KU LEUVEN

ARENBERG DOCTORAL SCHOOL
Faculty of Engineering Science
Department of Electrical Engineering

Development of a neonatal EEG monitor for automated brain analysis

Amir Hossein Ansari

Supervisor:
Prof. dr. ir. S. Van Huffel
Co-supervisor:
Prof. dr. G. Naulaers

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Engineering Science (PhD): Electrical Engineering

June 2018

Development of a neonatal EEG monitor for automated brain analysis

Amir Hossein ANSARI

Examination committee:
Prof. dr. ir. J. Berlamont, chair
Prof. dr. ir. S. Van Huffel, supervisor
Prof. dr. G. Naulaers, co-supervisor
Prof. dr. ir. J. Suykens
Prof. dr. L. Lagae
Prof. dr. ir. B. Puers
Prof. dr. P. Govaert
(Koningin Paola Kinderziekenhuis, Antwerp)
Prof. dr. ir. M. De Vos
(University of Oxford, Oxford, UK)

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Engineering Science (PhD): Electrical Engineering

June 2018

© 2018 KU Leuven – Faculty of Engineering Science
Uitgegeven in eigen beheer, Amir Hossein Ansari, Kasteelpark Arenberg 10, bus 2446, B-3001 Leuven (Belgium)

Alle rechten voorbehouden. Niets uit deze uitgave mag worden vermenigvuldigd en/of openbaar gemaakt worden door middel van druk, fotokopie, microfilm, elektronisch of op welke andere wijze ook zonder voorafgaande schriftelijke toestemming van de uitgever.

All rights reserved. No part of the publication may be reproduced in any form by print, photoprint, microfilm, electronic or any other means without written permission from the publisher.

Abstract

A neonatal intensive care unit (NICU) is a specialized unit for intensive care of the critically ill or premature newborn babies. This unit is equipped with incubators and different bedside systems to monitor and nurse the infants for several days or weeks. In the last decade, the electroencephalogram (EEG) has been considered as an important monitoring system in the NICUs. Different types of rhythms and patterns which manifest in the EEG signals can be related to brain development, sleep stages, and several brain abnormalities such as epileptic seizures. However, to detect these patterns, to differentiate them from artifacts, and to diagnose abnormalities, special expertise is required, which is not available around the clock. Furthermore, this analysis is very expensive, labor-intensive, and time-consuming. Thus, many NICUs only monitor limited number of neonates from whom brain monitoring is urgently needed. Therefore, developing a brain monitor equipped with automated algorithms and alarms alleviates the workload, decreases the costs, and gives the possibility to monitor more neonates simultaneously.

In order to develop such a neonatal brain monitor, this thesis focuses on two objectives: 1) developing new algorithms, or improving existing ones, for automated EEG analysis, 2) practical implementation of the algorithms for clinical use. For the former one, the research focuses on improving neonatal seizure detection, as well as neonatal sleep stage classification. To this end, first, the available recordings are characterized and the inter-rater agreement of several human EEG readers is studied. Then, a multi-stage seizure detector, which is an extension of a previously developed heuristic algorithm, is proposed. This multi-stage detector uses a machine learning technique as a post-processor and a novel set of features to discriminate seizures from artifacts. Next, an adaptive learning method is added to the detector in order to increase the reliability and accuracy of the algorithm through the use of an alarming system. Furthermore, two deep convolutional neural networks are proposed and applied to detect neonatal seizures and classify sleep stages. These networks are able to learn the required features automatically and consequently can classify raw

multi-channel EEG segments. Besides, a new framework for measuring the performance of algorithms with a multi-rated database and various degrees of high inter-rater disagreement is proposed. In this framework, 6 commonly used performance metrics - 3 epoch-based and 3 event-based metrics - are fuzzily extended and tested on the multi-rated seizure database and the results are compared with a majority voting technique. Finally, a learning platform is also developed based on the multi-rated seizure database annotated by 4 independent expert EEG readers.

For the second objective, several algorithms, developed by different researchers from our group, are implemented