, apart from maintaining a constant and correct level of brain oxygenation (through ventilation) as well as through hypothermia (brain cooling) [57]. This is a similar case for IVH (grades >II), although shunts can be surgically inserted to drain excess fluid. Medications are often instead provided to the mother early during pregnancy to limit the chances of having severe IVH in the first place [57].

2.3.4 The vital link to brain maturation

As stated in chapter 1, such cases of clear brain damage additionally affect general brain maturation. This can also lead to motor disabilities (including cerebral palsy in severe cases) as well as learning/behavioural problems (cognitive abnormalities)

that may only become clear once the baby is able to perform finer motor tasks and/or better communicate (i.e. by school-age) [43]. This said, the use of cUS for identifying brain damage can prove too sensitive, while not all cases of PVL (for instance) produce cysts that are easily recognisable on cUS scans [43].

Most significantly, deviations in brain maturation extend beyond clearly diagnosable brain damage (based on imaging modalities). The aforementioned treatments for the various diseases (surgical or otherwise) can also cause more subtle, yet adverse, alterations [43, 58]. In addition, recent research has pointed to the perception of pain during treatments as affecting cortical (in the outer layers of the brain) connectivity which, consequently, also adversely impacts brain maturation [59, 60]. Unfortunately, these subtler alterations remain difficult to detect in the NICU.

2.3.5 Treating abnormal brain maturation in the NICU

Due to the limitation of detecting abnormal brain maturation in the NICU, immediate treatments focus on more general approaches. This includes reducing disturbances to the baby, while maintaining an environment that is as constant as possible [43]. Furthermore, kangaroo care (skin-to-skin contact with the parent) is strongly encouraged due to clinical research suggesting improvements to neurodevelopment as a result [61–63].

If abnormal brain maturation could be identified much earlier in the NICU, one could exploit the increased plasticity of the preterm brain and its ability to better respond to treatments. This is the case for many, where further medication, physiotherapy, emotional bonding, and cognitive stimulation may correct milder deviations from normal maturation [64].

In the case of severe developmental deviations, it is often too late to completely prevent any mental and/or motor disabilities from surfacing later in life. Understanding what to expect early can help clinicians prevent neonatal death, and to appropriately counsel the parents during this difficult period [43]. Consequently, these babies remain the most critical cohort for clinicians to identify early.

Currently, the only accepted means by which to clinically assess brain maturation is at a later follow-up, typically from 9 months after NICU discharge. At this point, it is the resulting neurodevelopmental *outcome* of the baby that is measured, from which additional treatments are administered.

2.4 Assessing neurodevelopmental outcome

For a general and standardized way to assess neurodevelopmental outcome at follow-up, the Bayley Scales of Infant and Motor Development (BSID) were developed and remain the current clinical gold-standard.

2.4.1 Bayley Scales of Infant & Motor Development (BSID)

BSID consist of a series of tests that assess the expected mental, motor and behavioural proficiency of the baby, for their given age. It was first developed by. *N. Bayley* in 1969 [65], with an improved second version (BSID-II) released in 1993 [66]. Since 2006, a BSID-III framework has been introduced into NICUs, however, the majority of the data recorded in the NICU (at the time of writing) have continued with BSID-II such that this remains the focus here.

Clinical assessment focusses on mental and motor evaluation. BSID-II mental assessment is quantified using a normalized scored called the Mental Development Index (MDI). The MDI is designed to assess sensory perception, memory, learning, problem solving, early verbal communication (if old enough) and early number concepts [66]. Similarly, BSID-II motor assessment is quantified through the normalized Psychomotor Development Index (PDI), designed to assess body control and fine motor skills [66].

BSID-II tests come in a kit containing a series of intelligence- and physical-based tasks from which the MDI and PDI are calculated. In total, there are 178 test items to calculate MDI and 111 for PDI, with the scale applicable for corrected ages in the range 1-42 months [66]. Although each test is defined in detail and easily administered, they still require grading based on human interpretation.

Consequently, subjectivity is inevitable such that these tests are often undertaken by multiple independent raters with an associated inter-rater score of agreement.

A number of clinical studies have correlated BSID scores with certain abnormal outcomes (such as cerebral palsy) to assess inter-rater agreement. Results have proven reliable and reproducible, with normalized inter-rater scores typically >0.9 (out of a maximum of one) [67–69].

2.4.2 Treating abnormal outcome at follow-up

The late nature of identifying abnormal brain maturation, through assessing outcome at follow-up, limits the available solutions for treatment. Learning disabilities are difficult to completely cure, however, special education services and individualized education programs are now widely available to further aid the child's later education [70].

Motor deficiencies will often require a variety of disease-specific treatments, ranging from regular physiotherapy to invasive surgical procedures, speech therapy, toxin injections and other medications [64]. However, by this stage, the realistic aims are to reduce symptoms and halt further degeneration rather than to cure the disease completely.

2.5 Preterm sleep & the EEG

There remains a need to directly quantify brain maturation in the NICU, to identify developmental deviations earlier. As discussed in section 2.3.2, the EEG provides clinically-useful measures of acute behaviours (short duration events in need of urgent care) such as seizures. However, as stated in chapter 1, the ability of resting (background) EEG to capture brain maturation through sleep ontogenesis holds particular promise.

Preterm sleep states are clinically defined through a combination of EEG, behavioural, and cardio-respiratory criteria. The American Academy of Sleep Medicine (AASM) recently renewed the criteria for defining sleep and for visual sleep labelling in infants zero to two months of age [19]. However, there is no

standardized approach for defining sleep in preterm babies, with a common consesus generally decided between NICUs through collaborative research [41, 71].

In general, a maturing baby will fluctuate between periods of QS, AS, and wakefulness. QS is also known as non-REM sleep and is associated in preterm babies with various degrees of deep sleep. During deep sleep, there is heightened muscle atonia (as the baby is quiet and does not move), while the breathing and heart rates remain slow and regular. AS is also known as REM sleep and is characterized by REM movements, where the eyes move rapidly back and forth. This is the state where dreams occur, with enhanced movements and heightened vital signs evident [43].

2.5.1 EEG sleep ontogenesis

The changing characteristics representing QS and AS maturation with age, result in definitions of specific EEG sleep states. Figure 2.6 provides examples of these EEG states that develop with PMA and the associated characteristics are described in detail below.

<28 weeks PMA

At these very preterm ages, QS-, AS- and wakefulness-specific characteristics are not yet defined. Instead, more rudimentary EEG sleep states are present from as early as 24 weeks PMA that coincide with general REM and non-REM periods [72, 73]. During REM, specific characteristics include periods of predominantly high amplitude delta waves (0.5-3Hz frequency band), as shown in figure 2.6(a). These alternate with periods of strong suppression, known as Inter-Burst Intervals (IBIs), during non-REM. IBIs have typical durations (or lengths) of ≈30 seconds but can last up to 40-60 seconds [41, 43, 71, 74, 75].

28-29 weeks PMA

During this period, deep neuronal connections develop such that the EEG begins to reflect two distinct stages of QS and AS/wakefulness. However, these differences remain subtle, with IBI durations becoming shorter and slightly more variable

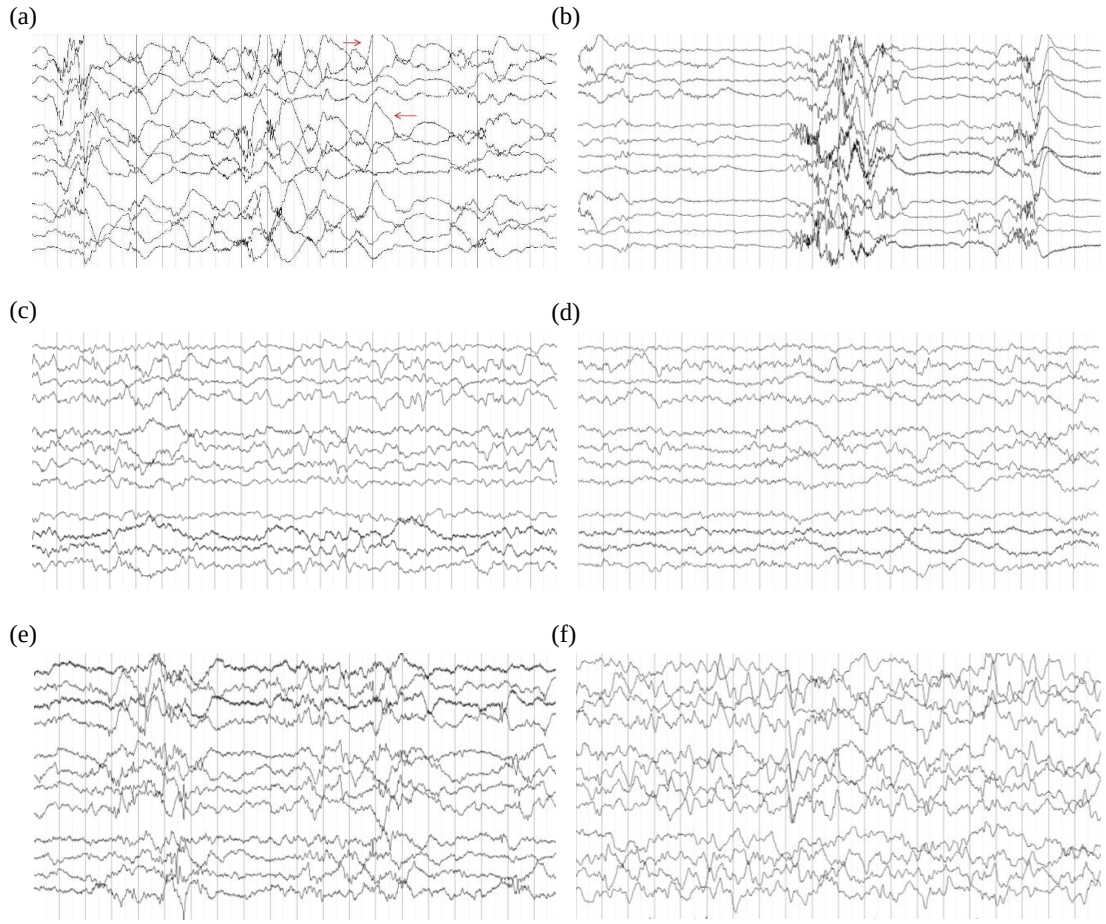


Figure 2.6: Examples of the different EEG sleep states that develop with PMA. (a) A rudimentary sleep state (reproduced with permission from [43]) (b) TD (QS and AS). (c) ASI (AS). (d) LVI (AS and wakefulness). (e) TA (QS). (f) HVS (QS).

between states. There also exists very high proportions of Indeterminate Sleep (IS). IS is defined as periods of sleep where the corresponding EEG characteristics and 'non-cerebral' measures (particularly heart and breathing rates) do not correspond to definite segments of QS, AS or wakefulness [43, 76, 77].

30-31 weeks PMA

It is at these ages that non-cerebral measures of REM, heart and breathing rates begin to synchronize with more specific EEG characteristics that represent QS, AS and wakefulness [17, 74, 76, 78]. EEG characteristics during QS are defined as 'discontinuous', with more distinct bursts of high amplitude delta activity (with durations $\geq$3 seconds) and increasing theta (3-8Hz frequency band) and alpha

(8-12Hz) behaviours, separated by shorter length IBIs (≤15-20 seconds). These periods of fluctuation between bursts and IBIs are henceforth referred to in this thesis as 'burst-inter-burst' behaviours and the resulting EEG state is known as Tracé Discontinu (TD). TD is illustrated in figure 2.6(b).

During both AS and wakefulness, similar TD behaviours are seen but IBI lengths are shorter (≤5-8 seconds). Non-cerebral measures of REM remain most distinctive of AS, while increased movement is indicative of wakefulness [71]. Towards 31 weeks, AS EEG becomes increasingly semi-continuous as burst amplitudes shorten and IBI amplitudes begin to increase [41, 43].

32-36 weeks PMA

From 32 weeks PMA, AS (and wakefulness) EEG is now defined as 'semi-continuous to continuous', with no discontinuous burst-inter-burst behaviours for ≥60 seconds. The overall amplitude of this behaviour decreases with increasing PMA. TD persists during QS but IBI lengths continue to decrease with increasing PMA.

Delta, theta and alpha activity continue to vary in amplitude and frequency across brain locations. Specifically between 32-34 weeks, delta waves with superimposed fast alpha rhythms (delta brushes) are seen across QS and AS/wakefulness and increase in prevalence during this period. Specific short-duration, high-frequency theta activity, known as temporal sawtooths, are also seen initially but decrease after 32 weeks in AS/wakefulness and after 34 weeks in QS. [41, 43].

By 35-36 weeks PMA, QS TD behaviour becomes increasingly semi-continuous (with IBI lengths now <10 seconds) and more variable (with higher amplitudes of 15-25μV). AS and wakefulness remain continuous but also with higher amplitude.

37-38 weeks PMA

It is at the onset of term-age that the growth of major cortical connections results in the differentiation of EEG into four distinct sleep states (two from QS and two from AS), as well as wakefulness. AS differentiates into a continuous state containing increased mixed-frequency information and is referred to as ASI (figure 2.6(c)). A further ASII state follows, consisting of a Low-Voltage Irregular (LVI)

behaviour (figure 2.6(d)) that contains intermixed delta and theta activity of lower voltage (when compared to other states at this age) [19, 43, 71, 74, 75]. A similar LVI pattern also constitutes wakefulness [71].

During QS, TD behaviour has sufficiently matured to an extent where the length of the burst periods (of amplitudes 50-150 μV) are equal to the length of the IBIs ($\approx$6 seconds). IBIs now also have high amplitudes extending to 25-50μV, with this state now known as Tracé Alternant (TA) and illustrated in figure 2.6(e). TA is preceded by a further QS High Voltage Slow-wave (HVS) state of predominantly continuous, high voltage delta activity (figure 2.6(f)) that appears from 38 weeks PMA. Delta brushes persist across all QS, AS and wakefulness states but decrease in propensity towards 38 weeks PMA [41, 43].

This sleep differentiation leads to a globally more continuous EEG pattern at term-age and the relative changes in discontinuity between QS, AS and wakefulness become less distinct [79].

>39 weeks

By well into term-age, delta brushes completely disappear from the AS states but remain present in QS, while HVS becomes more predominant with increasing PMA.

By 48 weeks PMA, HVS completely defines the QS state, with TA now completely disappearing. At this point, transitions from neonatal to infant sleep occurs, with AS now referred to as REM sleep and QS as non-REM sleep [80]. Sleep spindles (characteristic 10-12Hz frequency bursts of 0.5 second duration) begin to appear in non-REM which is distinctive of infant and adult sleep [19, 43, 81]. The clinical definitions of sleep are now more consolidated and standardized by the AASM [43].

Summary of EEG sleep maturation

Figure 2.7 summarizes the development of the EEG sleep states with PMA, highlighting changes in IBI lengths and the differentiation of the sleep states by term age. The changing proportions of QS and AS states with PMA are also illustrated and have a direct impact on sleep-wake cycling. Furthermore, the general

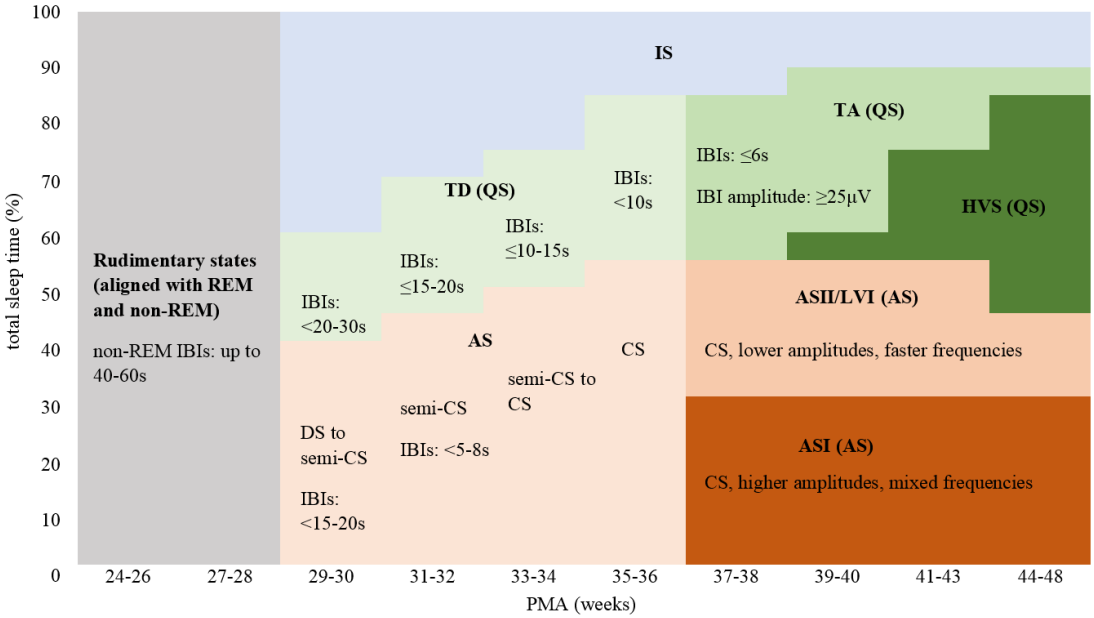


Figure 2.7: Summary of the changing nature of EEG sleep with PMA. Changes in total sleep time (which impacts sleep-wake cycling) and notable differences in particular EEG characteristics are described, focussing on the reduction in IBI lengths (in seconds (s)), and the transition from discontinuous (DS) to continuous (CS) patterns. Figure adapted with permission from [43].

reduction of IS contributions to sleep (with PMA) is shown and emphasizes the improved correlation between EEG and non-cerebral measurements.

2.5.2 Automated sleep staging

The detailed clinical description of EEG sleep staging in section 2.5.1 forms the basis for visual labelling by trained clinicians and EEG specialists. A number of automated methods that utilize both signal processing and machine learning have been subsequently developed to classify these states. These range from deriving biomarkers that reflect sleep state-specific EEG characteristics, to more complete supervised and unsupervised algorithms.

Sleep state biomarkers

Earliest sleep detection methods focussed on identifying individual clinically-motivated EEG features, as potential biomarkers to infer differences across sleep state. The most common features were derived by filtering the EEG into the distinct

frequency bands of delta, theta, alpha, and beta (12-30Hz) which are intrinsic to the clinical definitions of EEG sleep states. *Myers et al.* derived these bandpowers and identified early differences with sleep state in preterm babies at term-age [82].

In 2001, *Turnbull et al.* extracted similar measures to identify the TA sleep state. The evolution of the total theta bandpower was used to detect TA in full-term babies [83], however, while these methods proved reliable for TA detection, they were not sufficient to completely encompass the continuous HVS patterns also seen in QS by this age [71].

Paul et al. later applied a similar comparison at term-age (with both preterm and full-term babies), that extended the feature range beyond bandpowers to additional amplitude-based measures. These were derived from segmenting the EEG into 'quasi-stationary' segments (approximately stationary if of sufficiently short duration). Such measures revealed statistically significant differences across visually labelled periods of AS and QS [84, 85]. In 2009, *Piryatinska et al.* also introduced a significant non-linear feature, based on the fractal dimension. This quantified the level of smoothness, or 'complexity', in the segment between a value of one (a smooth, suppressed behaviour) and two (a very eccentricized, discontinuous behaviour) [86].

While these methods showed potential at term-age, utilizing such features for predicting sleep state across a wide range of preterm ages remained untested, until classification frameworks were developed to combine such biomarkers into a predictive model.

Supervised classification methods

A large number of methods have been developed to estimate periods of EEG bursts and suppressions in preterm babies [26, 28, 87–94]. However, as the EEG matures to more continuous patterns, such methods cannot classify the complete sleep state.

The first method of supervised sleep staging was developed as early as 1996 by *Scher et al.*, and applied to both preterm and full-term babies at term-age. Bandpowers from EEG and EMG, REM behaviour from EOG, and a variety of frequency-domain measures from nasal, thoracic and abdominal respiratory channels

were extracted as features and used to classify QS, AS or wakefulness. The study achieved an overall accuracy of >90% [95] but the technique remained reliant on certain manually calculated features (such as the number of wake events, or arousals) such that it was not fully automated.

From 2006, *Gerla et al.* did develop a fully automated method for term-age sleep detection of QS, AS, wakefulness (as well as movement) using a number of classifiers which incorporated knowledge of previous sleep state transitions. The method detected the sleep states by extracting features across different PSG signals. As well as EEG features, measures based on regularity of respiration, eye movements (from EOG), muscle motion (from EMG) and heart rate regularity (from ECG) were incorporated. These classifiers additionally modelled the transitions between sleep states, achieving peak accuracies of 70% [96, 97]. However, focus still remained on only term-age sleep staging without considering a wide preterm age range. Furthermore, these methods at best only detected the general states of QS, AS and wakefulness, but when focussing on term-age, there still remains no existing method that models the differentiated EEG states of ASI, ASII, HVS, and TA. This said, such transition-based classifiers remain promising candidates to achieve this.

Both the methods of *Scher et al.* and *Gerla et al.* also used multiple PSG modalities and did not focus specifically on EEG. In light of this, *Koolen et al.* developed a supervised classification method in 2017, which did focus solely on EEG and performed across a wide PMA range of 25-45 weeks. This method achieved an overall accuracy of 85% [34]. A standard classifier was trained, using EEG features that extended beyond simple bandpowers to include measures of burst dynamics and EEG functional connectivity (that model the relationship between different EEG channels and brain regions) [34]. Both GA and PMA were also included as features to account for maturational effects. However, it remains unclear if incorporating prior knowledge of age would limit the usefulness of this approach when used for quantifying brain maturation.

De Wel et al. dispensed with using PMA as a feature, training an alternative model to classify QS and non-QS (combining both the AS and wakefulness states

together). This method exclusively used entropy features that captured the complexity of the EEG [98] and achieved an accuracy of 90% over the range of 27-42 weeks PMA [29]. *Lavanga et al.* developed a similar method for QS detection using the same dataset, instead focussing on features based on fractal dimension. The method achieved a comparable accuracy of 89% [99].

Both studies revealed that with an appropriate choice of features, one could achieve accurate sleep staging performance across a wide PMA range while accounting for maturational effects. However, with a focus on only specific families of complexity features, it does beg the question of whether adopting a wider range of data-driven metrics could further improve performance and uncover additional information.

Finally, *Ansari et al.* most recently became the first to apply deep learning to preterm sleep staging. A deep neural network was trained with a series of hidden layers to extract abstract but meaningful features directly from the native EEG recordings. A final classification layer then estimated periods of QS and non-QS, with an accuracy of 92% [100].

Unsupervised classification methods

Unsupervised methods for identifying sleep stages have been around since 1991, first proposed by *Krajca et al.*. EEG was segmented into regions of similar characteristics by utilizing an adaptive segmentation algorithm [101]. A handful of traditional amplitude and bandpower features were extracted from each segment and segments with similar characteristics were grouped together using a clustering algorithm [102, 103].

After clustering, each EEG sample from the original recording was then replaced with its cluster label and plotted over time, deriving what *Krajca et al.* first introduced as a 'temporal profile' [102]. These temporal profiles were suggested to reflect the higher QS EEG discontinuity when compared to AS (and wakefulness). However, while successful when used as part of other methods (such as IBI detection [88]), the potential for automated sleep staging has never been validated.

Furthermore, the method showed vulnerability to high power artefacts which could result in misclassifications [101].

Arguably, the most widely used method in current preterm EEG research was developed by *Stevenson et al.* in 2014, and is based on detecting the EEG bursts which the authors referred to as 'Spontaneous Activity Transients (SATs)'. Multichannel EEG was first classified into general periods of high and low discontinuity by defining the bursts (SATs) and IBIs ('inter-SATs') using a specialised detection algorithm developed by *Palmu et al* [89]. By plotting the classification of SATs (and inter-SATs) over time, the resulting binary signal was then integrated over a sliding window to determine the evolution of SAT percentage (SAT%) with time. A threshold was applied to separate the EEG into periods of low activity (lower SAT% and greater discontinuity) and high activity (higher SAT% and more continuous EEG) [33].

The SAT% method proved very effective in detecting fluctuations in discontinuity and, from the point of view of sleep staging, it was accurate to assume these differing levels of EEG discontinuity as somewhat indicative of QS (and non-QS) periods, particularly when the preterm baby was <32 weeks PMA [33]. However, this narrow age range remains a limitation and is due to the assumption that the EEG always has a relatively low proportion of IBIs in high activity states, compared to low activity states. Beyond 32 weeks PMA, the gradual reduction of IBIs and increase in EEG continuity causes these assumptions to fail. Consequently, there remains no unsupervised method that is proven to classify sleep stages across a wide PMA range (specifically ages ≥32 weeks PMA).

2.6 Identifying EEG abnormalities

In cases of abnormal brain maturation, the normal development of the EEG states (as defined in section 2.5.1) is no longer followed. Alternative distinct EEG behaviours have been clinically identified based on the underlying cause of the abnormality. Abnormal brain development results in the EEG exhibiting either a 'dysmature' or disorganized pattern (or more commonly, combinations of both).

Both patterns are easily identified visually (by trained clinicians) according to the specific definitions below.

2.6.1 Dysmature patterns

With a distinct evolution of the background EEG characteristics with PMA, it is expected that normally developing babies will exhibit EEG patterns that reflect the actual PMA of the baby. Alternatively, one can say that the EEG 'brain-age' matches the PMA. For example, if a baby were born at 30 weeks GA with EEG recorded at 35 weeks PMA, the expected EEG characteristics and associated brain-age would be similar to that of a baby born at 35 weeks GA [43]. The allowable normal error between brain-age and PMA is clinically defined as ± 2 weeks for preterm babies at <37 weeks PMA, and one week for term-age babies [74, 104].

When larger deviations are noted, this is indicative of decelerated (or accelerated) rates of maturation and the EEG is deemed dysmature. Associated EEG behaviours include [20, 23, 104–107]:

- An abnormally high level of discontinuity for the given PMA.
- The persistence of theta activity (namely temporal sawtooths) beyond 32-34 weeks PMA.
- The persistence of high-voltage delta brushes (particularly during AS).
- Poor differentiation into multiple QS and AS sleep states near term-age.

2.6.2 Disorganized patterns

A disorganized pattern, unlike dysmaturity, has more distinctly unusual activity without any clear signs of the typical EEG sleep ontogeny. Instead, there are increased amounts of abnormal, sharp transients, as well as deformed 'cogwheel'-shaped delta brushes [21, 43]. An example of such an EEG pattern is shown in figure 2.8. The EEG often also exhibits severely depressed (low-amplitude) EEG behaviour before development into disorganized patterns [21].

Disorganized patterns are typically linked to white matter brain injuries or others that result in severe motor-based disabilities [43, 74, 106].

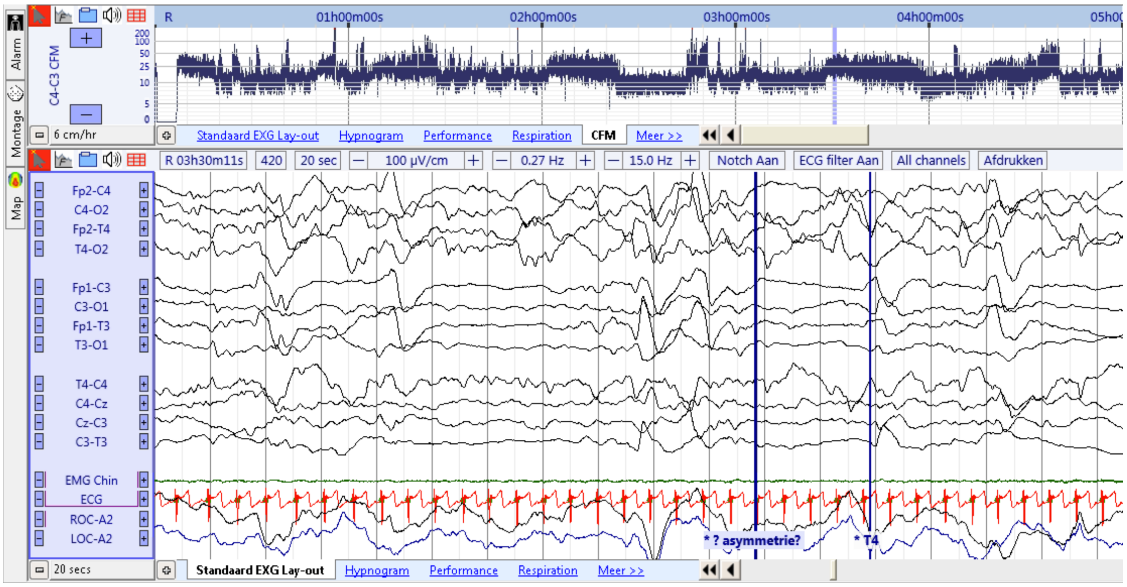


Figure 2.8: 20 seconds of a typical disorganized multichannel EEG pattern (Leuven NICU bipolar montage) from a preterm baby (31 weeks GA) recorded at 43 weeks PMA.

2.7 Quantifying brain maturation

Figure 2.7 summarized the age-dependent changes in EEG characteristics and sleep wake cycling (through changing sleep state proportions and differentiation). As first mentioned in chapter 1, early automated methods for quantifying brain maturation have focussed on both these aspects of EEG sleep as a potential means to identify abnormal brain maturation (due to both dysmature and disorganized patterns). This began with deriving novel biomarkers to capture the changing EEG characteristics with age, followed by methods for tracking brain maturation and quantifying sleep wake cycling.

2.7.1 Maturational biomarkers

Many of the sleep state biomarkers discussed in section 2.5.2 also have the potential to be important indicators of maturation. This began with amplitude and bandpower features (in delta, theta, alpha and beta) which showed significant trends with PMA across the full EEG and visually labelled periods of QS [4, 22, 108–115]. Similarly, maturational trends were demonstrated with measures of EEG complexity (namely entropy [29, 116, 117]) and through directly measuring changes in burst

and suppression behaviour [4, 92, 118]. However, there was no consideration for trends in these biomarkers across multiple sleep states.

Tolonen et al. first attempted to distinguish changes in bandpower across EEG states, applying a SAT-detection algorithm to recordings in the range 32-46 weeks PMA. The algorithm isolated periods of SATs and inter-SATs in the EEG, and assessed changes in SAT and inter-SAT bandpowers over the delta-beta range. Statistically significant trends with PMA were revealed [119]. However, as also discussed in section 2.5.2, SATs don't fully capture the sleep states as they transition to more continuous behaviours with age. Consequently, *Whitehead et al.* identified further trends in bandpower features within visually-labelled periods of QS, AS and wakefulness. In these cases, maturational trends were derived from preterm babies with normal neurodevelopmental outcome (at 9-24 months follow-up) and additional assessment across abnormal outcomes were not considered.

There are many clinical studies that have compared such biomarkers across cohorts of babies with *specific* abnormalities albeit over a narrow PMA range. For instance, persistence of low-frequency delta patterns were correlated with dysmature babies at term-age [16], and *Iyer et al.* also quantified EEG bursts in extremely preterm babies 72 hours after birth, linking specific alterations in burst behaviours to diseases such as IVH [25, 120]. However, there remained no consideration for the dependencies on sleep state.

Recent studies have identified biomarkers that reveal differences across both abnormalities *and* sleep state. *Lofhede et al.* found significant trends in delta and theta bandpowers in babies with asphyxia, as well as through quantifying burst and suppression patterns. This method incorporated an automated algorithm for the general separation of continuous and discontinuous EEG [87, 121]. *Shellhaas et al.* also identified changes in delta powers and higher entropies during visually labelled QS and AS periods in patients with and without HIE [24]. However, these methods focussed only on term-age babies.

Finally, *Cohen et al.* identified altered sleep-dependent trends across a wide PMA range using EEG bandpowers, focussing on abnormal preterms with restricted foetal

growth [122]. However, this approach was not fully automated, requiring visual labelling of the sleep states. Furthermore, with these methods designed for specific abnormalities, they were unable to assess general neurodevelopmental outcome.

While *Suppiej et al.* did directly link higher bandpowers in alpha and beta at 35 weeks PMA to normal and abnormal neurodevelopmental outcome [123], methods that achieve this remain few and far between and additionally do not incorporate both a wide PMA range and multiple sleep states.

Functional brain connectivity

The maturational biomarkers so far discussed have focussed on single EEG channels or were summarized across multiple channels by taking the mean or median value. However, another field of interest focusses on quantifying the synchrony between different channels (particularly within burst and suppression periods) as a means to measure functional brain connectivity [124–127]. The clinical potential of studying synchrony was illustrated in cases of HIE [128] and *Tokariev et al.* further identified changes in synchrony between babies with and without major cerebral lesions [129].

The majority of studies have aimed to define synchrony measures that correlate strongly over a wide range of PMA. Maturational trends in synchrony (and asynchrony) were identified based on statistical differences in features between channels, during QS and AS. Measures of bandpower coherence between channels, were identified as potential biomarkers [130–133]. Similar measures based on amplitude and phase were also shown to capture changes in synchrony [134–136]. However, the precise definition of synchrony and the allowable temporal delays between channels were strongly dependent on age and EEG duration, remaining ambiguously defined across studies [137].

In 2013, *Rasanen et al.* proposed the 'activation synchrony index' as a potentially more unifying and interpretable measure to examine temporal alignments over longer periods of EEG [137]. The activation synchrony index was based on measuring the statistical dependency between two channels. By quantizing both channels into discrete sequences $[a_1, a_2, \cdots, a_N]$ and $[b_1, b_2, \cdots, b_N]$ (with N the sequence length),

respectively, statistical dependency was first determined using the Energy-weighted Temporal Dependency Function (ETDF):

$$ETDF(\tau) = \sum_{i,j} E(a_i)E(b_j)\frac{p_\tau(a_i, b_j)^2}{p(a_i)p(b_j)} \tag{2.1}$$

$E(a_i)$ and $E(b_j)$ are the amplitudes corresponding to quantization levels a and b at location i and j, respectively. $p_\tau(a_i, b_j)$ denotes the joint probability of the pair (a_i, b_j) with a delay of τ sequence values between them, and $p(a_i)$ and $p(b_j)$ denote the individual quantization level probabilities.

$ETDF(\tau)$ was calculated across the lag range $\tau = -50$ to 50 and the values de-trended by subtracting the global minimum to achieve $ETDF_{norm}(\tau)$. The activation synchrony index was then robustly estimated by assuming that a peak would occur in $ETDF_{norm}(\tau)$ at zero lag (i.e. at $\tau = 0$, when the two sequences were perfectly aligned and maximally synchronous) [137]:

$$sync = ETDF_{norm}(\tau = 0) / \left(\frac{1}{101}\sum_{\tau=-50}^{50} ETDF_{norm}(\tau)\right) \tag{2.2}$$

Here, *sync* denotes the activation synchrony index. *Koolen, Dereymaeker et al.* subsequently confirmed this metric as a successful maturational biomarker by obtaining significant trends with PMA from visually-labelled QS EEG [30].

Most recently, *Lavanga et al.* modelled the connectivity between different brain regions using graph-based measures [138]. Regressing these measures with PMA, as well as seperately for visually-labelled QS and non-QS, revealed statistically significant trends [138].

Functional connectivity is clearly a strong maturational biomarker and has a proven ability to detect abnormal EEG patterns. While the clinical usefulness of such methods is unquestionable (with the activation synchrony index particularly more robust over longer EEG segments (epochs)), it remains unclear how such measures will perform in cases of adverse EEG quality. Due to channel specificity, EEG measurements must be of a reasonably high standard across the electrodes of interest. The presence of small artefacts could possibly result in strong asynchrony even in cases of truly synchronous background EEG. As a result, current methods carefully

pre-select the EEG for analysis. This is an important factor when considering such measures as features for more data-driven scenarios, as part of larger automated procedures, and when using data with limited visual pre-screening.

2.7.2 Estimated Maturational Age (EMA)

A number of approaches have previously aimed to classify particular preterm abnormalities, namely HIE, by using a combination of maturational biomarkers [30, 50, 88, 139–144]. However, *O'Toole et al.* was the first to define EMA as the expected age of the preterm baby from background EEG. EMA provided a means to combine multiple maturational biomarkers together and predict the general state of brain maturation over PMA.

The EEG was first divided into periods of low and high 'activity' based on SAT%, using the SAT detection algorithm of *Palmu et al.* [89]. A combination of the maturational biomarkers discussed in section 2.7.1 were computed from the full EEG recording, as well as from the high and low activity periods. Regression models were then trained to predict the baby's PMA from these biomarkers, resulting in an estimate of EMA. Training was possible as all recordings were visually pre-selected for normal outcome at follow-up (such that the EMA should equal the PMA) [145].

The most accurate EMA estimates were achieved using features derived from periods of low activity, with an error of 1.29 weeks [145]. However, only very preterm babies of 23-32 weeks PMA were assessed. *Stevenson et al.* later applied this method to a wider PMA range of 24-38 weeks, resulting in an error of 1.49 weeks. This study further defined each baby's developmental trajectory by plotting the EMA of the EEG recordings against PMA [146]. *De Wel et al.* performed a similar EMA prediction with only entropy features using visually labelled QS periods. The final regression yielded an error of 1.88 weeks [29].

Lavanga et al. alternatively utilized the graph-based functional connectivity biomarkers (first discussed in section 2.7.1) to train EMA regression models. Separate models were developed for both QS and non-QS, with QS achieving the best performance with an error of 2.11 weeks [138]. This study was subsequently

extended to include additional connectivity measures with an improved error of 1.51 weeks [147].

As described in section 2.6, particularly dysmature patterns are defined based on discrepancies between the PMA and the EEG brain-age. EMA is a measure of brain-age, however, these existing approaches have only used such methods as means to validate novel maturational biomarkers through achieving accurate estimates of PMA in normal outcome babies. Expanding this concept to include abnormal outcomes could transition such models from simply validation tools to legitimate diagnostic tools that can identify developmental deviations by tracking brain maturation.

Furthermore, while *O'Toole et al.* and *Stevenson et al.* accounted for different EEG characteristics in periods of low and high discontinuity, this does not fully encompass sleep states (especially near term-age). While *De Wel et al.* did focus on QS and *Lavanga et al.* developed models for both QS and non-QS, these were all visually labelled by clinicians. Expanding the concept of brain-age prediction models to a fully automated method will inevitably require integration with automated sleep staging algorithms.

Finally, methods that focussed on selecting a wider range of EEG biomarkers showed a tendency to produce better brain-age prediction performance than when using only specific families of biomarkers (such as complexity or functional connectivity). As also suggested for sleep staging in section 2.5.2, incorporation of larger data-driven feature sets may prove beneficial. In this context, it could extend brain-age prediction through potentially uncovering outcome-sensitive biomarkers after incorporating abnormal outcome data.

2.7.3 Quantifying sleep-wake cycling

The most common approaches have involved quantifying specific measures that summarize the overall sleep-wake cycling, some of which extend from the maturational biomarkers defined in section 2.7.1. These methods refer to sleep-wake cycling as Brain Activity Cycling (BAC), encoding the general fluctuation in 'vigilance states' of the baby through EEG [43]. Observing differences in BAC behaviour,

with respect to PMA, has previously been linked with adverse neurodevelopmental outcomes [77, 148–151]. The earliest forms of BAC was based on a visual assessment of aEEG, using measures of signal amplitudes and bandpowers [148, 149, 152, 153].

Automated methods have predominantly focussed on quantifying BAC, with the first introduced by *Niemarkt et al.*, who examined changes in the IBI-to-Bursts Ratio (IBR). As a first step, an automated algorithm identified IBIs based on the total bandpower and defined periods when the power fell outside minimum and maximum thresholds [31]. Within a sliding window, IBR was calculated based on the changing percentage of IBIs across the recording, with higher IBRs denoting more discontinuous EEG activity. The mean IBR over the full EEG recording was determined as a summarizing metric of BAC, showing statistically significant maturational trends [31]. *Navakatikyan et al.* proposed a similar approach for quantifying BAC, by developing an alternative algorithm to automate IBI detection. This was based on extracting maturational features from a derived peak-to-peak amplitude signal called the range-EEG (rEEG) [28].

In 2014, *Stevenson et al.* developed a method based on the SAT% signal discussed in section 2.5.2. The change in SAT% over time already provided a representation of BAC and by extracting a series of summarizing BAC measures from the signal's power, these were shown to correlate well with visual measures from aEEG [32, 33, 154].

Koolen, Dereymaeker et al. later developed a method in 2015 to quantify BAC based on the state of EEG suppression. These aptly named 'suppression curves' were derived from a measure called Line Length (LL), defined as the sum of absolute distances between EEG samples in successive one second epochs. Ratios of the median to mean LL values provided a measure of suppression, *supp*. Higher *supp* values denoted higher levels of suppression (reduced burst percentages and longer IBI durations) and computing *supp* for successive windows of EEG produced the suppression curve. The mean *supp* value across the curve formed a summary metric exhibiting strong maturational trends with PMA [27]. From LL, a measure of

the percentage of bursts (Burst%) could also be derived and *Koolen et al.* further demonstrated that this could additionally quantify BAC [26, 34].

Most recently, *Koolen et al.* further derived a measure of BAC by directly assessing changes in a 'sleep state probability index' after first performing automated sleep staging [34]. This remains the only approach that explicitly classifies QS and AS to derive a measure of BAC and showed similar performance when compared to previous measures of SAT%, *supp*, and Burst% [34].

Apart from the sleep-state probability index, existing BAC approaches focus on quantifying bursts and suppressions with single measures. While these measures reveal clear trends in ages where EEG discontinuity is present, they become potentially indistinguishable beyond term age as the EEG becomes increasingly continuous. This was demonstrated by *Dereymaeker et al.* who showed very little differences in *supp* >37 weeks PMA [27]. Consequently, sleep-wake cycling is not truly represented at older ages using the BAC approach.

An alternative approach is to assess changes in sleep state-specific sleep-wake cycling measures, such as sleep state proportions, durations, cycle lengths, and regularity. These have already shown great clinical potential [23, 24, 35, 36]. However, existing methods were limited to late preterm to term-age babies and all required prior visual labelling of the sleep states. The ability to perform automated sleep staging could further automate this clinically-driven approach to quantifying sleep-wake cycling. In particular, the differentiation of term-age sleep (an important indicator of dysmaturity) could be better captured as BAC measures are unable to fully represent this.

2.8 Summary & current limitations

- Continuous EEG monitoring captures sleep ontogenesis through the maturation of distinct EEG states that differentiate at term age.

- There remains no integrated or standardized approach to utilize this information in the NICU and identify abnormal brain development. Currently, neurodevelop-

mental outcome is only assessed 9 months later (at the earliest) through BSID-II scores.

- Very few methods exist to automatically classify EEG states over a wide PMA range, with most approaches all supervised and using specific sets of sleep state biomarkers. While unsupervised methods show promise, they are either not yet validated or are only effective over a narrow PMA range.

- Deviations in EEG sleep state development result in clinically-defined dysmature or disorganized EEG patterns (or combinations of both).

- Expanding to a larger, data-driven approach for feature extraction may uncover additional biomarkers sensitive to sleep state and brain maturation. Maturational biomarkers may also be identified that show sensitivity to neurodevelopmental outcome.

- EMA/brain-age prediction models show potential to track the changing EEG characteristics with PMA and quantify brain maturation. However, it is only assessed on normal outcome data with limited consideration for the effects across multiple sleep states. Incorporating abnormal outcomes and automated sleep staging is vital to develop these methods as diagnostic tools to identify developmental deviations (linked to abnormal neurodevelopmental outcome).

- Existing automated methods for quantifying maturation through assessing sleep-wake cycling involves using single summarizing metrics, with limited efficacy near term-age. No automated methods explicitly classify sleep to assess the changing sleep state proportions, durations, and overall quality of cyclicity with age. In particular, the differentiating sleep states by term age is an important checkpoint for identifying dysmaturity.

3
Datasets

3.1 Introduction

Chapter 2 included a clinical background on the preterm baby. Particularly, common physical and cognitive abnormalities were described, including standard diagnoses and treatments within the NICU and at follow-up.

In this chapter, the datasets used throughout this research are presented, based on the clinical definitions described in chapter 2. This includes the procedure for patient recruitment and data collection, as well as the NICU-specific labelling protocols for sleep and neurodevelopmental outcome.

All the data was obtained through collaborations with the NICU at the UZ Leuven Hospitals in Leuven, Belgium.

3.2 Patient recruitment & data collection

The study to recruit patients was conducted between September 2013 - September 2015 and approved by the ethical commission of the UZ Leuven hospitals. All patients without multiple congenital anomalies (birth defects) and/or dysmorphism (differences in body structure) were eligible for the study and infants were consecutively enrolled after informed parental consent. All preterm babies selected had a GA <32 weeks.

cUS scans were performed at least on days one, three, seven, 14, and repeated every subsequent two weeks until discharge. In cases of infants with brain lesions, this timing varied. Scans were then ranked into four categories based on the IVH and PVL grades (see section 2.3.1):

1. Normal: No signs of brain lesions
2. Mild: IVH grades I-II and/or PVL grade I
3. Severe: IVH grade III and/or PVL grades II-III
4. Other: Rare birth defects such as *corpus callosum agenesis* (complete or partial absence of the *corpus callosum* of the brain.)

The first EEG (and aEEG) recording was obtained as soon as the patient was clinically stable (between the first and third week of life). One EEG recording was then taken every two to three weeks up to a maximum of six recordings, before transfer or discharge from the NICU. All recordings were measured at 250Hz sampling frequency with a native bandpass filtering of 0.27-70Hz. A total of nine electrodes were used and placed according to the 10-20 electrode system and subsequently analysed using a single referential montage (Fp1, Fp2, C3, C4, T3, T4, O1, O2, and reference Cz), as described in section 2.3.2 and illustrated in figure 2.5 (BRAIN RT equipment, OSG, Mechelen, Belgium). Kangaroo Care was also allowed and encouraged during recording [43].

EEG recording length was decided to ensure that it covered at least two sleep cycles, with preferably a minimum of 3-4 hours when possible (based on the stability of the patient). This was recorded as part of a full PSG, with additional non-cerebral channels for respiratory activity (using a nasal thermistor) and ECG. For patients $\geq$36 weeks PMA, additional EMG, EOG (ROC and LOC), and piezo-electric belts were applied.

3.3 Grading neurodevelopmental outcome

All recruited patients were subsequently assessed at 9 months follow-up (and 24 months in some datasets) for neurodevelopmental outcome. BSID-II scores were determined for each patient (see section 2.4.1) and used as the primary measure, with the cUS categories used as a secondary measure. From these assessments, patients were organized into four neurodevelopmental outcome groups of varying severity. Normal outcome (group 1) was the most stringently defined according to the following criteria [43]:

- Normal BSID-II scores of MDI and PDI ≥85 [146] (defined as -1 Standard Deviation (SD) from a mean MDI/PDI score of 100).
- The absence of a severe grade brain lesion and no PVL.
- No use of any sedative or anti-epileptic medication during EEG recording.

The additional abnormal outcome groups were defined as follows:

1. Mild (group 2): Minimum BSID-II MDI or PDI >70 and <85.
2. Severe (group 3): Minimum BSID-II MDI or PDI ≤70 or presence of cerebral palsy.
3. Lost-to-follow-up/death (group 4): All patients who passed away before 9 months follow-up.

3.4 EEG visual labelling

All EEG recordings went through a visual pre-screening to identify any regions of extreme artefacts that presented across the majority of channels. These were predominantly due to large movements that corrupted the signal. A trained (primary) clinician then identified the longest possible continuous period of EEG where such artefacts were not present, while still including at least one complete sleep cycle. With the resulting recordings still of a long duration, some channel-specific deviations consisting of short duration high-power artefacts remained.

Some of the datasets were additionally visually labelled for sleep states, first by the primary clinician (rater), before a secondary rater double-checked the labelling of the first. A consensus was reached between the two raters for the final sleep labels. Labelling was performed using the EEG as well as additional PSG channels, according to the criteria outlined in section 2.5.1.

3.5 Overview of EEG data & demographics

Three separate datasets are defined in total. Although these were recorded within the same recruitment period and under the same conditions, they differ based on neurodevelopmental outcomes, timing of follow-up and the extent of visual

labelling. The specifics for each dataset are described below, including a summary of the clinical demographics.

3.5.1 Dataset 1

Dataset 1 includes the earliest patients recruited during the study period. All patients were preterm, with neurodevelopmental outcome assessed at both 9 months and 24 months follow-up. The dataset consists of the first 26 eligible normal outcome patients, with demographic and EEG information summarized in table 3.1. All EEG recordings were visually labelled for periods of QS (and non-QS).

3.5.2 Dataset 2

Dataset 2 includes preterm patients recruited later in the study period and, consequently, neurodevelopmental outcome is only provided at 9 months follow-up. This dataset is larger than dataset 1, and includes eligible patients across all four outcome groups. Visual sleep staging was not performed such that this dataset is unlabelled for sleep. The demographics and EEG information are summarized in table 3.2.

3.5.3 Dataset 3

Dataset 3 is the smallest, consisting of 16 patients and 16 recordings (one from each patient) measured at term-age, with PMA range 38-43 weeks. Mean EEG monitoring time is (hh:mm) 15:38 (range: 11:45-18:42) and mean EEG recording length is 07:53 (range: 01:58-17:48). Eight of these patients were preterm and recruited from the same population used to produce dataset 1 (with two patients included in both datasets 1 and 3). An additional eight healthy full-term patients were recruited and had no clinical complications related to preterm birth, such that the demographic information was not collected.

All EEG recordings were pre-screened for the highest available quality and a typical EEG behaviour for the given PMA. All recordings were visually labelled for the term-age EEG states of ASI, ASII, TA, HVS, IS, as well as ‘dubious’ periods (where the two raters did not agree). Channel-specific artefacts were also identified.

Table 3.1: Summary of patient demographics and EEG information for dataset 1. All patients were of normal neurodevelopment outcome at follow-up.
Values are expressed with % proportions, mean ± SD, and/or range (where appropriate). High-frequency oscillation ventilation: another form of mechanical ventilation. Bronchopulmonary dysplasia: need for oxygen/ventilation at 36 weeks PMA. Sepsis: positive or highly suspected based on blood tests, with antibiotic treatment >72 hours.

Patient demographics:	
Patient number:	26
Mean GA (weeks)	28.7 ± 2.3 (range: 24.6-32.0)
Male:Female	9:17 (35%:65%)
Mean birth weight (grams)	1267 ± 445 (range: 540-2540)
Mean respiratory support (days):	
Conventional mechanical ventilation	9 (range: 1-28)
High-frequency oscillation ventilation	3 (range: 8-13)
Non-invasive ventilation	25 (range: 1-63)
Patent ductus arteriosus:	
Drug treatment (Ibuprofen)	6 (23%)
Surgical ligation	2 (8%)
Sepsis (with antibiotics) >72 hours	16 (61%)
Necrotizing enterocolitis ≥Bell stage II	2 (8%)
cUS grades:	
Normal	16 (61%)
Grade I IVH	3 (12%)
Grade II IVH	2 (8%)
Flaring <14 days	3 (12%)
Retinopathy of prematurity (with laser therapy)	3 (12%)
Bronchopulmonary dysplasia	2 (8%)
Neurodevelopmental outcome (BSID-II):	
Mean MDI	109 ± 12
Mean PDI	112 ± 16
EEG recordings:	
Number of recordings	89
PMA range (weeks)	27-42
Mean EEG monitoring time (hh:mm)	06:57 (range: 01:31-19:58)
Mean EEG recording length (hh:mm)	05:09 (range: 01:30-16:59)
Visual sleep labelling?	Yes (QS and non-QS)

3.6 EEG pre-processing

Before further EEG analysis, all recordings were additionally bandpass filtered in the range 0.5-40Hz using a pair of standard 250-order Finite Impulse Response (FIR) high-pass and low-pass filters (see [155] for details).

Table 3.2: Summary of patient demographics and EEG information for dataset 2, separated into neurodevelopmental outcome groups (as defined in section 3.3). cUS categories are defined according to section 3.2.
Values are expressed with % proportions, mean ± SD, and/or range (where appropriate). Bronchopulmonary dysplasia: need for oxygen/ventilation at 36 weeks PMA. Sepsis: positive or highly suspected based on blood tests, with antibiotic treatment >72 hours.

	Group 1 (normal)	**Group 2 (mild)**	**Group 3 (severe)**	**Group 4 (death)**
Patient demographics:				
Patient number	21 (54%)	9 (23%)	6 (15%)	3 (8%)
Mean GA (weeks)	28.7 ± 2.2	29.2 ± 1.7	28.8 ± 2.9	29.0 ± 1.2
range:	24.6-31.9	26.3-31.3	25.0-31.7	24.9-27.3
Male (%)	71	100	50	67
Mean birth weight (grams)	1265 ± 371	1256 ± 354	1152 ± 529	796 ± 72
range:	570-1860	760-1700	485-1730	750-880
Mean respiratory support (days):				
Mechanical ventilation	5	4	10	20
range:	0-28	0-17	0-34	7-32
Non-invasive ventilation	15	23	31	45
range:	1-52	1-55	3-87	27-71
Patent ductus arteriosus	4 (19%)	2 (22%)	2 (33%)	2 (67%)
Sepsis (with antibiotics) >72 hours	7 (33%)	5 (56%)	2 (33%)	3 (100%)
Necrotizing enterocolitis ≥Bell stage II	3 (14%)	0 (0%)	0 (0%)	1 (33%)
cUS category:				
Normal	13	7	1	1
Mild	4	1	2	0
Severe	3	1	1	2
Other	1	0	2	0
Retinopathy of prematurity (with laser therapy)	3 (14%)	1 (11%)	2 (33%)	2 (67%)
Bronchopulmonary dysplasia	3 (14%)	2 (22%)	2 (50%)	2 (67%)
Neurodevelopmental outcome (BSID-II):				
Mean MDI	101 ± 10	84 ± 8	66 ± 12	N/A
Mean PDI	97 ± 10	94 ± 15	62 ± 11	N/A
EEG recordings:				
Number of recordings	62 (52%)	27 (23%)	20 (17%)	10 (8%)
PMA range (weeks)	31-48			
Mean EEG monitoring time (hh:mm)	06:50 (range: 01:19-24:36)			
Mean EEG recording length (hh:mm)	04:26 (range: 00:41-12:12)			
Visual sleep labelling?	No			

4
Essential Methodology & Technical Background

4.1 Introduction

Chapter 2 detailed the existing methods for automated sleep staging and quantifying brain maturation. Key limitations were identified which first motivated the goals of this research, as outlined in chapter 1. This chapter now presents the essential methodological background that has been identified to meet these goals.

To pinpoint novel sleep state and maturational biomarkers (that may also be outcome-sensitive), the set of features used in this research is first defined and facilitates a data-driven approach. These are collected from previously successful preterm, infant and adult EEG studies.

Identifying statistically significant biomarkers and developing more extensive brain-age prediction models requires regression-based methods. Consequently, a background to regression and some existing approaches are next presented.

Classification of sleep states to quantify sleep-wake cycling can be achieved through a variety of methods. Probabilistic classifiers are ideal candidates, as they provide the flexibility to incorporate multiple variables, model a variety of dependencies between said variables, as well as incorporating measures of uncertainty. An overview of probabilistic modelling is provided, with some traditional classifiers described.

This chapter concludes with a description of specific metrics for evaluating algorithm performance.

4.2 EEG feature extraction

The features defined have been collected from previously successful EEG studies on preterm EEG sleep staging and maturation (as discussed in chapter 2), neonatal seizure detection, as well as infant and adult sleep staging. A summary of these features also listed at the end of this section, with references to these applications.

Features were extracted from the native EEG, as well as from a variety of EEG decompositions, which separated the signal into specific frequency bands. Methods of signal decomposition are first described, followed by definitions of the features (separated by time-domain and frequency-domain).

Functional brain connectivity features (as described in section 2.7.1) were not included, due to the disadvantages of calculating such measures robustly over long duration EEG recordings (when channel-specific artefacts are present). This makes it challenging to reliably calculate such measures within a data-driven framework.

4.2.1 Signal decompositions

The most commonly used decomposition (or transform) is the Discrete Fourier Transform (DFT), which represents a signal in terms of its frequency components. It is defined in appendix A.2.1. The DFT has one disadvantage in that it assumes that the signal being transformed is stationary (i.e. statistical properties, such as the mean and variance, remain constant with time). However, most biomedical signals (including EEG) are inherently non-stationary. While the DFT is still predominantly used, decompositions that account for the non-stationarity of the signal may uncover additional information. To this end, features are also extracted from both the Discrete Wavelet Transform (DWT) and Empirical Mode Decomposition (EMD). These are defined in appendix A.2.2 and A.2.3, respectively.

In general, the DWT represents the signal as a series of Detail coefficients (Dl) and Approximation coefficients (Al) up to l levels of decomposition. The EMD represents the signal as a series of Intrinsic Mode Functions (IMFs), with an IMF defined if the following conditions hold:

Table 4.1: Frequencies extracted from/defined after performing the DFT, DWT, and EMD. Frequency bands (ranges) are shown for the DFT and DWT. EMD frequencies are approximated by calculating IMFs from a 30 second segment in QS and non-QS, across all the recordings comprising dataset 1. The mean of the median instantaneous frequencies is shown for each IMF.

DFT (range)		**DWT (range)**		**EMD (mean ± SD)**	
beta:	12-30Hz	D3:	15.6-31.3Hz	IMF1:	25 (±2.9)
alpha:	8-12Hz	D4:	7.8-15.6Hz	IMF2:	11 (±1.5)
theta:	3-8Hz	D5:	3.9-7.8Hz	IMF3:	5.6 (±1.0)
delta:	0.5-3Hz	A5:	0.5-3.9Hz	IMF4:	2.6 (±0.72)
				IMF5:	1.3 (±0.45)
				IMF6:	0.66 (±0.26)

1. The number of extrema and number of zero crossings differ by at most one.
2. At any point, the mean between the upper and lower envelopes are zero.

Comparing the decomposition frequency bands

In this research, the DWT is applied to five levels, with the coefficients D3-D5 and A5 retained for feature extraction [83, 156]. EMD is applied to six levels, resulting in the set IMF1-IMF6 [157, 158].

Table 4.1 summarizes the frequencies that are extracted from the DFT, and are represented within each decomposed band of the DWT and EMD. DWT frequency bands are inferred from the initial maximum signal frequency of 125Hz (the Nyquist rate for an EEG sampling frequency of 250Hz) [159].

IMF frequency content is less intuitive as the corresponding frequencies can vary from signal to signal. As an approximation (for illustrative purposes), IMFs were extracted from a clean 30 second segment within a QS and non-QS period (as defined in section 2.5), for each recording in dataset 1 (as defined in chapter 3). The median instantaneous frequency was calculated in each case (using equation (4.13) of section 4.2.2) and the mean frequency (± SD) for each of IMF1-IMF6 is given in table 4.1.

4.2.2 Time-domain features

Amplitude features are extracted directly from the EEG, including the absolute mean and SD of the amplitudes, differences between the maximum and minimum values, as well as similar measures from the first and second derivatives of the signal.

The first derivative, $\boldsymbol{x}' \in \mathbb{R}^{1\times N}$ (where N is the length), of a signal, $\boldsymbol{x} \in \mathbb{R}^{1\times N}$, is determined by the numerical approximation (for each i^{th} sample in the signal):

$$x'_i = \frac{x_{i+1} - x_{i-1}}{2\Delta t} \tag{4.1}$$

where Δt is the time division between successive signal samples. The second derivative, $\boldsymbol{x}''$, is calculated by further applying the above equation to $\boldsymbol{x}'$.

Additionally, the zero-crossing rate is determined by counting the proportion of zero-crossings in the signal over one second windows. The SD across windows is then taken to capture the variability of this feature over the full signal.

Coastline distance is defined as the sum of absolute differences between successive amplitudes in the signal [160]. Peaks are defined by identifying all the local maxima in the signal and, from the peaks, the 90^{th} percentile (90%) of peak amplitudes is determined as well as the 90% peak prominences and 90% peak widths. Prominence measures how much the peak stands out, relative to other peaks near that location [161]. Widths are calculated at the point of half the peak prominence [161].

Similar features are also determined for each of the delta, theta, alpha and beta bands, isolating these bands by filtering the EEG using 250-order FIR filters at the appropriate cut-off frequencies. In addition, amplitude-based measures are extracted from the wavelet coefficients D3-D5 and A5.

Skewness & kurtosis

Skewness measures the degree of asymmetry of the data distribution, when compared to a symmetrical normal distribution. Positive values indicate that the mass of the distribution is predominantly skewed to the left, while the inverse is true for negative values. It is defined for the signal, $\boldsymbol{x} \in \mathbb{R}^{1\times N}$, by [162]:

$$\text{skewness} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^3}{\left[\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2\right]^{3/2}} \tag{4.2}$$

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

Similarly, the kurtosis measures the level of 'tailedness' of the distribution, compared to the normal distribution. Values >3 indicate tails that exponentially

decay at a slower rate to the normal distribution (more outliers present) and the opposite holds true for values <3. Kurtosis is defined as [162]:

$$\text{kurtosis} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^4}{\left[\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2\right]^2} \tag{4.3}$$

with variables defined as for equation (4.2).

Hjorth parameters

Hjorth parameters were introduced by *B. Hjorth* in 1970 [163] and consist of the activity, mobility, and complexity. Hjorth activity is a measure of the signal power and is equivalent to the variance of the signal amplitudes.

Hjorth mobility is a representative measure of the signal's mean frequency and defined as [163]:

$$\text{mobility} = \sqrt{\frac{\text{var}(\boldsymbol{x}')}{\text{var}(\boldsymbol{x})}} \tag{4.4}$$

with $\boldsymbol{x} \in \mathbb{R}^{1xN}$ the signal and $\boldsymbol{x}'$ the first derivative. var$(\cdot)$ denotes the variance.

Hjorth complexity represents the change in frequency of the signal, and is derived from the Hjorth mobility [163]:

$$\text{complexity} = \frac{\text{mobility}(\boldsymbol{x}')}{\text{mobility}(\boldsymbol{x})} \tag{4.5}$$

where variables are defined as in equation (4.4), and mobility$(\cdot)$ denotes the Hjorth mobility.

Signal entropy measures

As first discussed in section 2.5.2, the entropy provides a measure of signal complexity. More specifically, it measures this by quantifying the regularity (or self-similarity) of the signal.

Approximate Entropy (ApEn) is a measure of signal entropy, where the signal is divided into consecutive segments, of length m, known as templates. Each i^{th} template is compared to all others and the total number of similar templates are determined, $C_i^{(m)}$. A pair of templates are defined as similar by first calculating the

Chebyshev distance (the maximum of the differences between each pair of respective template values) and comparing it to a threshold, r. A similar template is defined as having a Chebyshev distance below this threshold value [164].

The sum-of-logs of $C_i^{(m)}$ (for all i), $\Phi^{(m)}$, is then calculated and ApEn is defined as the difference between these values when calculated for both m and $m+1$ [164]:

$$\text{ApEn} = \Phi^{(m)} - \Phi^{(m+1)} \tag{4.6}$$

A small value of ApEn (tending to zero) denotes similar signal behaviour after m is increased by one, and suggests a greater regularity. The full mathematical definition of ApEn is provided by *Pincus et al.* in [164].

Sample Entropy (SampEn) was derived as a modification of ApEn and can be interpreted in the same way. As with ApEn, the total number of similar templates is first calculated. However, when comparing templates in ApEn, self-matches are included, while these are ignored for SampEn (reducing bias). A full mathematical definition is given by *Richman et al.* in [165].

Finally, *Costa et al.* proposed an extension of SampEn, called Multiscale Entropy (MSE). MSE is simply the application of SampEn to the signal at increasingly coarser scales. For each s^{th} scale, the original signal samples are grouped into non-overlapping windows, of length s, and the windowed samples are averaged. SampEn is then applied to this averaged signal [166].

As well as defining the first ten MSE scales as features, the study by *De Wel et al.* identified additional MSE-based entropy features as promising biomarkers for automated sleep-staging and maturation [29] (as previously discussed in sections 2.5.2 and 2.7.2). These are based on a multiscale curve of the MSE values versus scale (for scales one to 20). From this, the area under the multiscale curve, average slope of the curve (over scales one to five), average slope of the curve (over scales six to 20), and the maximum MSE value form additional features [29].

For ApEn, SampEn and MSE, the same parameters of m=2, and $r = 0.2 \cdot SD$ (where SD is the standard deviation of the signal amplitudes) were chosen in accordance with other preterm EEG studies [29, 117, 167].

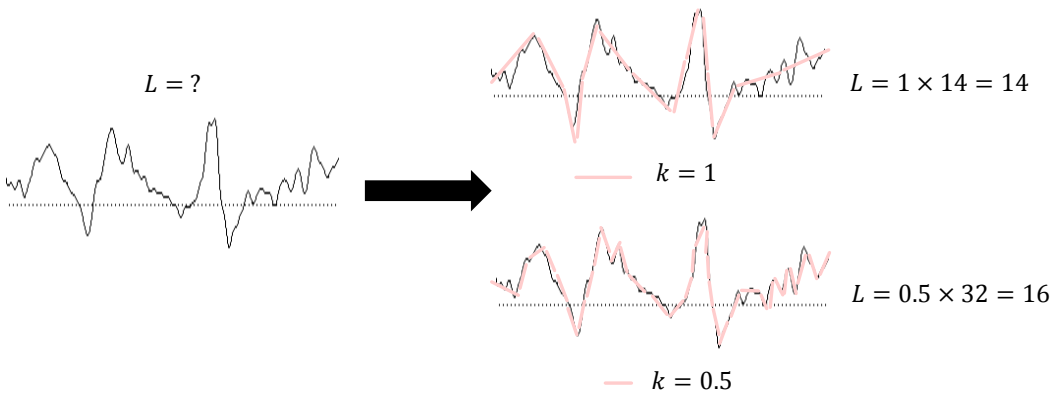


Figure 4.1: Illustration of how estimates of total signal length (L) change with the choice of length scale (k). In each case, L is defined as the product of k and the total number of length scales required to cover the signal. This relationship is used to define the fractal dimension.

Fractal dimension

The fractal dimension provides an alternative measure of signal complexity, specifically quantifying the space-filling capacity of the signal (as also discussed in section 2.5.2). The value ranges between one (a straight line) to two (filling an entire 2D plane) with values nearer two suggestive of higher signal fluctuations, amplitudes, and discontinuity. The two most common measures of fractal dimension were derived by *M. Katz* [168] and *T. Higuchi* [169]. Both measures derive the fractal dimension by estimating how the overall signal length changes with the specific length scale used (as illustrated in figure 4.1).

The Higuchi method achieves this graphically by plotting the estimated signal lengths against a range of scales up to a value of *kmax* [169]. Fractal dimension is then defined as the proportionality constant representing this relationship and was estimated for multiple *kmax* values of 20, 40 and 60.

Line Length (LL)

LL was previously shown by *Koolen et al* as a promising measure for assessing the sleep-wake cycling of preterm EEG, through measures of suppression and bursts, *supp* and Burst%, respectively (as described in section 2.7.3).

LL is inspired from the fractal dimension and is defined as the sum of absolute distances between each i^{th} sample, x_i, in a one second segment of signal [26]:

$$LL = \sum_{i=1}^{Fs-1} |x_{i+1} - x_i| \tag{4.7}$$

where Fs denotes the sampling frequency. This is calculated for each consecutive segment in the full signal, and again over multiple channels. After normalizing the values (per channel), median values across the channels result in the signal $\boldsymbol{LL^{(med)}} \in \mathbb{R}^{1\times M}$ (of length M). *supp* is defined as the ratio [26]:

$$supp = 1 - \frac{\text{median}(\boldsymbol{LL^{(med)}})}{\text{mean}(\boldsymbol{LL^{(med)}})} \tag{4.8}$$

Higher values of *supp* denote increasing levels of signal suppression.

Similarly, Burst% is calculated as the percentage of bursts in the signal. Bursts are classified using an adaptive threshold on $\boldsymbol{LL^{(med)}}$ [26]:

$$\text{Thr} = F \cdot \text{mean}(\boldsymbol{LL^{(med)}}) \tag{4.9}$$

A value of F=0.85 is experimentally determined in [26]. This same value is used here as it was chosen after training on a subset of the same dataset 1 used for this research.

The Teager-Kaiser Energy Operator (TKEO)

TKEO provides a measure of the signal energy over time. Although it is calculated directly in the time-domain, it incorporates both amplitude- and frequency-specific information by measuring local differences between adjacent signal samples [89]. It was first defined by *Kaiser et al.* for each i^{th} sample in the signal, $\boldsymbol{x} \in \mathbb{R}^{1\times N}$ (of length N) [170]:

$$TKEO_i = x_i^2 - x_{i-1}x_{i+1} \tag{4.10}$$

The mean and SD of the resulting TKEO signal are calculated as summary features to quantify the overall energy.

range-EEG (rEEG)

The rEEG can be thought of as an approximation of the aEEG and was also described in section 2.7.3 as a means to estimate IBIs and quantify sleep-wake cycling. It was developed by *Navakatikyan et al.* [28] and approximates the upper values of the EEG peak-to-peak amplitude. The difference between the maximum and minimum values in adjacent two second EEG segments are linearly connected and up-sampled to 8 Hz frequency. This produces a continuous signal that is sensitive to bursts and suppression [28]. Metrics of mean, median, SD, IQR, 5^{th} percentile (Lower Margin, LM) and 95^{th} percentile (Higher Margin, HM) are calculated as features to summarize the rEEG signal [34].

Variation coefficient & fluctuation index

The variation coefficient and fluctuation index are calculated for each of the IMFs derived from the EMD. Variation coefficient quantifies the level of amplitude variability for each IMF as [157]:

$$\text{variation coefficient} = \frac{\sigma^2}{\mu^2} \tag{4.11}$$

where σ and μ denote the SD and absolute mean of the IMF signal amplitudes, respectively.

Fluctuation index quantifies the intensity of the IMF signal changes. Denoting an IMF as $\boldsymbol{s} \in \mathbb{R}^{1\times N}$ (where N is the length) [157]:

$$\text{fluctuation index} = \frac{1}{N-1}\sum_{i=1}^{N-1}|s_{i+1} - s_i| \tag{4.12}$$

Hilbert-derived features

After performing the Hilbert-Huang transform (defined in appendix A.2.4), and representing each transformed IMF signal in the form of equation (A.12), the instantaneous amplitude signal is given by $\boldsymbol{a}$ and is equivalent to the signal envelope [171].

The instantaneous frequency signal represents the local IMF signal frequency over time. It is defined as the derivative of the (unwrapped) instantaneous phase,

$\mathbf{\Phi}$ (as also defined from equation (A.12)), and can be approximated for each timepoint by [171]:

$$\text{instantaneous frequency} = \frac{Fs}{2\pi}|\Phi_{i+1} - \Phi_i| \tag{4.13}$$

where Fs denotes the sampling frequency.

Median values of the instantaneous amplitude and instantaneous frequency signals are calculated as features from each IMF. These are also derived from the delta and theta bands, for comparison.

4.2.3 Frequency-domain features

The frequency-domain features are derived from the one-sided Welch periodogram (as defined in appendix A.2.1 and equation (A.5)), which represents the power spectral density of the signal as a function of frequency.

Relative bandpowers & average frequency

Relative bandpowers for the delta, theta, alpha, and beta bands are derived by integrating the periodogram across these frequency bands, using the rectangle method [172], and dividing by the total bandpower over the full delta-beta range.

The average frequency is calculated by first approximating the periodogram as a probability distribution. This is achieved by multiplying it by the frequency interval (to obtain the power within each bin), and then normalizing the result such that it sums to one. Average frequency is then defined as [161][1]:

$$\text{average frequency} = \sum_{f=0}^{Fs/2} f \cdot \tilde{p}_k \text{ and } f = \frac{k \cdot Fs}{2(N-1)} \tag{4.14}$$

where f is the frequency, $\tilde{p}_k$ denotes the normalized periodogram value at k (and f), N is the periodogram signal length, Fs is the sampling frequency, and $k \in \mathbb{N} : 0 \leq k \leq N-1$. This feature is calculated for the full signal, as well as for the delta, theta, alpha, and beta bands.

[1]For consistency with definitions in literature, the first element of the normalized periodogram is at k=0 here.

Power spectral entropy & 90% spectral edge frequency

Power spectral entropy is derived from the normalized periodogram, as for the average frequency in equation (4.14). It quantifies the regularity (or flatness) of the frequency spectrum [173][2]:

$$\text{spectral entropy} = -\frac{1}{\log(N)} \sum_{k=0}^{N-1} \tilde{p}_k \cdot \log(\tilde{p}_k) \tag{4.15}$$

with parameters defined as in equation (4.14).

The 90% spectral edge frequency is defined as the frequency below which 90% of the signal power lies. It is approximated by integrating the periodogram (using the rectangle method [172]) up to the frequency at which the proportion of the total power reaches 0.9 [86].

Audio-inspired features

Additional frequency-domain features are included that have gained popularity in EEG analysis, but are originally inspired from audio signal processing.

The spectral roll-off is equivalent to the spectral edge frequency evaluated at 85%. The spectral centroid is equivalent to the average frequency, as defined in equation (4.14). Finally, the spectral flux is a measure of how quickly the signal spectra changes in consecutive windows, by calculating the euclidean distance between the two normalized periodograms for each window [121].

As with the zero crossing rate, features were calculated over one second windows and the SD across windows taken to capture the variability over the full signal length.

4.2.4 Summary of extracted features

A total of 113 different features are calculated. Table 4.2 provides the complete list of features, including references to existing EEG applications (where possible). Procedures for feature extraction and pre-processing varied, depending on the application, and specifics are discussed in the relevant chapters.

[2]For consistency with definitions in literature, the first element of the normalized periodogram is at k=0 here.

Table 4.2: The full list of the 113 extracted EEG features, separated by time- and frequency-domain. References to previous EEG applications are also provided.

Category	Feature	References
TIME DOMAIN	SD and mean absolute amplitudes,	[87, 101, 156, 174, 175]
	Max. absolute amplitude and sum of first and second derivatives,	[101]
	Max.-min. difference of amplitudes,	[101]
	Skewness and Kurtosis,	[86, 92, 156, 160, 174, 175]
	Hjorth activity, complexity, mobility,	[156, 158, 160, 174, 176]
	ApEn, SampEn, and MSE (scales 1-10),	[156, 158, 176, 177]
	SD of MSE values,	
	Area under multiscale curve,	[29]
	Average slope of multiscale curve (scales 1-5),	[29]
	Average slope of multiscale curve (scales 6-20),	[29]
	Max MSE value,	[29]
	Katz and Higuchi Fractal Dimensions (kmax=20,40,60),	[86, 92, 118, 158]
	Zero crossing rate SD,	[121, 156, 158, 174, 176]
	Coastline distance,	[160]
	Mean and SD TKEO values,	[156, 176]
	90% peak value, width, prominence,	[160]
	LL suppression value (supp),	[26, 27, 34]
	LL Burst%,	[26, 34]
	Mean, median, LM (5th percentile), HM (95th percentile), SD, IQR, and skewness of the rEEG,	[28, 34]
	SD and mean absolute amplitudes in delta, theta, alpha, beta frequency bands,	[178]
	Hilbert median envelope amplitude in delta, theta bands,	
	Hilbert median instantaneous frequency in delta, theta bands,	
(*DWT*)	Mean absolute value and SD of amplitudes in D3, D4, D5, A5 bands,	[156, 158]
	Ratio of absolute mean values in adjacent bands,	
	Sum of squared coefficients in D3, D4, D5, A5 bands,	[158]
(*EMD*)	Variation coefficients for IMF1-IMF6,	[158, 177]
	Fluctuation index of IMF1-IMF6,	[158, 177]
	Mean absolute ratio of each pair of successive IMFs,	
	Hilbert median instantaneous frequency, IMF1-IMF6	[171, 179, 180]
FREQUENCY DOMAIN	Relative bandpower in delta, theta, alpha, beta bands,	[34, 86, 181, 182]
	Average frequency,	[101, 175]
	Average frequency in delta, theta, alpha, beta bands,	[175]
	Spectral roll-off SD, centroid SD, flux SD,	[121]
	Power spectral entropy,	[86, 92, 156, 158, 174, 176]
	90% spectral edge frequency,	[86, 87, 92, 121, 174, 176]
	EEG spectral beta/alpha ratio,	[23]

4.3 Regression methods

4.3.1 Linear regression

Identifying significant trends between continuous variables and developing predictive methods require fitting regression models. The simplest is linear regression, where one fits a model that assumes a linear relationship between a response variable, $\boldsymbol{y} \in \mathbb{R}^N$ (with N instances, or observations), and a series of multidimensional explanatory variables, $\boldsymbol{X} \in \mathbb{R}^{N \times P}$ (where P denotes the variable dimension). Many of the subsequent methods discussed stem from this simple model. For an i^{th} observation and denoting $x_{i,j}$ as an element in $\boldsymbol{X}$ [183]:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_P x_{iP} + \epsilon_i \tag{4.16}$$

where y_i is the i^{th} response in $\boldsymbol{y}$ and β values denote the regression coefficients (slopes) which explain the linear relationship between the explanatory and response. Additionally, β_0 denotes the constant intercept term. These are parameters that are fitted during training. ϵ_i denote the residuals that capture all the additional information to explain remaining discrepancies between the linear model estimates and the response.

Stacking all N observations together, the multiple regression model can be represented in matrix form [183]:

$$\mathbf{y} = \tilde{\boldsymbol{X}}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{4.17}$$

$\boldsymbol{\epsilon} \in \mathbb{R}^N$ is the vector of residuals. $\boldsymbol{\beta} \in \mathbb{R}^{P+1}$ is the vector of coefficients (including β_0 as the first value) and $\tilde{\boldsymbol{X}} \in \mathbb{R}^{N \times (P+1)}$ accounts for the multiplication with β_0 in $\boldsymbol{\beta}$ by adding a column of ones to the beginning of $\boldsymbol{X}$. Values of $\boldsymbol{\epsilon}$ are assumed to be drawn from a normal distribution with constant variance, $\mathcal{N}(0, \sigma^2)$.

Dummy variables & interactions

In cases where the data is divided into discrete groups, dummy variables can be introduced to account for the differences between groups. These variables can be binary or can take on multiple integers and are simply included as additional

explanatory variables in the regression model. As they are discrete, the overall effect is a shift in the intercepts of the regression lines for each group (with the slopes unchanged) [183].

To allow slopes to vary by group, interaction effects can also be introduced. These are defined by taking the product of the dummy variables and the chosen explanatory variables, resulting in additional interaction variables with associated coefficients [183].

Training a linear regression model

In order to train the model and estimate values for $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}$, the most commonly used method is Ordinary Least Squares (OLS). OLS models aim to find $\hat{\boldsymbol{\beta}}$ such that the difference between the observed response variable and that predicted by the linear model, is minimized [183]. Mathematically[3]:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left[S(\boldsymbol{\beta}) \right] \tag{4.18}$$

where:

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{N} |y_i - \sum_{j=1}^{P+1} \tilde{x}_{i,j}\beta_j|^2 = \sum_{i=1}^{N} e_i^2 \tag{4.19}$$

where $\tilde{x}_{i,j}$ is an element of $\tilde{\boldsymbol{X}}$. At $\hat{\boldsymbol{\beta}}$, the term e_i becomes ϵ_i and $S(\hat{\boldsymbol{\beta}})$ is known as the residual sum of squares, SS_{res}.

From this, a common measure to assess the tightness of the model fit and strength of the linear correlation is the coefficient of determination, R^2. This is defined as the ratio of SS_{res} to the total sum of squares, SS_{tot}, and measures the proportion of the overall variance in the data that is predictable from the trained model [183]:

$$R^2 = 1 - \frac{\mathrm{SS}_{res}}{\mathrm{SS}_{tot}}, \quad \text{where } \mathrm{SS}_{tot} = \sum_{i=1}^{N} |y_i - \bar{y}|^2 \tag{4.20}$$

where $\bar{y}$ is the mean of all y_i values.

[3]Here, the intercept coefficient defined previously (β_0) is the first element of $\boldsymbol{\beta}$ and indexed at j=1.

Prediction Intervals (PI)

When performing a linear regression, a confidence interval is often defined, to measure the confidence in the estimates of the underlying coefficients (at a particular percentage level). This takes into account any error in the fitted model.

The Prediction Interval (PI) provides a similar bound to measure the confidence in predicting a new value, $\hat{y}$, from the trained regression model, given a new value for the explanatory variable, $\hat{\boldsymbol{x}} \in \mathbb{R}^{P+1}$ ($\hat{\boldsymbol{x}}$ takes the same form as a row in $\tilde{\boldsymbol{X}}$). The PI accounts for both the error in the fitted model and the error associated with future estimates. For $\hat{y}$, this is calculated as follows [184]:

$$\hat{y}^{(PI)} \pm t_{N-(P+1)} \sqrt{s^2 \left(1 + \hat{\boldsymbol{x}}^\intercal \left(\tilde{\boldsymbol{X}}^\intercal \tilde{\boldsymbol{X}}\right)^{-1} \hat{\boldsymbol{x}}\right)} \tag{4.21}$$

where:

$$s^2 = \frac{SS_{res}}{N - (P+1)} \tag{4.22}$$

$t_{N-(P+1)}$ denotes the two-tailed draw from a t-distribution for a given significance level, with $d = N - (P+1)$ degrees of freedom (as defined in appendix A.1.2). To determine the 95% PI, values of $t_{N-(P+1)}$ are drawn for a p-value (p) of 0.05 [184].

Statistical significance between regression models

Regression models can be constructed with increasing numbers of variables and it is useful to understand which variables provide statistically significant improvements. In linear regression, models are typically compared using an F-test [185].

To perform an F-test, two models are compared with and without the inclusion of the additional explanatory variable. Model 1 is defined with this variable's regression coefficient fixed to zero, while model 2 allows the coefficient to be trained. The resulting F-statistic (F) is calculated from the SS_{res} for each model [185]:

$$F = \frac{\left(\frac{SS_{res}^{(1)} - SS_{res}^{(2)}}{d1}\right)}{\left(\frac{SS_{res}^{(2)}}{d2}\right)} = \frac{\left(\frac{SS_{res}^{(1)} - SS_{res}^{(2)}}{p2-p1}\right)}{\left(\frac{SS_{res}^{(2)}}{N-p2}\right)} \tag{4.23}$$

Superscripts 1 and 2 denote SS_{res} for model 1 and model 2, respectively. The terms $d1$ and $d2$ denote the degrees of freedom. In the case of regression, this can be

written in terms of the total number of parameters (coefficients) in each model ($p1$ and $p2$, respectively) and the total number of data points (N), as also shown in equation (4.23). F is compared to the critical value from the F-distribution (as defined in appendix A.1.3), to assess significance at the desired p-value [185].

Pearson & Spearman's rank correlation coefficients

Closely related to linear regression is the Pearson correlation coefficient, r, developed by *K. Pearson* [186]. It is used to quantify the strength of correlation between two datasets of N values, $\boldsymbol{x} \in \mathbb{R}^N$, and $\boldsymbol{y} \in \mathbb{R}^N$ and is defined as [186]:

$$r = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}} \tag{4.24}$$

The values $\bar{x}$ and $\bar{y}$ denote the sample means for each dataset, respectively. r has a range of minus one (perfect negative correlation) to one (perfect positive correlation) and values near zero denote weak or no correlation. In the case of an OLS linear regression model, if x_i is the explanatory variable and y_i the response, then r is related to R^2 by $|r| = \sqrt{R^2}$.

As an extension of r, which assumes linear correlations, *C. Spearman* [187] developed the Spearman's rank correlation coefficient, ρ, to assess the strength of a monotonic relationship between two datasets. To calculate ρ, each data point in both sets are ranked in ascending order and assigned integer values. In cases of ties, these ranks are recalculated and set equal to the average of the ranks that these data points span. Pearson correlation is then calculated from these ranks, with the resulting values falling in the same range as r and interpreted in a similar fashion.

4.3.2 The random-intercepts model

The standard linear regression model described in section 4.3.1 makes the underlying assumption that each data point is independent of one another. Furthermore, ϵ_i, are Independent and Identically Distributed (IID), meaning that they are drawn from the same underlying distribution (independently of one another). In cases involving repeated measures (such as multiple tests taken from the same group/subject),

there will likely be a further inter-dependency of the observations on the particular groups and the above assumptions are violated [188].

In these situations, a mixed-effects regression model is more accurate. Such models introduce group-specific differences in the parameters to ensure that correlations between the explanatory variables and the responses are not confounded by inter-group variability [188].

The simplest mixed-effects model is the random-intercepts model, which allows the intercepts for each group of data to uniquely vary. The model takes a similar form to equation (4.16). Denoting each i^{th} observation as additionally belonging to an m^{th} group [188]:

$$y_{im} = \overbrace{\beta_0 + \beta_1 x_{i1m} + \cdots + \beta_P x_{iPm}}^{\text{fixed-effect}} + \overbrace{u_m + \epsilon_{im}}^{\text{random-effect}} \tag{4.25}$$

The fixed-effect is equivalent to the standard linear-regression model, and the random-effect is introduced through the error terms u_m and ϵ_{im}. u_m denotes the additional group-specific variability, while ϵ_{im} accounts for the remaining error. Both these errors are assumed to be drawn from Normal distributions, $\mathcal{N}(0, \sigma_u^2)$ and $\mathcal{N}(0, \sigma_\epsilon^2)$, respectively [188].

Training this model requires estimation of both the fixed-effect parameters as well as σ_u^2 and σ_ϵ^2. The latter forms a combined covariance matrix that is trained using Maximum Likelihood (ML) estimation (where a probability distribution of the data fit over the parameter space is derived and maximized for a specific parameter choice [189]). This and the subsequent estimation of the fixed-effect coefficients involve a series of well-established linear transforms that first 'whitens' the data, with the full procedure comprehensively described in [188].

Once the model is trained, estimation of key metrics (namely R^2) and the assessment of model significance is equivalent to fixed-effects linear regression.

4.3.3 Regression trees & the Random Forest (RF)

While the concept of RFs have been around since the late '80s, it was *L. Breiman & A. Cutler* who first trademarked and developed the approach (in 1999) that is

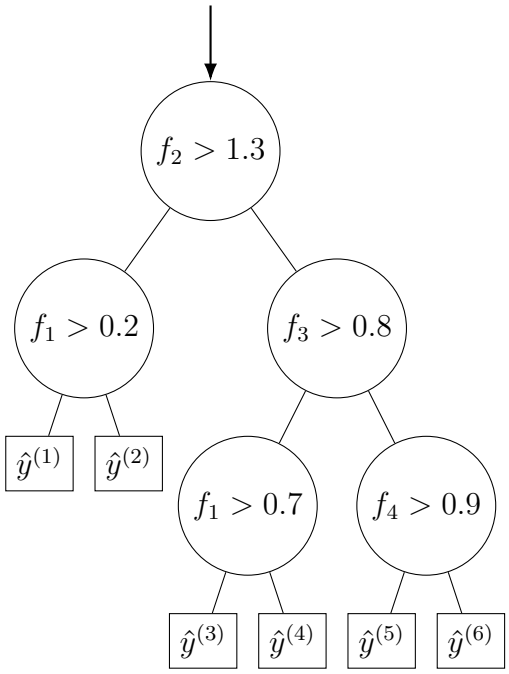


Figure 4.2: Illustration of a single regression tree which selects four different features: f_1, f_2, f_3, and f_4. Starting at the top (root node) and moving down, features are compared at a splitting criterion. Note that the same feature can be used multiple times (such as f_1 in this example). The tree terminates at leaf nodes where an estimate, $\hat{y}$, is then assigned.

most commonly used today [190, 191]. Unlike linear regression, RF regression is a non-linear method.

RFs consist of a large ensemble of regression trees. A single regression tree consists of a series of 'nodes' arranged top-down in a hierarchical structure. From a particular node (the 'parent' node), a binary split occurs producing two branches that lead to a pair of 'child' nodes which in turn become parent nodes that each have binary splits to further child nodes, and so on, until reaching a final set of terminating 'leaf' nodes. Every parent node in the tree represents a feature, f_j, and choosing which of its child nodes to traverse to is based on a threshold value (splitting criterion), s_j for f_j. The same f_j can be selected for multiple nodes in the tree (with different associated s_j values) [190, 191].

The feature space is divided into successively smaller regions until there is sufficiently few data points (over a small space) to form a single prediction, $\hat{y}^{(m)}$, for each m^{th} data partition. A simple tree is illustrated in figure 4.2 for six regions (and six predictions) [190, 191].

Training a regression tree & expanding to RFs

Predictions are derived from the means of the N training data points, y_i ($i \in \mathbb{N} : 1 \leq i \leq N$), within each partition, such that $\hat{y}^{(m)} = \bar{y}^{(m)}$. Training (or 'growing') such a regression tree involves systematically choosing the optimal f_j and s_j at each node that best divides the data. This involves minimizing the total Sum-of-Squares Error (SSE). Denoting each i^{th} instance (data point) of f_j as f_{ij}, SSE is minimized to split the data at each n^{th} node by [190]:

$$\min_{y}\left[\sum_{i:f_{ij}>s_j}(y-y_i)^2\right]+\min_{y}\left[\sum_{i:f_{ij}\leq s_j}(y-y_i)^2\right] \quad \forall(f_j,s_j) \tag{4.26}$$

One of the challenges of growing a regression tree is deciding when to terminate, and trees that are allowed to grow too large can lead to overfitting and poor performance. While methods such as pruning [190] were attempted, this limitation also motivated development of the RF regression model.

The RF attempts to combat overfitting by growing a large number of 'weaker' trees (typically >500) and generating a greater variability between these trees. This is achieved by bootstrap sampling (sampled with replacement) the data presented to each tree from the original dataset. Optimal features for each node are also chosen each time from a random subset. Consequently, an RF model requires prior selection of two parameters: the total number of trees in the forest (N_{trees}) and the size of the randomly selected feature subset at each node ($mtry$).

The final RF prediction is calculated as the mean of the predictions across all the trees.

Assessing RF feature importance

As the RF intrinsically selects features during training, the overall importance of each feature can be quantified directly from the trained model. A typical measure is the mean impurity difference, $\mathrm{Imp}(f_j)$, which measures the drop in an impurity measure at a node due to splitting on f_j. This is calculated by summing across all

nodes containing f_j in the tree, and averaging across all trees in the forest [191, 192]. Larger values denote higher feature importance. In general [192]:

$$\text{Imp}(f_j) = \frac{1}{N_{trees}} \sum_{t=1}^{N_{trees}} \sum_{n} p_n^{(t)} \Delta i_t(n) \quad \forall n \text{ where } f_j \text{ is used} \tag{4.27}$$

If $N_n^{(t)}$ denotes the total number of data points at the n^{th} node of the t^{th} tree, $p_n^{(t)} = N_n^{(t)}/N$ is the proportion of the total number of data points (N) at node n and $\Delta i_t(n)$ is the difference in the impurity measure. $\Delta i_t(n)$ is defined as [192]:

$$\Delta i_t(n) = i_t(n) - p_{n_L}^{(t)} i_t(n_L) - p_{n_R}^{(t)} i_t(n_R) \tag{4.28}$$

$i_t(n)$ denotes the impurity measure at parent node n. $p_{n_L}^{(t)} = N_{n_L}^{(t)}/N_n^{(t)}$ and $p_{n_R}^{(t)} = N_{n_R}^{(t)}/N_n^{(t)}$ denote the corresponding proportions of data points in the left (L) and right (R) child nodes, respectively.

For RF regression, the choice of $i_t(n)$ is typically the SSE, defined as [191, 192]:

$$i_t(n) = \sum_{i=1}^{N_n^{(t)}} \left(\hat{y}_i^{(n,t)} - y_i\right)^2 \tag{4.29}$$

$\hat{y}_i^{(n,t)}$ denotes the estimates of the t^{th} tree at node n. In this case, $\text{Imp}(f_j)$ is also referred to as the mean decrease in SSE [191].

4.4 Feature selection methods

Methods such as the RF in section 4.3.3 intrinsically deal with the problem of selecting features. However, many models require a separate algorithm for the pre-selection of features, in order to perform optimally. This is particularly true when there are large numbers of features (resulting in a dataset where the number of features $\gg$ number of data points) and/or when many features are strongly correlated with one another.

4.4.1 Least Absolute Shrinkage & Selection Operator (LASSO)

When performing linear regression (as described in section 4.3.1), such models tend to easily overfit when too many explanatory variables (features) are incorporated. The

method of LASSO is an extension of linear regression, which applies a regularization parameter, λ, that selects only a subset of the coefficients (and therefore explanatory variables) in the final model [193]. Performing LASSO involves modifying the linear regression optimization term of equation (4.19) [193]:

$$S_\lambda(\boldsymbol{\beta}) = \sum_{i=1}^{N} |y_i - \sum_{j=1}^{P+1} x_{i,j}\beta_j|^2 + \lambda \sum_{j=2}^{P+1} |\beta_j| \tag{4.30}$$

Variables are defined as in section 4.3.1 and $S_\lambda(\boldsymbol{\beta})$ is minimized through OLS. As λ is increased, the number of non-zero coefficients decreases.

4.4.2 Minimum Redundancy Maximum Relevance (mRMR)

mRMR, as with LASSO in section 4.4.1, aims to pre-select an optimal subset of features, $\mathcal{S}$, from the full set, $\mathcal{F}$. Unlike LASSO, which does not deal as well with highly correlated features, mRMR specifically aims to find a trade-off between features that are maximally relevant (in relation to a known target variable), while minimizing redundancy (selection of multiple relevant features that provide similar information). The general approach for mRMR is two-fold:

1. Maximum relevance: Begin with the empty set, $\mathcal{S} \in \varnothing$. A feature vector of N observations is denoted $\boldsymbol{f}_j \in \mathbb{R}^N$, with $j \in \mathcal{F}$. For all j, a specific measure to quantify the relevance between $\boldsymbol{f}_j$ and the target values, $\boldsymbol{y} \in \mathbb{R}^N$, is calculated and defined as $I(\boldsymbol{y}, \boldsymbol{f}_j)$. The index, j, contributing to the maximum value of $I(\boldsymbol{y}, \boldsymbol{f}_j)$, is the first element added to $\mathcal{S}$ (i.e. the most relevant feature) [194].

2. Minimum redundancy: For each subsequent addition to $\mathcal{S}$, the new candidate feature is compared to the current selected set. A measure between each feature pair is calculated, $I(\boldsymbol{f}_j, \boldsymbol{f}_s)$, where $s \in \mathcal{S}$. This quantifies the redundancy between the two features, and the next best feature added to $\mathcal{S}$ achieves the following trade-off [194]:

$$\underset{j \in \mathcal{F}-\mathcal{S}}{\arg\max} \left(I(\boldsymbol{y}, \boldsymbol{f}_j) - \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} I(\boldsymbol{f}_j, \boldsymbol{f}_s) \right) \tag{4.31}$$

$|\mathcal{S}|$ denotes the current size (or cardinality) of $\mathcal{S}$ and this step is repeated multiple times until the desired number of features are selected.

This method was first introduced by *Peng et al.* [194], which uses the Shannon entropy and mutual information measures for $I(\boldsymbol{y}, \boldsymbol{f}_j)$ and $I(\boldsymbol{f}_j, \boldsymbol{f}_s)$, respectively [194]. However, when assessing continuous features and target variables, this would require a discretization of the data. To avoid this, *Tsanas et al.* instead proposed an extension of mRMR using the Spearman's rank correlation coefficient (defined in section 4.3.1) as the measure for both $I(\boldsymbol{y}, \boldsymbol{f}_j)$ and $I(\boldsymbol{f}_j, \boldsymbol{f}_s)$ [195]. This approach is dubbed mRMR-Spearman.

4.5 Probabilistic models & classification

A probabilistic model represents the various states, events etc. as a series of random variables that comprise a larger inter-connected system. Each random variable is represented by specific probability distributions. Each possible value of a variable can be drawn from its distribution with a given probability and this quantifies the uncertainty (or plausibility) of that value. In the case of sleep states, for example, the random variable would be the set of all possible states and these would be modelled by a discrete probability distribution. Random variables that are continuous in nature would similarly be modelled by continuous probability distributions.

As mentioned in the beginning of this chapter, probabilistic models allow multiple random variables to be represented together, including the various relationships between them. In the case of clinical applications, modelling their uncertainties (as probabilities) is beneficial as it can account for the subjectivity/variability associated with clinical labelling (as in the case of sleep staging) and inter-patient variability.

Once the complete probabilistic model is defined, the value of a particular random variable in relation to any combination of the others, can be inferred. This is known as inference and the result is also a distribution representing the uncertainty in the solution.

In complex systems with many inter-connected random variables, such probabilistic models can be challenging to conceptualize. Consequently, the majority can be represented in a clearer graphical form.

4.5.1 Graphical form & Bayesian networks

The graphical form for representing probabilistic models provides a concise and intuitive means to understand the relationships (or dependencies) between the various random variables. As all these random variables are represented by probability distributions, the additional parameters that define these distributions can also be represented within this form. The use of graphical models is widespread but leads to variations in notation prompting a brief definition of the chosen notation for this thesis.

Figure 4.3(a) provides a summary of the chosen notation. Each random variable is represented as a node in the graph with clear and shaded nodes representing hidden (latent) and observed (directly measured) variables, respectively. Values not represented as nodes denote fixed parameters. Arrows (or edges) between a parameter and a node denotes that the associated random variable distribution is parametrized by that parameter. Edges between nodes denote a conditional dependency, meaning that the particular random variable is dependent (or conditioned) on the other.

Figure 4.3(b) presents a useful means to represent repeated instances of random variables in a model. The approach, known as plate notation, places these repeated structures (sub-graphs) within a plate and a number is included to denote how many instances occur. In the example of figure 4.3(b), both random variables a_i and b_i are repeated N times (with i denoting the particular instance). The associated parameters for each variable are also shown and, mathematically, one can define this model from its graphical form as[4]:

$$b_i|a_i \sim f(c,d) \tag{4.32}$$

$$a_i \sim \pi(e) \tag{4.33}$$

[4]Another common form in literature is to include the fixed parameters in the conditional random variable notation. In this example, the alternative would be $b_i|a_i, c, d \sim f(c,d)$ and $a_i|e \sim \pi(e)$. For this thesis, only the random variables are generally included in this notation for clarity (unless it is pertinent in certain definitions to include the parameters as well).

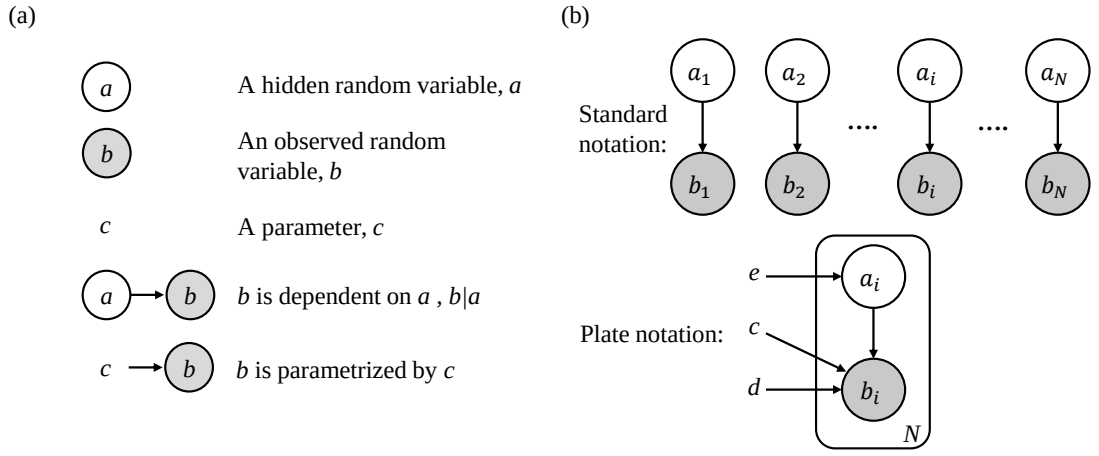


Figure 4.3: (a) The chosen notation for representing probabilistic graphical models. (b) An example conversion to plate notation, concisely representing N instances of the variables a_i and b_i (with i denoting the particular instance). Parameters c, d, and e of the random variable probability distributions are also shown.

where $b_i|a_i$ denotes that b_i is dependent on a_i. The symbol $\sim$ denotes that $b_i|a_i$ and a_i have the distributions $f(c, d)$ and $\pi(e)$, respectively. The density functions of these distributions form the conditional probability distributions for these variables, denoted by $p(b_i|a_i)$ and $p(a_i)$, respectively. As $p(a_i)$ does not have any conditional dependencies, it is also referred to as a prior probability.

Bayesian networks

In certain cases, the graphical form of the probabilistic model can be represented as a Directed Acyclic Graph (DAG). In DAGs, each edge is directed from one node to another and, specifically, there is no way to start at a particular node and return to that node by following any sequence of edges through the graph. Graphical models that are DAG are also known as Bayesian networks and all the random variables in the network are represented by their conditional probability distributions [189, 196].

Importantly, this leads to a simple means by which to directly represent the joint probability of all random variables, as the product of the known conditional probabilities in the network. This is called joint factorization or the 'chain rule'. For the two-variable example in figure 4.3(b):

$$p(a_i, b_i) = p(b_i|a_i)p(a_i) \tag{4.34}$$

In the more general case of a series of R random variables, $a_1, a_2, \cdots, a_r, \cdots, a_R$ [189, 196]:

$$p(a_1, a_2, \cdots, a_r, \cdots, a_R) = \prod_{r=1}^{R} p(a_r|\text{pa}_r) \tag{4.35}$$

where pa_r are the parent nodes of a_r (i.e. the random variables that a_r is dependent on). Deriving the joint probability greatly simplifies inference calculations, made even simpler by this factorized representation.

Performing inference in Bayesian networks

Performing inference in a Bayesian network involves determining the probability distribution across a particular variable (or variables) in the model, from the existing knowledge of the conditional dependencies in the network. As well as the chain rule, there are two additional rules that are utilized.

The most important is Bayes' rule, which allows any conditional distribution to be represented as a joint distribution (or vice-versa), through the following relationship [189]:

$$p(a|b) = \frac{p(b,a)}{p(b)} = \frac{p(b|a)p(a)}{p(b)} \tag{4.36}$$

This is shown for the general random variables a and b. In the wider scope of Bayesian theory, $p(b|a)$ is typically referred to as the likelihood, $p(a)$ the prior, and $p(b)$ the normalization term. The result, $p(a|b)$, is the posterior.

The second key rule is that of marginalisation, which allows particular variables to be eliminated, or 'marginalized out' [189, 196]:

$$p(b) = \sum_a p(b,a) = \sum_a p(b|a)p(a) \tag{4.37}$$

Here, a has been marginalized out, noting that this is another way of representing the denominator in equation (4.36). With these key rules, it is possible to perform inference on a number of models that can be represented as Bayesian networks.

Common Bayesian networks

Many traditional probabilistic models can be represented as Bayesian networks to facilitate inference. These include both regression and classification models such as naive Bayes, mixture models, Kalman filters, and Hidden Markov Models (HMMs) [189]. As Kalman filters and HMMs also incorporate transitions (temporal dependencies), these are sometimes referred to as dynamic Bayesian networks [196].

In this research, specific mixture models known as the Gaussian Mixture Model (GMM) and Dirichlet Process Mixture Model (DPMM), as well as the HMM, are of particular interest. These are described in more detail below.

4.5.2 The Gaussian Mixture Model (GMM)

The GMM is used to model a dataset using a mixture of Gaussian distributions. It has use as a method for probabilistic clustering (assigning a Gaussian to each clustered group of data in the space) and to generalize non-Gaussian (and often multi-modal) distributions [189]. Many other complex distributions could be used for the latter, such as kernel density estimation, log-normal distributions, exponential distributions etc. However, GMMs use Gaussian parameters (mean and variance) which simplify inference and typically capture clustered densities well. Training these parameters is also well established and tractable (forming a closed-form solution) [189].

The graphical model for a typical GMM is presented in figure 4.4. Each value of $\boldsymbol{x}_i \in \mathbb{R}^P$ (where P is the vector dimension) can be drawn from a set of K Gaussian distributions with parameters $\boldsymbol{\mu}_k \in \mathbb{R}^P$ and $\boldsymbol{\Sigma}_k \in \mathbb{R}^{P \times P}$ (where $k \in \mathbb{N} : 1 \leq k \leq K$).

For N instances of the random variable, $\boldsymbol{x}_i \in \mathbb{R}^P$, each i^{th} instance of $\boldsymbol{x}_i$ is conditioned on a component selection variable, $c_i \in \mathbb{N} : 1 \leq c_i \leq K$, which selects the Gaussian distribution with parameters $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ where $k{=}c_i$. c_i is itself drawn from a categorical distribution of length K, parametrized by the vector

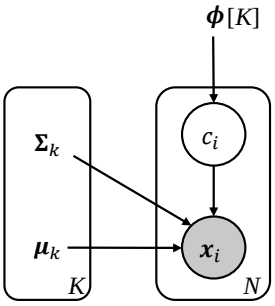


Figure 4.4: Plate notation for the GMM, with N instances of $\boldsymbol{x}_i \in \mathbb{R}^P$ (where P is the vector dimension). $\boldsymbol{x}_i$ is conditioned on a component assignment, $c_i \in \mathbb{N} : 1 \leq c_i \leq K$, and sampled from the corresponding c_i^{th} Gaussian distribution of which there are K in total. c_i is drawn from a prior distribution parametrized by $\boldsymbol{\phi}$ with the notation $[K]$ used to indicate that this is a vector of length K. $\boldsymbol{\mu}_k \in \mathbb{R}^P$ and $\boldsymbol{\Sigma}_k \in \mathbb{R}^{P \times P}$ denote each of the K Gaussian means and SDs.

$\boldsymbol{\phi} = [\phi_1, \phi_2, \cdots, \phi_k, \cdots, \phi_K]$. Mathematically, this model is defined as:

$$\boldsymbol{x}_i | c_i \sim \mathcal{N}(\boldsymbol{\mu}_{c_i}, \boldsymbol{\Sigma}_{c_i}) \tag{4.38}$$

$$c_i \sim \text{Cat}(\boldsymbol{\phi}) \tag{4.39}$$

In the case of using GMMs to fit non-Gaussian distributions, inference is typically performed to find $p(\boldsymbol{x}_i)$ by [189]:

$$p(\boldsymbol{x}_i) = \sum_{c_i=1}^{K} p(\boldsymbol{x}_i, c_i) = \sum_{c_i=1}^{K} p(\boldsymbol{x}_i | c_i) p(c_i) \tag{4.40}$$

$$= \sum_{c_i=1}^{K} \phi_{c_i} \cdot \mathcal{N}(\boldsymbol{x}_i; \boldsymbol{\mu}_{c_i}, \boldsymbol{\Sigma}_{c_i}) \tag{4.41}$$

$\mathcal{N}(\boldsymbol{x}_i; \boldsymbol{\mu}_{c_i}, \boldsymbol{\Sigma}_{c_i})$ denotes the value of $\mathcal{N}(\boldsymbol{\mu}_{c_i}, \boldsymbol{\Sigma}_{c_i})$ evaluated at $\boldsymbol{x}_i$. ϕ_{c_i} is the c_i^{th} value in $\boldsymbol{\phi}$ and also known as the component mixture weights. Equation 4.40 makes use of both marginalization and the chain rule, as defined in section 4.5.1.

Fitting a GMM model (an unsupervised procedure) involves identifying the parameters $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$, and $\boldsymbol{\phi}$. The most popular approach is the iterative Expectation Maximization (EM) algorithm.

Fitting a GMM using Expectation Maximization (EM)

The EM algorithm is applied to train a number of unsupervised probabilistic models (by estimating the model parameters). EM tackles the issue of finding parameters

by ML, where a likelihood function is difficult to derive and maximise in closed-form (due to the presence of latent variables).

In general, given a specific sequence of N observations, $[\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_i, \cdots, \boldsymbol{x}_N \in \mathbb{R}^P]$ (where P denotes the dimension) and $[z_1, z_2, \cdots, z_i, \cdots, z_N \in \mathbb{N} : 1 \leq z_i \leq K]$, which assigns each data point to one of K clusters (or components), EM is performed in two steps:

1. Expectation step (E-step): A posterior probability of each $\boldsymbol{x}_i$ belonging to each k^{th} component is calculated, $p(z_i = k|\boldsymbol{x}_i, \boldsymbol{\theta})$, given the current values of the model parameters, $\boldsymbol{\theta}$ (of arbitrary dimension). This assigns probabilistic (or soft) component assignments to each data point.

2. Maximisation step (M-step): Improved estimates of $\boldsymbol{\theta}$ are made from these soft assignments and the log-likelihood of the data fit to the current model is calculated, $\sum_{i=1}^{N} \log [p(\boldsymbol{x}_i|\boldsymbol{\theta})]$. Both steps are repeated until a maximum number of iterations are reached, or the log-likelihood improvement falls below a specified tolerance.

In the context of GMMs, z_i corresponds to c_i and $\boldsymbol{\theta}$ represents the parameters $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$, and ϕ_k. The EM algorithm for fitting a GMM is demonstrated in algorithm 4.1.

EM has a tendency to converge to local optima such that a choice of good parameter initializations are important. While it is common to initialize randomly, this method applies a version of the K-means algorithm (see appendix A.4) to better position the initial parameters. Furthermore, EM is often repeated multiple times with different initializations with the final model selected that generates the maximum log-likelihood.

4.5.3 The Hidden Markov Model (HMM)

An HMM models a system with transitions between states. However, unlike a standard Markov chain, these states are unobserved (or hidden) [197]. Instead, these hidden state transitions 'generate' the observable sequence. Such a model is

Algorithm 4.1 EM algorithm for fitting a GMM

input observations $[\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N \in \mathbb{R}^P]$ (where P denotes the dimension, and N the number of observations), number of components K.
initialize parameters $\boldsymbol{\mu}_k \in \mathbb{R}^P$, $\boldsymbol{\Sigma}_k \in \mathbb{R}^{P \times P}$ and ϕ_k, where $k \in \mathbb{N} : 1 \leq k \leq K$
repeat
  for $i = 1$ to N **do** ▷ E-step
    for $k = 1$ to K **do**

$$w_{ik} = p(c_i = k|\boldsymbol{x}_i, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \phi_k) = \frac{\phi_k \cdot \mathcal{N}(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{m=1}^{K} \phi_m \cdot \mathcal{N}(\boldsymbol{x}_i; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)}$$

      where $\mathcal{N}(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ denotes the value of $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ evaluated at $\boldsymbol{x}_i$
    end for
  end for
  for $k = 1$ to K **do** ▷ M-step

$$\phi_k^{(new)} = \frac{N_k}{N}, \text{ where } N_k = \sum_{i=1}^{N} w_{ik}$$

$$\boldsymbol{\mu}_k^{(new)} = \left(\frac{1}{N_k}\right) \sum_{i=1}^{N} w_{ik} \boldsymbol{x}_i$$

$$\boldsymbol{\Sigma}_k^{(new)} = \left(\frac{1}{N_k}\right) \sum_{i=1}^{N} w_{ik} (\boldsymbol{x}_i - \boldsymbol{\mu}_k^{(new)})^\intercal (\boldsymbol{x}_i - \boldsymbol{\mu}_k^{new})$$

    Set $\phi_k = \phi_k^{(new)}$, $\boldsymbol{\mu}_k = \boldsymbol{\mu}_k^{(new)}$, $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}_k^{(new)}$
  end for
  Calculate log-likelihood $\sum_{i=1}^{N} \log \left[\sum_{k=1}^{K} \phi_k \cdot \mathcal{N}(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right]$
until maximum iteration number or log-likelihood tolerance reached
output ϕ_k, $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$ $\forall k$

often used for biomedical time-series data as there is usually an underlying temporal dependency with associated states (as in the case of sleep staging) [197].

The typical HMM assumes a first-order Markov process, where the current state has a dependency on only the previous state. This is presented in figure 4.5 where the state at the i^{th} time point is denoted by z_i, and the observation by $\boldsymbol{x}_i \in \mathbb{R}^P$ (where P is the vector dimension).

From a state, $z_i \in \mathbb{N} : 1 \leq z_i \leq S$ (where S is the total number of states), each $\boldsymbol{x}_i$ is parametrized by state-dependent distributions with conditional (emission) probabilities $p(\boldsymbol{x}_i|z_i)$. The HMM is additionally parametrized by transition probabilities, $p(z_i|z_{i-1})$, which encode the probabilities of transitioning from each state z_{i-1} to z_i. This forms what is also commonly referred to as the transition matrix.

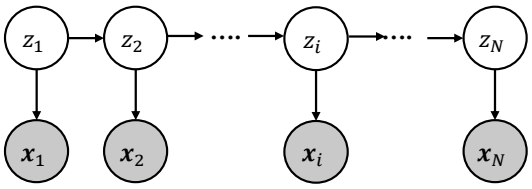


Figure 4.5: The typical graphical model for an HMM, representing the state transitions, z_i, and observations, $\boldsymbol{x}_i \in \mathbb{R}^P$ (where P denotes the dimension), for a sequence of length N.

In addition, there are the prior probabilities, $p(z_1)$, for the initial state z_1 [197].

These conditional probabilities are involved in three main HMM inference tasks [189]:

1. Model likelihood: This involves calculating the overall likelihood of an observation sequence originating from an HMM model, $p(\boldsymbol{x}_{1:N})$, where $\boldsymbol{x}_{1:N}$ denotes the sequence $[\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_i, \cdots, \boldsymbol{x}_N]$. This is most commonly performed using the recursive forward algorithm (described below).

2. Most likely state sequence: This involves identifying the most likely underlying state sequence $\hat{z}_{1:N}$, where $\hat{z}_{1:N}$ denotes the sequence of specific state assignments, $[\hat{z}_1, \hat{z}_2, \cdots, \hat{z}_i, \cdots, \hat{z}_N]$, used to generate $\boldsymbol{x}_{1:N}$. This requires maximizing $p(z_{1:N}, \boldsymbol{x}_{1:N})$ over all $z_{1:N}$ and is most commonly performed using the recursive Viterbi algorithm (described below).

3. Training: This involves estimating $\boldsymbol{\theta}$ (of arbitrary length), the parameters of the conditional probabilities in the model. In many cases, HMMs are unsupervised such that $\boldsymbol{\theta}$ is estimated in an iterative manner using a form of the EM algorithm called Baum-Welch. A detailed description of this algorithm is not included here, as training an unsupervised HMM is beyond the scope of this thesis.

The forward algorithm

Calculating $p(\boldsymbol{x}_{1:N})$ is challenging if one was to perform inference directly on the full sequence. This would require marginalizing over all possible state sequences, $z_{1:N}$,

which grows exponentially with length N. Consequently, the forward algorithm was developed to recursively solve the problem [189].

The forward algorithm defines the joint probability of observing the sequence $\boldsymbol{x}_{1:i}$, and ending in state z_i [189]:

$$\alpha_i(z_i) = p(\boldsymbol{x}_{1:i}, z_i) \tag{4.42}$$

Through marginalization and the chain rule, one can recursively compute $\alpha_i(z_i)$:

$$\alpha_i(z_i) = p(\boldsymbol{x}_i|z_i) \sum_{s=1}^{S} p(z_i|z_{i-1} = s)\alpha_{i-1}(s) \tag{4.43}$$

with:

$$p(\boldsymbol{x}_{1:N}) = \alpha_N \text{ where } \alpha_N = \sum_{s=1}^{S} \alpha_N(s) \tag{4.44}$$

The major issue with this implementation is the numerical underflow in $\alpha_i(z_i)$ that comes from repeatedly multiplying increasingly smaller quantities (all <1). To avoid this, the scaled conditional distribution is instead used, $\tilde{\alpha}_i(z_i) = p(z_i|\boldsymbol{x}_{1:i})$, and the final likelihood calculated in the log-domain. This log-scaled version is implemented as shown in algorithm 4.2.

The Viterbi algorithm

The Viterbi algorithm aims to maximize $p(z_{1:N}, \boldsymbol{x}_{1:N})$ to estimate the most probable state sequence, $\hat{z}_{1:N}$, that generates the observations $\boldsymbol{x}_{1:N}$. As with the forward algorithm, a recursive approach is required due to the exponentially large number of possible sequences as N increases [189].

The algorithm systematically calculates, $t_i(z_i)$, which is the probability of the most probable state sequence that generates the first i observations and finishes in the state z_i [189]:

$$t_i(z_i) = p(\hat{z}_{1:i-1}, z_i, \boldsymbol{x}_{1:i}) \tag{4.45}$$

Identifying $p(\hat{z}_{1:N}, \boldsymbol{x}_{1:N})$ involves first recursively calculating $t_i(z_i)$ by:

$$t_i(z_i) = \max_l \left[t_{i-1}(z_{i-1} = l) \cdot p(z_i|z_{i-1} = l) \cdot p(\boldsymbol{x}_i|z_i)\right] \tag{4.46}$$

Algorithm 4.2 Log-scaled forward algorithm

input sequence $[\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_i, \cdots, \boldsymbol{x}_N \in \mathbb{R}^P] \equiv \boldsymbol{x}_{1:N}$ (where P is the vector dimension and N denotes the number of observations), transition probabilities $p(z_i|z_{i-1})$, emission probabilities $p(\boldsymbol{x}_i|z_i)$, and prior probabilities $p(z_1)$, where $i \in \mathbb{N} : 1 \leq i \leq N$.

$c_1 = \sum_{s=1}^{S} p(\boldsymbol{x}_1|z_1 = s)p(z_1 = s)$ ▷ Basis

$\tilde{\alpha}_1(s) = \dfrac{p(\boldsymbol{x}_1|z_1 = s)p(z_1 = s)}{c_1}$

$\delta_1(s) = c_1\tilde{\alpha}_1(s)$

for $i = 2$ to N **do** ▷ Recursion

  for $s = 1$ to S **do**

    $\delta_i(s) = p(\boldsymbol{x}_i|z_i = s) \sum_{m=1}^{S} p(z_i = s|z_{i-1} = m)\tilde{\alpha}_{i-1}(m)$

  end for

  $c_i = \sum_{s=1}^{S} \delta_i(s)$

  for $s = 1$ to S **do**

    $\tilde{\alpha}_i(s) = \dfrac{\delta_i(s)}{c_i}$

  end for

end for

$\log[\, p(\boldsymbol{x}_{1:N})\,] = \sum_{i=1}^{N} \log[c_i]$ ▷ Termination

output $\log[\, p(\boldsymbol{x}_{1:N})\,]$

where $l \in \mathbb{N} : 1 \leq l \leq S$.

Back-pointers, $b_i(z_i)$, keep track of the state at $i-1$ from which a transition is made to each z_i during the recursion. When $i = N$, the recursion is complete and the most probable state sequence is uncovered by back-tracking through $b_i(z_i)$ to $i = 1$. The full approach is presented in algorithm 4.3 and implemented in the log-domain to avoid numerical underflow.

The HMM-GMM

Figure 4.6 presents the plate notation for one of the most common HMM variants where $\boldsymbol{x}_i$ is distributed by state-dependent GMMs (with parameters defined as in section 4.5.2) [198]. For clarity, the GMM model is collapsed into the node $\boldsymbol{x}_i$. Often, single Gaussian distributions are used, but this combined HMM-GMM model takes advantage of the GMM's ability to better generalize non-Gaussian

Algorithm 4.3 Log-Viterbi algorithm

input $[\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_i, \cdots, \boldsymbol{x}_N \in \mathbb{R}^P] \equiv \boldsymbol{x}_{1:N}$ (where P is the vector dimension and N denotes the number of observations), transition probabilities $p(z_i|z_{i-1})$, emission probabilities $p(\boldsymbol{x}_i|z_i)$, and prior probabilities $p(z_1)$, where $i \in \mathbb{N} : 1 \leq i \leq N$.

for j=1 to S **do** ▷ Basis
  $t_1(j) = \log[p(\boldsymbol{x}_1|z_1 = j)] + \log[p(z_1 = j)]$
  $b_1(j) = 0$
end for
for i=2 to N **do** ▷ Recursion
  for j=1 to S **do**
    $t_i(j) = \max_l \left(t_{i-1}(l) + \log[p(z_i = j|z_{i-1} = l)] + \log[p(\boldsymbol{x}_i|z_i = j)]\right)$
    $b_i(j) = \arg\max_l \left(t_{i-1}(l) + \log[p(z_i = j|z_{i-1} = l)] + \log[p(\boldsymbol{x}_i|z_i = j)]\right)$
  end for
end for
$\hat{z}_N = \arg\max_l [t_N(l)]$
for i=N to 2 **do** ▷ Back-tracking
  $\hat{z}_{i-1} = b_i(\hat{z}_i)$
end for
output $\hat{z}_{1:N}$

distributions. The HMM-GMM is mathematically defined by [198]:

$$\boldsymbol{x}_i|z_i \sim \sum_{k=1}^{K} \phi_{z_i,k} \cdot \mathcal{N}(\boldsymbol{\mu}_{z_i,k}, \boldsymbol{\Sigma}_{z_i,k}) \tag{4.47}$$

$$z_i|z_{i-1} \sim \text{Cat}(\boldsymbol{t}_{z_{i-1}}) \tag{4.48}$$

$$z_1 \sim \text{Cat}(\boldsymbol{\xi}) \tag{4.49}$$

With this notation, $p(z_i|z_{i-1})$ are parametrized by the probability row vectors, $\boldsymbol{t}_{z_{i-1}}$ (of length S), which encode the probabilities of transitioning from each state, z_{i-1}, to all other states. The prior, $p(z_1)$, is parametrized by the row vector, $\boldsymbol{\xi}$ (of length S), which encodes the probabilities of beginning in each state, z_1.

4.5.4 Dirichlet Process Mixture Models (DPMMs)

In the GMM and HMM-GMM models (discussed in sections 4.5.2 and 4.5.3, respectively), the number of mixture components, K, needed to be explicitly specified. However, in many cases, this choice of value is unknown and unintu-

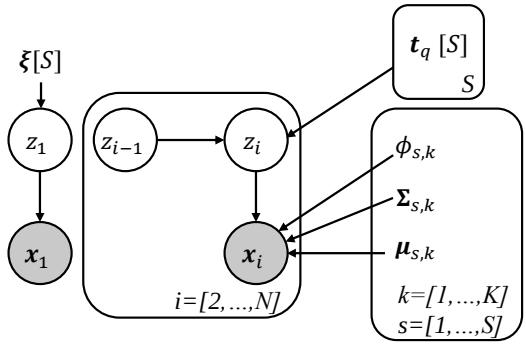


Figure 4.6: Plate notation for an HMM-GMM, using GMMs as the state-dependent observation distributions with parameters, $\boldsymbol{\mu}_{c_i,k} \in \mathbb{R}^P$, $\boldsymbol{\Sigma}_{c_i,k} \in \mathbb{R}^{P\times P}$, and $\phi_{c_i,k}$. $z_i \in \mathbb{N} : 1 \leq z_i \leq S$, $s \in \mathbb{N} : 1 \leq s \leq S$, and $k \in \mathbb{N} : 1 \leq k \leq K$ (P denotes the vector dimension, S the number of states, and K the number of components). HMM transitions from each state at $i-1$ is parametrized by the set of S vectors, $\boldsymbol{t}_q$ (each of length S), and the prior by $\boldsymbol{\xi}$ (also of length S).

itive. Alternatively, the DPMM extension can be used which implicitly selects K from the data.

While the GMM and HMM-GMM are known as Bayesian parametric models (with fixed distribution parameters), the DPMM is a Bayesian non-parametric model. This means that the distribution parameters can themselves be defined as random variables, with additional associated distributions. The parameters of these higher-level distributions are referred to as hyper-parameters.

Defining each i^{th} observation as $\boldsymbol{x}_i \in \mathbb{R}^P$ (where P is the vector dimension) and $\boldsymbol{\theta}_c$ (of arbitrary length) as one of κ component parameters[5] from which $\boldsymbol{x}_i$ is drawn, the DPMM model is defined as [199]:

$$\boldsymbol{x}_i|\boldsymbol{\theta}_c \sim F(\boldsymbol{\theta}_c) \tag{4.50}$$

$$\boldsymbol{\theta}_c|G \sim G \tag{4.51}$$

$$G \sim \mathcal{DP}(\delta, G_0) \tag{4.52}$$

The graphical model is shown in figure 4.7. Each $\boldsymbol{x}_i$ is drawn from a component distribution, $F(\boldsymbol{\theta}_c)$, parametrized by $\boldsymbol{\theta}_c$. In turn, $\boldsymbol{\theta}_c$ is drawn from a discrete

[5]It is important to note that, unlike K in the GMM and HMM-GMM, κ is a 'dynamic' variable reflecting that the number of components change according to the DPMM.

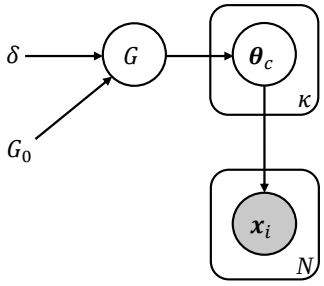


Figure 4.7: Plate notation for a DPMM. Each observation, $\boldsymbol{x}_i \in \mathbb{R}^P$ (where P is the vector dimension), is drawn from a component distribution parametrized by $\boldsymbol{\theta}_c$ (of arbitrary length) of which there are κ possible components. Each $\boldsymbol{\theta}_c$ is drawn from a discrete distribution, G, and κ changes according to the draw of G from a DP. The DP is parametrized by base distribution, G_0, and concentration parameter, δ.

distribution, G. Finally, G is drawn from a Dirichlet Process (DP) parametrized by a base distribution, G_0, and concentration parameter, δ [199][6].

The DP is a distribution over distributions, with each draw of G itself a discrete distribution. It can also be thought of as an extension of the Dirichlet distribution where the draw of G can have an infinite possible length [200]. Other possible extensions exist, such as the Pitman-Yor process or Poisson-Dirichlet process (specifically for grouped data) [201]. However, inferences using these more complex processes are challenging and the DP is usually a sufficient assumption. Furthermore, efficient inference procedures have been subsequently developed for the DP, as is later described.

In the context of a DPMM, drawing G from a DP can also be thought of as a discretized sampling of the parameter space. G_0 defines the continuous distribution over the parameter space, while δ defines how strong this discretization is. In the limit $\delta \to 0$, the draw of G (and consequently $\boldsymbol{\theta}_c$) would be a single value. On the other hand, as $\delta \to \infty$, the draw of G would become continuous and tend towards G_0. By controlling the level of discretization, δ controls the propensity of the DP to generate new components [199].

[6]Typically, this parameter is referred to as α in literature, however, δ is used here to avoid confusion with other measures used throughout this thesis.

A number of approaches have been developed to efficiently represent and draw from a DP. One such method is the Stick-breaking process, which provides a constructive way to directly draw G [199]. However, an alternative and arguably more efficient method is the colourfully named Chinese Restaurant Process (CRP) which clusters $\boldsymbol{x}_i$ and assigns $\boldsymbol{\theta}_c$ without explicitly calculating G [199].

The Chinese Restaurant Process (CRP)

The CRP clusters the data according to a DP using the analogy of a Chinese restaurant. An empty restaurant represents the infinite parameter space, with each table representing a particular component (with associated parameters) in the space. Each individual represents a data point and when sitting at a table, it is allocated to that particular component. The full algorithm is as follows [199]:

1. Begin with an empty restaurant of infinitely many tables (components).
2. The first customer ($\boldsymbol{x}_i$) enters the restaurant and always sits at the first table (component 1).
3. Every subsequent $(i+1)^{th}$ customer enters the restaurant and has two options:
   (a) Sit on any of the existing κ tables with probability $\frac{n_k}{n+\delta}$, where $k \in \mathbb{N} : 1 \leq k \leq \kappa$, n_k is the number of people currently occupying that table, and n is the total number of people.
   (b) Sit on the next unoccupied table with probability $\frac{\delta}{n+\delta}$.
4. For all occupied tables, draw parameters for each table from the base distribution, $\boldsymbol{\theta}_c \sim G_0$.

The CRP models the DP by employing a 'rich-gets-richer' scheme, whereby each new customer is more likely to sit at a table already occupied by a larger number of people. Each $\boldsymbol{\theta}_c$ is then drawn after the data points are clustered, however, $\boldsymbol{\theta}_c$ can also be assigned at the instance when a new component is created.

Employing the CRP to draw from the DP provides an efficient means for inference to obtain the DPMM parameters. Inference would typically involve maximising the posterior distribution, $p(\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_\kappa | \boldsymbol{x}_1, \cdots, \boldsymbol{x}_N)$. However, directly determining

this distribution is difficult and, since the DPMM is non-parametric (with changing κ), the EM algorithm (discussed in section 4.5.2) cannot be used. Instead, an iterative stochastic (probabilistic) method is adopted which incorporates the CRP.

Performing inference in the DPMM

In 2000, *R. Neal* published a series of stochastic methods for performing inference in a DPMM [200] that uses a form of Gibb's sampling. In general, a Gibb's sampler draws a new sample from a desired distribution by instead iteratively drawing from a related conditional distribution (that is easier to derive) [200].

In the context of the DPMM, for each $\boldsymbol{x}_i$, the assignment of $\boldsymbol{x}_i$ to a component, $z_i \in \mathbb{N} : 1 \leq z_i \leq \kappa$, is iteratively updated based on a distribution that is conditioned on the current assignments (and parameters $\boldsymbol{\theta}_c$). Calculating this distribution involves combining the CRP with the component distribution evaluated at $\boldsymbol{x}_i$, $F(\boldsymbol{x}_i; \boldsymbol{\theta}_c)$. $F(\boldsymbol{x}_i; \boldsymbol{\theta}_c)$ denotes $F(\boldsymbol{\theta}_c)$ from equation (4.50) evaluated at $\boldsymbol{x}_i$.

After all data assignments, values of $\boldsymbol{\theta}_c$ are then updated from a posterior distribution based on G_0 and the updated component assignments [200]. This process is then repeatedly iterated, reassigning all $\boldsymbol{x}_i$ and re-estimating all $\boldsymbol{\theta}_c$, until a predefined number of iterations are reached.

The earliest algorithms based on this procedure (such as Algorithm 2 in [200]) required calculating an integral over the entire parameter space, $\int_{\boldsymbol{\theta}} F(\boldsymbol{x}_i; \boldsymbol{\theta}) G_0 d\boldsymbol{\theta}$, to assign $\boldsymbol{x}_i$ to a new component. However, this is only feasible when G_0 is the conjugate prior of $F(\boldsymbol{x}_i; \boldsymbol{\theta})$ (i.e. it allows this integral to take the same form as the prior). Even then, this form can be mathematically complex to derive. As a solution, *R. Neal* instead proposed Neal's Algorithm 8 [200], which avoids explicitly computing this integral.

Neal's Algorithm 8 performs in a similar way as described above but introduces a finite set of empty 'auxiliary' components. In each iteration, m such components are defined, with each $\boldsymbol{x}_i$ only allowed to be allocated to either an existing component, or one of these new auxiliary components. Selecting from these limited components prevents the need to integrate over the entire parameter space [200]. This said, the

particular operation of assigning a data point to a new component and drawing new parameters remains computationally expensive when iterating through a large data size (N>1000). Consequently, an extension of Algorithm 8 is developed here and presented in algorithm 4.4. This 'stochastic Neal's Algorithm 8' allows the generation of new components using only a randomly selected sub-sample of the data points. The remaining data points are then only reassigned to one of the already existing components.

In this implementation, all data points are initialized as part of a single component. Although the algorithm is presented here for a general value of m, a choice of $m = 1$ is used throughout. This remains computationally fast and prevents the common problem of 'over-mixing', where too many components are generated and the data is constantly re-assigned without convergence [200].

The mixing coefficients, $[\phi_1, \cdots, \phi_\kappa]$, are also easily calculated as part of this inference and G_0 may often be parametrized with additional hyper-parameters. This depends on the chosen form which is also influenced by the choice of component distribution. As in the case of the GMM and HMM-GMM (in sections 4.5.2 and 4.5.3, respectively), one of the most commonly used component distributions is the Gaussian. In the DPMM framework, this specific combination is referred to as the DP-GMM.

The DP-GMM

A DP-GMM uses a Gaussian as the component distribution. $\boldsymbol{\theta}_c$ now refers to the component mean, $\boldsymbol{\mu}_c$ and covariance, $\boldsymbol{\Sigma}_c$. G_0 is most commonly selected as the Normal-inverse Wishart distribution, as it is the conjugate prior of the Gaussian distribution. This ensures that the posterior distribution for calculating new parameters in Neal's Algorithm 8 will also be Normal-inverse Wishart-distributed.

The DP-GMM model is a specific case of the DPMM illustrated in figure

Algorithm 4.4 Stochastic Neal's Algorithm 8

input observations $[\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N \in \mathbb{R}^P]$ (where P denotes the dimension, and N the number of observations), concentration parameter δ, base distribution G_0, number of auxiliary components m, number of iterations $numIter$, stochastic sample size $N^{(stoch)}$.

initialize component assignments for each data point as $[z_1, z_2, \cdots, z_N = 1]$, initial component parameters $\boldsymbol{\theta}_1$, $\kappa = 1$. ▷ All data points start in single component

for t=1 to numIter **do**
  Randomly draw $N^{(stoch)}$ of the total observations to form set $\mathcal{S}^{(stoch)}$.
  for each $\boldsymbol{x}_i \in \mathcal{S}^{(stoch)}$ **do**
    Let κ be number of distinct values of z_l for $l \neq i$ and set $h = \kappa + m$.
    if $z_i = z_l$ for some $l \neq i$ **then**
      Draw $\boldsymbol{\theta}_c$ independently from G_0 for $\kappa < c \leq h$.
    else if $z_i \neq z_l$ for all $l \neq i$ **then**
      Draw $\boldsymbol{\theta}_c$ independently from G_0 for $\kappa + 1 < c \leq h$.
    end if
    Draw a new value for z_i from set $\{1, \cdots, h\}$ with posterior probabilities:

$$p(z_i | z_{-i}, \boldsymbol{x}_i, \boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_h) = \begin{cases} b\dfrac{N_{-i,c}}{N-1+\delta} F(\boldsymbol{x}_i; \boldsymbol{\theta}_c) & \text{for } 1 \leq c \leq \kappa \\ \\ b\dfrac{\delta/m}{N-1+\delta} F(\boldsymbol{x}_i; \boldsymbol{\theta}_c) & \text{for } \kappa < c \leq h \end{cases}$$

    where b is a normalizing constant, z_{-i} denotes all values of z_l for $l \neq i$, and $N_{-i,c}$ is the number of $z_l = c$ for $l \neq i$.
    Relabel all z_l (including $l = i$) and $\boldsymbol{\theta}_c$ for only those components that contain at least one observation. Set κ to number of occupied components.
  end for
  for all $\boldsymbol{x}_i \notin \mathcal{S}^{(stoch)}$ **do**
    Draw a new value for z_i from set $\{1, \cdots, \kappa\}$ with posterior probabilities:

$$p(z_i | z_{-i}, \boldsymbol{x}_i, \boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_\kappa) = b\frac{N_{-i,c}}{N-1+\delta} F(\boldsymbol{x}_i; \boldsymbol{\theta}_c) \text{ for } 1 \leq c \leq \kappa$$

    if any components become empty **then** set this posterior probability to 0.
    end if
  end for
  Relabel all z_l (including $l = i$) and $\boldsymbol{\theta}_c$ for only those components that contain at least one observation. Set κ to number of occupied components.
  for all c in $\{1, \cdots, \kappa\}$ **do**
    Draw new $\boldsymbol{\theta}_c$ based on posterior distribution of G_0 and number of component observations, N_c.
    Calculate mixing coefficient $\phi_c = \frac{N_c}{N}$.
  end for
end for
output $[\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_\kappa]$, and $[\phi_1, \cdots, \phi_\kappa]$

4.7, and defined as:

$$\boldsymbol{x}_i|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c \sim \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) \tag{4.53}$$

$$\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c|G \sim G \tag{4.54}$$

$$G \sim \mathcal{DP}(\delta, G_0) \tag{4.55}$$

$$G_0 = \mathcal{NIW}(\boldsymbol{\mu}_0, \lambda_0, \boldsymbol{\Psi}_0, \nu_0) \tag{4.56}$$

$\mathcal{NIW}$ denotes the Normal-inverse Wishart distribution which has four additional hyperparameters. $\boldsymbol{\mu}_0$ and $\boldsymbol{\Psi}_0$ are the expected values of $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$, respectively. λ_0 and ν_0 are scaling factors (the latter also referred to as the degrees of freedom) and represent the extent of 'informativeness' in the prior. Typically, $\nu_0 > P - 1$ (where P is the dimension of the parameters) to ensure that the draw of Σ_c remains positive-definite [202].

A draw from $\mathcal{NIW}$ is achieved by first drawing $\boldsymbol{\Sigma}_c$ from the inverse Wishart distribution, $\boldsymbol{\Sigma}_c \sim \mathcal{W}^{-1}(\boldsymbol{\Psi}_0, \nu_0)$. $\boldsymbol{\mu}_c$ is then drawn from the prior Gaussian distribution, $\boldsymbol{\mu}_c \sim \mathcal{N}(\boldsymbol{\mu}_0, \frac{1}{\lambda_0}\boldsymbol{\Sigma}_c)$ [202, 203]. $\mathcal{W}^{-1}(\boldsymbol{\Psi}, \nu)$ has the general form [202, 203]:

$$\mathcal{W}^{-1}(\boldsymbol{\Psi}, \nu) = \frac{|\boldsymbol{\Psi}|^{\frac{1}{\nu}}}{2^{\frac{\nu P}{2}}\Gamma(\frac{\nu}{2})}|\boldsymbol{\Sigma}|^{\frac{\nu+P+1}{2}}e^{-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Psi}\boldsymbol{\Sigma}^{-1})} \tag{4.57}$$

where $\Gamma(\cdot)$ denotes the multivariate Gamma distribution and $\mathrm{tr}(\cdot)$ denotes the trace of the matrix (sum of the elements along the main diagonal).

The posterior distribution to update the component parameters is defined as [202, 203]:

$$\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c|\boldsymbol{x}_i\ (\forall i : z_i = c) \sim \mathcal{NIW}_c(\boldsymbol{\mu}_{c,N}, \lambda_{c,N}, \boldsymbol{\Psi}_{c,N}, \nu_{c,N}) \tag{4.58}$$

where:

$$\begin{aligned}
\mu_{c,N} &= \frac{\lambda_0\boldsymbol{\mu}_0 + N_c\bar{\boldsymbol{x}_c}}{\lambda_0 + N_c} \\
\lambda_{c,N} &= \lambda_0 + N_c \\
\nu_{c,N} &= \nu_0 + N_c \\
\boldsymbol{\Psi}_{c,N} &= \boldsymbol{\Psi}_0 + \boldsymbol{S}_c + \frac{\lambda_0 N_c}{\lambda_0 + N_c}(\bar{\boldsymbol{x}_c} - \boldsymbol{\mu}_0)(\bar{\boldsymbol{x}_c} - \boldsymbol{\mu}_0)^\intercal
\end{aligned}$$

$\boldsymbol{S}_c = (\boldsymbol{x}_i - \bar{\boldsymbol{x}_c})(\boldsymbol{x}_i - \bar{\boldsymbol{x}_c})^\intercal\ \forall i : z_i = c$ and $\bar{\boldsymbol{x}_c}$ denotes the sample mean of $\boldsymbol{x}_i\ \forall i : z_i = c$.

4.6 Assessing model performance

4.6.1 Cross-Validation (CV)

Training and testing machine learning models (particularly supervised models) requires a structured approach for splitting the data to avoid bias and/or over-fitting. Most often, a K-fold cross-validation (CV) is adopted where the dataset is split into K equal (or near-equal) partitions. Training is performed on the $K-1$ folds and testing on the remaining left-out fold. By repeating this for all K folds, the entirety of the data is used for both training and testing, further reducing possible bias if only a single, arbitrary split was chosen. Final model performance is typically summarized by calculating the mean performance across all the test folds [204].

In some cases, datasets may contain repeated measurements (such as multiple recordings) that originate from the same subject. In this case, it is important to split the folds by *subject*. This ensures that data from the same subject is not present in both the training and testing splits, as inter-subject variability may further induce a bias in the results. When there is limited data from S different subjects, a Leave-One-Subject-Out CV (LOSO CV) is effective to maximize the data available for training. LOSO CV is achieved by performing K-fold CV where $K = S$ [204].

Bias and over-fitting, however, is re-introduced when the test set is needed to further evaluate models for parameter tuning, or when introducing a feature selection method that requires training data. Under these conditions, a nested CV is more appropriate [205].

Nested CV

Nested CV utilizes two instances of CV. An outer-most K-fold CV first separates a test fold from the rest of the data. After the test split, a second inner-most L-fold CV further splits the remaining data into a training and validation set. Within this second split, multiple models can be trained with the training set and evaluated on the validation set to tune additional parameters, perform feature selection etc., while reserving the test fold for a final unbiased assessment of performance [205].

Algorithm 4.5 Nested CV

for each k^{th} fold, where $k \in \mathbb{N} : 1 \leq k \leq K$ **do**
  Separate the k^{th} fold from the data and reserve for testing.
  Perform feature selection using the remaining data and choose only the resulting selected features.
  for each l^{th} fold, where $l \in \mathbb{N} : 1 \leq l \leq L$ **do**
    Separate the l^{th} fold from the remaining data as a validation set.
    Denote the rest as the training set.
    for all parameter combinations in the grid **do**
      Train a model on the training set using the chosen parameters.
      Generate model estimates on the validation set and store.
    end for
  end for
  for all parameter combinations in the grid **do**
    Assess performance from the combined estimates across all L folds using an appropriate measure, $I(\cdot)$.
  end for
  Choose optimal parameter values yielding maximum performance (or minimum, depending on the choice of $I(\cdot)$).
  Re-train the model using all validation and training data (excluding the test set) and optimal parameter values.
  Generate model estimates on the test set and store.
end for
Assess overall performance from the combined estimates across all K folds, using $I(\cdot)$.

The precise structure of the nested CV varies, depending on the optimization method chosen for selecting parameters and if other routines (such as feature selection) are required. The method of nested CV chosen here is adapted from [34] and presented in algorithm 4.5. It incorporates both feature selection and a grid search for tuning parameters. A grid search exhaustively trains models for each parameter combination and assesses performance on the validation data, by using an appropriate performance measure.

Choices for $I(\cdot)$ are dependent on whether the problem is classification- or regression-based. A number of measures are presented in the subsequent sections.

4.6.2 Measures for classification

When performing classification, most measures compare performance between an algorithm and ground truth, with respect to a particular class (label). In the context of time-series signals (such as EEG), most measures are applied to each classified sample or windowed epoch in the signal.

In the simplest case of binary classification (label 0 or 1), when both the algorithm and ground truth agree with label 1, this is a 'hit' or true positive. Similarly, when the algorithm and ground truth both agree with label 0, this is a correct rejection or true negative. In cases of misclassification, an algorithm declaring 1 when the ground truth is 0 is a false alarm or false positive. Finally, the case of an algorithm declaring 0 when the ground truth is 1 is a 'miss' or false negative.

By summing the total number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), a number of summary measures are defined.

Sensitivity, specificity, precision & F1

The sensitivity (also known as the recall or true positive rate) measures the proportion of actual positives that are correctly identified by the algorithm, defined as [206, 207]:

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP+FN}} \tag{4.59}$$

Inversely, the specificity (also known as the selectivity or true negative rate) measures the proportion of the actual negatives that are correctly identified [206, 207]:

$$\text{specificity} = \frac{\text{TN}}{\text{TN+FP}} \tag{4.60}$$

As well as sensitivity, the precision (also known as the positive predicted value) measures the proportion of algorithm positives which are true positives:

$$\text{precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{4.61}$$

Precision is essentially a measure of algorithm 'exactness'. The harmonic mean of the sensitivity and precision results in the F1, which is a more informed measure

of a classifier's performance (compared to calculating the accuracy), as it can account for class imbalance [207]:

$$\mathrm{F1} = 2 \cdot \frac{\text{sensitivity} \cdot \text{precision}}{\text{sensitivity+precision}} \tag{4.62}$$

F1 is specifically penalized in cases of imbalanced data where there are higher proportions of negatives than positives in the dataset.

Event-based measures

In some cases, it is also useful to assess classification performance based on the precise number of overall periods (or 'events') that are correctly (or incorrectly) identified by an algorithm. Two additional measures that capture this information are the event-based True Positive Rate (eTPR) and the event-based False Discovery Rate (eFDR). eTPR measures the proportion of true events correctly detected by the algorithm. eFDR measures the proportion of all events detected by the algorithm that are misclassified [208]:

$$\mathrm{eTPR} = \frac{\text{No. correctly detected events}}{\text{Total no. true events}} \tag{4.63}$$

$$\mathrm{eFDR} = \frac{\text{No. incorrectly detected events}}{\text{Total no. events detected by algorithm}} \tag{4.64}$$

Higher values of eTPR and lower values of eFDR denote improving algorithm performance.

Receiver Operating Characteristic (ROC) curve

In cases where algorithms utilize a threshold to perform final classification, the aforementioned measures will vary depending on the choice of threshold. Consequently, in cases of binary classification, assessing performance using ROC curves has become popular as it dispenses with the need to commit to a single threshold value.

ROC curves plot the true positive rate (sensitivity) against the false positive rate (defined as 1-specificity), as the classification threshold is incrementally shifted over a wide range. This is illustrated in figure 4.8. Assessment of general classification

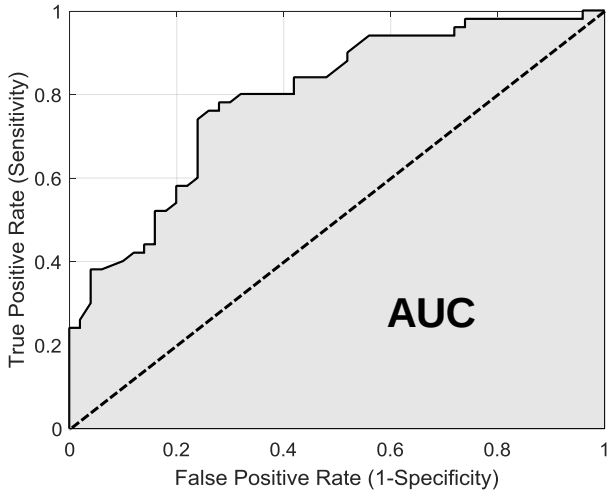


Figure 4.8: Illustration of an ROC curve. The grey shading denotes the AUC of the curve as a quantitative measure of overall algorithm performance (regardless of the classifier threshold chosen). A classifier should typically yield an AUC>0.5. The dashed line shows an ROC curve of a classifier performing at the chance level (with AUC=0.5).

performance can be seen by the extent to which the ROC curve is skewed towards the top left corner of the plot [207].

To summarize the ROC quantitatively, the Area Under the Curve (AUC) can be calculated by integrating the area under the ROC curve (shaded in figure 4.8). Perfect classification is achieved when AUC=1, while an AUC=0.5 is equivalent to chance agreement. This calculation can be approximated using the trapezoidal integration method [172].

Confusion matrices

The above classification metrics are affective in the case of binary states. However, when expanding to more than two states, other means to present classification performance is needed. A simple approach is the confusion matrix, which is a two-dimensional matrix that presents the number of state agreements (and disagreements) between predictions from the algorithm and the ground truth [209]. For binary classification, the general form can be written in terms of TP, FP, TN and FN, as previously defined for sensitivity, specificity, precision and F1. This is illustrated in figure 4.9. The dimensions of the confusion matrix are simply extended for cases of more than two states.

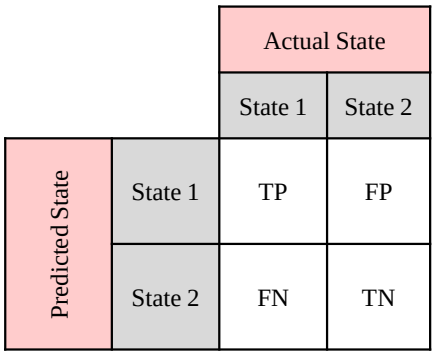


Figure 4.9: The general form of the confusion matrix in a binary classification problem. This summarizes the agreement between predictions from an algorithm and the ground truth.

Cohen's kappa (κ)

As with the F1 previously discussed, κ provides an alternative means to measure performance while taking into account imbalanced data. The added benefit of κ is the ability to assess performance when there are more than two states (calculable from the confusion matrix), while still accounting for chance agreement. It was first defined by *J. Cohen* [210] in 1960 to assess the inter-rater agreement between two human scorers and has since gained popularity to assess the performance of machine learning algorithms (particularly for sleep state classification). κ is defined as [210]:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{4.65}$$

p_o is the relative observed agreement between the two raters and is identical to the classification accuracy. p_e is the expected agreement (or the estimated probability of random agreement) and interpreted as the accuracy that a randomly assigning rater would be expected to achieve. This is estimated from the observed data by [210]:

$$p_e = \frac{1}{N^2} \sum_k n_k^{(1)} n_k^{(2)} \tag{4.66}$$

N is the total number of instances rated, while $n_k^{(1)}$ and $n_k^{(2)}$ denote the total number of predictions in each k^{th} state made by raters 1 and 2, respectively. In the case of balanced binary data, $p_e \approx 0.5$.

Unlike the aforementioned methods, κ is a more conservative measure and falls in the range $-1 \leq \kappa \leq 1$. Values near one denote almost perfect classification,

while values nearer zero tend towards chance agreement. If negative values are obtained, this suggests an agreement worse than random and is likely due to a detrimentally performing algorithm. While κ is interpreted differently depending on application, it is commonly regarded that values <0.4 are poor while values >0.75 are considered near-perfect [211, 212].

Krippendorff's alpha (α)

Krippendorffs α is an alternative measure of interrater agreement, first introduced by *K. Krippendorff* in 1970 [213, 214]. Unlike Cohen's κ, it is more easily generalized and can be calculated for multiple raters, adjust itself to small sample sizes, and can deal with missing data. It also assesses agreement for both nominal and ordinal values (amongst others) while κ is limited to only nominal data [213, 214].

To calculate α, each rater is denoted by $c_m \in \mathbb{N} : 1 \leq m \leq M$ (where M is the total number of raters) and each independently rates a series of units (i.e. epochs, events, etc.), $u_n \in \mathbb{N} : 1 \leq n \leq N$ (where N is the total number of units). Each unit is assigned a value from each rater, $v_{m,n} \in \mathbb{N} : 1 \leq v_{m,n} \leq V$ (where V is the largest possible value). The term, m_{u_j}, denotes the total number of values assigned to unit u_j and may be $< M$ if there are missing ratings.

The method assesses performance by counting the number of *pairable* values between raters across units. The number of pairable values is summarized in a $(V \times V)$ matrix of observed coincidences. Denoting each possible value pair as (v, v') and rater pair as (i, i'), the matrix is symmetrical about the main diagonal and each element, $o_{v,v'}$, is defined as [213, 214]:

$$o_{v,v'} = \sum_{j=1}^{N} \frac{\sum_{i:i\neq i'}^{M} I(v_{i,u_j} = v) \cdot I(v_{i',u_j} = v')}{m_{u_j} - 1} = o_{v',v} \tag{4.67}$$

The function $I(\cdot) = 1$ if $\cdot$ is true, and $I(\cdot) = 0$ otherwise. Additionally, the total number of each value in a pair can be determined from this matrix as $Q_v = \sum_{l=1}^{V} o_{v,l}$. The sum over the full matrix is denoted by Q.

Similar to Cohen's κ, α is calculated between observed and expected values to account for chance agreement. With observed coincidences determined, the

expected coincidence matrix estimates the probability of each pair occurring by chance [213, 214]:

$$e_{v,v'} = \frac{\sum_{j=1}^{N} \sum_{i:i \neq i'}^{M} I(v_{i,u_j} = v) \cdot I(v_{i',u_j} = v')}{Q - 1} \tag{4.68}$$

Finally, α is calculated by:

$$\alpha = 1 - \frac{\sum_{v=1,v'=1}^{V} o_{v,v'} \delta(v, v')}{\sum_{v=1,v'=1}^{V} e_{v,v'} \delta(v, v')} \tag{4.69}$$

The function $\delta(v, v')$ is defined according to the nature of the data. For nominal data, $\delta(v, v') = \delta_{nominal}(v, v')$ where [213, 214]:

$$\delta_{nominal}(v, v') = \begin{cases} 0 & \text{if } v = v' \\ 1 & \text{if } v \neq v' \end{cases} \tag{4.70}$$

For ordinal data, $\delta(v, v') = \delta_{ordinal}(v, v')$ where:

$$\delta_{ordinal}(v, v') = \left[Q_v + Q_{v'} - \left(\frac{Q_v + Q_{v'}}{2} \right) \right]^2 \tag{4.71}$$

As with κ, values of α are similarly interpretable and in the range $-1 \leq \alpha \leq 1$. It was suggested by *K. Krippendorff* that α values $\geq$0.67 are considered a good agreement while values $\geq$0.8 are considered excellent [214].

4.6.3 Measures for regression

As regression models produce continuous estimates, the previously described classification metrics of section 4.6.2 no longer apply. As was first defined in section 4.3.1, measures of R^2 provide one possible indication of regression performance through assessing the tightness of the fit. While useful, it still does not provide an explicit measure of how close the estimates are to the true values.

The RF regression in section 4.3.3 introduced the SSE which is often used to assess regression performance. However, this value cannot be compared between datasets of different sizes and is not always easily interpretable. Consequently, the Root Mean Squared Error ($\sqrt{MSE}$) is also defined.

Root Mean Squared Error ($\sqrt{MSE}$)

The $\sqrt{MSE}$ (also known as the root mean squared deviation) summarizes the total error between estimates and true values. Denoting $y_i^{(est)}$ as the i^{th} estimate and $y_i^{(true)}$ as the corresponding true value (in a dataset of N values) [215]:

$$\sqrt{MSE} = \sqrt{\frac{\sum_{i=1}^{N}\left(y_i^{(est)} - y_i^{(true)}\right)^2}{N}} \tag{4.72}$$

In the case of linear regression, this metric can also be derived from equation (4.19). Values close to zero indicate a near-perfect regression fit.

4.6.4 The Wilcoxon signed-rank test

The Wilcoxon signed-rank test is a non-parametric test which tests the null hypothesis, H_0, that two sets of measures (which are paired across tests, models, repeats, etc.) both originate from the same underlying distribution. If H_0 is rejected, the measures are deemed to originate from distinctly different distributions. This was first presented by *F. Wilcoxon* in 1965 [216] as a variant of the paired Student's t-test but does not make the same assumption that the measures are normally distributed.

The Wilcoxon signed-rank test assumes that the paired measures are calculated from the same underlying source and are randomly and independently drawn. Furthermore, values must be continuous and ordinal in nature. More specifically, it accepts or rejects the H_0 that the *ranked differences* of the paired measures originate from a symmetric distribution with mean zero [216]. To perform the test [216]:

1. Denoting the first i^{th} measure of a pair by $x_i^{(1)}$ and the second by $x_i^{(2)}$, the paired differences for each of N pairs is given by:

$$D_i = x_i^{(2)} - x_i^{(1)} \tag{4.73}$$

   The absolute values, $|D_i|$, are calculated and all pairs yielding a value of zero are excluded. The resulting reduced sample size is given by $N^{(r)}$.

2. All pairs are ranked in ascending order according to $|D_i|$ and each given an integer rank, $I_i \in \mathbb{N}$, starting at one. In cases of ties, the ranks are recalculated and set equal to the average of the ranks that these pairs span. This results in the augmented ranks, $R_i \in \mathbb{R}$.
3. The test statistic, W, is calculated as the sum of the signed ranks:

$$W = \sum_{i=1}^{N^{(r)}} \left(\frac{D_i}{|D_i|} \cdot R_i \right) \tag{4.74}$$

4. Under H_0, W is assumed to follow a symmetric distribution with mean, $\mu_W = 0$, and variance, σ_W^2, defined as:

$$\sigma_W^2 = \frac{N^{(r)}(N^{(r)}+1)(2N^{(r)}+1)}{6} \tag{4.75}$$

   To assess statistical significance, W is compared to a critical value, $W^{(critical)}$, dependent on $N^{(r)}$ and the required significance level. The two-sided test rejects H_0 if $|W| > W^{(critical)}$.
5. When $N^{(r)} \geq 10$, W converges to a z-distribution (as defined in appendix A.1.1). In this case, the z-score (z) is calculated as:

$$z = \frac{W}{\sigma_W}, \text{ where } \sigma_W = \sqrt{\frac{N^{(r)}(N^{(r)}+1)(2N^{(r)}+1)}{6}} \tag{4.76}$$

   This value can now be compared to the z-distribution, with H_0 rejected if it exceeds the critical z for a given p-value.

5
Unsupervised Quiet Sleep Detection Across Preterm Ages

Associated publication(s):

A. Dereymaeker*, K. Pillay* et al. "An Automated Quiet Sleep Detection Approach in Preterm Infants as a Gateway to Assess Brain Maturation". In: *Int. J. Neural Syst.* 27.6 (2017), pp. 1750023(1)-1750023(18). **Joint first authors.*

5.1 Introduction

Chapters 1 and 2 highlighted the importance of producing robust methods for automated EEG sleep staging, as a first step to develop fully automated methods to quantify brain maturation. This is additionally important in the context of this research, as the available abnormal outcome datasets are not visually labelled for sleep. To utilize this data in subsequent analyses, classification of sleep states must be achieved over a wide PMA range.

Chapter 2 identified a few supervised methods for detecting QS and non-QS periods over a wide PMA range. However, these methods require large training datasets (across PMA), to successfully generalize and avoid the effects of inter-patient variability. An unsupervised approach that uniquely adapts to each recording and avoids the need for large training sets would be an ideal solution. However, chapter 2 revealed limitations with existing approaches. The SAT% method [32, 33, 89] was limited to <32 weeks PMA, while the temporal profiling method of *Krajca*

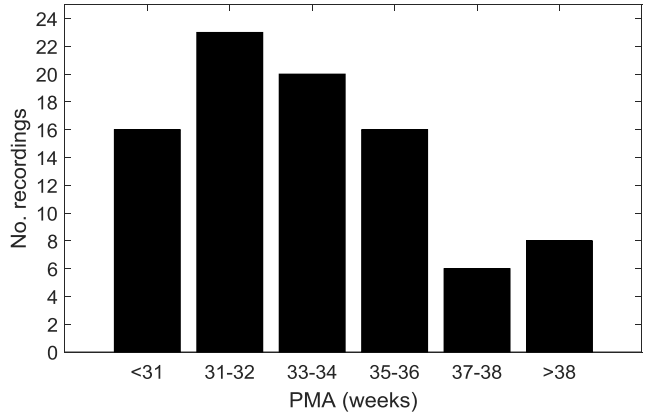


Figure 5.1: Histogram of the total number of EEG recordings used in the study, ordered by PMA. There were 89 recordings in total, ranging from 27 to 42 weeks PMA.

et al. [101–103] had potential but currently remains only a concept. Consequently, there is no unsupervised method to robustly detect sleep over a wide PMA range.

In this chapter, the method of *Krajca et al.* is expanded and validated, to evolve it into a robust and generalizable algorithm for detecting QS and non-QS, dubbed CLuster-based Adaptive Sleep Staging (CLASS). Performance of CLASS is assessed across a wide PMA range of preterm EEG, and compared to the SAT% algorithm. CLASS is also compared to the deep neural network of *Ansari et al.* [100] (also discussed in chapter 2). This uses the most 'modern' supervised deep learning techniques and was proposed after the development of CLASS.

5.2 Data

Dataset 1 was used for this study, as defined in chapter 3. The age distribution of this dataset is presented in the histogram of figure 5.1. 65% of the total data was labelled as non-QS and 35% as QS.

Based on the timing of data recruitment, the first 34 recordings that became available after screening were used for algorithm development and optimization and is referred to as the training set. The subsequent 55 recordings are referred to as the test set and were used solely to assess final algorithm performance. The patients used in each set were distinct to avoid any bias in the results.

5.3 The CLASS algorithm

The development of CLASS was based on the assumption that QS EEG was always *relatively* more discontinuous than non-QS and that this was maintained over a wide PMA range. Each stage of CLASS was developed from the originally proposed stages by *Krajca et al.* consisting of adaptive segmentation, clustering and QS classification (as defined in [101–103]), with an additional artefact removal routine introduced. A flowchart summarizing these stages is presented in figure 5.2(a). Each stage is further detailed below, with a series of key parameters defined throughout. The optimization method for tuning these parameters and the final values are presented subsequently in sections 5.4 and 5.6.1, respectively.

5.3.1 Artefact Subspace Reconstruction (ASR)

The method of *Krajca et al.* identified high power artefacts (predominantly due to movement) as potential sources of QS misclassification, due to producing unnatural periods of discontinuity in the EEG [101–103]. Thus, a rigorous artefact removal scheme was first required to ensure that CLASS remained robust.

Existing neonatal EEG pre-processing studies often exclusively use bandpass and notch filtering for artefact removal [89]. However, some artefacts cannot be sufficiently removed by this method alone. Other popular artefact removal methods include Independent Component Analysis [217] and Principal Component Analysis (PCA) [218], which involve transforming the EEG channels to a different component space that more clearly isolates the artefacts. However, such methods assume that EEG movement artefacts are stationary in nature, which is not the case.

An alternative technique called ASR, developed by *C. Kothe* [219], was used here. In general, the method applies PCA over a sliding window along the EEG channels, locally separating high-power artefacts from the clean signal [220] and reconstructing the signal from only the clean 'subspace'. ASR is illustrated in figure 5.3(a) on an epoch of EEG containing high-power, non-stationary artefacts.

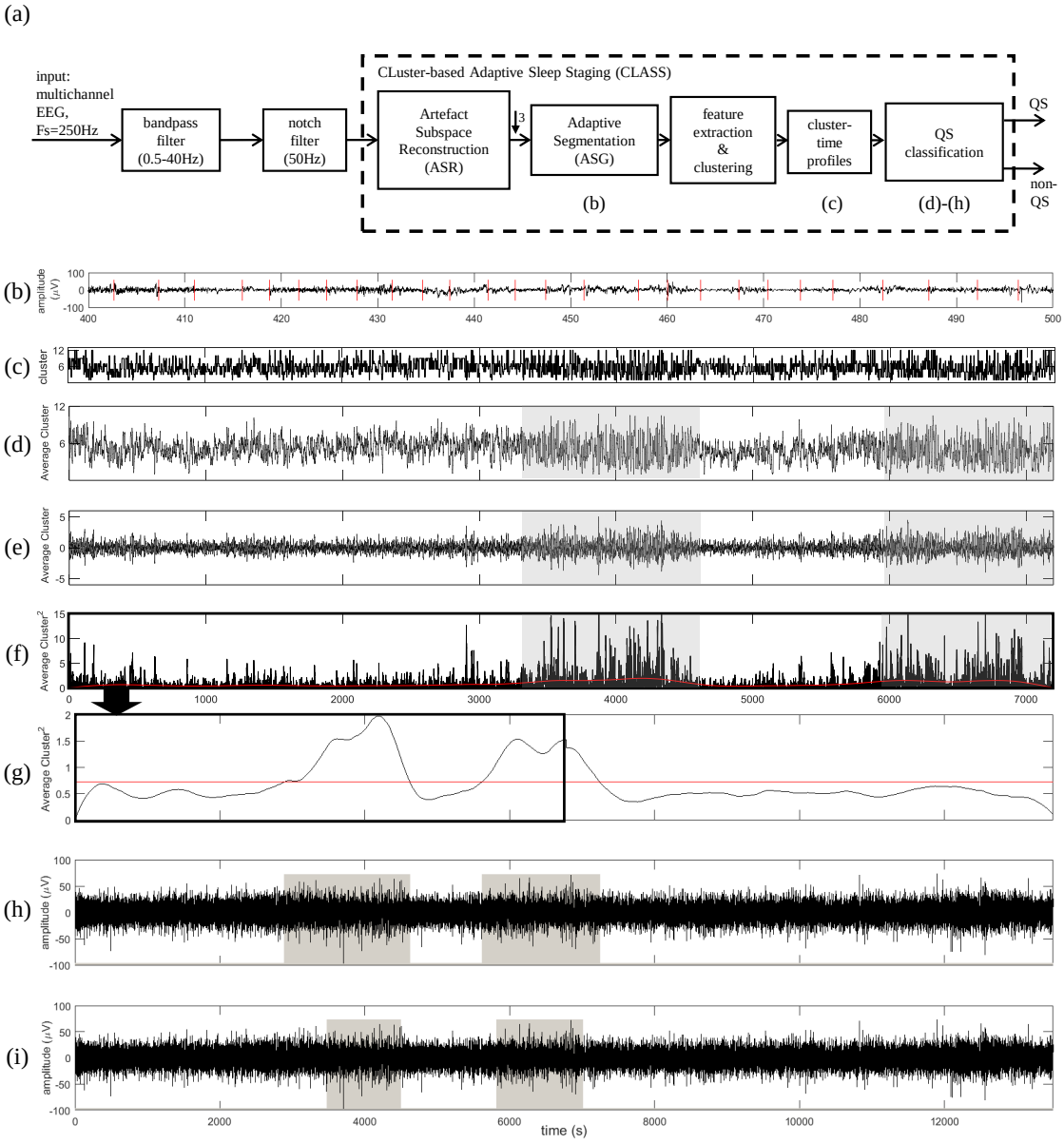


Figure 5.2: (a) Flowchart of the stages of EEG processing by CLASS. (b) Illustration of ASG for a 100 second period of EEG from a single channel. Red lines denote the ASG segment boundaries. (c) A cluster-time profile for a two hour epoch of EEG from a single channel. (d) The average cluster-time profile after taking the mean profile across all channels. Regions of increasing cluster fluctuation (shaded) correspond to higher EEG discontinuity and QS. (e) The de-trended signal after subtracting the running mean. (f) The square of the zeroed signal, with the signal envelope shown in red. (g) The stitched signal envelope for a four hour EEG recording, with the mean QS classification threshold shown in red. The first two hours of this stitched envelope correspond to the envelope in (f). (h) QS periods as estimated by CLASS (shaded) after thresholding. (i) The visually labelled QS periods (shaded).

This application of ASR used an efficient implementation developed for EEGlab (Swartz Center for Computational Neuroscience, San Diego, CA, USA) [221] and is applied in the following stages [219]:

1. Calibration PCA: A one minute epoch of artefact-free multichannel EEG (the 'calibration sequence') is selected to obtain thresholds for identifying clean signal and artefact subspaces within each sliding window of the 'raw' EEG. Thresholds are obtained to separate the clean and artefact subspaces by first applying a robust version of PCA [219] to the calibration sequence. As with conventional PCA (defined in appendix A.3), this involves calculating the covariance matrix, $\boldsymbol{Y} \in \mathbb{R}^{L \times L}$ (where L is the number of EEG channels).

2. Calculating clean and artefact subspace thresholds: Component thresholds for separating the clean and artefact subspaces are now defined in the calibration principal component space. Each component is divided into fixed-length segments and the Root-Mean-Square (RMS) power calculated for each segment. A Gaussian distribution is fitted to these powers and each c^{th} component threshold, t_c, defined based on the mean, μ_c, and standard deviation, σ_c, of the fitted distribution:

$$t_c = \mu_c + \sigma_c \cdot ASR_thresh \tag{5.1}$$

   where *ASR_thresh* is a pre-tuned parameter.

3. The reconstruction mixing matrix: Typical PCA artefact removal involves performing PCA on the raw EEG, eliminating those principal components that contain artefacts, and reconstructing the EEG directly from the remaining components. In ASR, each raw EEG recording is instead reconstructed using fixed statistics from the clean calibration sequence. To capture this information, the mixing matrix, $\boldsymbol{M} \in \mathbb{R}^{L \times L}$, is defined from the relationship $\boldsymbol{Y} = \boldsymbol{M}\boldsymbol{M}^{\intercal}$.

4. Cleaning the raw EEG: Within each sliding window of the raw EEG, $\boldsymbol{S} \in \mathbb{R}^{L \times N}$ (with N the EEG length), conventional PCA is applied to $\boldsymbol{S}$ to obtain the eigenvector, $\boldsymbol{V} \in \mathbb{R}^{L \times L}$, and eigenvalue, $\boldsymbol{\Lambda} \in \mathbb{R}^{L \times L}$, matrices. Reconstruction of

the EEG from the clean subspace is performed at a fixed time point (sample) within the EEG window, with this vector of amplitudes across channels denoted by $\boldsymbol{s} \in \mathbb{R}^L$. This is illustrated in figure 5.3(b). Cleaning of $\boldsymbol{s}$ to produce $\boldsymbol{s}^{(clean)} \in \mathbb{R}^L$ is achieved using a combined reconstruction matrix, $\boldsymbol{R} \in \mathbb{R}^{L \times L}$:

$$\boldsymbol{R} = \boldsymbol{M}(\boldsymbol{V}^\intercal \boldsymbol{M} \circ \boldsymbol{A})^+ \boldsymbol{V}^\intercal \tag{5.2}$$

$$\boldsymbol{s}^{(clean)} = \boldsymbol{R}\boldsymbol{s} \tag{5.3}$$

The Moore-Penrose pseudo-inverse is denoted by $+$ and $\boldsymbol{A}$ is a binary matrix that identifies which components in the EEG form the artefact subspace, by comparison of $\boldsymbol{\Lambda}$ with the calibration thresholds defined in equation (5.1) [219]. $\boldsymbol{R}$ also accounts for the various projections of the data between the calibration and raw EEG principal component spaces.

To clean the entire EEG, the sliding window is shifted sample-by-sample along the recording, with $\boldsymbol{R}$ recalculated each time and applied to the new $\boldsymbol{s}$ to obtain $\boldsymbol{s}^{(clean)}$. As $\boldsymbol{M}$ is derived from the (fixed) calibration sequence, it remains constant throughout.

A choice of calibration sequence at 40 weeks PMA was empirically found to best remove artefacts across the training set. The distinct burst-inter-burst periods in EEG below term-age (as described in chapter 2) closely resembled high-power artefacts in the more continuous term-age EEG, such that choosing calibration data from EEG below term-age did not remove artefacts well at the older ages. However, the inverse scenario of applying a term-age calibration sequence to EEG below term-age successfully removed artefacts without affecting the underlying burst-inter-burst behaviour.

5.3.2 Adaptive Segmentation (ASG)

After performing ASR, each EEG channel was then divided into varying length segments (typically one to five seconds long) by ASG, which segmented the signal non-uniformly such that each segment locally resembled a specific EEG pattern.

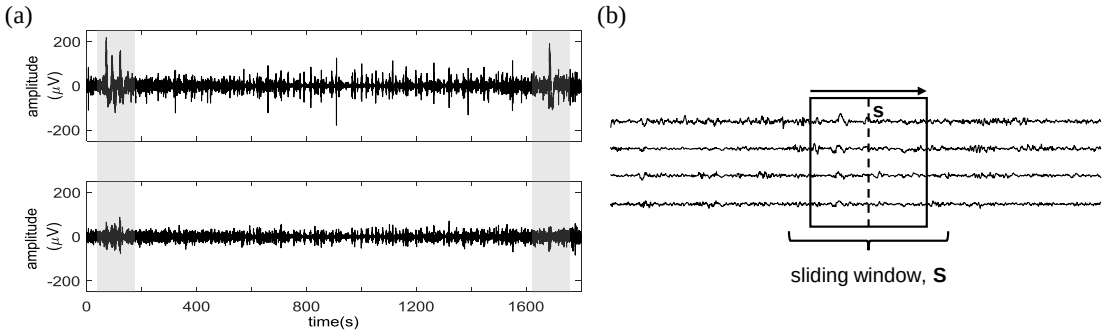


Figure 5.3: Illustration of the ASR method. (a) *Top:* A 30 minute epoch of bandpass filtered (0.5-40Hz) EEG in a single channel, before ASR is applied. High power artefacts are shaded. *Bottom:* The bandpass filtered signal after ASR is applied. The same shaded artefacts are now reduced while surrounding clean periods of the signal remain intact. (b) Illustration of the cleaning procedure of ASR on the EEG recording. Reconstruction metrics are calculated within the sliding window, $\mathbf{S} \in \mathbb{R}^{L \times N}$ (with N the EEG length and L the number of channels), in order to clean the sample of data along the dotted line, denoted by $\mathbf{s} \in \mathbb{R}^{L}$.

A number of ASG methods exist based on deriving an autocorrelation function [103], extracting power spectra information from the FFT [103], wavelet analyses [222], or simply based on identifying large differences when comparing the signals between two sliding, contiguous windows [102, 103]. Of all these methods, the latter was proven in [84, 85, 101–103] to be most efficient, accurate, and simple to implement. This method of ASG utilizes sliding contiguous windows to compare amplitude- and frequency-based measures between these windows as they slide along the recording. The locations of relatively large changes in these measures denote ASG segment boundaries as it reflects the onset of new segment patterns.

Each contiguous window has length *WIN* seconds which moves along the channel in steps of *SHIFT* samples, with both parameters requiring tuning. For each shift, the amplitude-based, $ADIF$, and frequency-based, $FDIF$, measures are calculated for each window. Denoting x_i as the i^{th} sample of EEG in the window, $ADIF$ and $FDIF$ are defined as [102, 103]:

$$ADIF = \sum_{i=1}^{WIN \cdot Fs} |x_i| \tag{5.4}$$

$$FDIF = \sum_{i=1}^{WIN \cdot Fs} |x_i - x_{i-1}| \tag{5.5}$$

where Fs is the sampling frequency. Differences between these measures in each window are combined to form a weighted value, G:

$$G = |ADIF^{(1)} - ADIF^{(2)}| + kF|FDIF^{(1)} - FDIF^{(2)}| \tag{5.6}$$

Superscripts 1 and 2 denote which window the measures originate from and kF is a pre-tuned parameter. Calculating G for every shift along the channel produces a signal for the full recording and ASG segment boundaries are estimated from the peaks of this signal.

It was noticed that over-segmentation could occur due to too many low-amplitude peaks, making the algorithm computationally expensive. Modifications to this method were previously attempted in [84, 85, 101–103] to solve this issue, such as incorporating a static or adaptive peak threshold. In this application, over-segmentation was efficiently avoided by using a pair of pre-tuned thresholds that set a minimum allowable height, *MINPEAKHEIGHT*, and distance, *MINPEAKDISTANCE*, between successive peaks.

ASG was applied to each EEG channel and figure 5.2(b) illustrates the method on 100 seconds of recording.

5.3.3 Feature extraction & cluster-time profiles

After segmenting the EEG by ASG, clustering was performed to group similar segments together based on [101–103]. The characteristics of each group of segments were represented using a number of time-domain and frequency-domain features (as defined in section 4.2) from each segment:

- SD of the amplitudes
- Maximum and minimum difference of amplitudes
- Maximum absolute amplitudes of first derivatives
- Maximum absolute amplitudes of second derivatives
- Average frequency
- Relative bandpower in delta band
- Relative bandpower in theta band

- Relative bandpower in alpha band
- Relative bandpower in beta band

The segment features across all channels were clustered together using the K-means algorithm (as defined in appendix A.4) into *cluster_num* distinct clusters. K-means was performed with 20 repetitions to ensure a good initialization and convergence.

The mean SDs of each clustered group of segments were used to relabel the clusters in increasing order of SD, with each sample in a segment replaced by its corresponding cluster label and plotted over time [101, 142]. Although these were referred to as 'temporal profiles' in [101–103], the term 'cluster-time profiles' was chosen here as a more informative description.

A cluster-time profile for a two hour EEG recording is shown in figure 5.2(c) (for a single EEG channel).

5.3.4 QS classification

By representing the evolving patterns of the segments using cluster-time profiles, periods indicating large changes in segment behaviour (relatively higher discontinuity associated with QS) were reflected by larger fluctuations in the cluster labels. Classification of these periods as QS (and non-QS) was performed based on [101–103]:

1. The cluster-time profiles were averaged across channels to form a single average profile. This was to accentuate periods of large fluctuation while smoothing out channel-specific deviations, as illustrated in figure 5.2(d).

2. The resulting average profile was de-trended by subtracting the running (time-varying) mean of the profile, as shown in figure 5.2(e). This eliminated any natural underlying transients in the signal that may affect subsequent QS classification. The running mean was calculated using a Moving Average (MA) filter of pre-tuned length *avg_win_length* samples and the resulting profile squared to further accentuate the peak regions (as shown in figure 5.2(f)).

3. A longer MA filter of pre-tuned length *smooth_win_length* samples was then further applied to produce a smooth envelope curve. QS periods were estimated using a threshold calculated as the mean of the envelope curve, as shown in figure 5.2(g).

In cases when long recordings >2 hours were processed, envelopes were derived separately for each two-hour segment and then stitched together before applying the classification threshold. This is also illustrated in figure 5.2(g).

A minimum of three consecutive minutes or three out of four consecutive minutes of the same sleep state are clinically required to identify AS and QS [95, 154]. Based on this scoring criteria, QS detections <3 minutes were removed as an additional post-processing step. Figure 5.2(h) shows the output of CLASS with the estimated QS periods shaded. The corresponding visually labelled QS estimates are shown in figure 5.2(i).

5.4 CLASS parameter optimization

The following CLASS parameters required tuning across the different stages:

- ASR: *ASR_thresh* (section 5.3.1)
- ASG: *WIN*, *SHIFT*, *kF*, *MINPEAKHEIGHT*, and *MINPEAKDISTANCE* (section 5.3.2)
- Cluster-time profiles: *cluster_num* (section 5.3.3)
- QS classification: *avg_win_length* and *smooth_win_length* (section 5.3.4)

Some of these parameters were also defined in [101–103], but were selected empirically and it was unclear how this was accomplished.

To ensure that CLASS performed as optimally as possible, a more structured parameter optimization approach was required. An exhaustive grid search was not feasible due to the computational intensity, with certain combinations of parameters also causing the algorithm to become detrimentally slow and inefficient. Many optimization methods have attempted to find local minima based on a loss function (through approaches such as gradient descent), however, there existed no such

function for optimizing CLASS. Instead, the alternative stochastic approach of Random Optimization (RO) was adopted.

RO was first introduced by *J. Matyas* [223] in 1965 and attempts to find a local optima by fitting Gaussian distributions over each parameter, centred on the current value. With each iteration, new parameter values are sampled from the distributions, with the algorithm then re-run and performance re-assessed. If performance does not improve, a new set of parameter values are drawn in the next iteration and the procedure is repeated. However, if the performance does improve, the Gaussian distributions are re-centred on the updated parameters before new draws are made [223].

This approach essentially performed a 'sensitivity analysis' on the CLASS parameters. However, it was important to first select informative initial values to ensure the algorithm was already in a reasonable region of the parameter space. Informative selections for the SDs of the Gaussian distributions were also important, to achieve appropriately sized shifts across the parameter space. These initializations were based on the previous methods of *Krajca et al.* and related studies [84, 85, 101–103, 140], as well as underlying clinical knowledge of the EEG behaviour. In some cases, empirical investigations using the training set were performed by visualizing how perturbations (of differing sizes) in the particular parameters affected the associated stage of CLASS.

RO was then performed on the training data for 200 iterations and the final parameter values used to assess CLASS performance on the test set.

5.5 Assessing CLASS performance

Based on those ages exhibiting similar EEG behaviours, six groups were defined according to PMA [71, 74]: <31 weeks, 31-32 weeks, 33-34 weeks, 35-36 weeks, 37-38 weeks, and >38 weeks. The agreement between visual labelling and CLASS-estimated QS periods was assessed using the binary classification measures of sensitivity, specificity and F1 (as defined in section 4.6.2). F1 more strongly accounted for the imbalance in the data towards non-QS.

While these measures assessed agreement between signal samples, this did not specify exactly how many QS periods (events) were correctly detected or the precise number of misclassifications. Consequently, the eTPR and eFDR were also calculated, as also defined in section 4.6.2. A correctly detected QS event was defined if the visually labelled and CLASS-estimated event overlapped by $\geq$50%. Although CLASS classified QS with a mean threshold value, ROC curves and AUC (defined in section 4.6.2) were also calculated for each test recording by varying this threshold.

These measures were additionally calculated for CLASS without applying ASR, to assess the importance of artefact removal. Similarly, to further quantify the contribution of ASG, CLASS was run alternatively using a uniform segmentation of the EEG signal at both one and five second durations (the typical duration range of the ASG-derived segments). CLASS was also compared when using only SD as a feature to derive the cluster-time profiles.

To assess performance against other QS detection algorithms, CLASS was compared to the SAT% algorithm. For a fair comparison, key parameters of this algorithm were also optimized by performing a grid search with the training set, before assessment of the optimized algorithm on the test set. Full details of the SAT% algorithm and parameter optimization is found in appendix A.5.

Finally, *Ansari et al.* made a direct comparison of CLASS with their state-of-art deep neural network. The proposed method passed the EEG directly through an 18-layer Convolutional Neural Network (CNN), with the first 14 layers performing feature extraction and the final four layers performing sleep stage classification [100]. In this study, CLASS was implemented using the same parameters outlined here. However, both the CNN and CLASS used fixed classification thresholds and these were tuned to better compare the two models. For CLASS, this was achieved by adding a weighted SD of the classification envelope to the original mean threshold. This weighting factor was optimized during training, as described in [100].

Table 5.1: Summary of the CLASS parameters. Initial and final values before and after RO are shown (after 200 iterations), with the resulting mean F1-errors. SDs of the RO Gaussian distributions are also given.

Parameter	CLASS stage	Summary	RO SD	Initial value	Final value
ASR_thresh	ASR	Threshold for separating the artefact and artefact-free subspaces in the EEG	2	10	15
WIN (s)	ASG	Length of the contiguous windows for identifying adaptive segment boundaries	0.02	0.75	0.74
SHIFT (samples)	ASG	The step shift size of the sliding contiguous windows	2	10	8
kF	ASG	The weighting used to determine the joint contributions of the frequency and amplitude measures for defining adaptive segment boundaries	1	10	13.8
MINPEAKHEIGHT (μV)	ASG	Minimum height between peaks for defining an adaptive segment boundary	5	100	108
MINPEAKDISTANCE (samples)	ASG	Minimum allowable distance between successive adaptive segment boundaries.	2	25	29
cluster_num	cluster-time profiles	Number of clusters for grouping the features to define the cluster-time profiles	2	12	18
avg_win_length (samples)	QS classification	Window length of MA filter to determine the running mean of the cluster-time profile	25	500	539
smooth_win_length (samples)	QS classification	Window length of MA filter to smooth the cluster-time profile	250	35000	35272
Mean F1-error				0.3402	0.3050

5.6 Results

5.6.1 Parameter selection by RO

The initial parameter values and Gaussian SDs for RO is summarized in table 5.1. To assess performance after each RO iteration, the F1-error was calculated (defined as F1-error $= 1 - \text{F1}$) and the mean taken across the training set.

Figure 5.4 illustrates the transitions in the mean F1-error and the corresponding parameter values, with each RO iteration. Shifts between values across iterations are shown by each graph and the general behaviour of the fluctuations in parameter values reflects the stochastic nature of the RO searching process. Updated parameter values (and re-centring of the Gaussian distributions) are highlighted by red points and this occurred every time the mean F1-error reached a new minimum value

(at the blue points).

A total of eight parameter updates were made by RO. Updates happened more frequently during the first 30 iterations as a clear direction in the parameter space for improved performance was found. Updates then became less frequent (with smaller changes in error) as an optimal pathway became more difficult to find at random. Starting with mean F1-error=0.3402, final parameters settled by iteration 144 (as also highlighted in figure 5.4) which corresponded to a mean F1-error=0.3050 (an improvement of 10.3%). The final parameter values are listed in the last column of table 5.1.

5.6.2 CLASS performance with PMA

Figure 5.5(a) shows the overall ROC performance of CLASS on the test set for each recording (in grey) and the median ROC curve (in black), with the latter producing an AUC=0.9743. Figure 5.5(b) shows the sensitivity, specificity, F1, eTPR, and eFDR results for CLASS across the PMA groups, with error bars denoting the median and Inter-Quartile Range (IQR).

Across the range 31-38 weeks PMA, CLASS distinguished QS with excellent sensitivity (median range 0.94-1.00), specificity (0.79-0.94), F1 (0.74-0.89), eTPR (1.00-1.00), and eFDR (0.00-0.25). At 35-36 weeks PMA, eFDR was optimal (with median=0) indicating very few misclassifications. The generally lower values of F1 (compared to sensitivity and specificity) was expected as it typically penalized data with more negatives (non-QS periods) than positives (QS periods).

At >38 weeks PMA, while eTPR, sensitivity, and specificity remained high, eFDR was also comparatively higher than at younger ages (with median=0.40). This suggested that QS periods were well detected but misclassifications became more prevalent.

At <31 weeks PMA, CLASS performance was most dubious showing comparatively worse results for all measures. These recordings corresponded to the poorer ROC curves seen in figure 5.5(a), although these were limited to only a few.

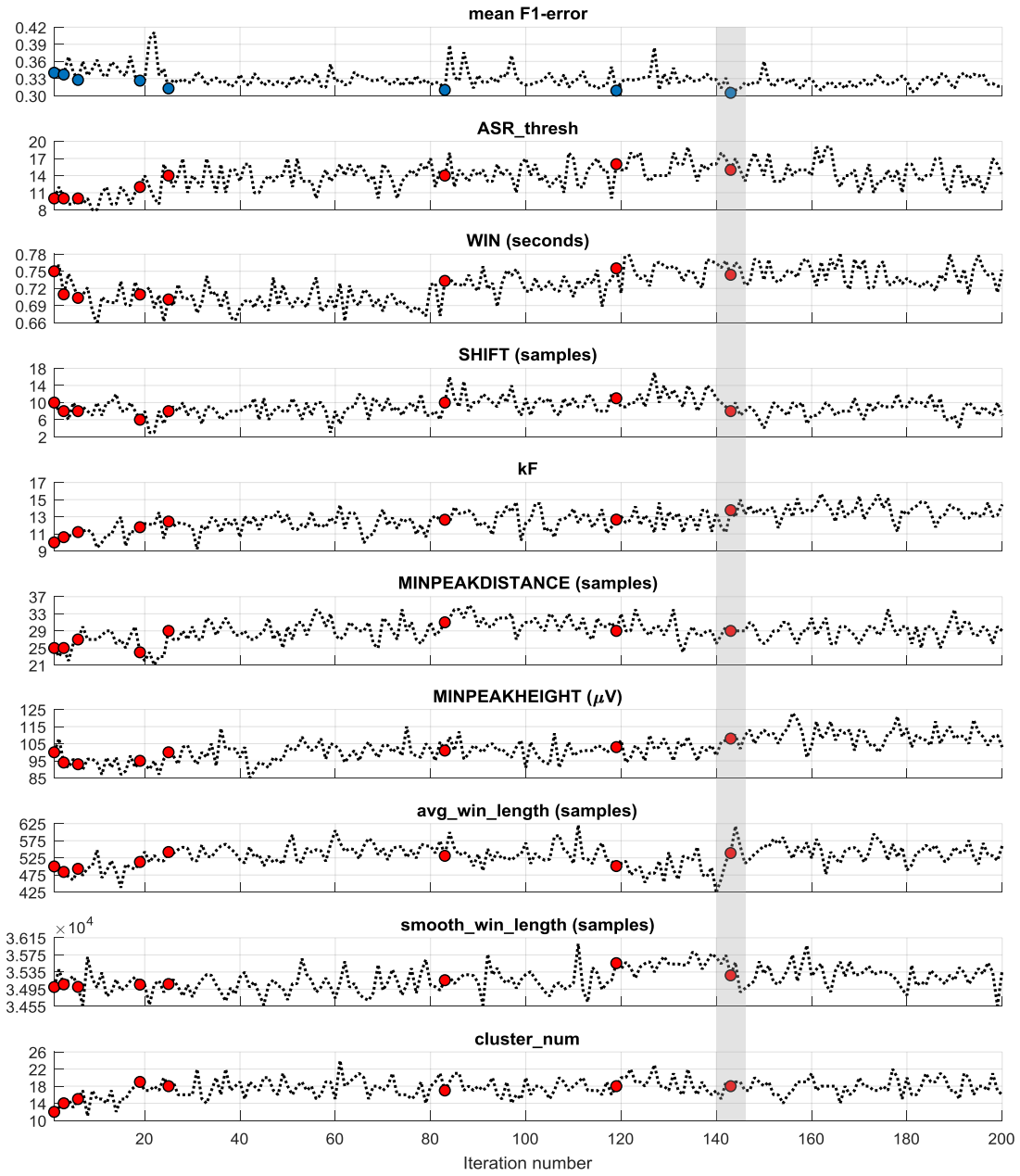


Figure 5.4: Illustration of the changes in mean F1-error and parameter values during RO, with 200 iterations, performed on the CLASS training set. Red points denote updated parameter values which improved CLASS performance and re-centred the RO Gaussian distributions for future iterations. These corresponded to a new minimum mean F1-error each time (shown by the blue points). The shaded region highlights the final mean F1-error and parameter values (at iteration 144).

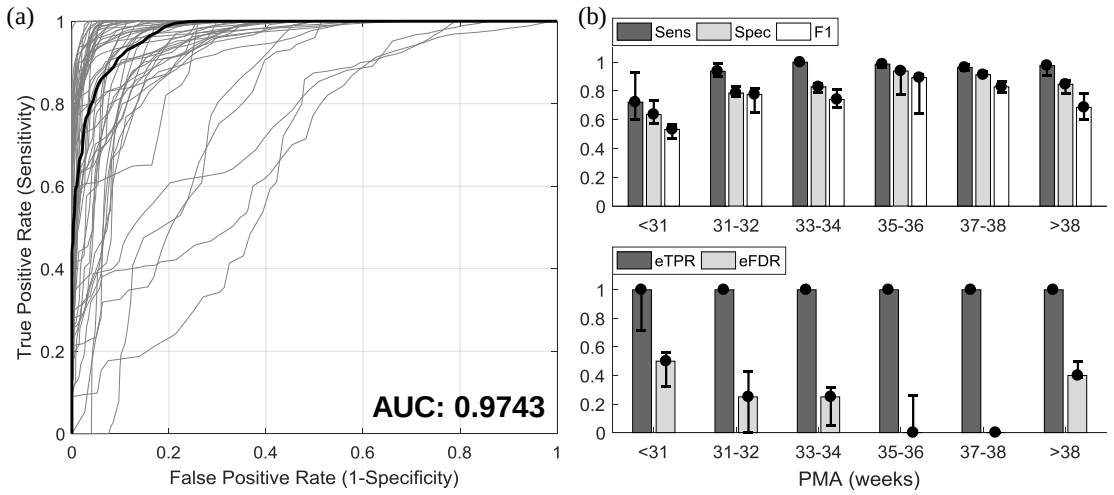


Figure 5.5: Assessing the performance of CLASS on the test set. (a) ROC curves of CLASS performance by varying the detection threshold. ROC curves for each recording in the test set are shown (in grey) and the resulting median ROC curve (in black). The AUC of the median ROC curve is also given. (b) CLASS performance with respect to PMA, denoting sensitivity (Sens), specificity (Spec), F1, eTPR, and eFDR. Error bars denote the medians and IQRs.

Overall, the results showed that CLASS performed best on EEG recordings at ages >31 weeks PMA.

5.6.3 Comparing CLASS algorithm stages

Table 5.2 lists the overall results across all recordings, comparing CLASS in its entirety to CLASS without ASR (denoted as CLASS-noASR), CLASS with uniform segmentation of one second (CLASS-USG1), CLASS with uniform segmentation of five seconds (CLASS-USG5), and CLASS with only SD as a feature (CLASS-SD).

The Wilcoxon signed-rank test (defined in section 4.6.4) was used to assess statistically significant differences between these model variants and CLASS in its entirety. This was performed across recording measures and those that showed significant differences at the $p<0.05$ level are also indicated in table 5.2.

When comparing CLASS to CLASS-noASR, all agreement measures showed statistically significant differences. Namely, CLASS-noASR had particularly poorer eFDR, F1, and AUC (medians of 0.44, 0.52 and 0.88, respectively), compared to CLASS (0.25, 0.76 and 0.97, respectively), although all measures were comparatively

Table 5.2: Comparing CLASS performance at different stages of the algorithm, using sensitivity (Sens), specificity (Spec), F1, eTPR, and eFDR. CLASS is compared to CLASS without ASR (CLASS-noASR), with uniform segmentation of one second (CLASS-USG1) and five seconds (CLASS-USG5), and using SD alone instead of multiple features (CLASS-SD). *denotes significant differences in the results when compared to CLASS in its entirety, with significance assessed at $p<0.05$ using the Wilcoxon signed-rank test. Medians are shown with IQRs in brackets.

Algorithm	Sens	Spec	F1	eTPR	eFDR	AUC
CLASS	0.97 (0.93-1.00)	0.82 (0.75-0.91)	0.76 (0.62-0.84)	1.00 (1.00-1.00)	0.25 (0.00-0.40)	0.97 (0.94-0.99)
CLASS-noASR	0.80* (0.66-0.96)	0.76* (0.66-0.85)	0.52* (0.39-0.72)	1.00* (0.67-1.00)	0.44* (0.25-0.65)	0.88* (0.75-0.97)
CLASS-USG1	1.00 (0.97-1.00)	0.79* (0.71-0.86)	0.72* (0.56-0.81)	1.00 (1.00-1.00)	0.25 (0.00-0.43)	0.98 (0.93-0.99)
CLASS-USG5	0.95* (0.88-0.98)	0.81 (0.73-0.91)	0.71* (0.58-0.82)	1.00* (1.00-1.00)	0.33* (0.04-0.50)	0.96* (0.91-0.98)
CLASS-SD	0.97 (0.94-1.00)	0.83 (0.76-0.91)	0.75 (0.63-0.85)	1.00 (1.00-1.00)	0.25 (0.00-0.40)	0.98 (0.95-0.99)
SAT%	0.77* (0.63-0.90)	0.76* (0.62-0.85)	0.53* (0.39-0.71)	1.00* (0.75-1.00)	0.67* (0.45-0.78)	0.85* (0.72-0.94)

worse. This showed the great importance of ASR in improving the quality and robustness of CLASS.

When comparing CLASS with CLASS-USG1, the eTPR, eFDR, and AUC produced very similar results, with no statistically significant differences. However, there were significant drops in F1 (0.72) and specificity (0.79), compared to CLASS (0.76 and 0.82, respectively). Additionally, the increased number of segments generated by CLASS-USG1 resulted in a drop in computational speed (timed after performance on the full test set) of ≈20%.

Increasing the length of uniform segmentation to five seconds, as in CLASS-USG5, produced an overall drop in performance. Apart from specificity, all other values had statistically significantly poorer results, particularly in F1 (0.71) and eFDR (0.33), when compared to CLASS (0.76 and 0.25, respectively).

Finally, performance comparison with CLASS and CLASS-SD revealed no statistically significant differences. In fact, CLASS-SD resulted in a marginally better AUC and specificity (0.98 and 0.83, respectively) compared to CLASS (0.97 and 0.82,

Table 5.3: Results taken from *Ansari et al.* [100]. Comparison of classification performance between CLASS and the CNN using AUC (taken from the median ROC curve), sensitivity (Sens), specificity (Spec), and Cohen's κ. Medians are shown with the IQRs in brackets (where appropriate).

Algorithm	AUC	Sens	Spec	κ
CLASS	0.92	0.72 (0.19)	0.97 (0.06)	0.75 (0.19)
CNN	0.92	0.88 (0.26)	0.93 (0.13)	0.74 (0.17)

respectively). However, F1 was also marginally lower (0.75) than with CLASS (0.76).

5.6.4 Comparing CLASS with SAT% and CNN

Comparison between CLASS and the optimized SAT% algorithm showed statistically significantly poorer results across all measures in the latter, as shown in table 5.2. Particularly, eFDR was much poorer (0.67) than CLASS (0.25). This indicated very large numbers of misclassifications by SAT% and confirmed that it is insufficient for accurately and robustly detecting QS.

Table 5.3 reproduces the results of *Ansari et al* [100]. In this study, results were presented slightly differently, with the AUC calculated once on the median ROC curve (and not individually per recording, as in table 5.2). As well as sensitivity and specificity, Cohen's κ was used (defined in section 4.6.2) as an additional measure of agreement between the algorithms and visual labelling.

Results from both algorithms were similar, with CLASS and the CNN producing exactly the same AUC performance of 0.92.

5.7 Discussion

From a physiological point of view, the performance of CLASS actually reflects the development of QS EEG behaviour in preterm infants, since it is based on the relative discontinuity caused mainly by the differences in EEG amplitude between QS and other states [71, 74]. The overall results confirm the ability of this automated algorithm to align with visual sleep labelling and identifies the best performance of the algorithm to classify QS at >31 weeks PMA.

At <31 weeks, there is still great uncertainty in classifying QS. As was detailed in chapter 2, a combination of cerebral and non-cerebral measures have been used for visual sleep labelling, since neither of these alone are considered a gold standard [19, 36, 74, 154, 224]. However, when all these criteria are required for state definition, increasingly premature infants have higher proportions of IS (as described in chapter 2). This indicates the immaturity of each cerebral and non-cerebral sleep measure to represent distinct sleep states in the very preterm baby [36] resulting in an increase of IS and lower levels of definite QS [18, 36, 77, 78]. The limitation remains true for visual labelling as well as for CLASS sleep classification.

With CLASS relying on discontinuity, IS can strongly resemble QS and be detected as such by the algorithm. Therefore, in preterm babies <31 weeks, CLASS more accurately captures the vigilance state cyclicity (which is made up of both QS and IS periods) rather than definite QS and should be interpreted as such [18, 32, 33, 77, 78].

As also described in chapter 2, in term ages >38 weeks PMA, the emergence of both HVS as well as TA QS patterns leads to a globally more continuous EEG and the relative changes in discontinuity between QS and non-QS become more subtle. CLASS still performs well here but the increased continuity further accentuates any remaining artefacts and explains the increase in misclassifications at this age.

Introduction of the artefact removal method, ASR, proves to remove a major proportion of artefacts. Of all the stages of CLASS, results point to the inclusion of ASR as contributing to the greatest improvement in performance. High power artefacts can skew the clustering by CLASS and 'appear' as discontinuous QS while also affecting the signal stitching; upon division into two hour segments, regions containing artefacts can be classified differently to regions that are relatively artefact-free. Envelope peak amplitudes may differ across stitched segments as a result, and the QS detection threshold becomes highly inaccurate.

With this said, the use of a single age calibration sequence might still be a limitation of ASR. In future applications of CLASS, the ability of ASR to

automatically and robustly select a sequence over PMA (and therefore adapt to age) may further improve the QS detection capabilities of CLASS at these extreme ages.

Implementing ASG as opposed to uniform segmentation also improves CLASS performance. The separation of the EEG into characteristic segments allows for a more structured and distinct clustering, unlike the selection of segments achieved by uniform segmentation. Such an arbitrary selection at fixed time points can lead to adjacent segments with elevated cluster labelling, resulting in increased misclassifications when compared to ASG. This is particularly evident when using longer five second segments.

Shorter segments of one second show less of a performance drop (though still slightly poorer) than with ASG but is more computationally expensive. The larger number of segments slows down the subsequent performance of the feature extraction and clustering steps of the algorithm. Overall, this indicates that ASG provides a more efficient and accurate segmentation of the EEG that will aid in the clinical usefulness of CLASS.

Using only SD as a feature for QS classification yields similar results to CLASS with multiple features. While SD is typically a very sensitive feature to high-power artefacts, with the inclusion of ASR in the preceding steps, SD proves sufficiently robust within this dataset to classify QS well. Therefore, this implementation of CLASS may be sufficient in most cases and is a more intuitive interpretation of the method. However, when dealing with very noisy recordings that may not be adequately filtered by ASR, SD may be more susceptible to artefacts than when using a combination of features for distinguishing QS and non-QS. Encountering noisy recordings is especially likely when assessing clinical outcomes from very large EEG datasets with limited pre-selection and prior screening.

Furthermore, by term age, the SD between recordings becomes less distinguishable across sleep states, with frequency-domain features (such as bandpowers) showing better separability in others studies [82, 83]. While SD may be sufficient

in this cohort limited to 42 weeks PMA, utilizing CLASS with multiple time-domain and frequency-domain features may hold it in better stead for classifying older babies at >42 weeks PMA.

Comparison of CLASS with SAT% further motivates the usefulness and novelty of CLASS in the clinical setting. SAT% depends on the detection of the suppressed periods of EEG (the IBIs) particularly in TD. However, as the preterm baby matures towards term age, the increase in IBI amplitude and the transition to continuous EEG makes the method only feasible at very young ages (as was first suggested in chapter 2). Therefore, while SAT% continues to show merit in other research, its use in explicit QS detection culminates in a poor performance across all ages.

CLASS also holds its own with a state-of-art deep-learning method. Although the CNN was developed subsequently, both achieve similar and very high performance results (noting that CLASS remains an unsupervised method unlike the CNN). This also confirms that, while it is difficult to obtain a global optimum for tuning the CLASS parameters (by RO), the chosen locally optimum values are acceptable and result in a robustly performing algorithm.

5.8 Conclusion

The key novel contributions outlined in this chapter are:

- Expansion of initial concepts by *Krajca et al.* to the first unsupervised method, CLASS, that automatically and robustly detects QS (and non-QS) periods from preterm EEG, over a wide PMA range (up to and including term age).

- An introduction to neonatal EEG analysis of the artefact-removal method, ASR, which aids the robustness of CLASS to artefacts in long duration EEG recordings. Most significantly, it is key to strengthening the direct practical applicability of CLASS for clinical care.

- Through the unsupervised nature of CLASS, the algorithm can uniquely predict sleep states for each recording, which reduces the effects of inter-patient variability without prior training on large datasets.

The CLASS algorithm also provides the required tool to perform sleep staging on the available unlabelled datasets in this research, including patients with abnormal neurodevelopmental outcome. Chapter 6 now focusses on incorporating this algorithm as part of a fully-automated approach to track brain maturation across PMA.

6

Deriving Developmental Trajectories Across Preterm Ages

Associated publication(s):

K. Pillay*, A. Dereymaeker* et al. "Deriving developmental trajectories from EEG to predict normal and abnormal brain maturation across preterm ages". Submitted to: *Sci. Rep.* (2018) **Joint first authors.*

6.1 Introduction

In chapter 5, the robust CLASS algorithm was developed for detecting QS from EEG, as a first step towards fully-automated approaches for quantifying brain maturation. More specifically, chapter 1 identified one approach as identifying appropriate biomarkers to track brain maturation over PMA.

Chapter 2 described previous examples of EMA/brain-age estimation as a solution to combine such biomarkers and derive developmental trajectories. However, these methods didn't fully utilize maturational information across multiple sleep states (albeit in a fully automated way) and focussed only on normal developmental outcome. This prompts the development of an expanded approach. Chapter 2 also suggested expansion to a data-driven framework that may uncover further outcome-sensitive biomarkers linked to abnormal dysmature and/or disorganized EEG patterns.

In this chapter, the CLASS algorithm is first used to automatically extract sleep states from the EEG. A large set of features is then extracted from these sleep states as candidate maturational biomarkers and incorporated within a new brain-age prediction model based on RF regression. This combined, fully automated procedure is then tested on both normal and abnormal neurodevelopmental outcome data to identify early normal and deviating trajectories.

6.2 Data

Dataset 1 was selected in this study, as defined in chapter 3, and was previously used to develop CLASS in chapter 5. Dataset 2 was additionally included (as also defined in chapter 3) to incorporate abnormal outcome data. Abnormal outcome data comprised the mild, severe, and lost-to-follow-up outcome groups.

As dataset 2 had no visual sleep labelling, CLASS was applied across both datasets 1 and 2 (for consistency) to extract periods of QS and non-QS. Due to the limitation of CLASS at ages <31 weeks PMA, these recordings were removed. This resulted in 73 recordings (26 patients) from dataset 1 while all recordings from dataset 2 were retained. From the estimates of CLASS, 33% of the EEG comprised QS, and 67% comprised non-QS.

Development of CLASS in chapter 5 successfully incorporated ASR to remove channel-specific, high power artefacts. A single calibration sequence was applied (from a term-age recording) to remove artefacts across a wide PMA range. This proved acceptable for automated QS and non-QS detection. However, adopting this method (with a fixed calibration sequence) for general EEG artefact removal, over a wide PMA range, may detrimentally affect the subtler EEG characteristics and impact the features extracted. Consequently, once CLASS was successfully applied (with ASR), features were extracted from the raw EEG.

6.3 Feature pre-processing

From each QS and non-QS period, the 113 features defined in section 4.2 were extracted from each channel and within equal-length epochs. Previous sleep staging

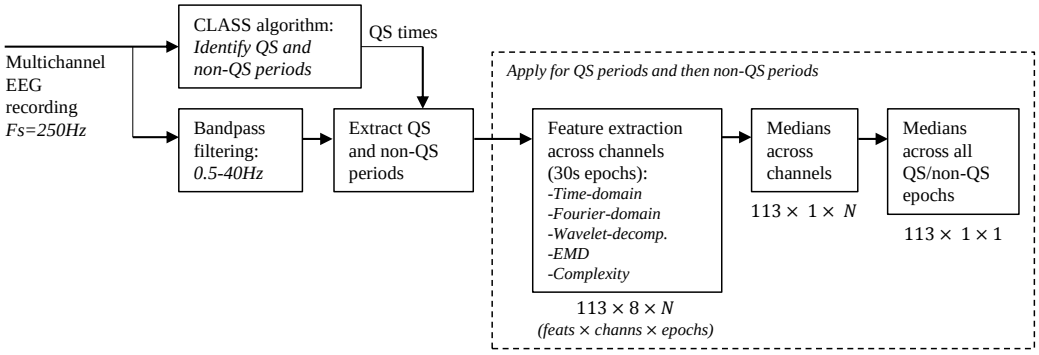


Figure 6.1: Pre-processing procedure for extracting a representative set of EEG features from CLASS-estimated QS and non-QS periods. Dimensions of the resulting outputs are provided at each stage (where applicable), based on the number of features (feats), channels (channs) and epochs in the recording (N).

studies in preterms used epoch lengths as long as 2.5 minutes [34], with certain features (such as the MSE) proving to be unreliable at window lengths as short as 10 seconds [29]. However, excessively long epoch lengths could miss brief state transitions, while lengths still needed to be sufficiently long to capture the burst-inter-burst behaviours observed during TD and TA (where suppression lengths are typically $\leq$6 seconds [41]). Therefore, a trade-off epoch length of 30 seconds was chosen. Median values were then taken across all channels and epochs belonging to each sleep state. Medians were chosen (as opposed to the mean) to limit the effects of channel-specific deviations [146].

The full procedure of feature extraction (including estimating QS and non-QS by CLASS) is summarized in figure 6.1. The result was a single set of 226 representative features per recording (113 from QS and 113 from non-QS).

6.4 Identifying maturational biomarkers

As a first step, a regression analysis was performed on dataset 1 to determine statistically significant correlations between each feature and PMA, and identify maturational biomarkers. Furthermore, by assessing correlations on features extracted from both QS and non-QS, these biomarkers could be quantified for multiple sleep states. The linear random intercepts model was used (defined in

section 4.3.2) to correct for inter-patient variability by grouping multiple recordings from the same patient together [146].

As a second step, these biomarkers were correlated with outcome using dataset 2. This involved extending the random intercepts model with an additional dummy variable (defined in 4.3.1) to represent normal, mild and severe/death groups (the latter combining the severe and lost-to-follow-up patients together). This identified particular outcome-sensitive biomarkers that could also capture subtler abnormal behaviours.

Each feature was also tested for interaction effects (defined in 4.3.1) with outcome and ignored if these proved significant at $p<0.05$. Testing for interactions avoided selecting features where different outcome groups exhibited separable intercepts but strongly overlapped (due to crossing trends).

6.5 The RF model for brain-age prediction

To construct a brain-age prediction model, an RF regression model was trained (as described in section 4.3.3) on dataset 1 before testing this model on dataset 2. As dataset 1 were all of normal outcome at follow-up, PMA provided a reasonable estimate of brain-age [145, 146].

RF regression was chosen as it is ideally suited to data where the number of features $\gg$ number of recordings. It also does not require feature standardization and intrinsically selects those features that are most predictive, without requiring an additional feature selection routine.

Typically, the choice of RF parameter (N_{trees}) is not very sensitive if set sufficiently high so a value of N_{trees}=1500 was chosen. Choosing the value of *mtry* affects the choice and variation of features used by the RF. As identifying abnormal deviations may be strongly dependent on the choice of features, models were trained using a wide range of *mtry* values in order to investigate this [225].

6.6 Model training & assessing performance

Before assessment on abnormal outcomes, the (non-linear) RF was first compared to a linear brain-age prediction method (such as that of *De Wel et al* [29], discussed in chapter 2). This approach applied LASSO feature selection (defined in section 4.4.1) followed by OLS linear regression (defined in section 4.3.1). The combined approach is referred to as LASSO+OLS. To improve the stability of LASSO, all features were initially excluded if there were no significant linear correlations with PMA at the $p<0.05$ level.

Both the RF and LASSO+OLS were trained and tested on dataset 1 using LOSO CV (defined in section 4.6.1). This was performed by removing all recordings belonging to the left-out patient (subject) each time. $\sqrt{MSE}$ (defined in section 4.6.3) was calculated as the performance metric (between the estimated brain-age and PMA) and evaluated per patient on each group of recordings.

Brain-age performance on both normal and abnormal outcomes was next assessed on dataset 2, after first training the RF on dataset 1. As shown in chapter 3, the maximum PMA of dataset 2 (48 weeks) extended beyond the maximum of dataset 1 (42 weeks). This could introduce large biases in the regression estimates >42 weeks so these recordings were removed from dataset 2 in this part of the analysis. The normal and mild outcome groups remained unchanged and three recordings were removed from the severe and lost-to-follow-up groups.

For each patient, the set of recordings in the plot of brain-age versus PMA formed the developmental trajectory [145, 146]. $\sqrt{MSE}$ was again used to compare the normal outcome and abnormal outcome trajectories in dataset 2. In addition, the 95% PI was derived by fitting a linear regression model to the normal outcome trajectories (as defined in section 4.3.1). This was used as a threshold to classify abnormal deviations (and patients) when any recording in the trajectory exceeded this value.

6.7 Results

6.7.1 Significant biomarkers for PMA

A total of 97 QS and 70 non-QS features showed statistically significant trends with PMA ($p<0.05$). Figure 6.2 shows the top twelve features for each sleep state ranked in descending order of R^2. To provide a better impression of the range of features that strongly evolved with PMA, those that were highly correlated with one another were removed based on a Spearman's rank correlation coefficient (defined in section 4.3.1) of $\rho>0.9$. The complete list of statistically significant features for QS and non-QS can be found in tables B.1 and B.2, respectively, in appendix B.1.

For QS, the most significant feature was 'MSE avg. lower slope (scales 1-5)'. The mean absolute ratios of IMF5/IMF4, IMF2/IMF1 and IMF6/IMF5 also proved strongly significant, as well as the average frequencies in the alpha, delta and theta bands. Across the top twelve features, R^2 range was 0.67-0.87.

For non-QS, the most significant feature was 'Line Length Burst%', showing the highest R^2 across both sleep states (0.91). The mean absolute IMF ratios proved dominant once more, along with the variation coefficients for IMF1 and IMF3. Hilbert-derived features were also prominent. Across these top twelve non-QS features, R^2 range was 0.52-0.91 and the minimum value dropped below that of the top twelve QS features.

6.7.2 Significant biomarkers for PMA & outcome

Figure 6.3(a) shows the top twelve of 13 features that showed significant differences between normal and mild outcomes (intercepts at $p<0.05$), while figure 6.3(b) lists the top twelve of a larger set of 42 features that showed significant differences between normal and severe/death outcomes. Data for different outcome groups are plotted along with their (fixed-effects) regression lines. Table B.3 (in appendix B.1) provides the full list of outcome-sensitive features.

Eight out of twelve of the top features separating normal and mild outcome originated from QS, including the amplitude features of '95% peak width' and 'zero crossing rate SD' as well as the more complex features of 'Line Length Burst%',

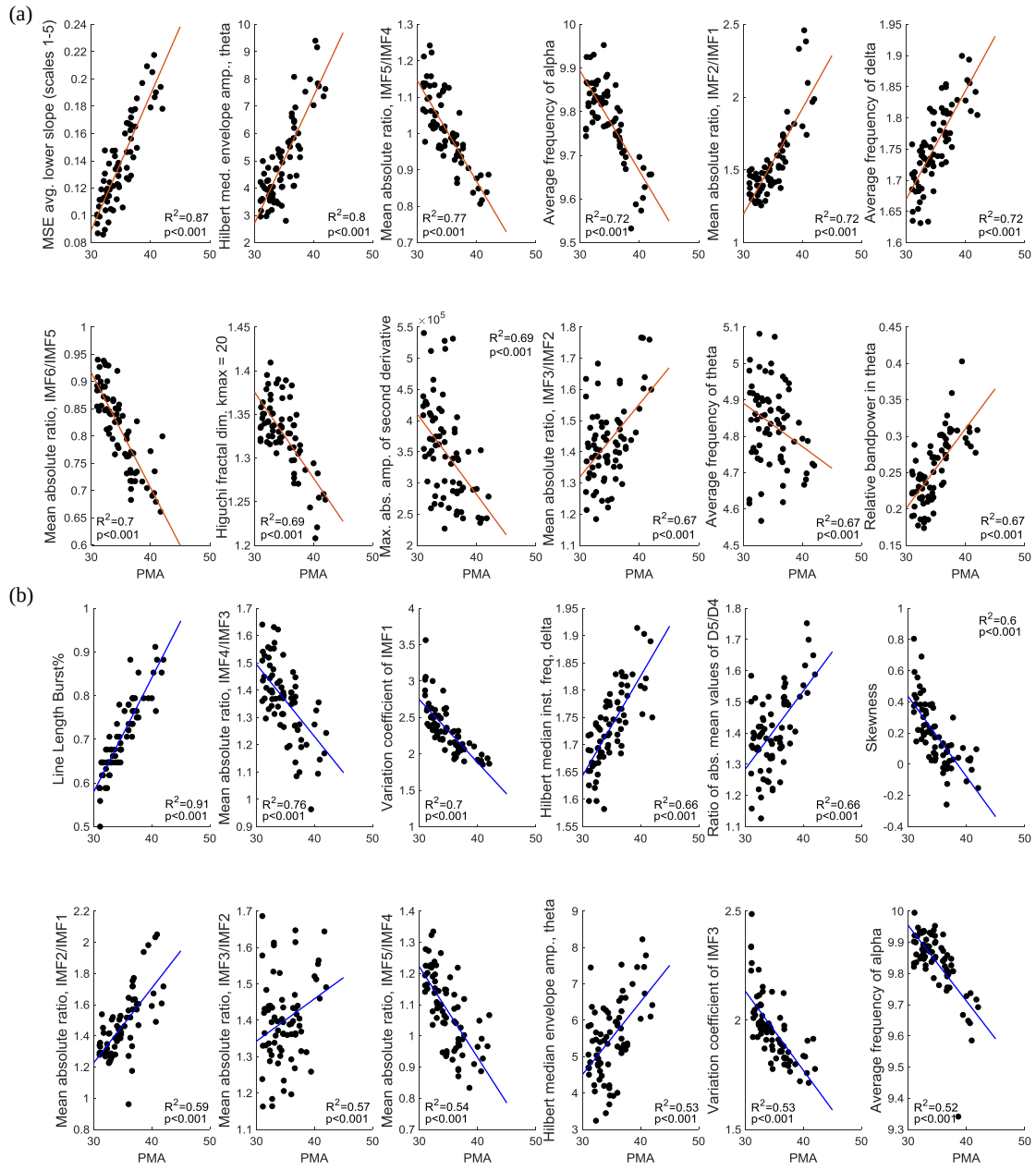


Figure 6.2: The top twelve statistically significant features (p<0.05) showing a linear correlation with PMA. Random intercept model were used to correct for inter-patient variability. The p-values and R^2 are also shown. (a) QS features. (b) non-QS features.

'MSE (scale 1)', and Higuchi fractal dimensions. 'Variation coefficient of IMF1' proved significant for both QS and non-QS.

When separating normal and severe/death outcomes, seven out of twelve of the top features originated from non-QS, with a strong representation of wavelet-based features. 'Variation coefficient of IMF1' was again significant while 'Hjorth activity' and 'SD of the rEEG' appeared strongly significant for QS. In addition, data from severe/death outcome had, visually, a noticeably greater spread in the feature distribution when compared to the mild and normal outcome trends in figure 6.3(a).

6.7.3 Comparing RF and LASSO+OLS

Figure 6.4 shows the performance of the RF and the LASSO+OLS algorithm after LOSO CV on dataset 1. Figure 6.4(a) shows the median $\sqrt{MSE}$ performance against *mtry*, with the shaded region denoting the IQRs across patients. Figure 6.4(b) shows the same plot for LASSO+OLS when varying λ (as defined in section 4.4.1) in the range zero (when all features were selected) to 1.6 (the maximum value before no features were selected). λ is plotted in reverse order to show the change in $\sqrt{MSE}$ as the number of selected features *increases*.

RF performance remained smooth, when varying *mtry*, and the highest error was seen for *mtry*<20 but then gradually improved. The error stabilized from *mtry*>20 at median $\sqrt{MSE}$=0.85. With LASSO+OLS, $\sqrt{MSE}$ was initially high (with values largely >1) due to a limited number of features selected for training. This then fell to a global minimum of $\sqrt{MSE} = 0.84$ at $\lambda = 0.25$ as a larger number of features were incorporated. However, once λ<0.25, $\sqrt{MSE}$ quickly increased due to overfitting.

LASSO+OLS provided a marginally better performance than RF, however, this was over a very short range of λ and the difference in $\sqrt{MSE}$ between the two models was small. The RF, on the other hand, performed consistently well as *mtry* was increased [226] suggesting a robustness to overfitting. This provided the necessary flexibility for assessing how changing the features in the model could impact brain-age prediction across both normal and abnormal outcomes.

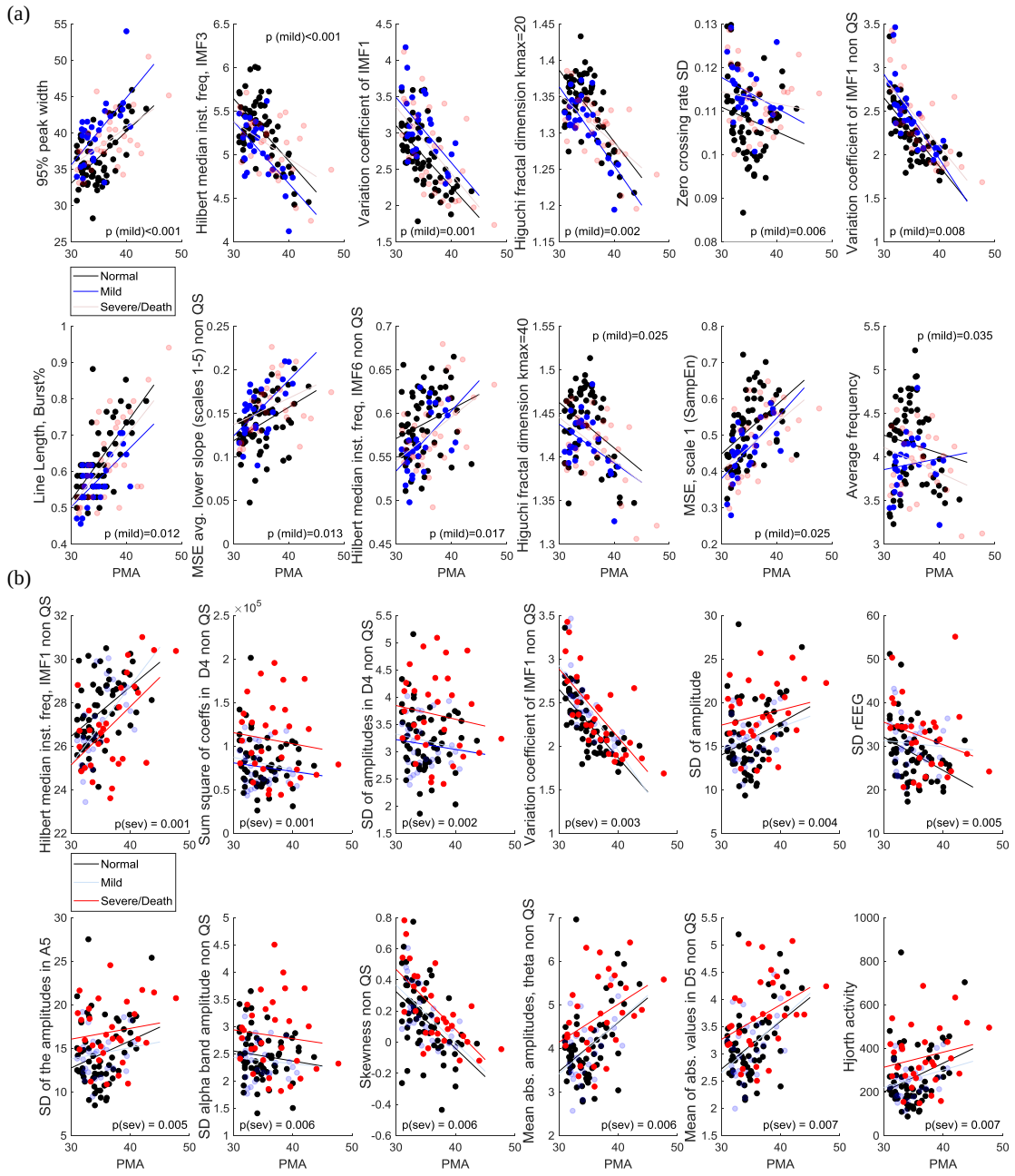


Figure 6.3: (a) The top twelve statistically significant features (intercepts at p<0.05) between normal and mild outcomes, using random intercepts models with a dummy variable for outcome. (Fixed effects) regression lines are also plotted for each outcome group. p (mild) denotes the p-value for the difference in intercepts between normal and mild outcomes. (b) The top twelve features showing statistically significant differences between normal and severe/death outcomes. p (sev) denotes the p-value for the difference in intercepts. In both cases, the non-significant outcome group is also plotted but faded for clarity.

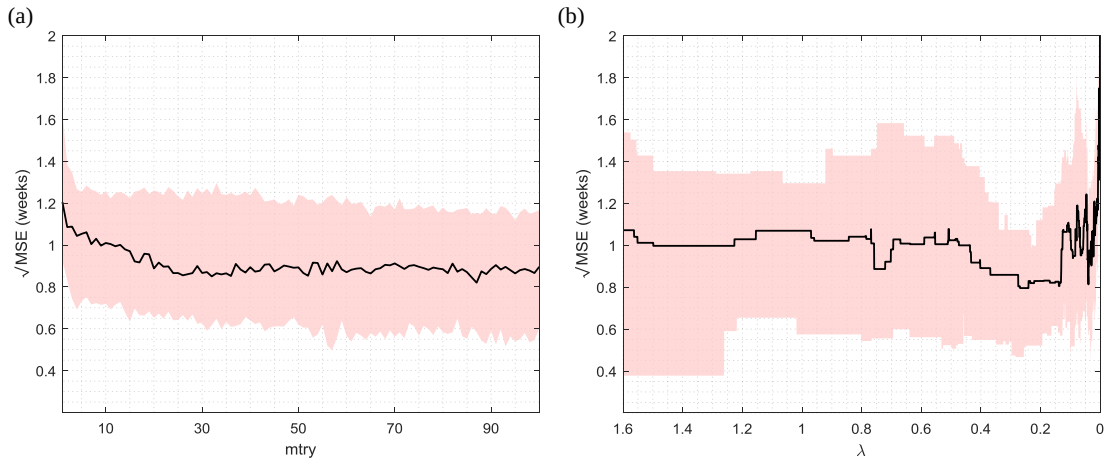


Figure 6.4: Comparison of the (non-linear) RF and (linear) LASSO+OLS regression models for brain-age prediction, using LOSO CV on dataset 1. (a) RF model performance assessed using $\sqrt{MSE}$ and varying *mtry*. (b) Corresponding performance for the LASSO+OLS model when varying λ (as defined in section 4.4.1). Medians are plotted, with shaded regions denoting the IQRs.

6.7.4 Assessing outcome across *mtry*

Figure 6.5(a) shows the $\sqrt{MSE}$ performance plots for dataset 2 after training the RF model on dataset 1. Median $\sqrt{MSE}$ for the normal and abnormal trajectories (combining mild, severe and lost-to-follow-up outcomes in the latter) are shown separately, with shaded regions denoting the IQRs across patients. After a period of variability, normal outcome error settled at a high *mtry* with median $\sqrt{MSE} = 0.8$. Abnormal outcome produced a consistently higher error than normal outcome across *mtry*. While normal outcome error remained consistent, abnormal outcome error continued to increase.

Figure 6.5(b) shows the patient classification accuracy for normal and abnormal outcome. After identifying deviating trajectories that exceeded the 95% PI of the normal outcome recordings, accuracy was measured as the percentage of patients that were correctly classified in each group.

Classification plots show some small erratic behaviour in general due to borderline classifications at the 95% PI boundaries. However, normal outcome classification generally remained above 81% (17/21 subjects) and peaked at 90% (19/21 subjects) for *mtry*>130.

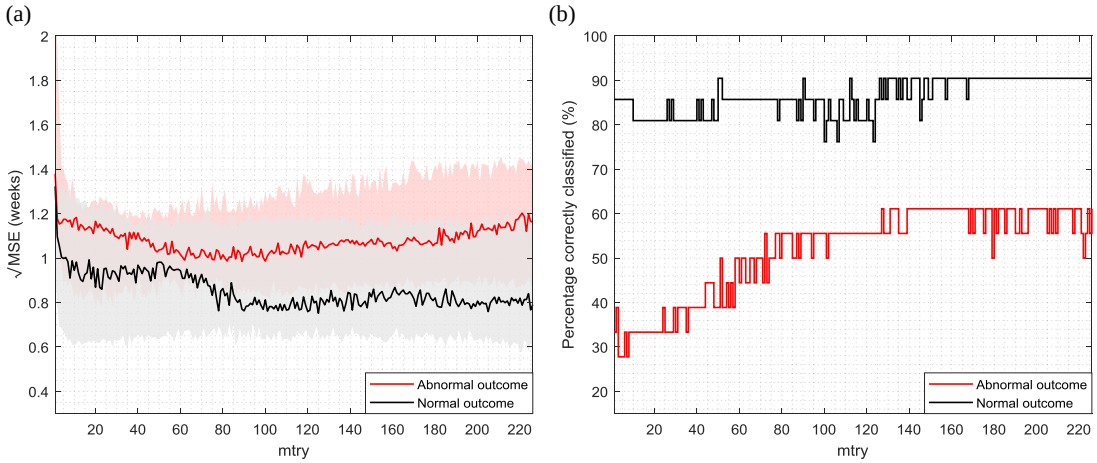


Figure 6.5: RF performance on dataset 2 (after training on dataset 1). (a) Trajectory $\sqrt{MSE}$ performance over *mtry* for normal and abnormal outcomes (combining mild, severe and lost-to-follow-up outcomes in the latter). Median values are shown with shaded regions denoting the IQRs. (b) Patient classification accuracy for normal and abnormal outcomes over *mtry*. Classification is made by identifying deviating trajectories that exceed the 95% PI defined from the normal outcome recordings. Accuracy is then measured as the percentage of patients that were correctly classified in each group.

Abnormal outcome classification started poorly at low *mtry*. However, there was a clear improvement as, *mtry* increased, peaking at a value of 61% (11/18 subjects) for *mtry*>130.

6.7.5 Developmental trajectories and *mtry*

Figure 6.6 provides examples of the trajectory plots for *mtry*=20 (figure 6.6(a)) and *mtry*=200 (figure 6.6(b)). Brain-age trajectories are plotted for each patient against PMA and separated for the normal, mild, severe and lost-to-follow-up outcomes. Shaded regions denote the normal outcome 95% PI and all trajectories that deviated at any point outside this interval are highlighted in red. The RF regression model showed a systematic bias towards younger ages in the overall trend, but this was accounted for by comparing to the 95% PI.

There was a clear increase in the number of deviant trajectories in the abnormal outcome groups, for *mtry*=200. To understand which outcome-sensitive features most contributed to the improved trajectory predictions at higher *mtry*, the top ten most commonly selected features by the RF model, that also previously proved significant for outcome (in section 6.7.2 and listed in table B.3), are shown in

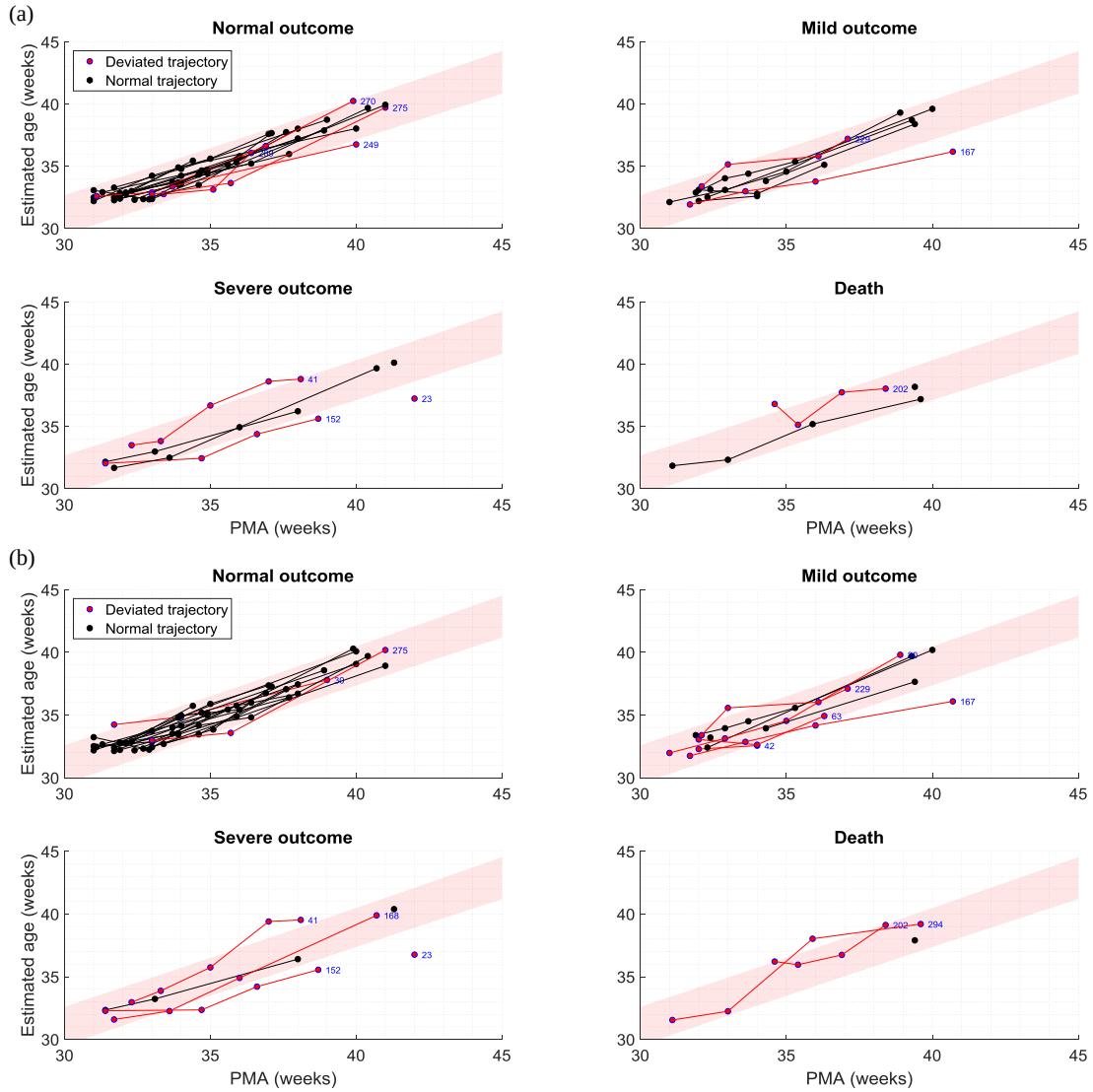


Figure 6.6: RF-derived trajectories of brain-age against PMA in dataset 2, separated for normal, mild, severe, and lost-to-follow-up (death) outcomes. Shaded regions denote the 95% PIs, based on the estimates of the normal outcome recordings. Any trajectories deviating from this threshold are deemed abnormal and highlighted in red. The (anonymous) patient IDs for these trajectories are also given in blue. (a) Trajectory plots for $mtry$=20. (b) $mtry$=200.

table 6.1. Feature importance in the RF was determined from the mean SSE decrease (defined in section 4.3.3).

At $mtry = 200$, across the top ten outcome-sensitive features, the top three features of 'Line Length Burst%', 'variation coefficient of IMF 1', and 'skewness' (all from non-QS) were most commonly selected by the model.

Table 6.1: The top ten most commonly selected features by RF for $mtry = 200$, which also showed statistically significant differences between normal and abnormal outcomes. The RF feature importance criterion, mean decrease in SSE, is also given.

No.	Feature	mean SSE decrease
1	Line Length, Burst% non QS	170
2	Variation coefficient of IMF1 non QS	44
3	Skewness non QS	21
4	Skewness rEEG	3.1
5	Variation coefficient of IMF1	2.2
6	Variation coefficient of IMF4	2.2
7	MSE avg. lower slope non QS	1.8
8	IQR rEEG non QS	0.99
9	LL suppression value non QS	0.97
10	Higuchi Fractal Dimension kmax=20	0.95

6.8 Discussion

6.8.1 Linking biomarkers and EEG characteristics

When considering individual feature correlations with PMA, the large number of significant features across both QS and non-QS suggests that the changing characteristics of EEG sleep states with age (as described in chapter 2) can be well quantified using this feature set.

For QS, 'MSE avg. lower slope (scales 1-5)' captures the changes in complexity and regularity within this sleep state. In cases of TD behaviour, regular burst-inter-burst behaviours result in lower MSE with this value increasing as the signal becomes less predictable and more complex during TA (and eventually HVS) [29].

The mean absolute IMF ratios provide a representation of the changing relative intensities of the EEG frequency bands. As summarized in table 4.1, lower IMFs correlate with the more oscillatory, higher frequency EEG components. The inverse is true for higher IMF values which capture low-frequency oscillations [157]. Consequently, the novel findings of strong positive correlations in the mean absolute ratios of IMF2/IMF1 and IMF3/IMF2 suggest a shift away from the highest oscillatory components. This is due to the very strong high amplitude burst periods of TD reducing as the EEG transitions to TA and HVS. Similarly, strong

negative correlations in IMF6/IMF5 and IMF5/IMF4 indicate a shift away from low frequency fluctuations as the amplitudes of the IBIs begin to increase with PMA [41].

Chapter 2 focussed on the clinically-defined shifts between the delta, theta, alpha and beta frequency bands across PMA and EEG states. Specific feature trends observed here further quantify shifts in activity *within* these defined bands. Increases in 'average frequency of delta' suggests a general transition away from low-frequency (slow) delta waves (as the baby matures) which is also clinically supported [22]. Decreases in the average frequencies of the alpha and theta bands further reflect an overall shift towards lower alpha and theta frequencies [17, 22]. This is partly due to the propensity of delta brushes at older ages and the disappearance in temporal sawtooths >34 weeks PMA.

A different representation of features is observed that reflect non-QS maturation. 'Line Length Burst%' provides the strongest positive correlation and even performs better than the corresponding values for QS (which were previously used to first demonstrate the efficacy of this feature in [26]). This reflects the strong transitions from burst-inter-burst non-QS characteristics to continuous LVI and ASI behaviours (from term-age). Furthermore, the reduction (and eventual disappearance) of IBIs results in a higher (mis-)classification of burst events [71]. The strength of this feature in non-QS is a new discovery and exemplifies the importance of identifying maturational biomarkers beyond only QS.

The 'Hilbert median instantaneous frequencies, delta' measures the characteristic frequency of delta as it varies with time, with the positive correlation reflecting the disappearance of slow delta, as in the case of QS [22]. Finally, the positive correlation of 'Hilbert median envelope amplitudes, theta' reflects an overall increase in theta amplitudes, possibly due to a previously defined 'rhythmic theta activity' that is more observable as the baby approaches term-age [178]. This could also be due to the maturation of TD burst-inter-burst behaviours into TA. The great significance of these Hilbert-based features is a novel finding and may account for a greater level of non-stationarity that may be present in non-QS EEG.

The larger number of features showing additional statistical differences between normal and severe/death outcome (when compared to mild outcome) is likely due to combinations of strong dysmature and disorganised EEG behaviours. This is also reflected by greater sparsity in the feature values. On the other hand, features for both normal and mild outcome groups generally present more organized and tighter trends with PMA while still remaining significantly different. Specifically, drops in 'Line Length Burst%', the Higuchi fractal dimensions, and 'MSE scale 1' indicate the presence of unusually long suppression periods (and fewer bursts).

Lower levels of complexity in abnormal patients (as reflected by 'MSE scale 1') also provides further support for the findings of *Lu et al.*, who discovered a reduced complexity at term-ages in babies who suffered from epilepsy later in life [167].

Finally, general higher-than-normal discontinuity is flagged by larger values of the 'variation coefficient of IMF1' (from QS and non-QS) and is indicative of greater fluctuations in the highest EEG frequency bands. This is also strongly reflected in non-QS by the wavelet-based 'SD of amplitudes in D4', with the D4 band approximately corresponding to the alpha-beta frequency ranges (as shown in table 4.1). Such unusually high levels of discontinuity is also associated with delayed dysmaturity [227, 228].

In general, the propensity of features derived from non-stationary decompositions express the benefit of accounting for the non-stationary nature of the EEG (as opposed to DFT-based features). This is particularly true when long epoch lengths are considered and the signal can no longer be approximated as stationary (typically at >1 second).

6.8.2 Linking developmental trajectories and outcome

When combining these features for brain-age prediction, the stability of RF regression as *mtry* is changed (unlike LASSO+OLS) is due to its unique architecture. As described in more detail in chapter 4, increasing *mtry* increases the size of the random feature subset for each tree node from which the best feature is chosen for that split. Achieving a stable $\sqrt{MSE}$=0.85 is an improvement on existing

methods [29, 145, 146] (of minimum error 1.49 weeks, as discussed in chapter 2) due to the larger size and variability of the feature set (although these datasets and assessment procedures vary across studies).

Increasing *mtry* reduces the error and improves the classification of abnormal outcome patients, while maintaining a consistently excellent performance for normal outcome patients. At low *mtry*, a fewer number of features are selected at each node and this increases the range of possible features represented across the forest. All these features exhibit strong correlations with PMA and are indicative of strong dysmature (and possibly some disorganized) patterns, due to disturbances in normal cortical development. This is exemplified when visualizing brain-age trajectories at *mtry*=20, with a few subjects in the mild and severe outcome groups exhibiting consistently deviant trajectories across PMA. This said, the larger variation in selected features at low *mtry* limits identification of subtler abnormalities. Many abnormal patients, particularly those with mild outcome, show no deviations at all. This is a similar problem with previous brain-age models which do not consider the sensitivity of each feature to outcome.

Assessing brain-age trajectories at *mtry*=200 shows why considering feature sensitivity to outcome becomes so important. At high *mtry*, the majority of features are presented to each tree node increasing the chance of selecting the same feature multiple times. This concentrates the model on a smaller subset of features and has the intriguing effect of improving the identification of abnormal patients. Additional abnormal trajectories (particularly mild outcome) are now largely identified by recording-specific deviations at <37 weeks PMA. This is particularly suggestive of an improved sensitivity to abnormal changes in discontinuity.

Quantifying the importance of these outcome-sensitive features to the RF reveal a particularly high prevalence of 'Line Length Burst% (non QS)'. It surfaces as a critical maturational biomarker that provides a trade-off between strong brain-age prediction performance and an improved sensitivity to outcome.

6.9 Conclusion

The key novel contributions outlined in this chapter are:

- The identification of novel features, largely from non-stationary EEG decompositions, as maturational biomarkers sensitive to both PMA *and* neurodevelopmental outcome. This is achieved using datasets of patients with both normal and abnormal outcomes.
- The incorporation of maturational biomarkers that capture the rich information found across multiple sleep states (QS and non-QS).
- The application of a (non-linear) RF regression model to combine this data-driven set of biomarkers. It successfully estimates brain-age trajectories and predicts developmental deviations (due to dysmature and disorganized EEG patterns). This is indicative of later abnormal neurodevelopmental outcomes (ranging from mild to severe).
- Through the use of CLASS for automated sleep staging, followed by brain-age prediction by RF regression, this is the first fully-automated approach that performs across a wide PMA range (to term age).

This method successfully demonstrates one of the possible approaches for quantifying brain maturation. Focus now shifts to the alternative approach of assessing sleep-wake cycling. Specifically, chapter 7 begins by presenting automated methods for classifying the differentiating EEG sleep states that develop near term age.

7

Multi-state Sleep Classification at Term-age

Associated publication(s):

K. Pillay et al. "Automated EEG sleep staging in the term-age baby using a Generative Modelling approach". In: *J. Neural Eng.* 15.3 (2018), p. 036004.

7.1 Introduction

Chapter 6 presented a successful approach for quantifying brain maturation, by identifying specific outcome-sensitive maturational biomarkers (across sleep states) and deriving developmental trajectories that could identify babies with normal and abnormal neurodevelopmental outcomes.

Chapter 1 also identified the assessment of sleep-wake cycling, as an alternative approach for quantifying brain maturation. Chapter 2 particularly highlighted the differentiation of EEG states by term age as a critical checkpoint to assess maturation, with the sleep-wake cycling continuing to change as the baby matures. Classification of these more intricate states from EEG could provide a means to automate and quantify this maturation which existing approaches (namely BAC measures) cannot fully capture.

Existing automated sleep staging methods in chapter 2 revealed none that have managed to classify these differentiating states, let alone from only EEG. The most similar approach was that of *Gerla et al.* [96] who applied an HMM to classify

QS, AS, wakefulness, and movement, using multiple PSG features. The HMM showed superior performance due to its ability to account for state transitions [96]. Chapter 4 further highlighted benefits of such probabilistic methods for modelling uncertainty which becomes increasingly important with limited data.

In this chapter, a traditional HMM-GMM classifier is proposed to perform automated sleep staging of the multiple differentiated states in term-age EEG, while also modelling state transitions. These models are trained using a data-driven approach for calculating the EEG features, to uncover meaningful sleep state biomarkers (with the approach also proving successful in chapter 6). To quantify the importance of modelling state transitions, HMM-GMM classification is also compared to classification with a state-dependent GMM (which does not model state transitions).

7.2 Data

The entirety of dataset 3 was used (as defined in chapter 3). With a limited set of visually labelled recordings, particular care was taken to select only clean EEG such that data quality remained high. Dubious periods where the clinicians didn't agree, as well as labelled artefacts and IS, were removed from further analysis. In total, IS contributed 0.9% of the total EEG, dubious periods contributed 15%, and artefacts contributed 40%.

As ASII and wakefulness both consisted of an LVI pattern and were difficult to distinguish from EEG alone, these were merged into a single LVI state, with the final four sleep states for classification defined as ASI (comprising 23% of the remaining EEG), HVS (15%), TA (33%) and LVI (29%).

7.3 Feature pre-processing

From each EEG recording, the 113 different features defined in section 4.2 were extracted from each channel and epoch. As in chapter 6, a trade-off epoch length of 30 seconds was chosen (described in section 6.3). In addition, an overlap of five seconds was introduced (to account for potential sleep state transitions at the

epoch boundaries). Also as in chapter 6, median features across channels were then taken to minimize the effects of channel-specific deviations.

Due to the aforementioned limitations of incorporating ASR for artefact removal (as discussed in section 6.2 of chapter 6), and the prior manual removal of such artefacts, feature extraction was performed on the raw EEG.

7.3.1 Patient-wise standardization

When assessing the median feature distributions across states and between recordings/patients, the scale and range of the data varied noticeably from one patient to the next. Figure 7.1(a) illustrates this variability between four patient recordings for the feature 'SD of theta band amplitudes' across the four EEG sleep states. To partially correct for this inter-patient variability (due to differing electrode placements and baseline changes), methods of patient-wise feature standardization have been proposed to shift the extracted features from different patients to the same scale [229]. This was applied here by standardizing the features from each recording, by subtracting the mean and dividing by the SD.

Figure 7.1(b) shows the same feature and recordings as in figure 7.1(a) after patient-wise standardization, demonstrating a reduced variability across recordings.

7.3.2 Feature selection

Feature selection was particularly important when fitting GMM distributions, as excessive features could prevent convergence or an optimal fit (due to the 'curse of dimensionality'). Moreover, with a large feature set calculated, high correlations between some features were inevitable. While the covariance of each Gaussian distribution could account for some correlation, it could still be problematic if too high.

The method of mRMR-Spearman (defined in section 4.4.2) was chosen here, maximizing the correlation between each feature and sleep state (from training data) while also reducing redundancies between highly correlated features.

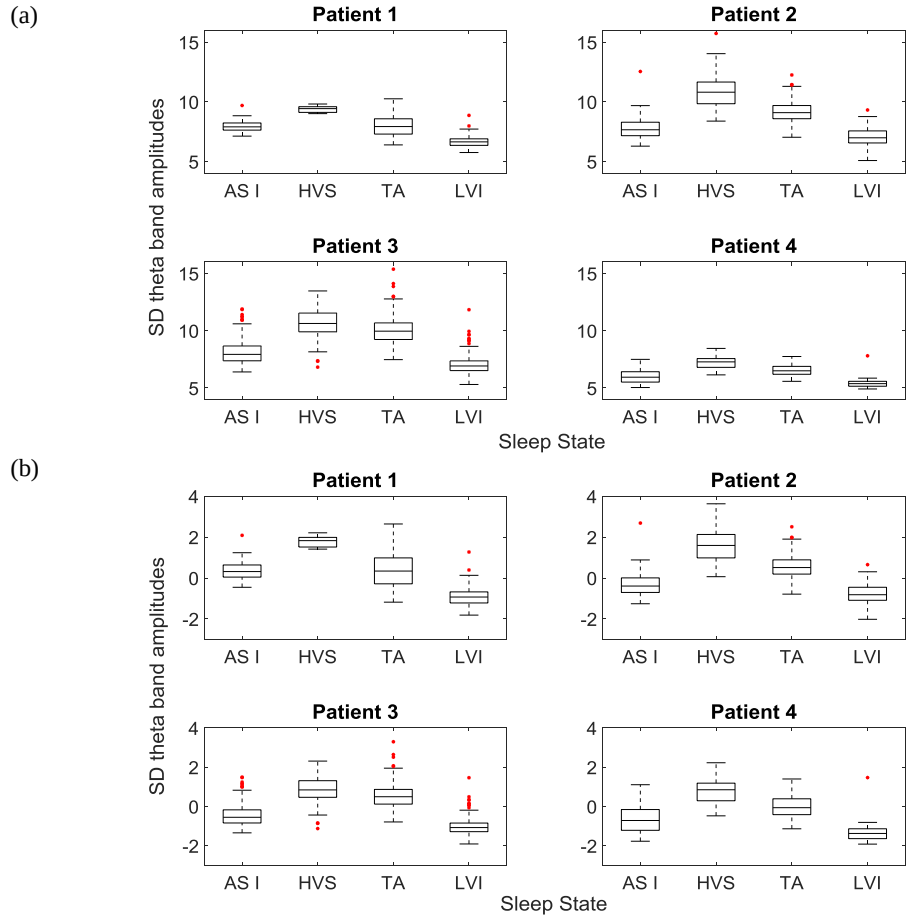


Figure 7.1: (a) Raw distributions of the feature 'SD of theta band amplitudes' for four patients/recordings across the four sleep states of ASI, HVS, TA and LVI. (b) The same distributions after patient-wise standardization.

7.4 Probabilistic sleep classification

To define the probabilistic classification models in this section, each i^{th} epoch of EEG features is denoted by $\boldsymbol{x}_i \in \mathbb{R}^P$ (where P is the number of selected features after mRMR-Spearman). Each $\boldsymbol{x}_i$ is conditioned by a sleep state, s_i (with S possible states).

7.4.1 Classification using the state-dependent GMM

The state-dependent GMM consists of a set of conventional GMMs (defined in section 4.5.2), one for each of the possible sleep states. GMMs were chosen as they could generalize arbitrarily complex and multi-modal distributions, especially in cases when features were not normally-distributed. Using the graphical model

notation defined in section 4.5.1, the resulting model is defined as:

$$\boldsymbol{x}_i|s_i \sim \sum_{k=1}^{K_{s_i}} \phi_{s_i,k} \cdot \mathcal{N}_{s_i,k}(\boldsymbol{\mu}_{s_i,k}, \boldsymbol{\Sigma}_{s_i,k}) \tag{7.1}$$

$$s_i \sim \text{Cat}(\boldsymbol{\xi}) \tag{7.2}$$

This encodes a likelihood, $p(\boldsymbol{x}_i|s_i)$, with each $\boldsymbol{x}_i$ drawn from a state-dependent GMM (with parameters defined in section 4.5.2). K_{s_i} denotes the number of Gaussian components for each sleep-state GMM and is allowed to vary with s_i. Due to the added dependency of each GMM on sleep-state, the sleep-state prior, $p(s_i)$, is also introduced and parametrized by the vector, $\boldsymbol{\xi}$ (of length S).

To perform inference on the sleep state, the posterior, $p(s_i|\boldsymbol{x}_i)$, is derived using Bayes' rule (defined in section 4.5.1):

$$p(s_i|\boldsymbol{x}_i) = \frac{p(\boldsymbol{x}_i|s_i)p(s_i)}{\sum_{s_i} p(\boldsymbol{x}_i|s_i)p(s_i)} \tag{7.3}$$

Finally, an estimate of the sleep state, $\hat{s_i}$, is determined by the maximum posterior (*maximum a posteriori*):

$$\hat{s_i} = \arg\max_{s_i} \left[p(s_i|\boldsymbol{x}_i)\right] \tag{7.4}$$

7.4.2 Classification using the HMM-GMM

The HMM-GMM (and associated parameters) are defined at the end of section 4.5.3, with the latent state variable, z_i, now denoting the sleep state, s_i. The model can also be thought of as extending the state-dependent GMM, with the transition parameter, $\boldsymbol{t}_{s_{i-1}}$. This adds an additional dependency of the current sleep state on the previous state.

Estimation of the most probable sleep state sequence, $\hat{s}_{1:N}$ (where N is the sequence length), is best performed recursively over the full sequence, $\boldsymbol{x}_{i:N}$, by the log-Viterbi algorithm (defined in section 4.5.3 and algorithm 4.3).

7.5 Model training & assessing performance

To train these probabilistic models and perform inference, nested CV was used (as defined in section 4.6.1 and algorithm 4.5) with outer-most 8-fold CV by patient (subject) and inner-most LOSO CV splits. This incorporated the mRMR-Spearman feature selection and the number of selected features to use in the model (P) was selected by grid search.

To train the state-dependent GMM, corresponding parameters, $\phi_{s_i,k}$, $\boldsymbol{\mu}_{s_i,k}$, and $\boldsymbol{\Sigma}_{s_i,k}$, were fitted directly using the EM algorithm (as defined in section 4.5.2 and algorithm 4.1) on each subset of labelled sleep data in the training set. Values for $\boldsymbol{\xi}$ were calculated from the total sleep state counts in the training data and K_{s_i} was also selected by grid search.

For the HMM-GMM, parameters were estimated as for the GMM, with $\boldsymbol{t}_{s_{i-1}}$ additionally estimated from the transitions counts in the training set.

The prior removal of IS, dubious and artefact data (as discussed in section 7.2) could introduce false transitions and a resulting bias in the model training (as well as the sleep estimates). Consequently, the model parameters were trained using only the data *within* the remaining clean regions of the recording. Furthermore, transitions that would have occurred into or out of a removed period, were ignored. Sleep state estimates were then performed separately for each remaining clean region and the results concatenated for the full recording.

To choose optimal values of P and K_{s_i} in the grid search and assess overall performance, mean Cohen's κ (as defined in section 4.6.2) was used as the performance measure and maximised across the grid. It provided means to assess performance by taking into account chance agreement as well as the data imbalance.

To assess final performance on the test recordings, κ was also used along with the sensitivity and specificity. Before performing multi-state classification, HMM and GMM models were trained for binary sleep classification, $s_i \in \{\text{QS, non-QS}\}$, as a benchmark to assess the performance of these probabilistic models when compared to the CLASS QS detection algorithm of chapter 5. Patient-wise

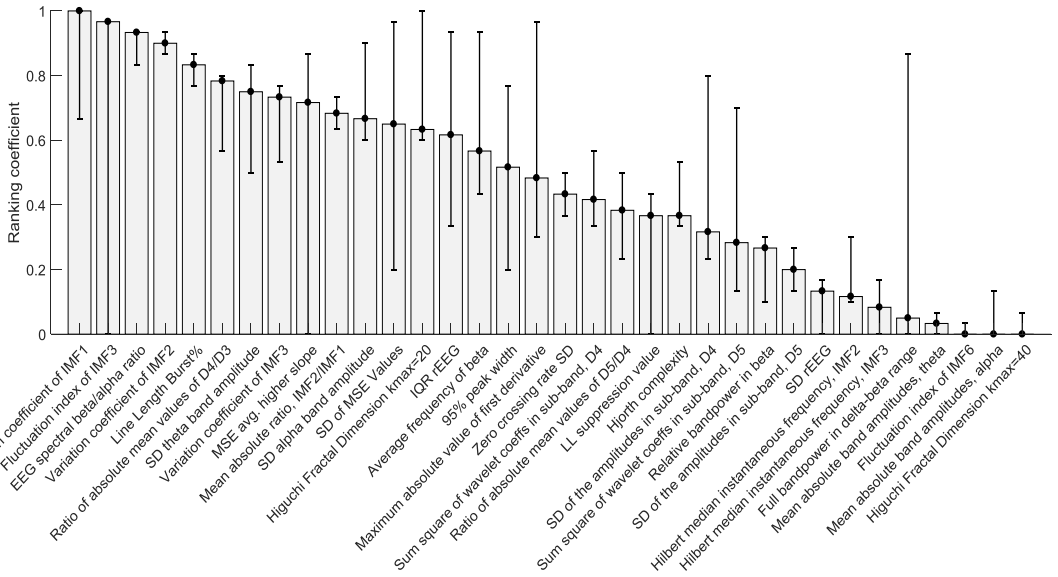


Figure 7.2: Median ranking coefficients of the top 30 features selected at each outer-most fold of nested-CV by mRMR-Spearman, for binary sleep classification. 34 different features were selected across folds. Error-bars denote the range.

standardization of recordings had not been previously explored for preterm EEG analysis so performance was also compared, with and without standardization, to better understand the importance of this step.

Finally, expansion to multi-state sleep classification was performed, $s_i \in \{\text{AS I, HVS, TA, LVI}\}$, and also compared with and without standardization.

7.6 Results

7.6.1 Binary classification performance

Nested CV data splits between folds were the same for each model, to ensure a fair comparison between methods. However, due to the varying training data between folds, selected features were ranked differently each time. To assess the variability of the rankings, the ranks of the top 30 features were determined at each fold, inverted (such that the most selected feature would have value '30' and the least selected a value '1') and finally normalized, to generate a 'ranking coefficient'. For binary classification, median ranking coefficients (across folds) are plotted in figure 7.2 for each feature. Error-bars denote the range.

34 different features appeared in the top 30 across folds showing a consistency between the nested-CV folds. The top five features were 'variation coefficient of IMF1', 'fluctuation index of IMF3', 'EEG spectral beta/alpha ratio', 'variation coefficient of IMF 2', and 'Line Length Burst %'. 'Variation coefficient of IMF1' was selected as the top feature in all folds apart from one. A high variability was seen in the 'fluctuation index of IMF3' (as indicated by the lower range of zero in figure 7.2) as it was not selected in three out of eight folds, but remained very prominent in five out of eight folds. Through grid search, a median of 7 (IQR 5.5-9.5) of these selected features were subsequently chosen (across folds) for the HMM-GMM, and 7 (IQR 6-7) for the GMM.

Table 7.1 shows the overall test performance for both the GMM and HMM-GMM. These are compared to models trained without patient-wise standardization, as well as the CLASS algorithm. As described in chapter 5, CLASS classifies QS periods as a whole event rather than per epoch. For accurate comparison of performance, the resulting estimated events were converted into the corresponding 30 second epochs that were classified by the GMM and HMM-GMM models.

Because of the limited number of recordings, applying statistical tests to compare models (such as the Wilcoxon signed-rank test used in chapter 5) becomes challenging, due to the dependency of p-values on sample size and the ambiguity associated with choosing an appropriate significance threshold in this case. Instead, each recording was scrutinized individually and overall performance compared using both the median and mean κ, sensitivity, and specificity, as shown in table 7.1. Discrepancies between the median and mean could identify potential outliers which become more significant with a smaller sample size. Such outliers are unexpected, due to the high quality of the resulting data used.

Results for binary classification showed mean and median κ values for all probabilistic models >0.8, with the HMM-GMM performing best (median κ=0.91). There was a small drop in performance ($\approx$2%) in the models when patient-wise standardization was not performed although it still remained very high.

Table 7.1: Comparing performance for binary sleep classification between the HMM-GMM and GMM, with and without patient-wise standardization (stand.), as well as the CLASS algorithm. Median (med) and mean κ, sensitivity (Sens) and specificity (Spec) are all shown.

Model	**med κ (IQR)**	**med Sens (IQR)**	**med Spec (IQR)**	**mean κ (±SD)**	**mean Sens (±SD)**	**mean Spec (±SD)**
HMM-GMM	0.91 (0.86-0.93)	0.95 (0.92-0.97)	0.97 (0.95-0.98)	0.89 (±0.07)	0.94 (±0.04)	0.96 (±0.04)
GMM	0.88 (0.80-0.93)	0.93 (0.89-0.97)	0.96 (0.88-0.97)	0.85 (±0.12)	0.92 (±0.06)	0.93 (±0.06)
HMM-GMM (no stand.)	0.89 (0.77-0.92)	0.96 (0.89-0.98)	0.97 (0.93-0.98)	0.82 (±0.19)	0.89 (±0.19)	0.94 (±0.09)
GMM (no stand.)	0.86 (0.74-0.95)	0.96 (0.89-0.98)	0.96 (0.88-0.98)	0.81 (±0.19)	0.89 (±0.19)	0.93 (±0.07)
CLASS	0.71 (0.53-0.76)	0.97 (0.91-100)	0.88 (0.83-0.91)	0.62 (±0.19)	0.95 (±0.06)	0.86 (±0.06)

These models also outperformed CLASS, which showed a lower performance across all measures (median κ=0.71). The larger drop in mean value (mean κ=0.62) further indicated that CLASS did not accurately detect the QS periods in certain cases. This was largely down to the comparatively lower specificity (median = 0.88).

Overall, these results suggested that the GMM and HMM-GMM models were excellent candidates for binary classification, with exceptionally high and consistent performance.

7.6.2 Multi-state classification performance

For multi-state classification, median ranking coefficients for the selected features are plotted in figure 7.3. 36 different features appeared in the top 30 across folds again showing a consistency with nested-CV, as for binary classification.

In contrast to binary classification, the top five selected features for multi-state classification were now quite different: 'coastline distance', 'mean absolute ratio, IMF6/IMF5', 'Hilbert median instantaneous frequency, IMF1', 'mean of absolute values in sub-band, D3', and 'average frequency of theta'. 'Coastline distance' was the most selected feature across all folds apart from one. Similarly, the 'average frequency of theta' was also consistently and highly selected across folds except for

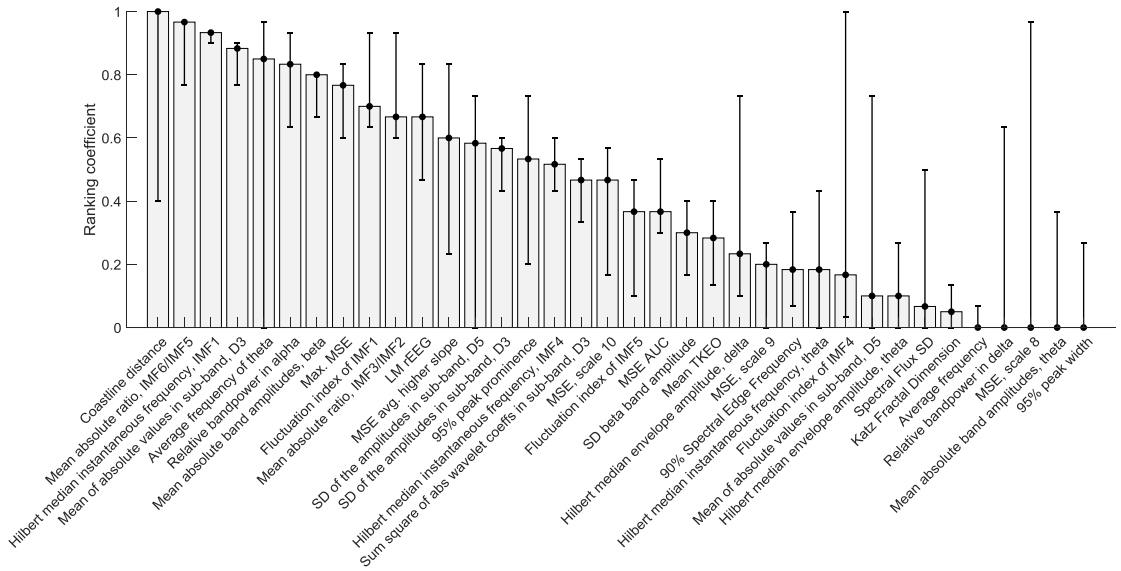


Figure 7.3: Median ranking coefficients of the top 30 features selected at each outer-most fold of nested-CV by mRMR-Spearman, for multi-state sleep classification. 36 different features were selected across folds. Error-bars denote the range.

one instance where it was replaced by the feature 'MSE, scale 8'. This explains the extreme ranges observed in figure 7.3 for both. Of these selected features, a median of 15.5 (IQR 13.5-17.5) were subsequently chosen (across folds) for the HMM-GMM, and 16 (15-20) for the GMM.

To assess the stability of the training procedure, figure 7.4 shows the κ performance curves over feature number (up to 30 features) for both the GMM (figure 7.4 (a)) and HMM-GMM (figure 7.4 (b)). Grey curves show the performance at each fold overlaid with the median performance curve across all folds. Curves were determined from the maximum mean validation performance across the grid search, as the number of features were increased.

Overall, a much higher κ performance was seen in the HMM-GMM compared to the GMM, with both models showing a smooth improvement as more features were included. Performance curves at each fold also followed a similar trend indicating a consistent and stable training of the models. In both cases, performance increased rapidly across the first five features and peaked around 15 features. Beyond this range, there was a plateau with no further improvement.

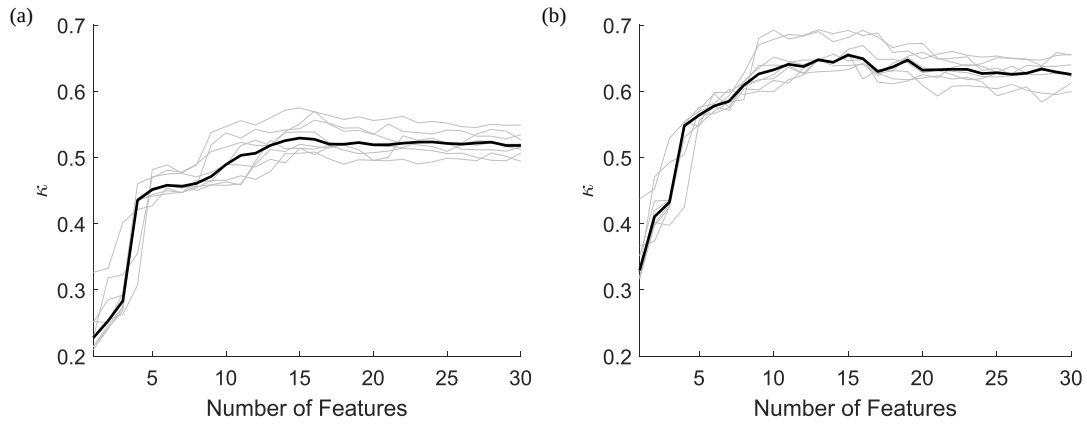


Figure 7.4: (a) κ performance curves for the GMM as the number of features are increased (up to 30), assessed on the validation data. Grey curves denote performance at each fold, with the black curve denoting the median performance across folds. (b) Corresponding performance curves for the HMM-GMM.

Figure 7.5(a) shows an example of the sleep classification hypnograms after visual labelling, for the GMM, and for the HMM-GMM, on one test recording. Figure 7.5(b) additionally shows the mean matrix of transition probabilities that were derived across the folds during training. Hypnogram estimates of the HMM-GMM were much smoother, compared to the GMM model, and better reflected the visual labelling. In contrast, the GMM hypnogram produced far more spurious, short-duration estimates.

Figures 7.5(c) and 7.5(d) show the combined confusion matrices (defined in section 4.6.2) of the GMM and HMM-GMM, respectively, compared to visual labelling. LVI was classified with similar performance between the two models. However, across the other three states, a higher proportion of epochs were correctly classified by the HMM-GMM. This was most notable for ASI, with 75.6% of epochs correctly classified in the HMM-GMM compared to 50.2% in the GMM. In general, the majority of misclassified epochs occurred between each pair of QS (HVS and TA) and non-QS (ASI and LVI) states.

Finally, table 7.2 shows the overall test performance of the models with and without scaling (as defined for binary classification). In this case, sensitivity and specificity measures were determined as a 'macro-average', by taking the mean of the measures when calculated for each of the four sleep states.

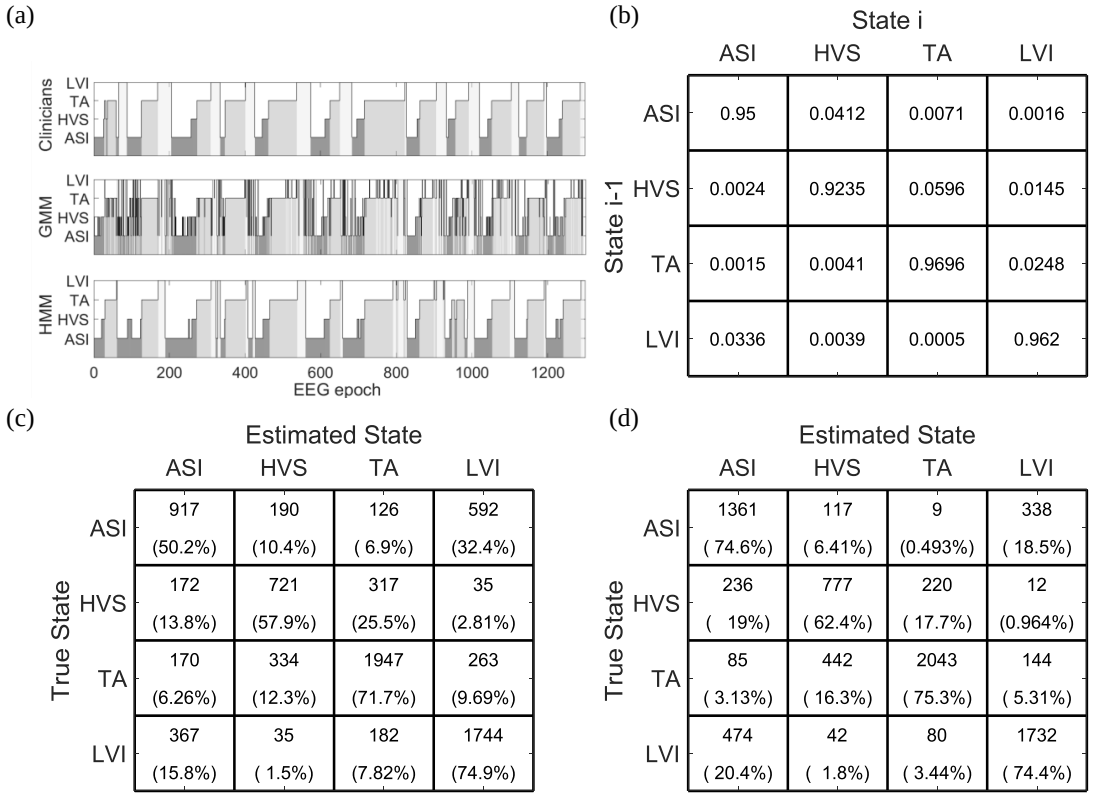


Figure 7.5: (a) An example hypnogram of multi-state sleep classification on one test recording, after visual labelling (*top*), the GMM (*middle*), and the HMM-GMM (*bottom*). (b) The mean HMM-GMM matrix of transition probabilities across folds. (c) Confusion matrix comparing the total number of correctly classified epochs between the true (visually labelled) states and estimates from the GMM for ASI, HVS, TA and LVI. The % proportion of epochs, with respect to the true state, are also given in brackets. (d) The corresponding confusion matrix for the HMM-GMM.

As first alluded to in figures 7.4 and 7.5(c)-(d), overall median and mean κ confirmed that the HMM-GMM outperformed the GMM, achieving a higher median performance (0.61) compared to the GMM (0.51). The HMM-GMM model also maintained a higher specificity (median 0.91) and sensitivity (median 0.72), compared to the GMM (medians 0.88 and 0.63, respectively). Furthermore, both mean and median values were similarly high, suggesting a consistency in the model performance and the presence of no extreme outliers.

In contrast, HMM-GMM performance without patient-wise standardization fell, with a median κ=0.58. This was largely caused by the drop in sensitivity (median 0.65). Interestingly, the GMM without standardization had variable results

Table 7.2: Comparing performance for multi-state sleep classification between the HMM-GMM and GMM, with and without patient-wise standardization (stand.). Median (med) and mean κ, sensitivity (Sens) and specificity (Spec) are shown and the Sens and Spec are macro-averaged across all four sleep states.

Model	**med κ (IQR)**	**med Sens (IQR)**	**med Spec (IQR)**	**mean κ (±SD)**	**mean Sens (±SD)**	**mean Spec (±SD)**
HMM-GMM	0.61 (0.52-0.67)	0.72 (0.70-0.74)	0.91 (0.90-0.93)	0.61 (±0.10)	0.71 (±0.06)	0.91 (±0.02)
GMM	0.51 (0.44-0.58)	0.63 (0.60-0.71)	0.88 (0.86-0.90)	0.51 (±0.11)	0.63 (±0.09)	0.87 (±0.06)
HMM-GMM (no stand.)	0.58 (0.49-0.69)	0.65 (0.53-0.71)	0.90 (0.88-0.93)	0.54 (±0.20)	0.62 (±0.13)	0.89 (±0.05)
GMM (no stand.)	0.57 (0.44-0.62)	0.64 (0.51-0.68)	0.88 (0.86-0.91)	0.51 (±0.16)	0.60 (±0.11)	0.87 (±0.07)

depending on whether median or mean performance values were compared. Through both the median and mean, the GMM without standardization performed worse than the HMM-GMM with standardization (across all measures). Median κ also suggested that the GMM without standardization performed better than the GMM with standardization (0.57 compared to 0.51, respectively). However, mean κ values suggested that both the models performed similarly (both 0.51). The large discrepancy between median and mean κ in the GMM without standardization, suggested the presence of extreme outliers. Upon further inspection, one particular recording had very poor κ=0.04. On the other hand, the GMM with standardization classified this same recording with κ=0.36, and the HMM-GMM with standardisation classified the recording with κ=0.61. These individual κ values for all patients and the different models are provided in table B.4 (in appendix B.2).

7.7 Discussion

EEG features derived from the DWT and EMD are prominently selected by mRMR-Spearman. As also discussed in chapter 2, this suggests a benefit of accounting for the non-stationary nature of the EEG, which is not considered by more traditional DFT-based metrics. Furthermore, such features derived from EEG decompositions reflect

existing knowledge (as discussed in chapter 2) that specific sleep states are defined by distinct behaviours that uniquely vary across the different frequency bands [41].

The larger number of features required for optimal performance in multi-state sleep classification, compared to binary classification, is also indicative of the greater complexity required by the models to distinguish between more subtly different states as opposed to broader QS and non-QS periods.

Particularly, the features 'coastline distance' and 'Hilbert median instantaneous frequency, IMF1' are regularly selected in multi-state classification and seem to identify the subtler differences between these sleep states well. The 'coastline distance' is sensitive to large fluctuations, such as bursts and suppression, which is most discriminative between TA and other states [19, 41, 71] (as described in chapter 2). 'Hilbert median instantaneous frequency, IMF1' encodes the shifts in frequency within the more oscillatory high frequency bands which is affected by the presence (or lack thereof) of IBIs. LVI, ASI and HVS all contain different proportions of mixed frequency content [71] which this feature also proves sensitive to. Additionally, other prominent features such as 'Mean absolute ratio, IMF6/IMF5' are clearly sensitive to such variations.

When comparing these features to those linked with maturation in chapter 6 (see tables B.1-B.3 in appendix B), 'coastline distance' also showed significant correlations with PMA (when extracted from non-QS). Furthermore, the 'Hilbert median instantaneous frequency, IMF1' and 'Mean absolute ratio, IMF6/IMF5' (both extracted from non-QS) showed additional sensitivity to outcome. This reveals the presence of novel features that can behave as both sleep state (across four states) and maturational biomarkers.

Fluctuations in feature importance across folds is indicative of the effect of limited data on feature selection and CV. With very different features occasionally selected in certain folds, without causing detriment to performance, there is likely a number of possible feature subsets that could prove optimal. Moreover, the choice of feature selection plays a big part. Due to the potential for high redundancy between features, mRMR-Spearman was chosen. However, the use of alternative

metrics/methods for feature selection may well identify a different subset that achieves similar classification performance.

Both the GMM and HMM-GMM show consistently excellent binary sleep classification performance even with a model, such as a GMM, that does not account for state transitions. Both models also outperform the CLASS algorithm. While CLASS was previously shown in chapter 5 to perform term-age classification well, it performs less well here due to a larger number of term-age recordings $\geq$40 weeks PMA. As described in chapter 5, the globally more continuous EEG that appears at term-age results in a reduction in the relative difference in discontinuity between QS and non-QS. CLASS relies on detecting this relative discontinuity but, at $\geq$40 week PMA, the increasing proportion of continuous HVS (compared to TA) makes the separation between QS and non-QS less distinct. This results in greater misclassification and lower specificity. The supervised nature of the GMM and HMM-GMM, on the other hand, allows both the HVS and TA portions of QS to be classified well.

When compared to other supervised methods in literature, as discussed in chapter 2, the method of *Gerla et al.* reports a lower QS accuracy (versus AS) of 70% at term-age [96]. The MSE-based approach of *De Wel et al.* presents an improved QS (and non-QS) classification sensitivity and specificity of 90% (based on the ROC curve) [29]. Each of these approaches are performed with different features, training-testing procedures, and datasets, preventing a perfect comparison. However, the high classification performance of particularly the method of *De Wel et al.*, is similar to the GMM and HMM-GMM. This further indicates that there remains a strong distinction between overall QS and non-QS states at term-age.

It is in multi-state classification of ASI, HVS, TA and LVI, that a clearer improvement is seen with the HMM-GMM (compared to the other models), with the introduction of patient-wise standardization resulting in a consistently good classification across all recordings. The stronger distinction between QS and non-QS limits the problem of inter-patient variability in binary classification. However, the subtler differences between states in multi-state classification are more compromised

by this variability and the overlap between different states, from recording to recording, reduces the discriminatory power of the features. Specific cases when the mean performance is lower than the median (due to very poor classification of certain recordings) are indicative of such high variability.

By introducing transition probabilities, classification is further improved by accounting for the propensity of certain sleep state transitions to occur more often than others. As discussed in chapter 2, HVS predominantly precedes TA (if QS is sustained [19, 71]) and this is reflected by the higher transition probability seen from HVS to TA than from TA to HVS (see figure 7.5(b)). In tandem with the very high probabilities of self-transitions, this explains the improved classification performance compared to the GMM. However, even by accounting for transitions and correcting for some of the inter-patient variability, similarities between QS and non-QS state pairs make classification at this level still challenging. As first touched upon in chapter 2, this is because term-age marks a period of transitionary sleep that is maturing into states that resemble infant to adult sleep [19, 71, 80]. This said, these models are the first to show that it is indeed possible to successfully classify these multiple, differentiated states entirely from EEG.

7.8 Conclusion

The key novel contributions outlined in this chapter are:

- The first method to classify the multiple, differentiated EEG sleep states at term-age. Assessing the quality and rate of this differentiation provides potential means to assess sleep-wake cycling during this period and quantify brain maturation.
- The unique application of an HMM-GMM, which models state transitions to provide smoother and more reliable estimates of sleep state (even with a small dataset). Reliability is improved by applying a patient-wise feature standardization to reduce the effects of inter-patient variability.
- The ability to classify these multiple sleep states from only EEG (without requiring additional PSG signals) based on a data-driven extraction of features.

The HMM-GMM forms part of a larger family of probabilistic models and can be extended in novel ways to incorporate an age-dependency. This can be useful for assessing the sleep-wake cycling over a wider age range, particularly accounting for extremely dysmature (undifferentiated) sleep-wake cycling at term-age. The development of such extensions is the subject of chapter 8.

8

End-to-end Models for Assessing Sleep-Wake Cycling

Associated publication(s):

K. Pillay et al. "A Bayesian parametric model for quantifying brain maturation from sleep-EEG in the vulnerable newborn baby". In: *2018 IEEE Eng. Med. Biol. 40th Annu. Conf.*, IEEE, 2018, pp. 1-4.

K. Pillay et al. "A unifying Bayesian approach for preterm brain-age prediction that models EEG sleep transitions over age". In: *2017 NIPS workshop on Machine Learning for Health (ML4H)*, NIPS, 2017. arXiv:1809.07102.

8.1 Introduction

In chapter 7, a probabilistic HMM-GMM model was successfully applied to EEG recordings to classify the multiple, differentiating sleep states at term-age, as a potential means to assess sleep-wake cycling and quantify maturation.

As described in chapter 4, probabilistic models form part of a larger family of Bayesian networks which can be extended to incorporate additional variables. This provides a platform to build on chapter 7 and develop expanded models with an explicit dependency on age, that could accurately classify EEG sleep states and sleep-wake cycling across a wider PMA range. This age-dependency can also incorporate brain-age (as used previously in chapter 6), classifying cases of sleep-wake cycling where the EEG remains dysmature. As discussed in chapter 2, this could specifically

identify sleep-wake cycling at term-age that remains undifferentiated.

This chapter presents a series of 'end-to-end' solutions to simultaneously classify both sleep states and brain-age. The first is an age-varying GMM (aGMM) which assumes that each EEG epoch is independent from one another. The natural extension of this model is an age-varying HMM-GMM (aHMM-GMM) that additionally considers the state transitions between successive EEG epochs. Finally, a further extension is proposed which incorporates a DP-GMM (as defined in chapter 4) to avoid explicitly selecting the number of GMM components. This model is referred to as an aDP-HMM-GMM and is the first application of Bayesian non-parametrics to preterm EEG analyses.

8.2 Data

Datasets 1 and 3, as defined in chapter 3, were combined to train and validate these models. Dataset 1 was previously used to train and test the CLASS algorithm in chapter 5 and to train the brain-age prediction models in chapter 6. Dataset 3 was used to train the term-age probabilistic models of chapter 7. This resulted in a combined 40 patients (32 preterm, eight full-term) and 103 recordings, aged 27-43 weeks PMA.

Recordings were organized into six age groups for classifying brain-age. As also discussed in chapter 6, PMA provided a reasonable estimate of brain-age in normal outcome recordings [145, 146]. It was important to separate term-age recordings at 38 weeks PMA, as this marked the onset of sleep differentiation, including the emergence of HVS. Groupings of recordings <38 weeks PMA were less physiologically distinct and chosen to ensure sufficient data was available within each group for model training. The resulting groups were defined as 1: <30 weeks, 2: 30-31 weeks, 3: 32-33 weeks, 4: 34-35 weeks, 5: 36-37 weeks, and 6: $\geq$38 weeks. The corresponding distribution of recording numbers are presented in figure 8.1.

Although chapter 7 focussed on explicitly classifying the differentiated EEG states, classification here focussed on the binary sleep states of QS and non-QS as this is consistently defined across the full PMA range. From the visual labels,

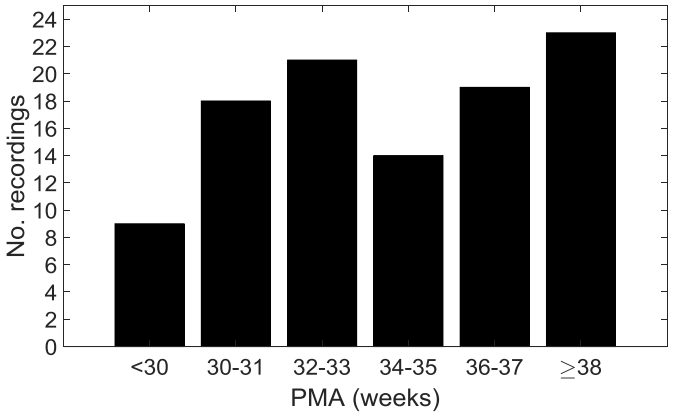


Figure 8.1: Histogram of the total number of EEG recordings separated into age groups. There were 103 recordings ranging from 27 to 42 weeks PMA.

74% of the data was estimated from non-QS and 26% from QS. While similar across the majority of age groups, proportions differed considerably in group 1 (<30 weeks) with 59% from non-QS and 41% from QS.

8.3 Feature pre-processing

As first motivated in chapter 7, each recording was divided into 30 second epochs with five second overlap. The 113 features defined in section 4.2 were extracted from each epoch and channel, with median values taken across channels.

Also in chapter 7, patient-wise feature standardization had been previously used. However, this was not applied here as standardizing by patient was likely to effect feature correlations with age. Furthermore, very little differences were observed with and without standardization for binary sleep classification in term-age data (as shown in table 7.1). Finally, features were extracted from the raw EEG due to the current limitations in artefact removal (namely ASR) discussed in chapter 6.

Feature selection was required to obtain a relevant subset for training the models. mRMR-Spearman (defined in section 4.4.2) was previously chosen in chapter 7 and proved successful when used in tandem with probabilistic models. However, conventional mRMR-Spearman (as well as other feature selection algorithms) only correlate feature relevance with one known label. To develop models that performed both sleep staging and brain-age prediction, feature selection had to

be discriminative for both (i.e. two labels). This prompted an extension of mRMR-Spearman to achieve this.

The maximum relevance step in mRMR-Spearman correlates each feature to a single label using the Spearman's rank correlation coefficient (defined in section 4.3.1). To incorporate both sleep and age, a new 'aggregate' Spearman's rank correlation coefficient was derived, taking into account the dependencies between each feature, age, and sleep state. The procedure below was now used in lieu of the conventional maximum relevance step in mRMR-Spearman:

1. A linear regression model is fitted between the j^{th} feature vector, $\boldsymbol{f}_j \in \mathbb{R}^N$ (where N is the total number of observations) and the corresponding vector of age outcomes, $\boldsymbol{y}^{(a)} \in \mathbb{N}^N$:

$$\boldsymbol{f}_j = \beta_0 + \beta_1 \boldsymbol{y}^{(a)} + \boldsymbol{\epsilon}_j^{(s)} \tag{8.1}$$

   The resulting residuals, $\boldsymbol{\epsilon}_j^{(s)} \in \mathbb{N}^N$, are decorrelated with *age* and only correlated with sleep state.
2. The sleep-dependent Spearman's rank correlation coefficient, $\rho_j^{(s)}$, is determined between $\boldsymbol{\epsilon}_j^{(s)}$ and the vector of sleep outcomes, $\boldsymbol{y}^{(s)} \in \mathbb{N}^N$.
3. A linear regression is again performed to de-correlate $\boldsymbol{f}_j$ with *sleep state*, resulting in the residuals $\boldsymbol{\epsilon}_j^{(a)} \in \mathbb{N}^N$.
4. The age-dependent Spearman's rank correlation coefficient, $\rho_j^{(a)}$, is determined between $\boldsymbol{\epsilon}_j^{(a)}$ and $\boldsymbol{y}^{(a)}$.
5. The aggregate Spearman's rank correlation coefficient, $\rho_j^{(agg)}$, is calculated as the mean of $\rho_j^{(s)}$ and $\rho_j^{(a)}$.
6. Steps 1-5 are repeated for all j.

These resulting aggregate feature coefficients, $\rho_j^{(agg)}$, were then carried forward to the usual minimum redundancy step, completed as described in section 4.4.2 (where $\rho_j^{(agg)} = I(\boldsymbol{y}, \boldsymbol{f}_j)$).

8.4 The aGMM

The graphical model for the aGMM is presented in figure 8.2(a), and is an extension of the sleep-dependent GMM applied in chapter 7. Specifically, the model is derived by extending equations (7.1)-(7.2) for age.

An i^{th} epoch of EEG features is denoted by the observation vector, $\boldsymbol{x}_i \in \mathbb{R}^P$ (where P is the number of selected features after mRMR-Spearman). Each $\boldsymbol{x}_i$ is conditioned on sleep state, $s_i \in \{\text{QS}, \text{non-QS}\}$, and age, $a \in \{1, 2, \cdots, 6\}$. The overall model is defined as:

$$\boldsymbol{x}_i | s_i, a \sim \sum_{k=1}^{K_{s_i}} \phi_{a,s_i,k} \cdot \mathcal{N}(\boldsymbol{\mu}_{a,s_i,k}, \boldsymbol{\Sigma}_{a,s_i,k}) \tag{8.2}$$

$$s_i | a \sim \text{Cat}(\boldsymbol{\xi}_a) \tag{8.3}$$

$$a \sim \text{unif}\{1, 6\} \tag{8.4}$$

a is drawn from a discrete uniform distribution as there is no physiological reasoning for one age being more likely than another. s_i is modelled as an age-dependent (binary) categorical distribution with parameters, $\boldsymbol{\xi}_a$, encoding the probabilities of being in each sleep state at a given age. Conditioning this on age accounts for the observed changes in sleep proportions as the EEG matures. $\boldsymbol{x}_i$ is distributed by sleep- and age-dependent GMMs, with component parameters defined as in section 4.5.2.

The total number of GMM components, K_{s_i}, is allowed to differ between sleep states (as in chapter 7). In these models, K_{s_i} could also be varied with age. However, the requirement to select these values by grid search (as later discussed in section 8.7) makes it computationally infeasible, with the inclusion of each additional parameter resulting in an exponential increase in computational time.

8.4.1 Inference

Equations (8.2)-(8.4) define the forms for the known conditional density distributions, $p(\boldsymbol{x}_i|s_i, a)$, $p(s_i|a)$, and $p(a)$, respectively. The first goal is to predict the age given a full recording, $p(a|\boldsymbol{x}_{1:N})$, where $\boldsymbol{x}_{1:N}$ denotes the sequence of epochs $[\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N]$ for convenience. This can be represented in terms of the known conditional probabilities using the mathematical relationships defined in section 4.5.1.

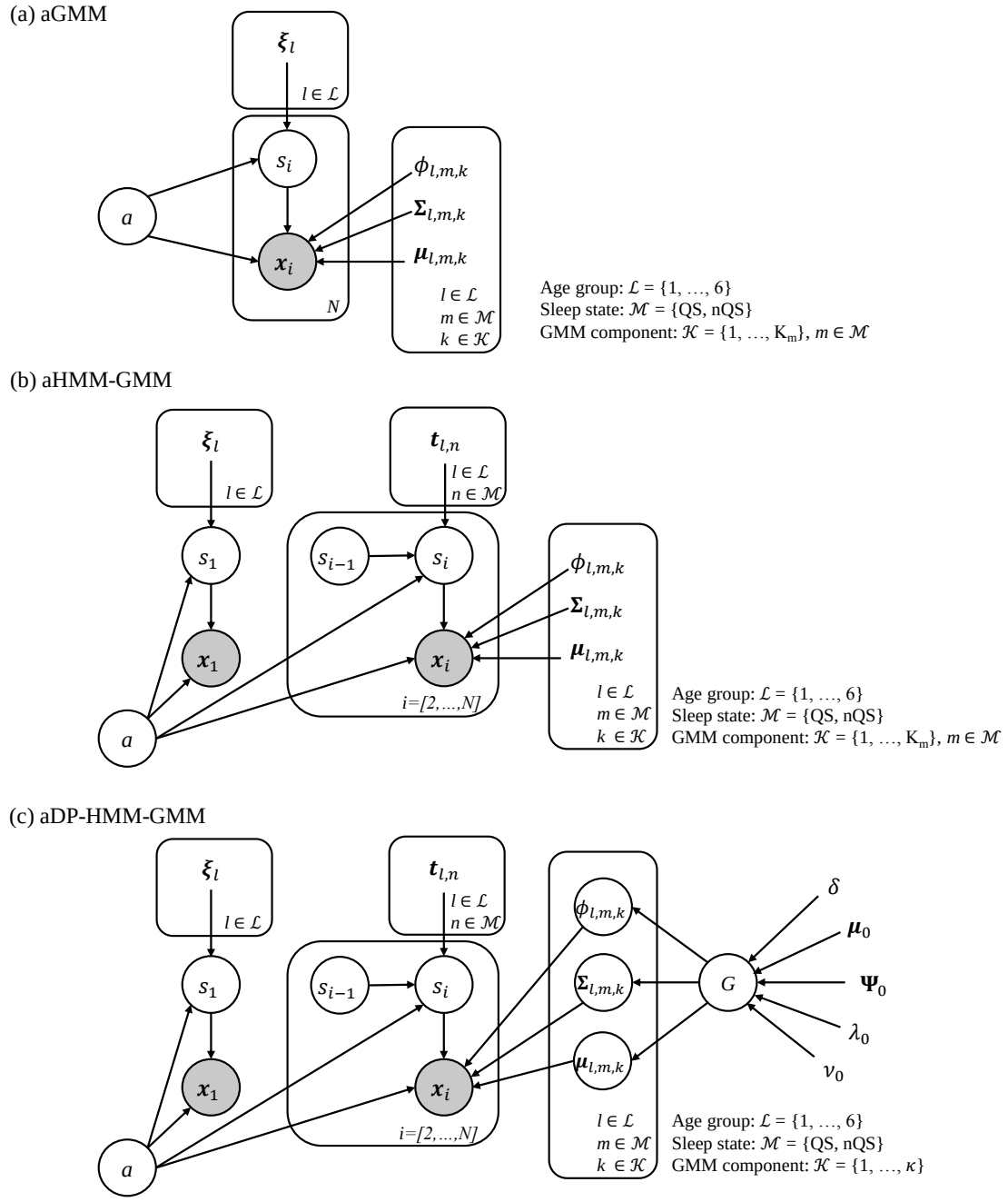


Figure 8.2: (a) The aGMM model. Each i^{th} epoch of EEG features, $\boldsymbol{x}_i \in \mathbb{R}^P$ (where P is the number of selected features), is conditioned by sleep state, $s_i \in \{\text{QS}, \text{non-QS}\}$, and age, $a \in \{1, 2, \cdots, 6\}$. $\boldsymbol{x}_i$ is distributed by sleep- and age-dependent GMMs and the number of components, K_m, is allowed to vary with sleep-state. (b) The aHMM-GMM model. Transitions are additionally modelled, with s_i conditioned on s_{i-1} as well as a. (c) The aDP-HMM-GMM. The number of components and GMM parameters are now implicitly selected from the data by incorporating a DP-GMM. Parameters are Normal-inverse Wishart-distributed, with additional hyper-parameters $\boldsymbol{\mu}_0$, λ_0, $\boldsymbol{\Psi}_0$, ν_0 and concentration hyper-parameter, δ (as defined in section 4.5.4).

The posterior, $p(a|\mathbf{x}_{1:N})$, is first described using Bayes' rule. As the *maximum a posteriori* is chosen to predict age, the normalization term can be dropped to approximate $p(a|\mathbf{x}_{1:N})$ by:

$$p(a|\boldsymbol{x}_{1:N}) = \frac{p(a)p(\boldsymbol{x}_{1:N}|a)}{\sum_a p(a)p(\boldsymbol{x}_{1:N}|a)} \propto p(a)p(\boldsymbol{x}_{1:N}|a) \tag{8.5}$$

As shown in figure 8.2(a), each $\boldsymbol{x}_i$ is conditionally independent of one another given a. This assumption can be applied to equation (8.5) by treating the overall likelihood as a product of the individual epoch likelihoods and converting the result to a joint probability:

$$p(a|\boldsymbol{x}_{1:N}) \propto p(a)\prod_{i=1}^{N} p(\boldsymbol{x}_i|a) = p(a)\prod_{i=1}^{N}\left[\frac{p(\boldsymbol{x}_i,a)}{p(a)}\right] \tag{8.6}$$

$p(\boldsymbol{x}_i, a)$ is expressed in terms of s_i, by marginalizing over s_i and applying the chain rule to represent this in terms of the known conditional probabilities in the model:

$$p(a|\boldsymbol{x}_{1:n}) \propto p(a)\prod_{i=1}^{N}\left[\frac{\sum_{s_i} p(\boldsymbol{x}_i,s_i,a)}{p(a)}\right] = p(a)\prod_{i=1}^{N}\left[\sum_{s_i} p(\boldsymbol{x}_i|s_i,a)p(s_i|a)\right] \tag{8.7}$$

To prevent underflow, equation (8.7) is evaluated as a log-posterior:

$$\log\left[p(a|\boldsymbol{x}_{1:N})\right] \propto \log\left[p(a)\right] + \sum_{i=1}^{N}\log\left[\sum_{s_i} p(\boldsymbol{x}_i|s_i,a)p(s_i|a)\right] \tag{8.8}$$

Finally, the estimate of the age, $\hat{a}$, is obtained by the *maximum a posteriori*:

$$\hat{a} = \arg\max_a\left(\log\left[p(a|\boldsymbol{x}_{1:N})\right]\right) \tag{8.9}$$

Given $\hat{a}$, a further inference equation is derived using Bayes' rule to estimate the sleep state, $\hat{s_i}$:

$$p(s_i|\boldsymbol{x}_i,\hat{a}) \propto p(\boldsymbol{x}_i|s_i,\hat{a})p(s_i|\hat{a}) \text{ and } \hat{s_i} = \arg\max_{s_i}\left[p(s_i|\boldsymbol{x}_i,\hat{a})\right] \tag{8.10}$$

8.5 The aHMM-GMM

The aHMM-GMM builds on the HMM-GMM model (defined in section 4.5.3 and previously used in chapter 7) for age-dependency, by extending equations (4.47)-(4.49). Unlike the aGMM assumption of independent epochs, transitions are now

modelled, with inference performed on the full recording sequence, $\boldsymbol{x}_{1:N}$. The graphical model is shown in figure 8.2(b) and is defined as follows:

$$\boldsymbol{x}_i|s_i, a \sim \sum_{k=1}^{K_{s_i}} \phi_{a,s_i,k} \cdot \mathcal{N}(\boldsymbol{\mu}_{a,s_i,k}, \boldsymbol{\Sigma}_{a,s_i,k}) \tag{8.11}$$

$$s_i|s_{i-1}, a \sim \text{Cat}(\boldsymbol{t}_{a,s_{i-1}}) \tag{8.12}$$

$$s_1|a \sim \text{Cat}(\boldsymbol{\xi}_a) \tag{8.13}$$

$$a \sim \text{unif}\{1, 6\} \tag{8.14}$$

When compared to the aGMM, the aHMM-GMM additionally includes the age-dependent transition parameter, $\boldsymbol{t}_{a,s_{i-1}}$, to define the transitions probability distribution, $p(s_i|s_{i-1}, a)$. Parameter, $\boldsymbol{\xi}_a$, represents the initial age-dependent sleep state probabilities at $i = 1$, which defines the distribution, $p(s_1|a)$.

8.5.1 Inference

By modelling transitions, the inference for $p(a|\boldsymbol{x}_{1:N})$ can be derived by incorporating the recursive methods typically used for HMM inference. Introducing age dependencies on both the prior and transition probabilities allows the inference to be structured as an HMM model selection problem. In short, the age-conditioned HMM likelihood, $p(\boldsymbol{x}_{1:N}|a)$, can be calculated for each age group using the log-scaled forward algorithm (defined in section 4.5.3 and algorithm 4.2) and the *maximum a posteriori* of $p(a|\boldsymbol{x}_{1:N})$ calculated from these likelihoods to estimate age, $\hat{a}$. This is summarized in algorithm 8.1.

Given $\hat{a}$, the most probable sleep state sequence, $\hat{s}_{1:N}$, is determined by the log-Viterbi algorithm (as defined in section 4.5.3 and algorithm 4.3) using the model parameters when $a = \hat{a}$.

8.6 The aDP-HMM-GMM

So far, both the aGMM and the aHMM-GMM require an explicit selection for K_{s_i}, which is unintuitive. The aDP-HMM-GMM provides an alternative non-parametric means for implicitly selecting the number of components from the data. The graphical model is shown in figure 8.2(c).

Algorithm 8.1 aHMM-GMM inference for $p(a|\boldsymbol{x}_{1:N})$

input sequence $\boldsymbol{x}_{1:N} \in \mathbb{R}^{P\times N}$ (where P denotes the vector dimension and N the number of observations), transition probabilities $p(s_i|s_{i-1}, a)$, emission probabilities $p(\boldsymbol{x}_i|s_i, a)$, prior sleep probabilities $p(s_1|a)$, and prior age probabilities $p(a)$, where $i \in \mathbb{N} : 1 \leq i \leq N$.

for a from 1 to 6 **do** ▷ Calculate for each age group
  Calculate $\log[p(\boldsymbol{x}_{1:N}|a)]$ using log-scaled forward algorithm ▷ Algorithm 4.2
end for

$\log[p(a|\boldsymbol{x}_{1:N})] \propto \log[p(a)] + \log[p(\boldsymbol{x}_{1:N}|a)]$

$\hat{a} = \arg\max_a \left(\log[p(a)] + \log[p(\boldsymbol{x}_{1:N}|a)]\right)$

output $\hat{a}$

The aDP-HMM-GMM further extends the aHMM-GMM, with the GMM parameters now drawn as in the case of a DP-GMM (described in section 4.5.4). The full model is defined as[1]:

$$\boldsymbol{x}_i|s_i, a, \boldsymbol{\mu}_{a,s_i,\bullet}, \boldsymbol{\Sigma}_{a,s_i,\bullet}, \phi_{a,s_i,\bullet} \sim \sum_{k=1}^{\kappa} \phi_{a,s_i,k} \cdot \mathcal{N}(\boldsymbol{\mu}_{a,s_i,k}, \boldsymbol{\Sigma}_{a,s_i,k}) \tag{8.15}$$

$$\boldsymbol{\mu}_{a,s_i,k}, \boldsymbol{\Sigma}_{a,s_i,k}, \phi_{a,s_i,k}|G \sim G \tag{8.16}$$

$$G \sim \mathcal{DP}(\delta, G_0) \tag{8.17}$$

$$G_0 = \mathcal{NIW}(\boldsymbol{\mu}_0, \lambda_0, \boldsymbol{\Psi}_0, \nu_0) \tag{8.18}$$

$$s_i|s_{i-1}, a \sim \text{Cat}(\boldsymbol{t}_{a,s_{i-1}}) \tag{8.19}$$

$$s_1|a \sim \text{Cat}(\boldsymbol{\xi}_a) \tag{8.20}$$

$$a \sim \text{unif}\{1, 6\} \tag{8.21}$$

Here, the notation, $\boldsymbol{\mu}_{a,s_i,\bullet}$, $\boldsymbol{\Sigma}_{a,s_i,\bullet}$, and $\phi_{a,s_i,\bullet}$ denote all the GMM component parameters for a given a and s_i. The total number of components is now a dynamic value that can change for each a and s_i combination. The Gaussian parameters are drawn from a Normal-inverse Wishart distribution, with additional hyper-

[1] These equations and the overall graphical model in figure 8.2(c) illustrate how each of the sleep- and age-dependent GMM parameters are further drawn from a DP. Strictly speaking, this means that the DP variables of G, G_0 (and therefore the G_0 hyperparameters) also have dependencies on s_i and a. However, these additional dependencies are not included here for clarity.

parameters, $\boldsymbol{\mu}_0$, λ_0, $\boldsymbol{\Psi}_0$, ν_0, and a concentration hyper-parameter, δ, as defined for the DP-GMM in section 4.5.4.

8.6.1 Inference

With the parameters trained, inference in the aDP-HMM-GMM is exactly the same as the aHMM-GMM, discussed in section 8.5.1. Estimates for $\hat{a}$ are derived from the *maximum a posteriori*, incorporating the log-scaled forward algorithm, and $\hat{s}_{1:N}$ is determined by the log-Viterbi algorithm.

8.7 Model training & assessing performance

To train model parameters and perform inference, nested CV (defined in section 4.6.1 and algorithm 4.5) was used (as in chapter 7), with outer-most 10-fold and inner-most 18-fold splits by patient (subject). This incorporated the extended mRMR-Spearman feature selection, with the subsequent number of selected features to train the model (P) selected by grid search.

For training the aGMM, parameters for each age- and sleep-dependent GMM were determined by fitting GMMs to the corresponding age and sleep partitions of the training data, using the EM algorithm (defined in section 4.5.2 and algorithm 4.1). Values for K_{s_i} were also selected by grid search and $\boldsymbol{\xi}_a$ were calculated from the total sleep state counts in the training data for each age group.

For the aHMM-GMM, the GMM parameters, $\boldsymbol{\xi}_a$, and K_{s_i}, were determined as for the aGMM. In addition, $\boldsymbol{t}_{a,s_{i-1}}$ were calculated from the transition counts in the training data for each age group.

To train the aDP-HMM-GMM, $\boldsymbol{t}_{a,s_{i-1}}$ and $\boldsymbol{\xi}_a$ were determined as for the aHMM-GMM. GMM parameters were calculated using stochastic Neal's Algorithm 8 (defined in section 4.5.4 and algorithm 4.4). This required the additional assignment of the hyper-parameters. Values for $\boldsymbol{\mu}_0$ were determined by choosing the mean of the training data (for each s_i and a partition) to ensure that resulting draws of $\boldsymbol{\mu}_{a,s_i,k}$ from the Normal-inverse Wishart distribution remained close to the data.

For a similar reason, $\mathbf{\Psi}_0$ was estimated as the diagonal matrix of the data variances (for each s_i and a partition) to ensure a similar case held true for $\mathbf{\Sigma}_{a,s_i,k}$.

As was noted by *Schurrman et al.* [230], when fitting $\boldsymbol{\mu}_0$ and $\mathbf{\Psi}_0$ in this manner, data is also used to estimate the prior. This can exaggerate the posterior estimates and cause over-fitting. It is suggested to keep the values of ν_0 and λ_0 as low as possible to limit this exaggeration. Consequently, typical values of $\lambda_0 = 1$ and $\nu_0 = P + 1$ were chosen [231].

As the number of components were implicitly selected from the data (controlled by δ), only the value of P required explicit selection by grid search. Selecting the optimal value of δ was not intuitive and including it as part of the grid search would by computationally expensive and counter-productive. Instead, full aDP-HMM-GMM models were trained over a candidate set of δ values (covering a wide range) to initially assess the sensitivity of this parameter to overall model performance (as is later presented in section 8.8.2).

To assess the validation and test performances in nested CV, Krippendorff's α was chosen (defined in section 4.6.2) as it provided a general measure to determine both the (nominal) sleep-state agreements for each recording, $\alpha^{(s)}$, and the (ordinal) age agreement, $\alpha^{(a)}$, across all recordings. To ensure models that performed well for both age and sleep state prediction were favoured, a combined assessment measure was derived for selecting the specific parameters by grid search, P and K_{s_i} (the latter for the aGMM and aHMM-GMM models):

$$I = \alpha^{(a)} - \frac{1}{N^{(v)}} \sum_{j=1}^{N^{(v)}} \left(1 - \alpha_j^{(s)}\right) \tag{8.22}$$

where $N^{(v)}$ denotes the number of validation recordings and $\alpha_j^{(s)}$ denotes the value of $\alpha^{(s)}$ for the j^{th} validation recording.

Final test performance was compared across the three models using $\alpha^{(a)}$, $\alpha^{(s)}$, sensitivity, and specificity (with the latter two used as additional sleep state measures and defined in section 4.6.2). The brain-age prediction model of chapter 6 was also used as a comparison to the end-to-end models. As this approach utilized the CLASS algorithm (of chapter 5) to perform sleep staging, followed by an RF

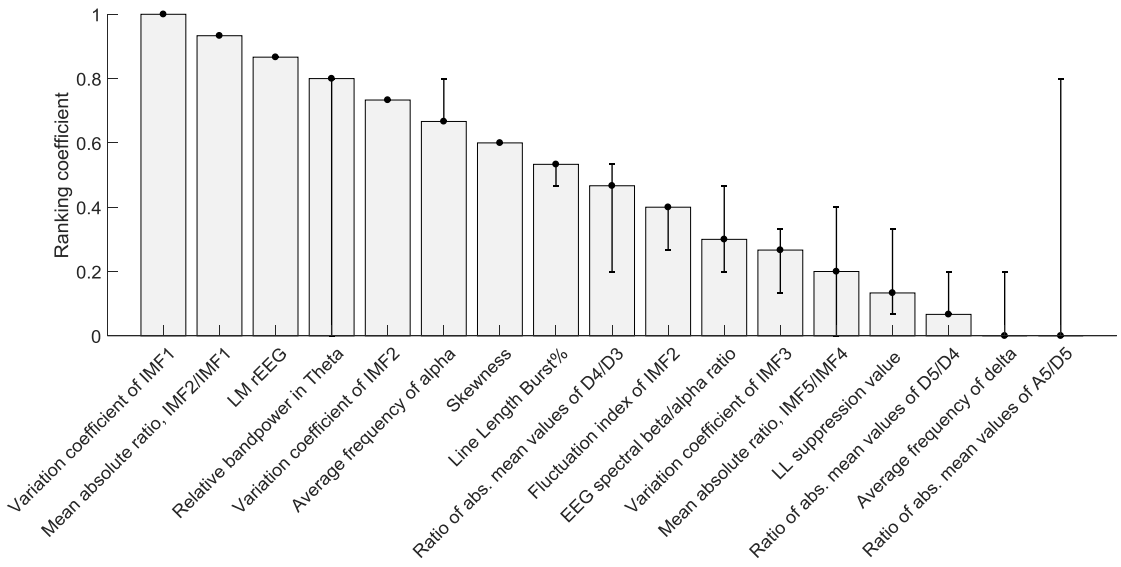


Figure 8.3: Median ranking coefficients for the top 15 features selected at each outer-most fold by mRMR-Spearman. Error-bars denote the range.

regression model for brain-age prediction, the combined method is referred to as 'CLASS+RF' here. Sleep-state performance was also directly compared to CLASS.

Finally, the ability to classify normal and abnormal sleep-wake cycling was demonstrated on a few normal and abnormal recordings at term-age. These recordings were selected from dataset 2 (defined in chapter 3) by the same clinicians who visually labelled the data for sleep. Recordings were selected for good quality and minimal artefacts. This was the same dataset used to identify abnormal outcome trajectories in chapter 6.

8.8 Results

8.8.1 Feature performance

A total of 15 features were selected for each outer-most fold using mRMR-Spearman. As was first calculated in chapter 7 (see section 7.6.1), the median ranking coefficients across folds for each selected feature was calculated and plotted in figure 8.3. Error-bars denote the range.

17 different features appeared across folds, showing a consistency in the features selected during nested CV. Furthermore, the majority of the first eight features

showed small errors suggesting a consistent ranking across folds. In particular, the features 'variation coefficient of IMF 1', 'mean absolute ratio in IMF2/IMF1', and 'LM rEEG', were selected as the top three in every fold. The fourth most prominent feature, 'relative bandpower in theta', exhibited an unusually large error as it was either highly selected (in eight out of ten folds) or not at all. In these cases, the feature 'ratio of absolute mean values of A5/D5' was selected instead.

Through the grid search, the aGMM performed best with a median P=9 (IQR 6-9) features taken across folds, and a median of 3.5 components (IQR 3-4) taken across both folds and sleep states. The aHMM-GMM performed best with median P=10 (IQR 9-12) features and median 3.5 components (IQR 2-4).

Figure 8.4 shows the maximum validation performance for both the aGMM and aHMM-GMM, as the number of features are varied, using the combined assessment measure (I) defined in equation (8.22). This was calculated as previously described in chapter 7. The aGMM showed a clear peak performance between six to eight features. However, there was great variability between folds and performance dropped after eleven features. On the other hand, the aHMM-GMM plateaued, with a similar peak performance from six features onwards and consistent performance between folds.

The number of features and components required by the aDP-HMM-GMM varied based on the choice of δ and is described below.

8.8.2 Varying δ in the aDP-HMM-GMM

Figure 8.5 shows the performance for five instances of training the aDP-HMM-GMM with δ values of 0.01, 0.1, 1, 10, and 100. Figure 8.5(a) shows both the $\alpha^{(a)}$ and $\alpha^{(s)}$ agreements across δ. As $\alpha^{(s)}$ is determined for each test recording, medians and IQRs are presented here. In figure 8.5(b), the resulting number of components across δ are shown for each sleep state, with medians taken across the folds and age groups.

As can be seen, increasing δ resulted in larger numbers of generated components, particularly for δ>1. Towards δ=100, the number of components for QS increased in an almost exponential fashion. However, performance for both brain-age and sleep state prediction remained relatively robust and insensitive to such a large

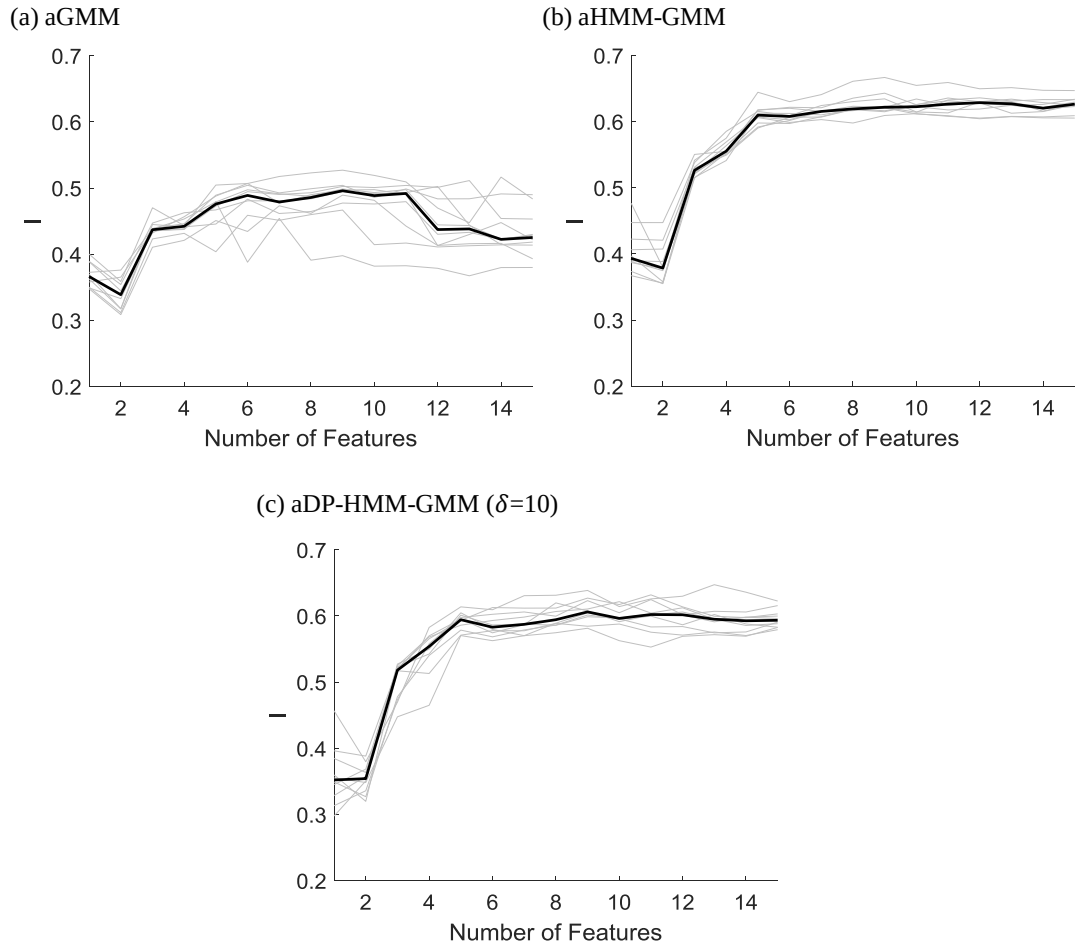


Figure 8.4: (a) I performance curves for the aGMM as the number of features are increased (up to 15), assessed on the validation data. Grey curves denote performance at each fold, with the black curve denoting the median performance across folds. (b) Corresponding performance curves for the aHMM-GMM. (c) Corresponding performance curves for the aDP-HMM-GMM.

variation in δ. As δ=10 showed highest performance, this model was chosen for subsequent comparisons. This is more clearly shown from the data underpinning figure 8.5(a), presented in table B.5 (of appendix B.3).

At δ=10, the aDP-HMM-GMM performed best, with a median P=9 (IQR 8-9) features across folds, and a median of six components (IQR 5-9.5) across folds, sleep states and age groups. Figure 8.4(c) shows the performance of this model as the number of features vary. This proved similar to the aHMM-GMM in figure 8.4(b), showing stability and consistency across folds.

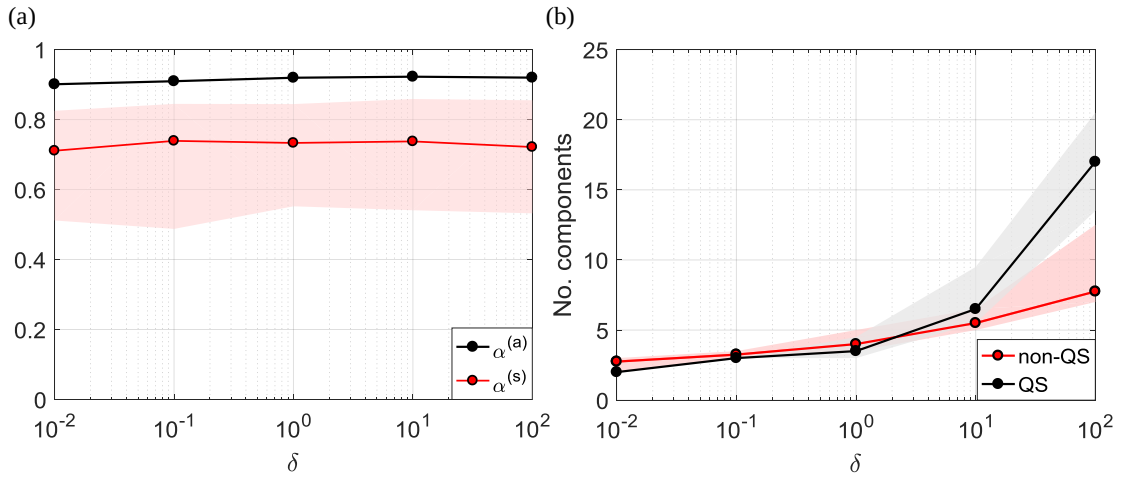


Figure 8.5: (a) $\alpha^{(a)}$ and $\alpha^{(s)}$ performance of the DP-HMM-GMM for different values of δ. As $\alpha^{(s)}$ is determined for each test recording, medians and IQRs are shown for this measure. (b) The change in the number of components with δ for each sleep state. Medians and IQRs are taken across the folds and age groups.

8.8.3 Fitting GMMs by EM & Neal's Algorithm 8

To further illustrate the differences between fitting data for the aDP-HMM-GMM by Neal's Algorithm 8 and the aHMM-GMM/aGMM models by the EM algorithm, the density distributions, $p(\boldsymbol{x}_i|s_i, a)$, for one CV fold is shown in figure 8.6. Figure 8.6(a) shows the distributions after aDP-HMM-GMM training for QS, across all six age groups. Figure 8.6(b) shows the corresponding distributions for the aHMM-GMM. In both cases, this is plotted for the top two selected features of 'variation coefficient of IMF 1' and 'mean absolute ratio in IMF2/IMF1'.

Both methods identified the same dense central regions where the majority of the data was located. However, components were fitted quite differently between methods to the sparser regions of data. In the aDP-HMM-GMM model, components were generally more tightly fitted to the extremer data points, while the aHMM-GMM had a tendency to fit single, broader components to these regions.

8.8.4 Comparing overall model performances

Table 8.1 presents the overall age and sleep state performance for the aGMM, aHMM-GMM and aDP-HMM-GMM. For accurate comparison, CLASS+RF was trained on the same dataset using LOSO CV and the continuous brain-age estimates

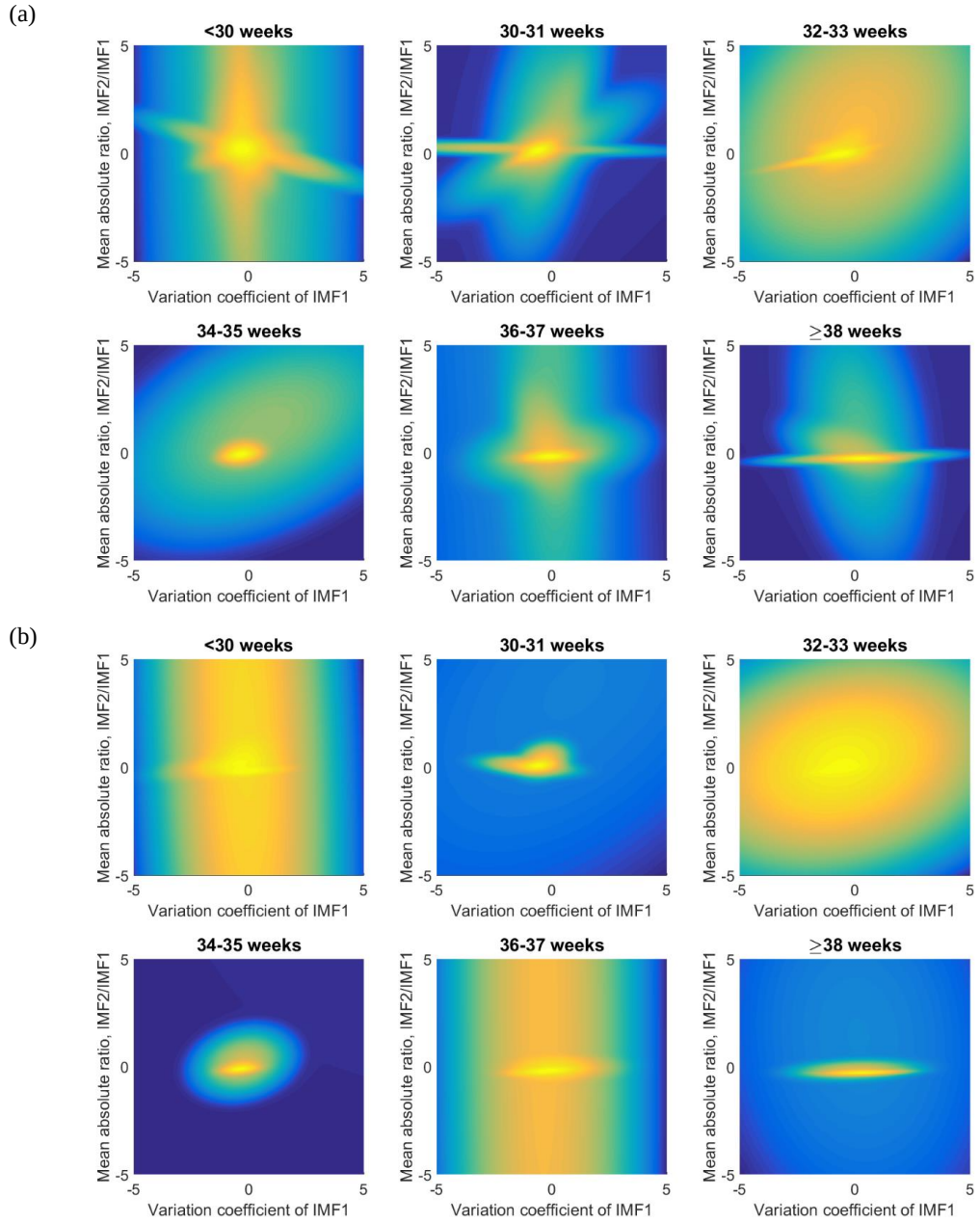


Figure 8.6: (a) Density distributions, $p(\boldsymbol{x}_i|s_i, a)$, for the aDP-HMM-GMM using Neal's Algorithm 8. Each plot shows one of the six age groups during QS from one CV fold, and plotted using the top two features: 'Variation coefficient of IMF 1' and 'mean absolute ratio in IMF2/IMF1'. (b) The corresponding density distributions in the aHMM-GMM using the EM algorithm. A log-range colourbar is applied.

Table 8.1: Comparing overall brain-age and sleep state performance between the aDP-HMM-GMM ($\delta = 10$), aHMM-GMM, aHMM-GMM (K_{s_i}=1), aGMM, and CLASS+RF ($mtry = 200$) models. Assessment measures compared are $\alpha^{(a)}$, $\alpha^{(s)}$, sensitivity (Sens) and specificity (Spec). Medians are shown with IQR in brackets. †denotes that sleep staging performance was compared to the CLASS algorithm. *denotes statistical significant differences at p<0.05 between the aDP-HMM-GMM and other models using the Wilcoxon signed-rank test. For brain-age, this was performed directly on the predicted values and, for sleep states, between the assessment measures for each recording.

Model	$\boldsymbol{\alpha^{(a)}}$	$\boldsymbol{\alpha^{(s)}}$	**Sens (sleep)**	**Spec (sleep)**
aDP-HMM-GMM ($\delta = 10$)	0.92	0.74 (0.54-0.86)	0.90 (0.80-0.96)	0.94 (0.85-0.97)
aHMM-GMM	0.93	0.76 (0.54-0.85)	0.91 (0.73-0.96)	0.94 (0.87-0.97)
aHMM-GMM (K_{s_i}=1)	0.87	0.41* (0.00-0.69)	0.97 (0.82-0.99)	0.74* (0.35-0.94)
aGMM	0.91	0.60* (0.38-0.74)	0.74* (0.54-0.84)	0.92 (0.86-0.96)
CLASS+RF ($mtry = 200$)†	0.92	0.69* (0.50-0.79)	0.98* (0.92-1.00)	0.86* (0.78-0.91)

of the regression categorized to match the same groupings used here. Based on the optimal model performance seen in chapter 6, a parameter value of $mtry = 200$ was selected. When comparing sleep performance to the CLASS algorithm, the sample-based estimates of CLASS were converted to the corresponding 30 second epochs (as was also performed when comparing CLASS to the term-age models in chapter 7).

As an additional comparison, the aHMM-GMM was also applied when using a single component for both sleep states (K_{s_i}=1). The model now simply fits single Gaussian distributions to the data and these results are also listed in table 8.1.

The Wilcoxon signed-rank test (defined in section 4.6.4) was used to assess statistically significant differences between the aDP-HMM-GMM and the other models. For brain-age, this was performed directly on the predicted values. For sleep states, this was performed on the assessment measures for each recording. Measures that showed significant differences at the p<0.05 level are also indicated in table 8.1.

When comparing the aDP-HMM-GMM and the aHMM-GMM models (in their entirety), there were no statistically significant differences between the two models, and both were the best performing for brain-age prediction ($\alpha^{(a)}$=0.92 and 0.93, respectively). Sleep state performance also performed best with these two models ($\alpha^{(s)}$=0.74 and 0.76, respectively). However, when comparing both models to the aHMM-GMM for K_{s_i}=1, much poorer performance was seen, particularly for

sleep classification, with $\alpha^{(s)}$=0.41 and statistically significantly different. This comparison highlights the importance of developing and extending models that focus on fitting multiple Gaussian components to the data.

Across all models, there was in fact no statistically significant differences in $\alpha^{(a)}$ performance. However, the aGMM did have statistically significantly poorer sleep-state performance, compared to the aDP-HMM-GMM, for $\alpha^{(s)}$ (0.60 compared to 0.74) and sensitivity (0.74 compared to 0.90).

All sleep measures for CLASS were also statistically significantly different. CLASS had a higher sensitivity (0.98) than the aDP-HMM-GMM (0.90), but a lower specificity (0.86 compared to 0.94) and $\alpha^{(s)}$ (0.69 compared to 0.74). The comparison with $\alpha^{(s)}$ indicated that the aDP-HMM-GMM out-performed CLASS.

8.8.5 Comparing brain-age classification performance

Figures 8.7(a)-(d) show the age confusion matrices (between estimated brain-age and PMA) for the aDP-HMM-GMM, aHMM-GMM, aGMM, and CLASS+RF models, respectively.

Misclassifications into adjacent age groups were deemed acceptable due to similarities in EEG properties near the age group boundaries. These acceptable regions are shaded in grey in figure 8.7. Any misclassifications beyond adjacent regions are shaded in red.

All four models classified the majority of data within the acceptable region. There was one recording at <30 weeks PMA that was consistently misclassified. Upon further inspection, the quality of this recording was found to be particularly poor, with comparatively fewer clean epochs and a higher number of missing channels, when compared to the remaining recordings.

The aGMM and CLASS+RF additionally misclassified two recordings and one recording, respectively, outside the acceptable region. Overall, CLASS-RF classified recordings better than the other models at younger ages, but performed worse when approaching term age. Performance reflected what was also seen in

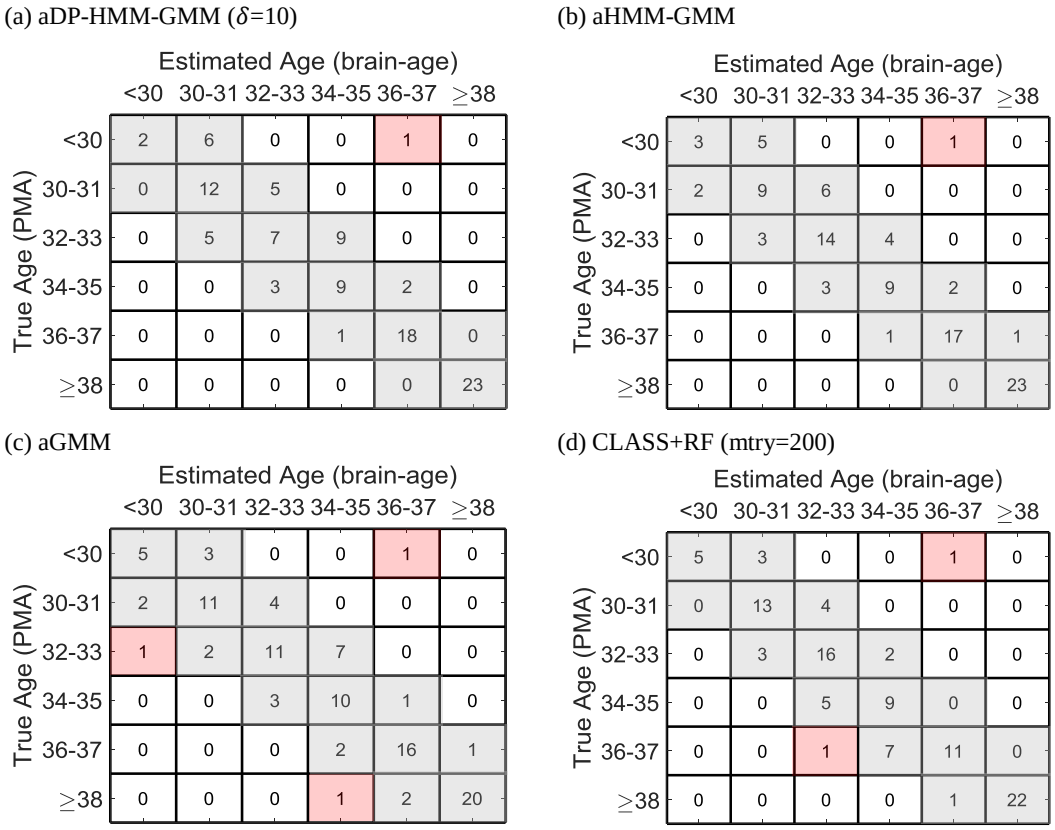


Figure 8.7: Confusion matrices for brain-age classification of recordings across PMA groups. The acceptable classification region includes the adjacent age groups and are shaded in grey. Any misclassified recordings outside this region are shaded in red. (a) The aGMM. (b) The aHMM-GMM. (b) The aDP-HMM-GMM ($\delta = 10$). (d) The CLASS+RF model with $mtry = 200$. The continuous regression estimates are categorized by age group.

chapter 6, with the RF regression model showing a slight bias towards younger age estimates across PMA.

Only the aDP-HMM-GMM and aHMM-GMM models classified term-age recordings perfectly at ≥38 weeks PMA and there was very little difference in performance between these two models from 34 weeks PMA onwards. At 32-33 weeks PMA, the aHMM-GMM classified more recordings perfectly (along the main diagonal) with the aDP-HMM-GMM over- and under-estimating brain-age into adjacent groups (although still within the acceptable region). On the other hand, the aDP-HMM-GMM was superior to the aHMM-GMM in classifying recordings at 30-31 weeks PMA.

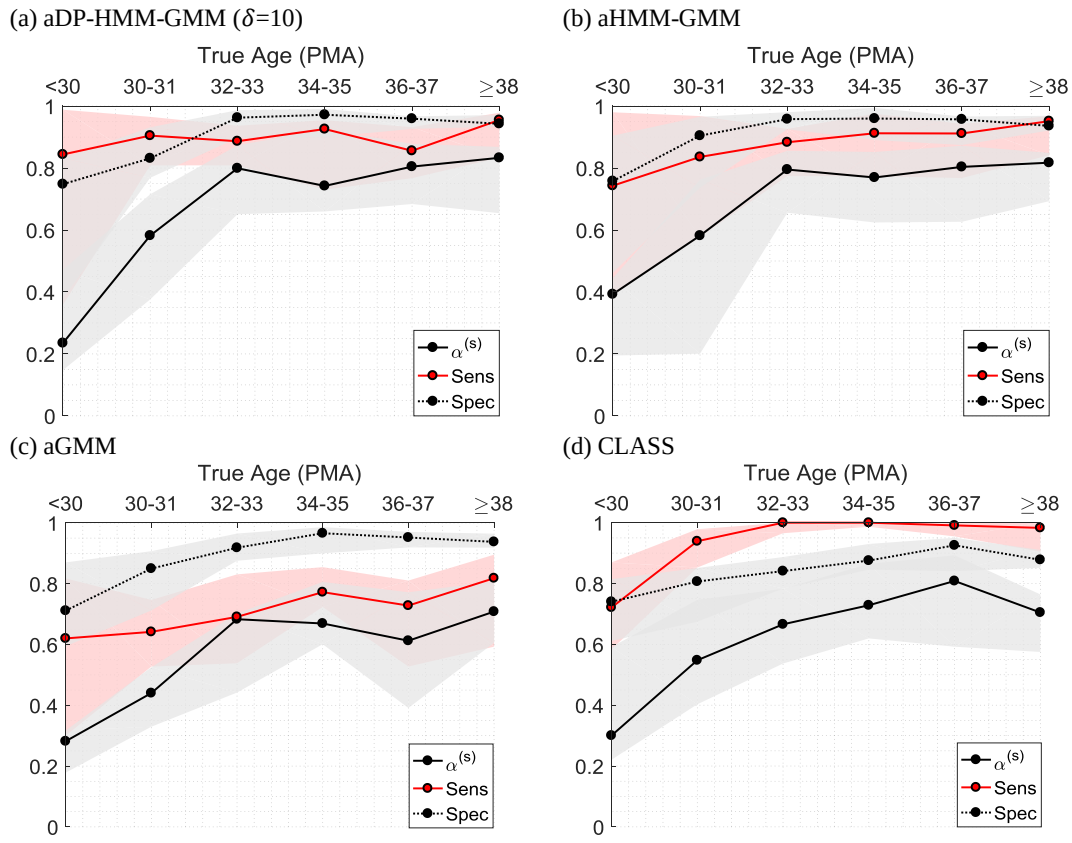


Figure 8.8: Sleep state classification performance across PMA groups (in weeks), assessed using $\alpha^{(s)}$, sensitivity, (Sens) and specificity (Spec). Median values are shown, with shaded regions denoting the IQRs. (a) The aGMM. (b) The aHMM-GMM. (b) The aDP-HMM-GMM ($\delta = 10$). (d) The CLASS algorithm.

8.8.6 Comparing sleep state classification performance

Figures 8.8(a)-(d) present the sleep state classification performance for each model, including the CLASS algorithm. Median results are compared for $\alpha^{(s)}$, sensitivity, and specificity across PMA, with shaded regions denoting the IQRs.

Across all models, poorest performance was observed in the youngest age group of <30 weeks PMA. Performance then improved from 30-31 weeks PMA onwards, with this improvement higher in the aDP-HMM-GMM and aHMM-GMM models. In both these models, median sensitivity and specificity remained consistently >0.8, with both values peaking at ≥38 weeks PMA (above 0.9).

In comparison, both the aGMM and CLASS had poorer overall performance at ≥38 weeks PMA ($\alpha^{(s)}$=0.71 and 0.70, respectively), compared to the aDP-HMM-GMM and aHMM-GMM ($\alpha^{(s)}$=0.83 and 0.82, respectively). Furthermore, the

comparatively higher sensitivity and lower specificity in CLASS agreed with the expected performance observed in chapter 5.

The aHMM-GMM showed a smoother general performance over PMA than the aDP-HMM-GMM (although both remained consistently high). However, at 30-31 weeks PMA, the $\alpha^{(s)}$ IQR was clearly greater in the aHMM-GMM (0.20-0.76) than the aDP-HMM-GMM (0.38-0.72), while median values remained at both 0.58. This was also the age group where the aHMM-GMM showed poorer brain-age prediction performance, as shown in figure 8.7.

8.8.7 Assessing sleep-wake cycling

As the aDP-HMM-GMM model avoids the need to explicitly select K_{s_i} and achieved similarly optimal performance to the aHMM-GMM, it was the ideal model to further illustrate the potential for classifying normal and abnormal sleep-wake cycling.

The aDP-HMM-GMM was first re-trained once with all the data previously used for the nested CV. A value of P=9 was chosen, as this was most commonly selected across the CV folds (in four out of ten cases). The trained model generated a median 5.5 components (IQR 6-8) and was then tested using the additional term-age recordings from dataset 2. A total of six recordings were selected, three with normal outcome and three with abnormal outcome. Figure 8.9 shows the estimated hypnograms of the sleep-wake cycling for these additional recordings.

All three normal-outcome recordings in figures 8.9(a)-(c) showed a regular sleep-wake cycling pattern between QS and non-QS. The recordings in figures 8.9(a) and (b) were also correctly classified by the model as brain-age group 6 ($\geq$38 weeks), indicating a differentiated sleep-wake cycling pattern. Figure 8.9(c) was classified in brain-age group 5 (36-37 weeks), however, this was still acceptable due to the regularity of the sleep-wake cycling and the PMA of 38 weeks falling near the boundary between groups 5 and 6.

In the case of the abnormal recordings, figure 8.9(d) was correctly classified in brain-age group 6, suggesting that the sleep states were now differentiated. However, the sleep-wake cycling remained poor, with reduced periods (proportions

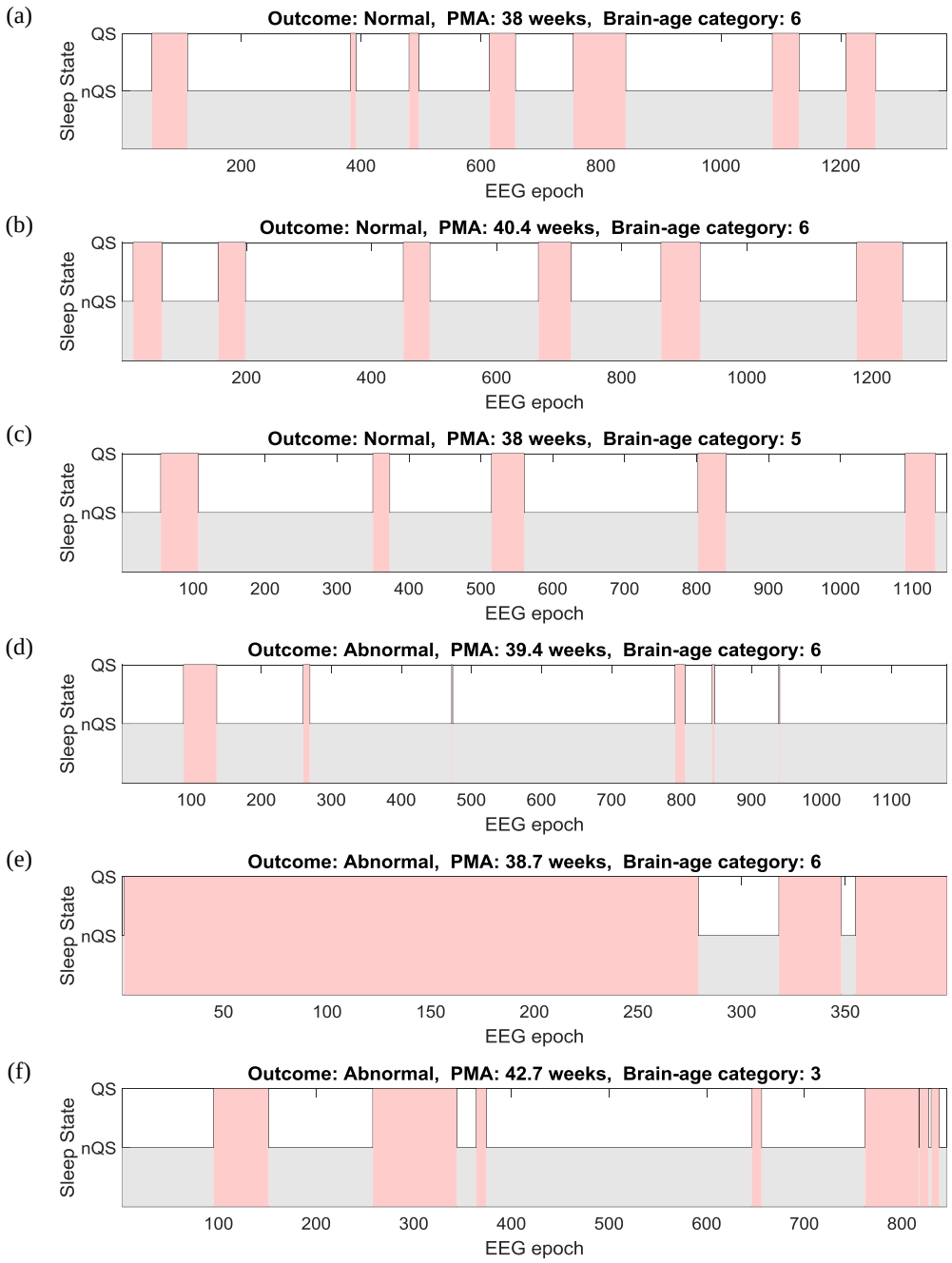


Figure 8.9: Estimated hypnograms by the aDP-HMM-GMM for QS and non-QS (nQS) in six EEG recordings. In each case, the PMA and estimated brain-age group (category) is also given. Recordings in (a)-(c) are from patients with normal outcome and (d)-(f) from patients with abnormal outcome.

and durations) of QS. Figure 8.9(e) was similar, however, it instead exhibited poor sleep-wake cycling through unusually protracted periods of QS. Figure 8.9(f) also showed an irregular (but less extreme) sleep-wake cycling pattern. Most noticeably, this recording was classified as brain-age group 3 (32-33 weeks), suggestive of a strongly undifferentiated (dysmature) EEG behaviour. In fact, this particular baby was lost-to-follow-up before 9 months of age.

8.9 Discussion

The consistent selection of the three features 'variation coefficient of IMF1', 'mean absolute ratio of IMF2/IMF1', and 'LM rEEG', reveals the existence of robust metrics that can simultaneously behave as both sleep state and maturational biomarkers. The prominence of features based on IMFs 1 and 2 shows the benefit of accounting for the non-stationary nature of the EEG. The 'variation coefficient of IMF1' represents the variability of higher frequency EEG oscillations and maintains strong differences over PMA between the burst-inter-burst states of QS, and the relatively more continuous EEG of AS [41], as described in chapter 2.

'Mean absolute ratio of IMF2/IMF1' also showed strong maturational trends in chapter 6 (when extracted for QS), as did the 'variation coefficient of IMF1' and 'LM rEEG' (from both QS and non-QS). 'Variation coefficient of IMF1' from both QS and non-QS were also strong outcome-sensitive maturational biomarkers. This is evident from tables B.1-B.3 in appendix B. Furthermore, 'LM rEEG' was prominent for the multi-state classification of term-age sleep, as shown in chapter 7, with a sensitivity to changes in IBI behaviour across states. It also proves affective here for detecting the reducing IBI lengths with increasing PMA [71]. These comparisons further build on chapter 7, showing that specific features can behave as both sleep-state and maturational biomarkers. These developed end-to-end models are dependent on such features that can simultaneously capture both these effects.

The best performing models for both sleep state and brain-age estimation were the aDP-HMM-GMM and the aHMM-GMM. When comparing model fitting, there remains no significant difference in overall performance between these two models,

even as δ is increased in the aDP-HMM-GMM (resulting in larger numbers of components). This is due to the combined density distributions capturing the majority of the data well, regardless as to whether components are fitted by Neal's Algorithm 8 or by the EM algorithm.

This said, there remains some fundamental differences in model fitting using these two methods. EM fits components in an ML sense with no prior knowledge. On the other hand, Neal's Algorithm 8 assigns components and parameters from prior distributions. Inference in the aDP-HMM-GMM becomes increasingly confident as more data is included. However, when there is limited data or a presence of outliers, parameter draws are more strongly influenced by the uninformative prior and can result in the generation of new components with large variances [200, 230]. This seems to have a limited effect overall, but explains the slightly poorer brain-age estimation of the aDP-HMM-GMM at 32-33 weeks PMA, when compared to the aHMM-GMM.

However, the ability of the aDP-HMM-GMM to implicitly choose the number of components from the data allows greater flexibility in the model fitting across both age and sleep state. This is more difficult when fitting the aHMM-GMM by EM, as components are chosen as part of a grid search and only varies with sleep state. This can underfit or overfit the model at certain age groups, and explains the slightly poorer performance of the aHMM-GMM at 30-31 weeks PMA, compared to the aDP-HMM-GMM.

In terms of sleep state classification, the excellent performance of both models (particularly at term-age) is due to the incorporation of state transitions. As was the case in chapter 7, probabilities of self-transitions are very high and have a resulting 'smoothing' effect on the sleep-state estimates, reducing cases of spurious estimates. Without considering transitions, the subtle differences in the fitted distributions between QS and non-QS increase the chance of misclassification. Even with transitions, the persistently low performance at <30 weeks is due to the rudimentary nature of sleep states at this age which remains difficult to classify from EEG alone [19, 41, 71] (as discussed in chapter 2).

The comparatively worse performance of the aGMM is due to the independent-epoch assumption, whereby sleep state transitions are not considered. While the effect is less detrimental for brain-age prediction, poorer sleep state performance across PMA is clearly affected by the presence of spurious estimates.

CLASS+RF brain-age prediction is expectedly high, as it is based on a regression model. However, this model requires a larger proportion of the feature set to achieve this prediction performance (see chapter 6). In contrast, these end-to-end models use no more than eleven features in any fold. Even by incorporating more features, the inherent bias in the RF towards underestimating brain-age causes misclassification at the oldest age groups. Reduced reliability at these ages by CLASS+RF is also due to first performing sleep staging using CLASS, which doesn't classify sleep in term-age recordings as well as the aDP-HMM-GMM and aHMM-GMM.

Applying the aDP-HMM-GMM to the additional dataset of normal and abnormal outcomes, reveals that this model could potentially classify sleep-wake cycling, while avoiding the need to explicitly select the number of components for each sleep state and age group (unlike with the aHMM-GMM). Analysing sleep-wake cycling across normal and abnormal outcomes remains only an illustration, as there is no visually labelled ground truth for these additional recordings (see chapter 3). However, it still shows examples of clear differences in sleep quality between normal and abnormal outcomes, as well as identifying cases of undifferentiated (dysmature) sleep development at term age. While future work will require visually labelled abnormal data to more stringently quantify the sleep-wake cycling, this model already shows promise as a flexible, end-to-end diagnostic tool for the NICU setting.

8.10 Conclusion

The key novel contributions outlined in this chapter are:

- The development of novel end-to-end probabilistic models that simultaneously predict sleep state and brain-age from preterm EEG recordings (over a wide PMA range) to assess sleep-wake cycling and quantify brain maturation.

- Specifically, a Bayesian non-parametric aDP-HMM-GMM model that can more easily be generalized to new datasets by avoiding the need to explicitly pre-select the number of GMM components for each sleep state and age group.
- Particularly for term-age recordings $\geq$38 weeks PMA, a method where brain-age can additionally classify the sleep-wake cycling to identify cases where sleep remains undifferentiated (dysmature).

With this chapter demonstrating the great potential of these models as diagnostic tools, future work will aim to expand these approaches for greater generalizability and robustness, with further clinical validation using larger labelled and abnormal outcome datasets. Potential avenues are discussed as part of chapter 9.

9
Conclusions & Future Work

The overarching aim of this research was to develop fully automated methods to quantify brain maturation in the preterm baby, from EEG sleep, specifically focussing on developing diagnostic tools for the early prediction of neurodevelopmental outcome. This thesis presented a series of solutions for meeting this aim, focussing on the specific goals that were first outlined in chapter 1 and reiterated here:

1. To develop robust methods for automated EEG sleep staging, essential to achieve fully automated approaches for quantifying brain maturation.

2. To identify relevant EEG maturational biomarkers to estimate sleep-dependent developmental trajectories and identify early deviations linked to long-term abnormal neurodevelopmental outcomes.

3. To develop automated solutions to assess EEG sleep-wake cycling, as an alternative means to quantify brain maturation and identify developmental deviations.

4. To develop such methods using both normal and abnormal outcome data, allowing a more objective and practical assessment of their clinical potential as diagnostic tools in the NICU.

This chapter first summarizes the key results and contributions of this thesis, in relation to these goals. Avenues for future work are then presented which stem from the remaining limitations of this research.

9.1 Summary of results & contributions

9.1.1 Automated sleep staging

In chapter 5, existing concepts were expanded and validated on a large preterm EEG dataset to produce the robust QS (and non-QS) detection algorithm, CLASS. Specifically, the introduction of the artefact removal method, ASR, ensured that this model performed consistently well at >31 weeks PMA. Overall performance resulted in a median sensitivity, specificity and AUC of 0.97, 0.82, and 0.97, respectively, outperforming the existing unsupervised SAT% algorithm (of median AUC=0.85). Particularly, the SAT% algorithm more accurately captured vigilance state cyclicity, and was unable to fully detect QS at older ages. CLASS also performed similarly to the state-of-art supervised CNN of *Ansari et al.* [100], but allowed the classification of sleep without pre-training on large datasets. This ensured that data could be carried forward (without bias) in subsequent analysis to quantify brain maturation (as in chapter 6) and that the methods developed in chapter 6 were fully-automated from start to finish.

Chapters 7 and 8 also presented supervised methods to classify sleep at term-age and across wider preterm ages, respectively. Particularly, chapter 7 presented the first method for multi-state sleep classification of the differentiated sleep states at term-age. Most importantly, it demonstrated that this could be achieved using only EEG. Traditional probabilistic GMM and HMM-GMM models were proposed, with the HMM-GMM achieving the highest overall median κ performance of 0.61, through modelling the transitions between sleep states.

9.1.2 EEG biomarkers & trajectory estimation

By incorporating abnormal data, chapter 6 utilized a data-driven approach to feature extraction, uncovering novel maturational biomarkers including those that showed additional sensitivity to neurodevelopmental outcome. With the large number of features extracted, the different chapters focussed on interpreting the most strongly significant few with respect to sleep state and/or age. When comparing across the larger lists of statistically significant features and those prominently identified

by feature selection, there were consistent trends in the biomarkers chosen for maturation, sleep state, or both.

Features based on complexity (namely MSE), rEEG, LL Burst%, as well as non-stationary EMD- and DWT-based measures, were particularly strong candidates. For example, 'mean absolute ratio of IMF2/IMF1' captured shifts away from high oscillatory periods as the EEG transitioned from TD to TA states with PMA (defined in chapter 2), while also a prominent biomarker to assess sleep-wake cycling in chapter 8. EMD and DWT features were also most prominent in showing differences between normal and abnormal outcome groups (including mild outcomes). This research uncovered the importance of better capturing non-stationary EEG behaviours instead of relying solely on well-established features that were derived from the DFT (using the incorrect assumption that the EEG is a stationary signal).

Combining novel (sleep-dependent) maturational biomarkers next resulted in a fully-automated brain-age prediction model, based on RF regression, that could provide estimates of the developmental trajectory for each patient. When validated on normal outcome data, it produced a peak performance of $\sqrt{MSE}$=0.85 weeks, which was a great improvement on existing methods [29, 145, 146] (as discussed in chapter 2, with minimum error of 1.49 weeks). Furthermore, the method was able to predict abnormal outcome from the deviating trajectories (due to dysmature and/or disorganised patterns), with a peak accuracy of 61%. Incorporating abnormal outcome for the first time develops this approach towards a practical diagnostic tool.

9.1.3 Assessing sleep-wake cycling

The automated sleep staging of differentiated term-age sleep in chapter 7 opened the possibility of assessing sleep-wake cycling during this period. Chapter 2 identified how the quality of sleep differentiation, particularly the increasing proportions of HVS, were important indicators of maturation. By successfully classifying these sleep states at >37 weeks PMA in chapter 7, the developed HMM-GMM model provides a tool to clinically assess changes in sleep state duration, proportion, and the general quality of cyclicity. This was particularly crucial as existing automated

sleep-wake cycling methods focus on single BAC measures which could not fully capture this information at term-age.

By extending these GMM and HMM-GMM models for age-dependency, chapter 8 proposed novel, end-to-end approaches that could simultaneously classify sleep states while also classifying the overall sleep-wake cycling (through estimating brain-age). Specifically, the aDP-HMM-GMM was the first Bayesian non-parametric application to preterm EEG analyses and provided means to generalize the models to new datasets (without explicitly selecting the number of GMM components). Overall, it produced a near-perfect median $\alpha^{(a)}$=0.92 for brain-age classification, and median $\alpha^{(s)}$, sensitivity, and specificity of 0.74, 0.90, and 0.94, respectively, for sleep state classification.

Furthermore, the aDP-HMM-GMM was illustrated on some clinically pre-selected EEG recordings of both normal and abnormal outcome. This demonstrated the ability of the model to identify regular sleep-wake cycling and brain-age in normal outcome recordings. It also demonstrated cases where sleep-wake cycling (proportions, durations, etc.) were very poor in abnormal outcome recordings, as well as cases where the brain-age was lower than expected. As all these recordings were of term-age, unusually low brain-age estimates were indicative of an undifferentiated (dysmature) sleep pattern.

9.2 Future work

9.2.1 Data quality

The datasets included in this research and collected from the UZ Leuven Hospitals, Leuven, Belgium, are currently one of the largest preterm EEG datasets available in Europe. Care was taken through this research to minimize the removal of EEG recordings, with none subsequently excluded (after clinical pre-screening) on the grounds of adverse data quality. Very long recordings were incorporated (compared to many previous studies), however, limitations consequently arose with respect to data quality.

Development of CLASS in chapter 5 successfully incorporated ASR to remove channel-specific high power artefacts using a single calibration sequence (from a term-age recording). However, as first described in chapter 6, using a single calibration sequence to remove artefacts across a wide PMA range may alter the subtler EEG behaviours. Consequently, ASR was not included as part of feature extraction in chapters 6, 7 and 8. Adapting ASR to robustly select calibration sequences, specific to the PMA of the recording, may improve it as a method for general preterm EEG artefact removal. This could also improve the quality of feature extraction and general performance of the models outlined in this thesis.

In chapter 7, particular care was taken to select only high quality data for classifying the differentiated sleep states at term age. The additional removal of channel-specific artefacts during pre-processing resulted in an idealized dataset. However, in the NICU, such artefacts will often contaminate the recordings and prevent successful classification. This is particularly true when dealing with a transition-based approach such as the HMM-GMM. Again, it would be advantageous to incorporate artefact removal and including such methods here is the next step towards a robust approach that is suitable for the clinical setting.

Successful artefact removal would also improve the robustness of EEG feature extraction *across channels*, opening the possibility of incorporating functional brain connectivity measures into this data-driven framework. As discussed in chapter 2, such measures also demonstrate strong links with sleep staging and maturation.

9.2.2 Labelling sleep state & outcome

As first stated in the chapter 2, the AASM recently renewed their recommendations for EEG sleep labelling in infants zero to two months of age. However, strict rules for scoring sleep from EEG in premature infants are still lacking and based on expert opinion [19] making the ground truth ambiguous.

To develop the algorithms presented in this thesis, accurate visual labelling was required. Cerebral and non-cerebral signals were used in combination to increase this accuracy (as described in chapter 2) but the ambiguity of the ground truth

and the use of only a single primary rater limits reliability. This is a common issue seen in other neonatal sleep classification studies [32, 86, 227] and incorporating multiple independent EEG raters will further improve the reliability of preterm (and full-term) EEG sleep interpretation. Recently, the NeoGUARD group (coordinated by the NICU in UZ Leuven) has developed a method to improve the inter-rater scoring of seizure detection in neonates [232]. There is hope that this will pave the way for a similar approach to EEG sleep labelling.

The time and expertise required to visually label each recording, particularly for more than two EEG sleep states (as in chapter 7), results in a limited data size. A larger labelled dataset is likely to further improve classifier performance and generalizability, though this limitation underpins why there is a need for automated methods in the first place. Furthermore, with the present lack of visual sleep labelling in abnormal outcome datasets, only an illustration of the aDP-HMM-GMM for assessing sleep-wake cycling was possible in chapter 8. Additional visual labelling of these critical datasets can lead to a more in-depth validation of these models, through specifically quantifying and comparing differences in sleep state proportions, durations etc. This would further require time-matched studies between cohorts and has been previously achieved in the NICU [23, 24, 35, 36].

Finally, the BSID-II scores used for measuring neurodevelopmental outcome remains an indirect ground truth. While it is the clinical gold standard, the delay between NICU discharge and assessment at nine months follow-up (at the earliest) misses a critical period where additional environmental stresses may have an adverse impact on brain maturation. To further improve classification accuracy for identifying abnormal babies in chapter 6, labelling at the recording level is necessary to categorize the precise levels of dysmature and/or disorganized patterns (partly explaining the abnormal outcome classification performance achieved). As discussed in chapter 2, such studies have previously attempted this, predominantly for grading HIE [30, 50, 88, 139–143].

9.2.3 Incorporating additional modalities

As discussed in chapter 1, focussing only on EEG as a modality is beneficial, as it is affordable, non-invasive, and feasible for real-time monitoring in the NICU. Incorporating additional PSG modalities would be counter-productive in this respect, with the exception of ECG which typically accompanies EEG monitoring in the NICU (even with reduced PSGs at younger PMAs).

The present datasets did not include reliable ECG recordings across the full duration of the EEG, however, future incorporation of these signals could facilitate the separation of non-QS into AS and wakefulness, as well as IS (as described in chapter 2). This provides further means to quantify maturation [18] through expanding the developed models to additional states and incorporating novel ECG maturational biomarkers (such as those recently defined for preterm sleep by *Werth et al.* [14]).

9.2.4 Extending the probabilistic models

Achieving a completely generalizable model through the aDP-HMM-GMM (in chapter 8) still remains limited by the selection of P features by grid search. However, a further extension of this Bayesian non-parametric model would be to incorporate the Indian Buffet Process (IBP) [233]. While the CRP (defined in chapter 4) allows for the number of GMM components to be selected from an infinite number of possible values, the IBP models the feature space in a similar fashion. By incorporating an IBP, the number and choice of features would become a further random variable, with an accompanying probability distribution and associated hyperparameter(s) [233]. In that vein, the present hyper-parameters $\boldsymbol{\mu}_0$, λ_0, $\boldsymbol{\Psi}_0$, ν_0, and δ could be further modelled using additional distributions (at the cost of computational time). However, including further hyper-parameters to train may require increasingly larger datasets than what is currently available.

In addition, apart from $\geq$38 weeks PMA, all other brain-age categories for these models were selected arbitrarily to ensure sufficient data for training. However, this inevitably leads to the overlapping boundaries between the age groups and adjacent

misclassifications. Due to the similarity in behaviours at these boundaries, this will always be a drawback of categorizing age. However, a larger data size may further allow the definition of groups with smaller age intervals to reduce this error. With additional data, future work could also aim to group ages beyond 38 weeks PMA to classify the continuing differentiation of the sleep states that occurs up to 48 weeks PMA [41, 71] (as described in chapter 2). This opens the possibility of identifying both delayed *and* accelerated sleep differentiation.

Finally, the extreme extension would be to dispense with age groups completely and develop a probabilistic model based on the 'continuous' PMA (in days), as was used for regression in chapter 6. This would effectively combine the trajectory estimation (in chapter 6) with sleep-wake cycling assessment to achieve a truly unified approach. Such a model would require a shift from using DPMMs for clustering data to a regression-based approach, now estimating the sleep state distributions over a continuous age variable. The Gaussian Process (GP) [234], another Bayesian non-parametric model, could be one potential candidate to achieve this. However, this would again require large datasets that generalize well across the continuous PMA range.

9.2.5 Further avenues for machine learning

Novelty detection

The identification of normal and abnormal developmental trajectories (as in chapter 6) required longitudinal monitoring of the EEG in the NICU. However, EEGs were measured at irregular PMA intervals, dependent on the delicate balance between accurate diagnostics and avoiding unnecessary disturbances to the vulnerable preterm baby. If more regular, time-spaced recordings were achieved in the NICU, combined with a recording-specific labelling of EEG abnormalities (as suggested in section 9.2.2), novelty detection methods could be applied [235].

Novelty detection is a field of machine learning that focusses on identifying abnormal data when such data is limited. Specifically, regression-based methods focus on identifying larger-than-normal trajectory deviations based on specific

'probablistic-based', 'distance-based' or 'domain-based' methods, among others (as defined in [235]). The existing RF-based approach of chapter 6, or aforementioned GP-based extensions, could both be used within this framework.

Is there a role for deep learning?

Deep learning applications have exploded over the past ten years, particularly in the context of computer vision and imaging. In biomedical engineering, these methods have rapidly become the go-to techniques for medical imaging (namely MRI) and is gaining popularity for analysing time-series signals, particularly in EEG sleep staging [236]. It therefore seems pertinent to briefly discuss its potential for quantifying brain maturation.

As discussed in section 2, *Ansari et al.* [100] demonstrated automated QS detection of preterm EEG using CNNs. This confirmed that preterm EEG sleep staging can indeed by classified successfully by deep learning methods, while accounting for the maturation of the EEG characteristics with PMA [100]. Deep learning typically requires large datasets (which are reliably labelled) and this is becoming increasingly achievable for preterm EEG.

With PMA also defined for each EEG recording, deep learning approaches could also focus on estimating brain-age to derive developmental trajectories. However, it is unclear if there would be sufficient data available to perform a continuous estimate of brain-age using such methods. Furthermore, the major drawback of deep learning is the loss of interpretation as the models become progressively 'deeper', with the data transformed in increasingly abstract ways. On the other hand, explicit calculation of EEG features (as in this research) and knowledge of their mathematical definitions allows easier clinical interpretation. Probabilistic models also explicitly describe each random variable where the conditional dependencies can be directly inferred and understood. For these reasons, there remains key advantages of pursuing the approaches developed in this thesis.

Appendices

A

Additional Distributions & Algorithms

A.1 Statistical distributions

A.1.1 The z-distribution

Figure A.1 illustrates the z-distribution. This is simply a standard Normal distribution, $\mathcal{N}(0,1)$, over the z-score (z). z are the standardized values [237]:

$$z = \frac{x - \mu}{\sigma} \tag{A.1}$$

μ denotes the mean and σ the standard deviation of the data.

To assess significance, the p-value (p) is determined which defines the cut-off (critical value) of z where the probability is less than a pre-defined significance threshold. Typically, this threshold is defined at the 95% level, where p=0.05. For p<0.05, critical z values for a two-tailed test are ±1.96, as shown in figure A.1.

A.1.2 The student's t-distribution

The t-distribution is a function, $f(t)$, over a t-statistic (t) and defined as [238, 239]:

$$f(t) = \frac{\Gamma\left(\frac{d+1}{2}\right)}{\sqrt{d\pi} \cdot \Gamma\left(\frac{d}{2}\right)} \cdot \left(1 + \frac{t^2}{d}\right)^{-\frac{d+1}{2}} \tag{A.2}$$

The parameter, d, defines the degrees of freedom and $\Gamma(\cdot)$, the Gamma function. The t-distribution is similar in form to the z-distribution in section A.1.1 (symmetrical and bell-shaped) but has heavier tails that are dependent on the choice of d. Figure A.2(a) shows the t-distributions for various choices of d. When d=100, figure A.2(b) shows the critical values of t that achieves p<0.05 for a two-tailed test.

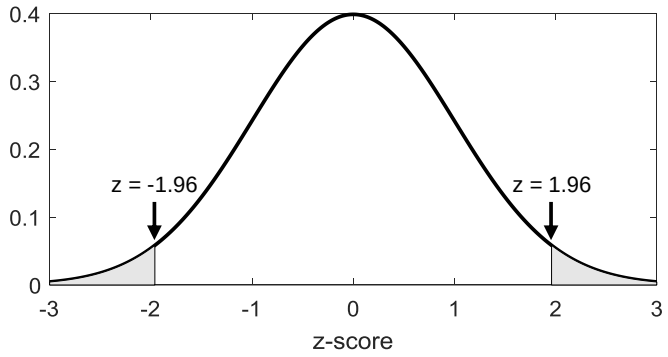


Figure A.1: The standard z-distribution. To assess if a test is statistically significant, the z-score is compared to a critical value for a given p-value. For p<0.05, critical z-scores are z<-1.96 and z>1.96 (as shown).

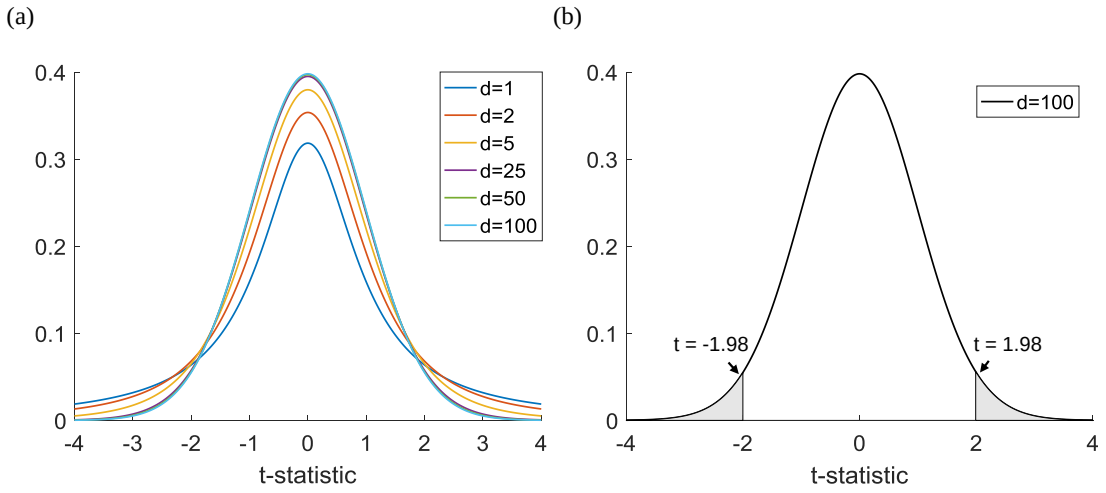


Figure A.2: (a) The t-distribution for various values of d. To assess if a test is statistically significant, the t-statistic is compared to critical values for a given significance level. (b) The t-distribution with d=100. The critical t-statistics for p<0.05 are t<-1.98 and t>1.98 (as shown).

A.1.3 The F-distribution

The F-distribution is a function, $f(F)$, over an F-statistic (F) and defined as [185]:

$$f(F) = \frac{\Gamma\left(\frac{d1+d2}{2}\right)}{\Gamma\left(\frac{d1}{2}\right)\Gamma\left(\frac{d2}{2}\right)} \cdot \left(\frac{d1}{d2}\right)^{\frac{d1}{2}} \cdot \frac{F^{\frac{d1-2}{2}}}{\left(1+\frac{d1}{d2}F\right)^{\frac{d1+d2}{2}}} \tag{A.3}$$

Parameters, $d1$ and $d2$, define the degrees of freedom and $\Gamma(\cdot)$ denotes the Gamma function. Unlike the z- and t-distributions (in sections A.1.1 and A.1.2, respectively), the F-distribution is non-symmetrical (depending on $d1$ and $d2$) and can assume a wider variety of forms, including those similar to exponential decay and beta-

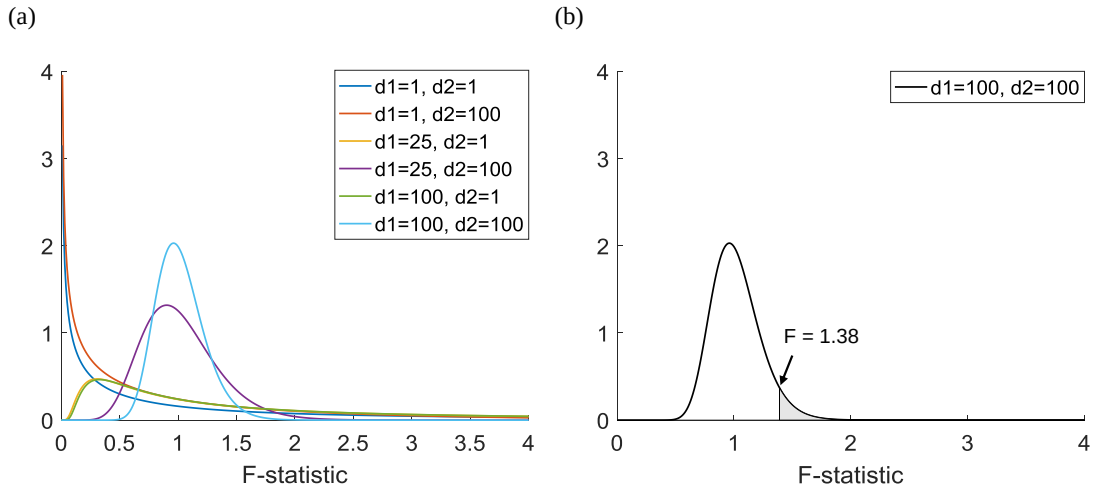


Figure A.3: (a). The F-distribution for various values of $d1$ and $d2$. To assess if a test is statistically significant, the F-statistic is compared to a critical value for a given significance level. (b) The F-distribution with $d1$=100 and $d2$=100. The critical F-statistic for p<0.05 is F>1.38 (as shown).

distributions. It is also positive-definite, suitable for only one-tailed tests.

Figure A.3(a) shows F-distributions for various choices of $d1$ and $d2$. When $d1$=100 and $d2$=100, figure A.3(b) shows the critical value of F that achieves p<0.05.

A.2 Signal decompositions

A.2.1 The Discrete Fourier Transform (DFT)

The DFT represents a signal as a spectrum of its frequency content, through approximating the signal as a sum of potentially infinite sinusoids of varying amplitude, frequency and phase. For a signal, $\boldsymbol{x} \in \mathbb{R}^{1\times N}$ (where N is the length), the DFT, $\mathcal{FT}\{\boldsymbol{x}\}$, is given by [155][1]:

$$\mathcal{FT}\{\boldsymbol{x}\} \rightarrow X_k = \sum_{n=0}^{N-1} x_n \mathrm{e}^{-j\frac{2\pi}{N}kn} \tag{A.4}$$

where x_n is the n^{th} sample in $\boldsymbol{x}$, $j = \sqrt{-1}$, and $k \in \mathbb{N} : 0 \leq k \leq N-1$. This is efficiently implemented in software using the Fast Fourier Transform representation [155].

[1]For consistency with definitions in literature, the first element of $\boldsymbol{x}$ is at n=0 and the DFT of $\boldsymbol{x}$ is at k=0 here.

From the DFT, the periodogram can be defined which is an approximation of the power spectral density of $\boldsymbol{x}$ [155][2]:

$$p_k = \frac{1}{N \cdot Fs}|X_k|^2 \tag{A.5}$$

where Fs is the sampling frequency, X_k is the value of the DFT of $\boldsymbol{x}$ at k, and $k \in \mathbb{N} : 0 \leq k \leq N-1$.

Calculating the periodogram for an entire signal is inherently noisy due to its finite length. As an alternative, the Welch periodogram is calculated to reduce such noise through windowing the signal, deriving the periodogram for each window, and averaging the resulting periodograms together [240]. Most commonly, a Hamming window is selected with 50% overlap and a selected length such that a minimum of eight windows can be fitted over the full signal length [161]. The Welch method additionally converts the periodogram to the one-sided spectrum, covering the applicable frequency range of 0 to $Fs/2$.

A.2.2 The Discrete Wavelet Transform (DWT)

The DFT in section A.2.1 represents a signal as a series of sinusoidal 'basis functions' of varying frequency. Similarly, the DWT represents the signal as a series of wavelet basis functions which are spatially localized (non-zero over a finite period), and derived from dilated, translated, and scaled versions of a 'mother wavelet'. This produces local temporal representations of the signal's frequency content and results in an overall frequency- *and* time-varying signal spectrum [159].

The most efficient approach to apply the DWT and obtain this spectrum is through a cascading filter bank of multiple levels, as illustrated in figure A.4.

At each level, the signal is passed through an orthogonal pair of low-pass and high-pass filters, which evenly split the frequency content into low- and high-frequency components at that level. This is achieved by convolution between the signal and

[2]For consistency with definitions in literature, the first element of the periodogram is at k=0 here.

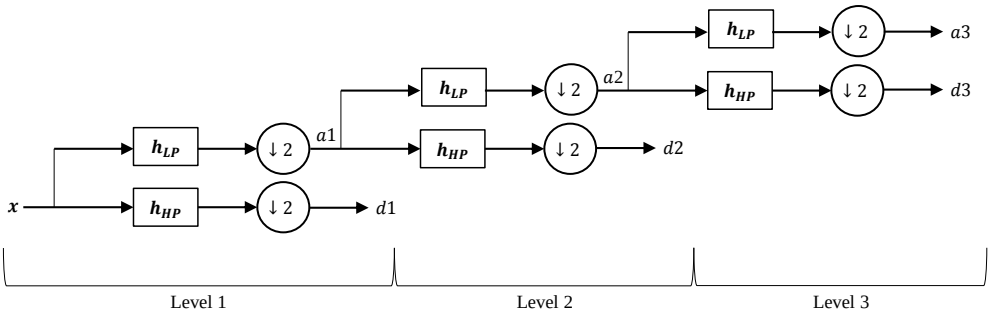


Figure A.4: The cascading DWT filter bank for decomposing a non-stationary signal, $\boldsymbol{x} \in \mathbb{R}$ (of any given length), into frequency bands. Three levels of decomposition are shown here, resulting in the final coefficients, d1-d3 and a3.

filter. Denoting the signal as $\boldsymbol{x} \in \mathbb{R}^{1\times N}$ (of any given length, N) and the filter as $\boldsymbol{h} \in \mathbb{R}^{1\times M}$ (of length $M < N$), each i^{th} sample of the resulting signal is calculated as:

$$y_i = \sum_{k=-\infty}^{\infty} x_k h_{i-k} \tag{A.6}$$

For the low-pass filter, this output is referred to as the Detail coefficient, Dl, for the l^{th} level. Similarly, the output for the high-pass filtered signal is referred to as the Approximation coefficient, Al.

By then downsampling the Approximation coefficient by a factor of two, the above filtering operations are then repeated on the resulting signal to produce Detail and Approximation coefficients for the next level and so on. The result is the decomposition of $\boldsymbol{x}$ into multiple frequency bands. After each level, the frequency resolution is doubled while the time resolution is halved [159].

The choice of low-pass and high-pass filters are determined from the mother wavelet. Various wavelets exist and depend on the application. The Debauchies-4 (db4) wavelet is chosen here due to previous successes in other preterm EEG studies [156, 176]. The db4 form of the orthogonal low-pass ($\boldsymbol{h}_{LP}$) and high-pass ($\boldsymbol{h}_{HP}$) filters are defined as [159]:

$$\boldsymbol{h}_{LP} = \left[\frac{1+\sqrt{3}}{4\sqrt{2}}, \frac{3+\sqrt{3}}{4\sqrt{2}}, \frac{3-\sqrt{3}}{4\sqrt{2}}, \frac{1-\sqrt{3}}{4\sqrt{2}}\right] \tag{A.7}$$

$$\boldsymbol{h}_{HP} = \left[\frac{1-\sqrt{3}}{4\sqrt{2}}, \frac{-3+\sqrt{3}}{4\sqrt{2}}, \frac{3+\sqrt{3}}{4\sqrt{2}}, \frac{-1-\sqrt{3}}{4\sqrt{2}}\right] \tag{A.8}$$

A.2.3 Empirical Mode Decomposition (EMD)

The EMD is similar to the DWT (section A.2.2) in that it decomposes non-stationary signals into a series of components called Intrinsic Mode Functions (IMFs), which represent different frequency bands. This decomposition is conducted entirely in the time-domain through an approach called 'sifting'. This is performed as follows [171]:

1. Connect all the local maxima of the signal together using a cubic spline fit to produce an upper envelope. Similarly, produce the lower envelope by connecting all the local minima together.

2. Determine the mean between the two envelopes and subtract from the original signal to produce the component, $\boldsymbol{h}_1 \in \mathbb{R}^{1\times N}$ (where N is the length of the signal).

3. Check if $\boldsymbol{h}_1$ satisfies the requirements for an IMF, which is defined as:

   (a) Across the whole of $\boldsymbol{h}_1$, the number of extrema and number of zero crossings must differ by at most one.

   (b) At any point, the mean between the upper and lower envelopes of $\boldsymbol{h}_1$ are zero.

4. If $\boldsymbol{h}_1$ satisfies the above conditions, it becomes the first IMF, IMF1. Otherwise, $\boldsymbol{h}_1$ is deemed a 'proto-IMF' and the sifting procedure of steps 1-3 is repeated, with $\boldsymbol{h}_1$ now the signal and a subsequent component, $\boldsymbol{h}_{11}$, derived. Sifting is recursively repeated (producing $\boldsymbol{h}_{12}$, $\boldsymbol{h}_{13}$ etc.) until the resulting component meets the conditions for an IMF (as in step 3).

   The maximum number of sifting steps is limited by a stoppage criterion. The most commonly used was defined by *Huang et al.* [171] where the sum of differences, $SDiff_k$, is defined for the k^{th} sifted component by:

$$SDiff_k = \sum_{i=1}^{N} \frac{|h_{1(k-1)i} - h_{1ki}|^2}{h_{1(k-1)i}^2} \tag{A.9}$$

   where h_{1ki} denotes the i^{th} sample in $\boldsymbol{h}_{1k}$. Sifting is stopped once $SDiff_k$ exceeds a threshold of 0.3 [171] in which case IMF1 is set to the value of $\boldsymbol{h}_{1k}$.

5. With IMF1 defined, this is subtracted from the original signal and the above steps repeated to obtain the next decompositions, IMF2, and so on until the desired number of IMFs are defined. IMF1 contains the highest frequency information, while higher numbered IMFs contain progressively lower frequency information.

A.2.4 The Hilbert-Huang Transform

The Hilbert-Huang transform is an extension of the EMD (defined in section A.2.3) where the Hilbert Transform is applied to the resulting IMFs. The Hilbert Transform of a general signal, $\boldsymbol{s} \in \mathbb{R}^{1\times N}$ (of any length, N), can be applied in the time-domain through convolving $\boldsymbol{s}$ with the function $\frac{1}{\pi i}$ [171]:

$$\tilde{s}_i = \sum_{k=-\infty}^{\infty} s_k \frac{1}{\pi(i-k)} \tag{A.10}$$

With $\tilde{\boldsymbol{s}} \in \mathbb{R}^{1\times N}$ (of same dimensions as $\boldsymbol{s}$) calculated, each i^{th} sample of the analytic signal, $\boldsymbol{s}^{(A)} \in \mathbb{C}^{1\times N}$ (also of same dimensions) is defined as:

$$s_i^{(A)} = s_i + j\tilde{s}_i \tag{A.11}$$

which can also be represented in the exponential form:

$$s_i^{(A)} = a_i \mathrm{e}^{j\Phi_i} \tag{A.12}$$

where $j = \sqrt{-1}$. The samples, a_i, for all i, result in $\boldsymbol{a} \in \mathbb{R}^{1\times N}$ which is the instantaneous amplitude (envelope) of the signal. Similarly, the samples, Φ_i, result in $\boldsymbol{\Phi} \in \mathbb{R}^{1\times N}$ which is the instantaneous phase (with both $\boldsymbol{a}$ and $\boldsymbol{\Phi}$ having the same dimensions as $\boldsymbol{s}$) [171].

A.3 Principal Component Analysis (PCA)

PCA was first invented by *K. Pearson* in 1901 [241] and is a procedure to represent a multivariate dataset in a new multidimensional principal component space. The directions of these new principal components are orthogonal to one another and correlated, with the data transformed in such a way that the first principal component is oriented in the direction of the largest possible variance. To perform PCA:

1. The data is first standardized and represented by the matrix, $\boldsymbol{X} \in \mathbb{R}^{P \times N}$ (with P and N the vector dimension and number of observations, respectively).

2. The covariance, $\boldsymbol{\Sigma} \in \mathbb{R}^{P \times P}$, is calculated by:

$$\boldsymbol{\Sigma} = \frac{1}{N-1} \sum_{i=1}^{N} (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\intercal \tag{A.13}$$

   Each column vector, $\boldsymbol{x}_i \in \mathbb{R}^P$, corresponds to the i^{th} observation (column) in $\boldsymbol{X}$ and $\boldsymbol{\mu} \in \mathbb{R}^P$ denotes the data mean.

3. Eigenvalues and eigenvectors are derived from $\boldsymbol{\Sigma}$ by performing the eigendecomposition:

$$\boldsymbol{\Sigma} = \boldsymbol{V} \boldsymbol{\Lambda} \boldsymbol{V}^{-1} \tag{A.14}$$

   This is possible as $\boldsymbol{\Sigma}$ is square, symmetric, and invertible. Each column of $\boldsymbol{V} \in \mathbb{R}^{P \times P}$ is an eigenvector representing the direction of a principal component. $\boldsymbol{\Lambda} \in \mathbb{R}^{P \times P}$ is a diagonal matrix with corresponding eigenvalues. Each eigenvalue is proportional to the variance along the particular principal component and the eigenvector corresponding to the largest eigenvalue represents the first principal component.

4. Eigenvectors are rearranged in descending order, using the corresponding eigenvalues, to provide the principal components in order from first (most variance) to last (least variance).

A.4 K-means clustering

The K-means algorithm remains one of the simplest and most intuitive methods for clustering data points into K distinct clusters [242].

After random initialization of each j^{th} cluster mean (centroid), $\boldsymbol{\mu}_j \in \mathbb{R}^P$ (where P denotes the vector dimension), each i^{th} data point, $\boldsymbol{x}_i \in \mathbb{R}^P$, is iteratively assigned a cluster label, $c_j \in \mathbb{N} : 1 \leq c_j \leq K$ ($j \in \mathbb{N} : 1 \leq j \leq K$). In each iteration, all data points are reassigned to the nearest cluster based on the minimum Euclidian distance between the data point and each centroid. The centroids are then recalculated after

Algorithm A.1 K-means clustering

input $[\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N \in \mathbb{R}^P]$ (where P denotes the vector dimension and N the number of observations), number of clusters K.
initialize random cluster centroids $[\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \cdots, \boldsymbol{\mu}_k \in \mathbb{R}^P]$
repeat
  for $i = 1$ to N **do**
    $c_i = \arg\min_j \left(||\boldsymbol{x}_i - \boldsymbol{\mu}_j|| \right)$
    where $||\cdot||$ is the Euclidian distance.
  end for
  for $j = 1$ to K **do**
    $\boldsymbol{\mu}_j = \frac{\sum_{i=1}^{N} 1\{c_i = j\}\boldsymbol{x}_i}{\sum_{i=1}^{N} 1\{c_i = j\}}$
    where $1\{c_i = j\}$ denotes a value of 1 when $c_i = j$ and 0 otherwise.
  end for
until no further cluster reassignments

the new assignments and the process repeated until convergence, when the data points are no longer re-assigned [242]. K-means is summarized in algorithm A.1.

A.5 The SAT% algorithm

The SAT% algorithm begins with a classification of SAT periods using the method developed by *Palmu et al.* [89]:

1. After filtering the EEG signal in the range 0.5-10Hz to achieve the filtered signal, $\boldsymbol{x} \in \mathbb{R}^N$ (where N is the signal length), the Non-Linear Energy Operator (NLEO) is calculated for each i^{th} sample, x_i, by:

$$g_i = |x_i x_{i-3} - x_{i-1} x_{i-2}| \tag{A.15}$$

2. The processed signal is then smoothed by averaging using a sliding window centred on each i^{th} sample. This results in an average for each original signal sample, $g_i^{(avg)}$. The sliding window length, *SAT_WIN*, requires tuning.

3. To remove continuous artefacts that may result in an elevated baseline level, a baseline correction is performed by subtracting from each $g_i^{(avg)}$, the minimum

value of the signal from the one minute epoch preceding the current sample:

$$x_i^{(NLEO)} = g_i^{(avg)} - \min\left(g_j^{(avg)} \ : \ j \in \{i - 60 \cdot Fs, \cdots, i\}\right) \tag{A.16}$$

where Fs denotes the EEG sampling frequency.

4. Periods of SATs are then classified from this resulting signal if values exceed a threshold level, *SAT_THR*, for longer than a time limit, *MIN_DUR*. The parameter *SAT_THR* requires specific tuning and clinical knowledge of typical SAT durations results in setting *MIN_DUR* to one second [33, 89].

5. SAT classification produces a binary signal (with estimated SAT periods assigned a binary value of '1'). This is then converted to a SAT% signal by calculating the SAT percentage over a sliding window of length 20 seconds [32]. Finally, estimates of QS (and non-QS) are obtained by taking the mean SAT% signal as a threshold, with values *below* this threshold classified as QS [33].

For the comparison of QS classification by SAT% and the CLASS algorithm outlined in chapter 5, the SAT% algorithm was applied to the same EEG dataset (with Fs=250Hz) as with CLASS. As SAT% only uses a single channel, channel C3 was selected here to correspond best with previous studies [33, 89].

To tune the parameters *SAT_WIN* and *SAT_THR*, a grid search was performed using the CLASS training set. Training performance was assessed for each parameter combination using the mean F1-error (as defined in section 5.6.1 for optimizing the CLASS parameters). Results are shown in figure A.5, with optimal parameters corresponding to *SAT_WIN*=0.5 and *SAT_THR*=0.2. These optimized parameters were then used to run the algorithm on the test data to produce the final results presented in chapter 5.

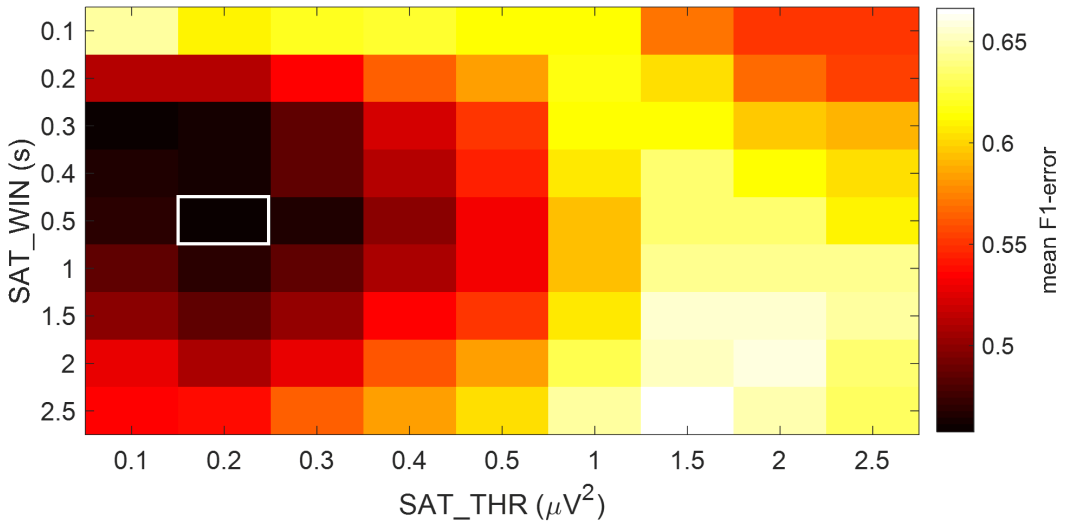


Figure A.5: Grid search of the SAT% parameters, *SAT_WIN* and *SAT_THR*, on the same training dataset used to optimize the CLASS algorithm (defined in chapter 5). Performance is assessed using the mean F1-error ($= 1 - F1$), with the minimum value shown by the white box and corresponding to *SAT_WIN*=0.5 and *SAT_THR*=0.2.

B
Supplementary Tables

B.1 Chapter 6: Significant feature correlations

Table B.1: Significant QS features (at $p<0.05$) using a random intercepts model between each feature and PMA. Values for the slope coefficients, p-values, and R^2 are shown. Random intercepts account for inter-patient variability due to multiple recordings from the same patient. A total of 97 QS features were significant and are arranged in descending order of R^2.

Feature	**Coefficient (slope)**	***p* (slope)**	$\boldsymbol{R^2}$
MSE avg. lower slope (scales 1-5)	0.0099	6.00E-29	0.87
MSE, scale 7	0.064	2.20E-26	0.85
MSE, scale 6	0.062	1.40E-25	0.84
MSE, scale 9	0.065	6.20E-26	0.84
MSE, scale 8	0.065	7.70E-26	0.84
MSE, scale 5	0.057	4.90E-25	0.84
MSE, scale 10	0.064	3.50E-25	0.83
MSE, scale 4	0.052	5.60E-23	0.81
MSE AUC	1.1	3.00E-23	0.81
Skewness rEEG	-0.073	1.90E-22	0.81
Hilbert median envelope amplitude, theta	0.47	3.40E-23	0.80
MSE, scale 3	0.044	1.30E-20	0.78
Variation coefficient of IMF4	-0.060	2.80E-21	0.77
Mean absolute ratio, IMF5/IMF4	-0.028	1.10E-20	0.77
Max. MSE	0.065	1.00E-20	0.76
Spectral flux SD	-0.13	1.30E-19	0.74
Mean absolute amplitudes, theta	0.29	2.40E-18	0.73
Variation coefficient of IMF5	-0.042	3.10E-20	0.73
Variation coefficient of IMF3	-0.059	1.50E-18	0.73
Average frequency of Alpha	-0.023	4.40E-18	0.72
Mean absolute ratio, IMF2/IMF1	0.073	3.70E-20	0.72
Mean of absolute values in sub-band, D5	0.22	4.10E-18	0.72
Average frequency of Delta	0.017	1.10E-18	0.72
Line Length, Burst%	0.026	5.90E-20	0.72
Fluctuation index of IMF3	0.026	4.30E-18	0.72
Mean absolute ratio, IMF6/IMF5	-0.021	7.70E-20	0.70
MSE, scale 2	0.033	8.50E-16	0.70
95% peak prominence	2.2	4.00E-16	0.70
Higuchi Fractal Dimension kmax=20	-0.0099	7.80E-15	0.69

Continued on next page

Continued from previous page

Feature	**Coefficient (slope)**	***p* (slope)**	R^2
Maximum absolute amplitude of second derivative	-13000	1.20E-08	0.69
Variation coefficient of IMF6	-0.025	8.60E-18	0.69
Hilbert median envelope amplitude, delta	0.59	5.20E-16	0.68
LM rEEG	2.6	7.80E-19	0.68
Kurtosis	-0.34	7.90E-16	0.68
Mean absolute ratio, IMF3/IMF2	0.023	2.00E-09	0.67
Average frequency of theta	-0.012	1.20E-04	0.67
Relative bandpower in theta	0.011	5.90E-14	0.67
Maximum absolute amplitude of first derivative	-84	3.10E-05	0.65
Fluctuation index of IMF4	0.014	3.00E-14	0.64
Variation coefficient of IMF1	-0.10	5.80E-12	0.64
Hilbert median instantaneous frequency, IMF3	-0.064	6.30E-09	0.64
IQR rEEG	-1.1	2.30E-06	0.63
MSE, scale 1 (SampEn)	0.018	1.10E-12	0.63
Ratio of absolute mean values of D4/D3	0.041	6.00E-15	0.62
SD of MSE Values	0.010	6.60E-15	0.62
95% peak width	0.91	6.90E-11	0.62
SD theta band amplitude	0.31	7.50E-13	0.61
SD rEEG	-0.91	5.50E-06	0.61
EEG spectral beta/alpha ratio	-0.048	4.80E-15	0.60
Mean absolute band amplitudes, delta	0.34	1.10E-11	0.60
Median rEEG	2.7	1.00E-11	0.59
Higuchi Fractal Dimension kmax=40	-0.0058	2.70E-07	0.59
SD of the amplitudes in sub-band, D5	0.25	4.40E-12	0.58
LL suppression value	-0.011	1.10E-12	0.58
Skewness	-0.056	1.50E-13	0.57
Sum square of wavelet coeffs in sub-band, D5	19000	2.50E-12	0.56
Fluctuation index of IMF2	0.025	6.80E-12	0.56
Hilbert median instantaneous frequency, delta	0.014	2.90E-13	0.55
Average frequency	-0.038	1.40E-03	0.55
Mean of absolute values in sub-band, D4	0.094	3.50E-11	0.55
Variation coefficient of IMF2	-0.079	1.70E-12	0.55
SD delta band amplitude	0.35	3.60E-08	0.54
Full bandpower in delta-beta range	12	3.10E-09	0.54
Relative bandpower in beta	-0.0028	4.10E-08	0.53
Ratio of absolute mean values of D5/D4	0.035	7.40E-11	0.50
Mean absolute ratio, IMF4/IMF3	-0.031	2.60E-12	0.50
Hilbert median instantaneous frequency, IMF4	-0.017	2.10E-04	0.50
Hilbert median instantaneous frequency, IMF2	-0.16	3.10E-09	0.50
Mean rEEG	2.2	5.50E-08	0.50
Relative bandpower in delta	-0.0049	1.30E-03	0.49
Higuchi Fractal Dimension kmax=60	-0.0025	1.60E-02	0.49
90% spectral edge frequency	-0.18	4.40E-05	0.49
SD TKEO	-0.13	1.30E-04	0.48
Mean absolute band amplitudes, alpha	0.070	4.10E-09	0.48
Approximate entropy (ApEn)	0.0077	3.00E-06	0.43
Zero crossing rate SD	-0.00059	1.70E-02	0.40
SD of the amplitudes in sub-band, D4	0.072	3.20E-04	0.40
Coastline distance	300	1.80E-08	0.39
Spectral centroid SD	-0.017	1.70E-03	0.39
Katz fractal dimension	0.0022	2.00E-07	0.38

Continued on next page

Continued from previous page

Feature	**Coefficient (slope)**	***p* (slope)**	**R^2**
SD of amplitude	0.31	1.70E-03	0.38
SD alpha band amplitude	0.053	1.50E-03	0.37
Sum square of wavelet coeffs in sub-band, D4	3900	5.50E-04	0.36
95% peak value	0.73	1.30E-03	0.34
SD of the amplitudes in sub-band, A5	0.25	8.90E-03	0.33
Fluctuation index of IMF5	0.0031	2.30E-04	0.27
Hjorth activity	0.38	6.80E-03	0.23
Mean absolute amplitude	0.44	6.00E-04	0.21
Sum of absolute amplitudes	3300	6.00E-04	0.21
MSE avg. higher slope (scales 6-20)	-0.00058	3.00E-04	0.19
Mean of absolute values in sub-band, D3	0.027	5.70E-04	0.18
Mean absolute band amplitudes, beta	0.033	1.50E-03	0.17
Ratio of absolute mean values of A5/D5	-0.11	2.90E-04	0.17
Mean of absolute values in sub-band, A5	0.38	3.80E-03	0.16
Sum of squared amplitudes	110000	2.30E-02	0.14
Hjorth complexity	-0.0041	5.20E-03	0.12
Fluctuation index of IMF1	0.011	2.00E-02	0.077

Table B.2: Significant non-QS features (at $p<0.05$) using a random intercepts model between each feature and PMA. Values for the slope coefficients, p-values, and R^2 are shown. Random intercepts account for inter-patient variability due to multiple recordings from the same patient. A total of 70 non-QS features were significant and are arranged in descending order of R^2.

Feature	**Coefficient (slope)**	***p* (slope)**	**R^2**
Line Length, Burst%	0.026	3.20E-32	0.91
Mean absolute ratio, IMF4/IMF3	-0.026	2.20E-12	0.76
Variation coefficient of IMF1	-0.087	1.70E-17	0.70
Hilbert median instantaneous frequency, delta	0.018	6.80E-16	0.66
Ratio of absolute mean values of D5/D4	0.025	6.70E-11	0.66
MSE avg. higher slope (scales 6-20)	-0.0017	2.10E-13	0.64
Average frequency of delta	0.017	2.50E-15	0.64
Relative bandpower in Theta	0.012	1.10E-11	0.63
Skewness	-0.051	4.40E-14	0.60
Mean absolute ratio, IMF2/IMF1	0.048	1.20E-12	0.59
Mean absolute ratio, IMF3/IMF2	0.012	9.60E-04	0.57
Relative bandpower in Delta	-0.0088	4.90E-06	0.57
Mean absolute ratio, IMF5/IMF4	-0.029	1.10E-12	0.54
Spectral entropy	0.016	1.20E-03	0.54
Hilbert median envelope amplitude, theta	0.20	1.90E-08	0.53
Variation coefficient of IMF3	-0.036	2.80E-12	0.53
Average frequency of alpha	-0.024	1.20E-12	0.52
LM rEEG	1.0	1.30E-07	0.52
Zero crossing rate SD	-0.0014	3.00E-07	0.47
Higuchi Fractal Dimension kmax=60	0.0028	1.50E-02	0.46
Median rEEG	-0.88	1.40E-02	0.44

Continued on next page

Continued from previous page

Feature	**Coefficient (slope)**	***p* (slope)**	R^2
Kurtosis	-0.17	1.50E-10	0.43
LL suppression value	-0.0049	7.80E-10	0.42
Variation coefficient of IMF2	-0.047	3.80E-10	0.42
Variation coefficient of IMF4	-0.033	8.70E-10	0.41
Hilbert median instantaneous frequency, theta	0.012	1.90E-03	0.38
Hilbert median instantaneous frequency, IMF5	0.012	9.10E-06	0.37
SD of the amplitudes in sub-band, D4	-0.039	3.90E-02	0.36
SD Alpha band amplitude	-0.031	3.20E-02	0.34
IQR rEEG	-2.0	7.30E-08	0.34
Relative bandpower in Beta	-0.0016	4.10E-03	0.34
Mean absolute band amplitudes, theta	0.090	7.30E-04	0.34
Full bandpower in delta-beta range	-4.0	2.60E-02	0.34
Mean of absolute values in sub-band, D5	0.071	5.20E-04	0.33
Fluctuation index of IMF3	0.0093	1.50E-04	0.32
Sum square of wavelet coeffs in sub-band, D4	-2400	1.80E-02	0.31
Ratio of absolute mean values of D4/D3	0.026	7.30E-06	0.27
EEG spectral beta/alpha ratio	-0.046	3.10E-05	0.27
SD beta band amplitude	-0.071	3.40E-04	0.27
Mean rEEG	-1.1	1.90E-02	0.26
Fluctuation index of IMF5	-0.0032	2.90E-03	0.26
MSE, scale 6	0.035	4.90E-06	0.25
MSE, scale 7	0.035	6.00E-06	0.25
MSE, scale 8	0.034	7.50E-06	0.24
MSE, scale 5	0.034	8.20E-06	0.24
Hilbert median instantaneous frequency, IMF6	0.0044	1.10E-03	0.23
Higuchi Fractal Dimension kmax=20	-0.0047	1.50E-03	0.23
Variation coefficient of IMF5	-0.016	1.30E-05	0.23
Fluctuation index of IMF6	-0.0024	7.90E-05	0.22
MSE, scale 9	0.032	2.20E-05	0.22
MSE, scale 4	0.031	2.50E-05	0.22
MSE avg. lower slope	0.0048	2.50E-05	0.22
Skewness rEEG	-0.030	2.90E-05	0.21
SD rEEG	-1.3	3.50E-05	0.21
Hilbert median instantaneous frequency, IMF1	0.23	2.50E-04	0.21
MSE, scale 10	0.029	7.30E-05	0.20
MSE, scale 3	0.026	8.90E-05	0.19
MSE, scale 2	0.020	1.70E-04	0.18
MSE, scale 1 (SampEn)	0.012	5.10E-04	0.18
SD of the amplitudes in sub-band, D3	-0.052	1.80E-03	0.17
MSE AUC	0.47	3.00E-04	0.17
HM rEEG	-3.6	9.60E-04	0.16
Sum square of wavelet coeffs in sub-band, D3	-1800	1.50E-03	0.15
Maximum absolute value of first derivative	-110	1.20E-03	0.14
Maximum absolute value of second derivative	-14000	2.80E-03	0.12
Mean absolute ratio, IMF6/IMF5	-0.0062	3.20E-02	0.091
Average frequency of Beta	0.049	2.40E-02	0.089
Max. MSE	0.018	1.40E-02	0.079
Maximum and minimum difference of samples	-3.0	3.40E-02	0.070
Spectral flux SD	-0.029	2.60E-02	0.068

Table B.3: Significant features (at $p<0.05$) using a random intercepts model between each feature, PMA and a dummy variable for mild and severe (sev) outcomes. Lost-to-follow-up recordings are also included in the severe outcome group. Random intercepts account for inter-patient variability due to multiple recordings from the same patient. Mild and severe outcome p-values (for intercepts compared to normal outcome) are shown and the statistical significance for each group is indicated by ✓.

Feature	***p* (mild)**	***p* (sev)**	**Mild sig.**	**Sev sig.**
MSE, scale 1 (SampEn)	0.025	0.039	✓	✓
Zero crossing rate SD	0.0065	0.017	✓	✓
Variation coefficient of IMF1 non QS	0.0083	0.0029	✓	✓
95% peak width	0.000044	0.21	✓	
Hilbert median instantaneous frequency, IMF3	0.00018	0.20	✓	
Variation coefficient of IMF1	0.0012	0.057	✓	
Higuchi Fractal Dimension kmax=20	0.0019	0.10	✓	
Line Length, Burst%	0.012	0.14	✓	
MSE avg. lower slope non QS	0.013	0.073	✓	
Hilbert median instantaneous frequency, IMF6 non QS	0.017	0.050	✓	
Higuchi Fractal Dimension kmax=40	0.025	0.086	✓	
Average frequency	0.035	0.073	✓	
Fluctuation index of IMF1	0.050	0.17	✓	
Hilbert median instantaneous frequency, IMF1 non QS	0.056	0.0012		✓
Sum square of wavelet coeffs in sub-band, D4 non QS	0.57	0.0013		✓
SD of the amplitudes in sub-band, D4 non QS	0.60	0.0022		✓
SD of amplitude	0.91	0.0039		✓
SD rEEG	0.38	0.0046		✓
SD of the amplitudes in sub-band, A5	0.85	0.0051		✓
SD alpha band amplitude non QS	0.48	0.0057		✓
Skewness non QS	0.40	0.0062		✓
Mean absolute band amplitudes, theta non QS	0.75	0.0065		✓
Mean of absolute values in sub-band, D5 non QS	0.72	0.0071		✓
Hjorth activity	0.86	0.0071		✓
IQR rEEG	0.32	0.0088		✓
Sum of squared amplitudes	0.84	0.0094		✓
95% peak value	0.85	0.011		✓
Mean of absolute values in sub-band, A5	0.85	0.011		✓
Sum of absolute amplitudes	0.77	0.011		✓
Mean absolute amplitude	0.77	0.011		✓
Full bandpower in delta-beta range	0.58	0.012		✓
Line Length, Burst% non QS	0.066	0.013		✓
Maximum absolute value of second derivative	0.50	0.015		✓
Hilbert median instantaneous frequency, IMF5 non QS	0.053	0.015		✓
SD delta band amplitude	0.50	0.015		✓
LL suppression value non QS	0.17	0.018		✓
Hilbert median envelope amplitude, theta non QS	0.69	0.018		✓
SD beta band amplitude non QS	0.29	0.019		✓

Continued on next page

Continued from previous page

Feature	p (mild)	p (sev)	Mild sig.	Sev sig.
Maximum absolute value of first derivative non QS	0.31	0.020		✓
SD TKEO	0.23	0.020		✓
Skewness	0.37	0.021		✓
Mean absolute band amplitudes, delta	0.61	0.023		✓
Mean rEEG	0.88	0.026		✓
IQR rEEG non QS	0.48	0.029		✓
Median rEEG non QS	0.23	0.030		✓
Sum square of wavelet coeffs in sub-band, D3 non QS	0.34	0.031		✓
Fluctuation index of IMF5 non QS	0.13	0.037		✓
SD of the amplitudes in sub-band, D3 non QS	0.20	0.039		✓
Fluctuation index of IMF3 non QS	0.44	0.039		✓
Skewness rEEG	0.28	0.042		✓
Mean absolute ratio, IMF6/IMF5 non QS	0.20	0.043		✓
Variation coefficient of IMF4	0.31	0.044		✓

B.2 Chapter 7: Recording-specific performance

Table B.4: Individual patient κ scores for the HMM-GMM and GMM models, with and without personalized feature standardization (stand.). Unusually low values (at κ<0.2) are shown in bold.

Patient	HMM-GMM	GMM	HMM-GMM (no stand.)	GMM (no stand)
1	0.61	0.50	0.49	0.53
2	0.58	0.58	0.50	0.62
3	0.46	0.46	**0.14**	0.30
4	0.47	0.32	0.36	0.40
5	0.66	0.53	0.62	0.50
6	0.67	0.55	0.56	0.62
7	0.79	0.70	0.72	0.67
8	0.47	0.50	0.70	0.64
9	0.72	0.66	0.70	0.61
10	0.58	0.46	0.75	0.61
11	0.61	0.36	**0.080**	**0.040**
12	0.65	0.42	0.53	0.45
13	0.47	0.35	0.66	0.47
14	0.64	0.67	0.59	0.64
15	0.74	0.58	0.68	0.62
16	0.60	0.53	0.54	0.43

B.3 Chapter 8: aDP-HMM-GMM δ performance

Table B.5: $\alpha^{(a)}$ and $\alpha^{(s)}$ for different values of δ in the aDP-HMM-GMM. As $\alpha^{(s)}$ is calculated per recording, the median is given, with IQR in brackets. The highest model performance is shown in bold and corresponds to δ=10.

$\boldsymbol{\delta}$	$\boldsymbol{\alpha}^{(a)}$	$\boldsymbol{\alpha}^{(s)}$
0.01	0.90	0.71 (0.51-0.83)
0.1	0.91	0.74 (0.49-0.84)
1	0.92	0.73 (0.55-0.84)
10	**0.92**	**0.74 (0.54-0.86)**
100	0.92	0.72 (0.53-0.86)

References

[1] H. Blencowe et al. “Born too soon: The global epidemiology of 15 million preterm births understanding the data preterm birth what is it? Defining preterm birth”. In: *Reprod. Health* 10.Suppl 1 (2013), pp. 1–13.

[2] S. L. Bonifacio et al. “A new neurological focus in neonatal intensive care”. In: *Nat. Rev. Neurol.* 7.9 (2011), pp. 485–494.

[3] G. Ball et al. “The Effect of Preterm Birth on Thalamic and Cortical Development”. In: *Cereb. Cortex* 22.5 (2012), pp. 1016–1024.

[4] E. Biagioni et al. “Background EEG activity in preterm infants: correlation of outcome with selected maturational features”. In: *Electroencephalogr. Clin. Neurophysiol.* 91.3 (1994), pp. 154–162.

[5] L. J. Woodward et al. “Neonatal MRI to Predict Neurodevelopmental Outcomes in Preterm Infants”. In: *N. Engl. J. Med.* 355.7 (2006), pp. 685–694.

[6] L. A. Ramenghi et al. “Magnetic resonance imaging assessment of brain maturation in preterm neonates with punctate white matter lesions”. In: *Neuroradiology* 49.2 (2007), pp. 161–167.

[7] E. Biagioni et al. “Maturation of cerebral electrical activity and development of cortical folding in young very preterm infants”. In: *Clin. Neurophysiol.* 118.1 (2007), pp. 53–59.

[8] V. Chau et al. “Abnormal brain maturation in preterm neonates associated with adverse developmental outcomes”. In: *Neurology* 81.24 (2013), pp. 2082–2089.

[9] J. Vinall et al. “Slower Postnatal Growth Is Associated with Delayed Cerebral Cortical Maturation in Preterm Newborns”. In: *Sci. Transl. Med.* 5.168 (2013), 168ra8–168ra8.

[10] M. Cao et al. “Early Development of Functional Network Segregation Revealed by Connectomic Analysis of the Preterm Human Brain”. In: *Cereb. Cortex* 27.3 (2016), bhw038.

[11] S. Vanhatalo and P. Fransson. “Advanced EEG and MRI Measurements to Study the Functional Development of the Newborn Brain”. In: *Prenat. Postnatal Determ. Dev. Neuromethods, vol 109.* Humana Press, New York, NY, 2016. Chap. 4, pp. 53–68.

[12] R. O. Lloyd et al. “Predicting 2-y outcome in preterm infants using early multimodal physiological monitoring”. In: *Pediatr. Res.* 80.3 (2016), pp. 382–388.

[13] J. R. Isler et al. “An automated method for coding sleep states in human infants based on respiratory rate variability”. In: *Dev. Psychobiol.* 58.8 (2016), pp. 1108–1115.

[14] J. Werth et al. “Unobtrusive assessment of neonatal sleep state based on heart rate variability retrieved from electrocardiography used for regular patient monitoring”. In: *Early Hum. Dev.* 113 (2017), pp. 104–113.

[15] P. B. Marschik et al. “A Novel Way to Measure and Predict Development: A Heuristic Approach to Facilitate the Early Detection of Neurodevelopmental Disorders”. In: *Curr. Neurol. Neurosci. Rep.* 17.5 (2017), p. 43.

[16] A. Le Bihannic et al. “Prognostic value of EEG in very premature newborns”. In: *Arch. Dis. Child. Fetal Neonatal Ed.* 97.2 (2012), pp. 2010–2013.

[17] M. S. Scher. “Ontogeny of EEG-sleep from neonatal through infancy periods”. In: *Sleep Med.* 9.6 (2008), pp. 615–636.

[18] S. Graven and J. Browne. “Sleep and Brain Development. The Critical Role of Sleep in Fetal and Early Neonatal Brain Development”. In: *Newborn Infant Nurs. Rev.* 8.4 (2008), pp. 173–179.

[19] M. Grigg-Damberger. “The visual scoring of sleep in infants 0 to 2 months of age”. In: *J. Clin. Sleep Med.* 12.3 (2016), pp. 429–445.

[20] E. Pavlidis, R. O. Lloyd, and G. B. Boylan. “EEG - A Valuable Biomarker of Brain Injury in Preterm Infants”. In: *Dev. Neurosci.* 39.1-4 (2017), pp. 23–35.

[21] F. Hayakawa et al. “Disorganized Patterns: Chronic-Stage EEG Abnormality of the Late Neonatal Period Following Severely Depressed EEG Activities in Early Preterm Infants”. In: *Neuropediatrics* 28.05 (1997), pp. 272–275.

[22] H. J. Niemarkt et al. “Maturational changes in automated EEG spectral power analysis in preterm infants”. In: *Pediatr. Res.* 70.5 (2011), pp. 529–534.

[23] M. S. Scher et al. “Physiologic Brain Dysmaturity in Late Preterm Infants”. In: *Pediatr. Res.* 70.5 (2011), pp. 524–528.

[24] R. Shellhaas et al. “Quantitative sleep stage analyses as a window to neonatal neurologic function”. In: *Neurology* 82.5 (2014), pp. 390–395.

[25] K. K. Iyer et al. “Cortical burst dynamics predict clinical outcome early in extremely preterm infants”. In: *Brain* 138 (2015), pp. 2206–2218.

[26] N. Koolen et al. “Line length as a robust method to detect high-activity events: automated burst detection in premature EEG recordings.” In: *Clin. Neurophysiol.* 125.10 (2014), pp. 1985–94.

[27] A. Dereymaeker et al. “The suppression curve as a quantitative approach for measuring brain maturation in preterm infants”. In: *Clin. Neurophysiol.* 127.8 (2016), pp. 2760–2765.

[28] M. A. Navakatikyan, D. O’Reilly, and L. J. Van Marter. “Automatic measurement of interburst interval in premature neonates using range EEG”. In: *Clin. Neurophysiol.* 127.2 (2016), pp. 1233–1246.

[29] O. De Wel et al. “Complexity Analysis of Neonatal EEG Using Multiscale Entropy: Applications in Brain Maturation and Sleep Stage Classification”. In: *Entropy* 19.10 (2017), p. 516.

[30] N. Koolen et al. “Interhemispheric synchrony in the neonatal EEG revisited: activation synchrony index as a promising classifier.” In: *Front. Hum. Neurosci.* 8 (2014), p. 1030.

[31] H. J. Niemarkt et al. “Quantitative analysis of maturational changes in EEG background activity in very preterm infants with a normal neurodevelopment at 1 year of age.” In: *Early Hum. Dev.* 86.4 (2010), pp. 219–24.

[32] K. Palmu et al. “Sleep wake cycling in early preterm infants: Comparison of polysomnographic recordings with a novel EEG-based index”. In: *Clin. Neurophysiol.* 124.9 (2013), pp. 1807–1814.

[33] N. J. Stevenson et al. “Measuring brain activity cycling (BAC) in long term EEG monitoring of preterm babies.” In: *Physiol. Meas.* 35.7 (2014), pp. 1493–508.

[34] N. Koolen et al. “Automated classification of neonatal sleep states using EEG”. In: *Clin. Neurophysiol.* 128.6 (2017), pp. 1100–1108.

[35] I. Fagioli and P. Salzarulo. “Sleep states development in the first year of life assessed through 24-h recordings”. In: *Early Hum. Dev.* 6.2 (1982), pp. 215–228.

[36] A. A. dos Santos et al. “Behavior and EEG concordance of active and quiet sleep in preterm very low birth weight and full-term neonates at matched conceptional age”. In: *Early Hum. Dev.* 90.9 (2014), pp. 507–510.

[37] K. Pillay et al. “A Bayesian parametric model for quantifying brain maturation from sleep-EEG in the vulnerable newborn baby”. In: *2018 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* 2018, pp. 1–4.

[38] K. Pillay et al. “Automated EEG sleep staging in the term-age baby using a Generative Modelling approach”. In: *J. Neural Eng.* 15 (2018). arXiv: 1809.07102v1.

[39] K. Pillay and M. De Vos. “A unifying Bayesian approach for preterm brain-age prediction that models EEG sleep transitions over age”. In: *NIPS Work. Mach. Learn. Heal.* Los Angeles, 2017. arXiv: 1809.07102.

[40] A. Dereymaeker et al. “An Automated Quiet Sleep Detection Approach in Preterm Infants as a Gateway to Assess Brain Maturation”. In: *Int. J. Neural Syst.* 27.06 (2017), 1750023(1)–1750023(18).

[41] A. Dereymaeker et al. “Review of sleep-EEG in preterm and term neonates”. In: *Early Hum. Dev.* 113 (2017), pp. 87–103.

[42] “Age Terminology During the Perinatal Period”. In: *Pediatrics* 114.5 (2004), pp. 1362–1364.

[43] A. Dereymaeker. “Automated EEG analysis to quantify brain function in preterm and term neonates”. PhD thesis. Catholic University Leuven, 2017.

[44] L. Andruscavage and D. J. Weissgold. “Screening for retinopathy of prematurity.” In: *Br. J. Opthamology* 86 (2002), pp. 1127–1130.

[45] M. J. Bell et al. “Neonatal Necrotizing Enterocolitis: Therapeutic Decisions Based upon Clinical Staging”. In: *Ann. Surg.* 187.1 (1978), pp. 1–7.

[46] A. M. Fujii et al. “Neonatal necrotizing enterocolitis with intestinal perforation in extremely premature infants receiving early indomethacin treatment for patent ductus arteriosus”. In: *J. Perinatol.* 22.7 (2002), pp. 535–540.

[47] K. E. Gregory et al. “Necrotizing Enterocolitis in the Premature Infant”. In: *Adv. Neonatal Care* 11.3 (2011), pp. 155–164.

[48] K. Kristof, E. Kocsis, and K. Nagy. “Clinical microbiology of early-onset and late-onset neonatal sepsis, particularly among preterm babies”. In: *Acta Microbiol. Immunol. Hung.* 56.1 (2009), pp. 21–51.

[49] W. Deng, J. Pleasure, and D. Pleasure. “Progress in Periventricular Leukomalacia”. In: *Arch. Neurol.* 65.10 (2008), pp. 1291–1295.

[50] I. Korotchikova et al. “Quantitative EEG analysis in neonatal hypoxic ischaemic encephalopathy”. In: *Clin. Neurophysiol.* 122.8 (2011), pp. 1671–1678.

[51] P. Ballabh. “Intraventricular Hemorrhage in Premature Infants: Mechanism of Disease”. In: *Pediatr. Res.* 67.1 (2010), pp. 1–8.

[52] S. Niedermaier and A. Hilgendorff. “Bronchopulmonary dysplasia - an overview about pathophysiologic concepts”. In: *Mol. Cell. Pediatr.* 2.1 (2015), p. 2.

[53] R. G. Fisher et al. “Patent ductus arteriosus in adults-long-term follow-up: Nonsurgical versus surgical treatment”. In: *J. Am. Coll. Cardiol.* 8.2 (1986), pp. 280–284.

[54] L. S. De Vries, M. J. Benders, and F. Groenendaal. “Imaging the premature brain: Ultrasound or MRI?” In: *Neuroradiology* 55.SUPPL. 2 (2013), pp. 13–22.

[55] M. Fumagalli et al. “Cranial ultrasound findings in late preterm infants and correlation with perinatal risk factors”. In: *Ital. J. Pediatr.* 41.1 (2015), pp. 1–7.

[56] R. K. Malhotra and A. Y. Avidan. “Sleep Stages and Scoring Technique”. In: *Atlas Sleep Med.* Second Edi. Elsevier Inc., 2014. Chap. 3, pp. 77–99.

[57] N. Evans and M. Kluckow. “Early ductal shunting and intraventricular haemorrhage in ventilated preterm infants.” In: *Arch. Dis. Child. Fetal Neonatal Ed.* 75.3 (1996), F183–6.

[58] S. Vanhatalo. “Playing music to preemies: boosting of soothing the brain?” In: *Acta Paediatr.* 107.4 (2018), pp. 549–550.

[59] R. Slater et al. “Cortical pain responses in human infants.” In: *J. Neurosci.* 26.14 (2006), pp. 3662–6.

[60] R. E. Grunau. "Neonatal pain in very preterm infants: long-term effects on brain, neurodevelopment and pain reactivity." In: *Rambam Maimonides Med. J.* 4.4 (2013), e0025.

[61] M. S. Scher et al. "Neurophysiologic assessment of brain maturation after an 8-week trial of skin-to-skin contact on preterm infants". In: *Clin. Neurophysiol.* 120.10 (2009), pp. 1812–1818.

[62] J. Baley. "Skin-to-Skin Care for Term and Preterm Infants in the Neonatal ICU." In: *Pediatrics* 136.3 (2015), pp. 596–599.

[63] F. Kaffashi et al. "An analysis of the kangaroo care intervention using neonatal EEG complexity: A preliminary study". In: *Clin. Neurophysiol.* 124.2 (2013), pp. 238–246.

[64] A. Spittle et al. *Early developmental intervention programmes provided post hospital discharge to prevent motor and cognitive impairment in preterm infants*. Tech. rep. 11. 2015.

[65] A. P. Matheny, A. B. Dolan, and R. S. Wilson. "Bayley's Infant Behavior Record: Relations between behaviors and mental test scores." In: *Dev. Psychol.* 10.5 (1974), pp. 696–702.

[66] L. Nellis and B. E. Gridley. "Review of the Bayley Scales of Infant Development-Second edition". In: *J. Sch. Psychol.* 32.2 (1994), pp. 201–209.

[67] L. Deroma et al. "[Interrater reliability of the Bayley III test in the Italian Northern-Adriatic Cohort II]." In: *Epidemiol. Prev.* 37.4-5 (2013), pp. 297–302.

[68] J. H. Lee et al. "Reliability and Applicability of the Bayley Scale of Infant Development-II for Children With Cerebral Palsy." In: *Ann. Rehabil. Med.* 37.2 (2013), pp. 167–74.

[69] S. R. Manandhar et al. "Acceptability and Reliability of the Bayley Scales of Infant Development III Cognitive and Motor Scales among Children in Makwanpur". In: *J Nepal Heal. Res Counc* 14.32 (2016), pp. 47–50.

[70] S. Johnson et al. "Academic attainment and special educational needs in extremely preterm children at 11 years of age: the EPICure study." In: *Arch. Dis. Child. Fetal Neonatal Ed.* 94.4 (2009), F283–9.

[71] A. M. Husain. "Review of neonatal EEG." eng. In: *Am. J. Electroneurodiagnostic Technol.* 45.1 (2005), pp. 12–35.

[72] M. S. Scher, M. W. Johnson, and D. Holditch-Davis. "Cyclicity of Neonatal Sleep Behaviors at 25 to 30 Weeks' Postconceptional Age". In: *Pediatr. Res.* 57.6 (2005), pp. 879–882.

[73] M. F. Vecchierini, A. M. D'Allest, and P. Verpillat. "EEG patterns in 10 extreme premature neonates with normal neurological outcome: Qualitative and quantitative data". In: *Brain Dev.* 25.5 (2003), pp. 330–337.

[74] M. André et al. “Electroencephalography in premature and full-term infants. Developmental features and glossary”. In: *Clin. Neurophysiol.* 40.2 (2010), pp. 59–124.

[75] T. N. Tsuchida et al. “American Clinical Neurophysiology Society Standardized EEG Terminology and Categorization for the Description of Continuous EEG Monitoring in Neonates”. In: *J. Clin. Neurophysiol.* 30.2 (2013), pp. 161–173.

[76] M. Mirmiran, Y. G. Maas, and R. L. Ariagno. “Development of fetal and neonatal sleep and circadian rhythms”. In: *Sleep Med. Rev.* 7.4 (2003), pp. 321–334.

[77] H. Kidokoro et al. “What does cyclicity on amplitude-integrated EEG mean?” In: *J. Perinatol.* 32.8 (2012), pp. 565–569.

[78] L. Curzi-Dascalova et al. “Sleep state organization in premature infants of less than 35 weeks’ gestational age.” In: *Pediatr. Res.* 34.5 (1993), pp. 624–628.

[79] S. Vanhatalo and K. Kaila. “Development of neonatal EEG activity: From phenomenology to physiology”. In: *Semin. Fetal Neonatal Med.* 11.6 (2006), pp. 471–478.

[80] M. Grigg-Damberger et al. “The visual scoring of sleep and arousal in infants and children.” In: *J. Clin. sleep Med.* 3.2 (2007), pp. 201–240.

[81] M. H. Silber et al. “The visual scoring of sleep in adults”. In: *J. Clin. Sleep Med.* 3.2 (2007), pp. 121–131.

[82] M. M. Myers et al. “A novel quantitative measure of Trace-alternant EEG activity and its association with sleep states of preterm infants”. In: *Dev. Psychobiol.* 31.3 (1997), pp. 167–174.

[83] J. Turnbull et al. “Automated detection of tracé alternant during sleep in healthy full-term neonates using discrete wavelet transform”. In: *Clin. Neurophysiol.* 112.10 (2001), pp. 1893–1900.

[84] K. Paul et al. “Comparison of quantitative EEG characteristics of quiet and active sleep in newborns”. In: *Sleep Med.* 4.6 (2003), pp. 543–552.

[85] K. Paul et al. “Quantitative topographic differentiation of the neonatal EEG”. In: *Clin. Neurophysiol.* 117.9 (2006), pp. 2050–2058.

[86] A. Piryatinska et al. “Automated detection of neonate EEG sleep stages.” In: *Comput. Methods Programs Biomed.* 95.1 (2009), pp. 31–46.

[87] J. Löfhede et al. “Classification of burst and suppression in the neonatal electroencephalogram.” In: *J. Neural Eng.* 5.4 (2008), pp. 402–10.

[88] V. Matic et al. “Automated EEG inter-burst interval detection in neonates with mild to moderate postasphyxial encephalopathy”. In: *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*. 2012, pp. 17–20.

[89] K. Palmu et al. “Optimization of an NLEO-based algorithm for automated detection of spontaneous activity transients in early preterm EEG.” en. In: *Physiol. Meas.* 31.11 (2010), pp. 85–93.

[90] K. Palmu et al. “Detection of ’EEG bursts’ in the early preterm EEG: visual vs. automated detection.” In: *Clin. Neurophysiol.* 121.7 (2010), pp. 1015–22.

[91] P. E. Chauvet et al. “Evaluation of automatic feature detection algorithms in EEG: Application to interburst intervals”. In: *Comput. Biol. Med.* 54 (2014), pp. 61–71.

[92] K. Murphy et al. “Automated analysis of multi-channel EEG in preterm infants”. In: *Clin. Neurophysiol.* 126.9 (2014), pp. 1692–1702.

[93] P. Mirzaei and G. Azemi. “A new NLEO based technique for the detection of burst–suppression patterns in multichannel neonatal EEG signals”. In: *Analog Integr. Circuits Signal Process.* 92.2 (2017), pp. 255–262.

[94] X. Navarro et al. “Multi-feature classifiers for burst detection in single EEG channels from preterm infants”. In: *J. Neural Eng.* 14.4 (2017), p. 046015.

[95] M. S. Scher et al. “Computer classification of sleep in preterm and full-term neonates at similar postconceptional term ages”. In: *Sleep* 19.1 (1996), pp. 18–25.

[96] V. Gerla et al. “Multichannel Analysis of the Newborn EEG Data”. In: *Itab 2006* (2006), pp. 1–6.

[97] V. Gerla et al. “Newborn sleep stage classification using hybrid evolutionary approach”. In: *Int. J. Bioelectromagn.* 9.1 (2007), pp. 25–26.

[98] L. Fraiwan and K. Lweesy. “Newborn sleep stage identification using multiscale entropy”. In: *2nd Middle East Conf. Biomed. Eng.* IEEE, 2014, pp. 361–364.

[99] M. Lavanga et al. “Automatic quiet sleep detection based on multifractality in preterm neonates: Effects of maturation”. In: *2017 39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* IEEE, 2017, pp. 2010–2013.

[100] A. H. Ansari et al. “Quiet Sleep Detection in Preterm Infants using Deep Convolutional Neural Networks”. In: *J. Neural Eng.* (in press) (2018).

[101] V. Krajca et al. “Neonatal EEG Sleep Stages Modelling by Temporal Profiles”. en. In: *Lect. notes Comput. Sci.* Ed. by R. M. Díaz, F. Pichler, and A. Q. Arencibia. Vol. 4739. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2007, pp. 195–201.

[102] V. Krajca et al. “Automatic identification of significant graphoelements in multichannel EEG recordings by adaptive segmentation and fuzzy clustering”. In: *Int J Biomed Comput.* 28.1–2 (1991), pp. 71–89.

[103] V. Krajca et al. “Automatic Detection of Sleep Stages in Neonatal EEG Using the Structural Time Profiles”. In: *2005 IEEE Eng. Med. Biol. 27th Annu. Conf.* 2005, pp. 6014–6016.

[104] G. Holmes and C. Lombroso. “Prognostic value of background patterns in the neonatal EEG”. In: *J. Clin. Neurophysiol.* 10.3 (1993), pp. 323–52.

[105] S. N. T. Tich et al. “Pathological features of neonatal EEG in preterm babies born before 30 weeks of gestationnal age”. In: *Neurophysiol. Clin.* 37.5 (2007), pp. 325–370.

[106] N. Hayashi-Kurahashi et al. “EEG for Predicting Early Neurodevelopment in Preterm Infants: An Observational Cohort Study”. In: *Pediatrics* 130.4 (2012), e891–e897.

[107] M. Perivier et al. “Neonatal EEG and neurodevelopmental outcome in preterm infants born before 32 weeks”. In: *Arch. Dis. Child. - Fetal Neonatal Ed.* 101.3 (2016), F253–F259.

[108] V. Havilicek, R. Childiaeva, and V. Chernick. “Ontogeny of EEG Power Characteristics of Quiet Sleep Periodic Cerebral Rhythm in Preterm Infants”. In: *Neuropaediatrics* 6.2 (1975), pp. 151–161.

[109] A. H. Bell et al. “Variation in power spectral analysis of the EEG with gestational age.” In: *J. Clin. Neurophysiol.* 8.3 (1991), pp. 312–9.

[110] A. H. Bell et al. “Spectral edge frequency of the EEG in healthy neonates and variation with behavioural state.” In: *Biol. Neonate* 60.2 (1991), pp. 69–74.

[111] K. Holthausen et al. “Brain dysmaturity index for automatic detection of high-risk infants.” In: *Pediatr. Neurol.* 22.3 (2000), pp. 187–91.

[112] S. Vanhatalo et al. “DC-EEG discloses prominent, very slow activity patterns during sleep in preterm infants”. In: *Clin. Neurophysiol.* 113.11 (2002), pp. 1822–1825.

[113] A. Okumura et al. “Amplitude spectral analysis of maturational changes of delta waves in preterm infants.” In: *Brain Dev.* 25.6 (2003), pp. 406–10.

[114] A. Okumura et al. “Amplitude Spectral Analysis of Theta/Alpha/Beta Waves in Preterm Infants”. In: *Pediatr. Neurol.* 34.1 (2006), pp. 30–34.

[115] E. M. Schumacher et al. “Feasibility of long-term continuous EEG monitoring during the first days of life in preterm infants: An automated quantification of the EEG activity”. In: *Pediatr. Res.* 69.5 (2011), pp. 413–417.

[116] S. Janjarasjitt, M. Scher, and K. Loparo. “Nonlinear dynamical analysis of the neonatal EEG time series: The relationship between neurodevelopment and complexity”. In: *Clin. Neurophysiol.* 119.4 (2008), pp. 822–836.

[117] D. Zhang et al. “Neurodevelopment in newborns: a sample entropy analysis of electroencephalogram”. In: *Physiol. Meas.* 30.5 (2009), pp. 491–504.

[118] M. Carrozzi, A. Accardo, and F. Bouquet. “Analysis of sleep-stage characteristics in full-term newborns by means of spectral and fractal parameters.” In: *Sleep* 27.7 (2004), pp. 1384–93.

[119] M. Tolonen et al. “Development of the spontaneous activity transients and ongoing cortical activity in human preterm babies”. In: *Neuroscience* 145.3 (2007), pp. 997–1006.

[120] K. K. Iyer et al. “Early Detection of Preterm Intraventricular Hemorrhage From Clinical Electroencephalography”. In: *Crit. Care Med.* 43.10 (2015), pp. 2219–2227.

[121] J. Löfhede et al. “Automatic classification of background EEG activity in healthy and sick neonates.” In: *J. Neural Eng.* 7.1 (2010), p. 16007.

[122] E. Cohen et al. “EEG power spectrum maturation in preterm fetal growth restricted infants”. In: *Brain Res.* 1678 (2018), pp. 180–186.

[123] A. Suppiej et al. “Spectral analysis highlight developmental EEG changes in preterm infants without overt brain damage”. In: *Neurosci. Lett.* 649 (2017), pp. 112–115.

[124] C. T. Lombroso. “Quantified electrographic scales on 10 pre-term healthy newborns followed up to 40-43 weeks of conceptional age by serial polygraphic recordings.” In: *Electroencephalogr. Clin. Neurophysiol.* 46.4 (1979), pp. 460–74.

[125] P. G. Grieve et al. “EEG functional connectivity in term age extremely low birth weight infants”. In: *Clin. Neurophysiol.* 119.12 (2008), pp. 2712–2720.

[126] M. Wacker et al. “A processing scheme for time-variant phase analysis in EEG burst activity of premature and full-term newborns in quiet sleep: a methodological study.” In: *Biomed. Eng. / Biomed. Tech.* 57.6 (2012), pp. 491–505.

[127] K. Schiecke et al. “Quadratic phase couplings in the EEG of premature and full-term newborn during quiet sleep”. In: *Methods Inf. Med.* 54.3 (2015), pp. 205–208.

[128] B. H. Walsh, D. M. Murray, and G. B. Boylan. “The use of conventional EEG for the assessment of hypoxic ischaemic encephalopathy in the newborn: A review”. In: *Clin. Neurophysiol.* 122.7 (2011), pp. 1284–1294.

[129] A. Tokariev et al. “Phase synchrony in the early preterm EEG: Development of methods for estimating synchrony in both oscillations and events”. In: *Neuroimage* 60.2 (2012), pp. 1562–1573.

[130] E. Pereda et al. “Topography of EEG complexity in human neonates: Effect of the postmenstrual age and the sleep state”. In: *Neurosci. Lett.* 394.2 (2006), pp. 152–157.

[131] J. J. Gonzalez et al. “Assessment of electroencephalographic functional connectivity in term and preterm neonates”. In: *Clin. Neurophysiol.* 122.4 (2011), pp. 696–702.

[132] E. J. Meijer et al. “Functional connectivity in preterm infants derived from EEG coherence analysis”. In: *Eur. J. Paediatr. Neurol.* 18.6 (2014), pp. 780–789.

[133] E. J. Meijer et al. “Interhemispheric connectivity estimated from EEG time-correlation analysis in preterm infants with normal follow-up at age of five”. In: *Physiol. Meas.* 37.12 (2016), pp. 2286–2298.

[134] A. Zwanenburg et al. “Automatic detection of burst synchrony in preterm infants”. In: *2012 Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* IEEE, 2012, pp. 4720–4723.

[135] A. Tokariev et al. “Functional Brain Connectivity Develops Rapidly Around Term Age and Changes Between Vigilance States in the Human Newborn”. In: *Cereb. Cortex* 26.23 (2015), pp. 4540–4550.

[136] A. Tokariev et al. “Preterm Birth Changes Networks of Newborn Cortical Activity”. In: *Cereb. Cortex* (in press) (2018), pp. 1–13.

[137] O. Räsänen, M. Metsäranta, and S. Vanhatalo. “Development of a novel robust measure for interhemispheric synchrony in the neonatal EEG: Activation Synchrony Index (ASI)”. In: *Neuroimage* 69 (2013), pp. 256–266.

[138] M. Lavanga et al. “Monitoring Effective Connectivity in the Preterm Brain: A Graph Approach to Study Maturation”. In: *Complexity* 19.10 (2017), pp. 1–13.

[139] N. J. Stevenson et al. “An automated system for grading eeg abnormality in term neonates with hypoxic-ischaemic encephalopathy”. In: *Ann. Biomed. Eng.* 41.4 (2013), pp. 775–785.

[140] V. Matic et al. “Holistic approach for automated background EEG assessment in asphyxiated full-term infants”. In: *J. Neural Eng.* 11.6 (2014), p. 66007.

[141] V. Matic et al. “Objective differentiation of neonatal EEG background grades using detrended fluctuation analysis”. In: *Front. Hum. Neurosci.* 9 (2015), pp. 1–14.

[142] V. Matic et al. “Improving reliability of monitoring background EEG dynamics in asphyxiated infants”. In: *IEEE Trans. Biomed. Eng.* 63.5 (2016), pp. 973–983.

[143] R. Ahmed et al. “Grading hypoxic-ischemic encephalopathy severity in neonatal EEG using GMM supervectors and the support vector machine”. In: *Clin. Neurophysiol.* 127.1 (2016), pp. 297–309.

[144] A. Dereymaeker et al. “Automated EEG background analysis to identify neonates with hypoxic-ischemic encephalopathy treated with hypothermia at risk for adverse outcome: A pilot study”. In: *Pediatr. Neonatol.* (in press) (2018).

[145] J. O’Toole et al. “Estimating functional brain maturity in very and extremely preterm neonates using automated analysis of the electroencephalogram”. In: *Clin. Neurophysiol.* 127.8 (2016), pp. 2910–2918.

[146] N. J. Stevenson et al. “Functional maturation in preterm infants measured by serial recording of cortical activity.” In: *Sci. Rep.* 7.1 (2017), p. 12969.

[147] M. Lavanga et al. “A brain-age model for preterm infants based on functional connectivity”. In: *Physiol. Meas.* 39.4 (2018), p. 044006.

[148] K. Klebermass et al. “Amplitude-Integrated EEG Pattern Predicts Further Outcome in Preterm Infants”. In: *Pediatr. Res.* 70.1 (2011), pp. 102–108.

[149] S. Wikström et al. "Early single-channel aEEG/EEG predicts outcome in very preterm infants." In: *Acta Paediatr.* 101.7 (2012), pp. 719–26.

[150] G. Natalucci et al. "Delayed cyclic activity development on early amplitude-integrated EEG in the preterm infant with brain lesions". In: *Neonatology* 103.2 (2013), pp. 134–140.

[151] I. Benavente-Fernández et al. "Low-voltage pattern and absence of sleep-wake cycles are associated with severe hemorrhage and death in very preterm infants". In: *Eur. J. Pediatr.* 174.1 (2015), pp. 85–90.

[152] J. Sisman, D. E. Campbell, and L. P. Brion. "Amplitude-Integrated EEG in Preterm Infants: Maturation of Background Pattern and Amplitude Voltage with Postmenstrual Age and Gestational Age". In: *J. Perinatol.* 25.6 (2005), pp. 391–396.

[153] L. Hellstrom-Westas et al. "Amplitude-integrated EEG Classification and Interpretation in Preterm and Term Infants". In: *Neoreviews* 7.2 (2006), e76–e87.

[154] I. Korotchikova et al. "Sleep–wake cycle of the healthy term newborn infant in the immediate postnatal period". In: *Clin. Neurophysiol.* 127.4 (2016), pp. 2095–101.

[155] J. Proakis and D. Manolakis. *Digital Signal Processing. Principles, Algorithms, and Applications.* Third. Upper Saddle River, NJ, USA: Prentice-Hall, Inc, 1996.

[156] B. Şen et al. "A comparative study on classification of sleep stage based on EEG signals using feature selection and classification algorithms". In: *J. Med. Syst.* 38.3 (2014), p. 18.

[157] S. Li et al. "Feature extraction and recognition of ictal EEG using EMD and SVM". In: *Comput. Biol. Med.* 43.7 (2013), pp. 807–816.

[158] K. Aboalayon et al. "Sleep Stage Classification Using EEG Signal Analysis: A Comprehensive Survey and New Investigation". In: *Entropy* 18.9 (2016), p. 272.

[159] A. Graps. "An Introduction to Wavelets". In: *IEEE Comput. Sci. Eng.* 2.2 (1995), pp. 50–61.

[160] B. D. Yetton et al. "Automatic detection of rapid eye movements (REMs): A machine learning approach". In: *J. Neurosci. Methods* 259 (2016), pp. 72–82.

[161] The MathWorks. *Signal Processing Toolbox User's Guide R2018b.* Tech. rep. Natick, MA, USA: The MathWorks Inc., 2018.

[162] D. N. Joanes and C. A. Gill. "Comparing Measures of Sample Skewness and Kurtosis". In: *J. R. Stat. Soc.* 47.1 (1998), pp. 183–189.

[163] B. Hjorth. "EEG analysis based on time-domain properties". In: *Electroencephalogr. Clin. Neurophysiol.* 29 (1970), pp. 306–310.

[164] S. M. Pincus, I. M. Gladstone, and R. A. Ehrenkranz. "A regularity statistic for medical data analysis". In: *J. Clin. Monit.* 7.4 (1991), pp. 335–345.

[165] J. S. Richman and J. R. Moorman. "Physiological time-series analysis using approximate entropy and sample entropy." In: *Am. J. Physiol. Heart Circ. Physiol.* 278.6 (2000), H2039–H2049.

[166] M. Costa, A. L. Goldberger, and C. K. Peng. "Multiscale entropy analysis of biological signals". In: *Phys. Rev. E* 71.2 (2005), p. 021906.

[167] W. Y. Lu et al. "Multiscale entropy of electroencephalogram as a potential predictor for the prognosis of neonatal seizures". In: *PLoS One* 10.12 (2015), pp. 1–11.

[168] M. J. Katz. "Fractals and the analysis of waveforms". In: *Comput. Biol. Med.* 18.3 (1988), pp. 145–156.

[169] T. Higuchi. "Approach to an irregular time series on the basis of the fractal theory". In: *Phys. D Nonlinear Phenom.* 31.2 (1988), pp. 277–283.

[170] J. F. Kaiser. "On a simple algorithm to calculate the energy of a signal". In: *Acoust. Speech, Signal Process. 1990. ICASSP-90., 1990 Int. Conf.* 2.10 (1990), pp. 381–384.

[171] N. E. Huang et al. "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis". In: *Proc. R. Soc. A Math. Phys. Eng. Sci.* Vol. 454. 1971. The Royal Society, 1998, pp. 903–995.

[172] K. E. Atkinson. *An introduction to numerical analysis*. New York, NY, USA: John Wiley & Sons, 2008.

[173] D. Abasolo et al. "Entropy analysis of the EEG background activity in Alzheimer's disease patients." In: *Physiol. Meas.* 27.3 (2006), pp. 241–53.

[174] A. Temko et al. "EEG-based neonatal seizure detection with Support Vector Machines". In: *Clin. Neurophysiol.* 122.3 (2011), pp. 464–473.

[175] R. S. Huang, L. L. Tsai, and C. J. Kuo. "Selection of valid and reliable EEG features for predicting auditory and visual alertness levels." In: *Proc. Natl. Sci. Counc.* 25.1 (2001), pp. 17–25.

[176] B. R. Greene et al. "A comparison of quantitative EEG features for neonatal seizure detection". In: *Clin. Neurophysiol.* 119.6 (2008), pp. 1248–1261.

[177] C. Li. "Complexity Analysis of Physiological Time Series with Applications to Neonatal Sleep Electroencephalogram Signals". PhD thesis. Case Western Reserve University, 2013.

[178] E. Biagioni et al. "Distribution of sleep and wakefulness EEG patterns in 24-h recordings of preterm and full-term newborns." In: *Early Hum. Dev.* 81.4 (2005), pp. 333–9.

[179] Y. Li et al. "Sleep stage classification based on EEG Hilbert-Huang transform". In: *2009 4th IEEE Conf. Ind. Electron. Appl.* IEEE, 2009, pp. 3676–3681.

[180] L. Fraiwan et al. "Automated sleep stage identification system based on time-frequency analysis of a single EEG channel and random forest classifier". In: *Comput. Methods Programs Biomed.* 108.1 (2012), pp. 10–19.

[181] P. A. Estévez et al. "Polysomnographic pattern recognition for automated classification of sleep-waking states in infants". In: *Med. Biol. Eng. Comput.* 40.1 (2002), pp. 105–113.

[182] A. Temko et al. "Inclusion of temporal priors for automated neonatal EEG classification". In: *J. Neural Eng.* 9 (2012), p. 046002.

[183] C. Rao. *Linear statistical inference and its applications*. 2nd. New York, NY, USA: John Wiley & Sons, 1973.

[184] J. J. Faraway. *Linear Models with R*. 2nd. Boca Raton, FL, USA: Chapman & Hall/CRC, 2014.

[185] N. L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous Univariate Distributions, Volume 2*. Second. New York, NY, USA: Wiley, 1995.

[186] K. Pearson. "Notes on regression and inheritance in the case of two parents". In: *Proc. R. Soc. London*. 1985, pp. 240–242.

[187] C. Spearman. "The Proof and Measurement of Association between Two Things". In: *Am. J. Psychol.* 15.1 (1904), p. 72.

[188] D. Ruppert, M. Wand, and R. Carroll. "Mixed Models". In: *Semiparametric Regres.* New York, NY, USA: Cambridge University Press, 2003. Chap. 4, pp. 91–111.

[189] J. Stuart and N. Peter. *Artificial Intelligence a Modern Approach*. Vol. 72. 1-2. 2003, pp. 81–138.

[190] L. Breiman et al. *Classification and Regression Trees*. The Wadsworth and Brooks-Cole statistics-probability series. Taylor & Francis, 1984.

[191] L. Breiman. "Random forests". In: *Mach. Learn.* 45.1 (2001), pp. 5–32.

[192] G. Louppe et al. "Understanding variable importances in forests of randomized trees". In: *Adv. Neural Inf. Process. Syst. 26 (NIPS 2013)*. 2013, pp. 431–439.

[193] R. Tibshirani. "Regression Shrinkage and Selection via the Lasso". In: *J. R. Stat. Soc.* 58.1 (1996), pp. 267–288.

[194] H. Peng, F. Long, and C. Ding. "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 27.8 (2005), pp. 1226–1238.

[195] A. Tsanas, M. A. Little, and P. E. Mcsharry. "A methodology for the analysis of medical data". In: *Handb. Syst. Complex. Heal.* New York, NY, USA: Springer Science+Business Media, 2013. Chap. 7, pp. 113–125.

[196] K. P. Murphy. "Dynamic Bayesian Networks: Representation, Inference and Learning". PhD thesis. University of California, Berkeley, 2002.

[197] A. Cohen. “Hidden Markov models in biomedical signal processing”. In: *20th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* Vol. 20. 3. 1998, pp. 1145–1150.

[198] M. Gales and S. Young. “The Application of Hidden Markov Models in Speech Recognition”. In: *Found. Trends Signal Process.* 1.3 (2007), pp. 195–304.

[199] N. L. Hjort et al. *Bayesian Nonparametrics*. New York, NY, USA: Cambridge University Press, 2010.

[200] R. M. Neal. “Markov Chain Sampling Methods for Dirichlet Process Mixture Models”. In: *J. Comput. Graph. Stat.* 9.2 (2000), pp. 249–265.

[201] J. Pitman and M. Yor. “The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator”. In: *Ann. Probab.* 25.2 (1997), pp. 855–900.

[202] A. Gelman. *Bayesian data analysis*. Boca Raton, FL, USA: Chapman & Hall/CRC, 2014, p. 73.

[203] C. M. Bishop. *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer Science+Business Media, 2006.

[204] S. Arlot and A. Celisse. “A survey of cross-validation procedures for model selection”. In: *Stat. Surv.* 4 (2010), pp. 40–79.

[205] D. Krstajic et al. “Cross-validation pitfalls when selecting and assessing regression and classification models”. In: *J. Cheminform.* 6.1 (2014), p. 10.

[206] D. G. Altman and J. M. Bland. “Diagnostic tests. 1: Sensitivity and specificity.” In: *BMJ* 308.6943 (1994), p. 1552.

[207] D. Powers. “Evaluation: From Precision, Recall and F-Measure To ROC, Informedness, Markedness & Correlation”. In: *J. Mach. Learn. Technol.* 2.1 (2011), pp. 37–63.

[208] R. M. Rangayyan. *Biomedical Signal Analysis*. 2nd. Hoboken, New Jersey: John Wiley & Sons inc., 2015.

[209] S. V. Stehman. “Selecting and interpreting measures of thematic classification accuracy”. In: *Remote Sens. Environ.* 62.1 (1997), pp. 77–89.

[210] J. Cohen. “A Coefficient of Agreement for Nominal Scales”. In: *Educ. Psychol. Meas.* 20.1 (1960), pp. 37–46.

[211] J. N. Mandrekar. “Measures of interrater agreement.” In: *J. Thorac. Oncol.* 6.1 (2011), pp. 6–7.

[212] K. Gwet. *Handbook of inter-rater reliability: the definitive guide to measuring the extent of agreement among multiple raters*. Third. Gaithersburg, MD, USA: Advanced Analytics, LLC, 2012, pp. 121–128.

[213] A. Hayes and K. Krippendorff. “Answering the Call for a Standard Reliability Measure for Coding Data”. In: *Commun. Methods Meas.* 1.1 (2007), pp. 77–89.

[214] K. Krippendorff. “Agreement and information in the reliability of coding”. In: *Commun. Methods Meas.* 5.2 (2011), pp. 93–112.

[215] R. J. Hyndman and A. B. Koehler. "Another look at measures of forecast accuracy". In: *Int. J. Forecast.* 22.4 (2006), pp. 679–688.

[216] F. Wilcoxon. "Individual comparisons of grouped data by ranking methods." In: *J. Econ. Entomol.* 80-83.6 (1946), p. 269.

[217] M. De Vos et al. "Automated artifact removal as preprocessing refines neonatal seizure detection". In: *Clin. Neurophysiol.* 122.12 (2011), pp. 2345–2354.

[218] T. D. Lagerlund, F. W. Sharbrough, and N. E. Busacker. "Spatial filtering of multichannel electroencephalographic recordings through principal component analysis by singular value decomposition." In: *J. Clin. Neurophysiol.* 14.1 (1997), pp. 73–82.

[219] C. A. Kothe. *The Artifact Subspace Reconstruction method*. San Diego, CA, California, 2013.

[220] T. Mullen et al. "Real-time modeling and 3D visualization of source dynamics and connectivity using wearable EEG." In: *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng. Med. Biol. Soc. Annu. Conf.* 2013, pp. 2184–7.

[221] C. A. Kothe and S. Makeig. "BCILAB: A platform for brain-computer interface development". In: *J. Neural Eng.* 10.5 (2013), p. 056014.

[222] H. Azami et al. "An intelligent approach for variable size segmentation of non-stationary signals". In: *J. Adv. Res.* 6.5 (2015), pp. 687–698.

[223] J. Matyas. "Random optimization". In: *Autom. Remote Control* 26.2 (1965), pp. 246–253.

[224] M. S. Scher. "Understanding sleep ontogeny to assess brain dysfunction in neonates and infants." In: *J. Child Neurol.* 13 (1998), pp. 467–474.

[225] R. Genuer, J. Poggi, and C. Tuleau-malot. *Random Forests : some methodological insights*. Tech. rep. INRIA, 2008.

[226] R. Genuer, J. Poggi, and C. Tuleau-malot. "Variable selection using Random Forests". In: *Pattern Recognit. Lett.* 31.14 (2012), pp. 2225–2236.

[227] M. S. Scher. "Neurophysiological assessment of brain function and maturation: I. A measure of brain adaptation in high risk infants". In: *Pediatr. Neurol.* 16.3 (1997), pp. 191–198.

[228] K. Watanabe, F. Hayakawa, and A. Okumura. "Neonatal EEG: A powerful tool in the assessment of brain damage in preterm infants". In: *Brain Dev.* 21.6 (1999), pp. 361–372.

[229] S. Wang, J. Gwizdka, and W. A. Chaovalitwongse. "Using Wireless EEG Signals to Assess Memory Workload in the n-Back Task". In: *IEEE Trans. Human-Machine Syst.* 46.3 (2016), pp. 424–435.

[230] N. K. Schuurman, R. P.P. P. Grasman, and E. L. Hamaker. "A Comparison of Inverse-Wishart Prior Specifications for Covariance Matrices in Multilevel Autoregressive Models". In: *Multivariate Behav. Res.* 51.2-3 (2016), pp. 185–206.

[231] I. Alvarez, J. Niemi, and M. Simpson. *Bayesian inference for a covariance matrix*. 2014. arXiv: 1408.4050.

[232] A. Dereymaeker et al. "Interrater agreement in visual scoring of neonatal seizures based on majority voting on a web-based system: The Neoguard EEG database." In: *Clin. Neurophysiol.* 128.9 (2017), pp. 1737–1745.

[233] T. L. Griffiths and Z. Ghahramani. "The Indian Buffet Process: An Introduction and Review". In: *J. Mach. Learn. Res.* 12 (2011), pp. 1185–1224.

[234] C. E. Rasmussen. "Gaussian Processes in Machine Learning". In: *Adv. Lect. Mach. Learn.* Berlin Heidelberg: Springer-Verlag, 2004, pp. 63–71.

[235] M. A. Pimentel et al. "A review of novelty detection". In: *Signal Processing* 99 (2014), pp. 215–249.

[236] A. Supratak et al. "DeepSleepNet: A Model for Automatic Sleep Stage Scoring Based on Raw Single-Channel EEG". In: *IEEE Trans. Neural Syst. Rehabil. Eng.* 25.11 (2017), pp. 1998–2008.

[237] E. Kreyzig. *Advanced Engineering Mathematics*. 4th. New York, NY, USA: Wiley, 1979.

[238] R. Hogg and A. T. Craig. *Introduction to Mathematical Statistics*. New York, NY, USA: Macmillan, 1978.

[239] S. Senn and W. Richardson. "The first t-test". In: *Stat. Med.* 13.8 (1994), pp. 785–803.

[240] P. Welch. "The Use of Fast Fourier Transform for the Estimation of Power Spectra: A Method Based on Time Averaging Over Short, Modified Periodograms". In: *IEEE Trans. Audio Electroacoust.* 15.2 (1967), pp. 70–73.

[241] K. Pearson. "LIII. On lines and planes of closest fit to systems of points in space". In: *London, Edinburgh, Dublin Philos. Mag. J. Sci.* 2.11 (1901), pp. 559–572.

[242] D. MacKay. "An Example Inference Task: Clustering". In: *Inf. Theory, Inference Learn. Algorithms*. Cambridge: Cambridge University Press, 2003. Chap. 20, pp. 284–292.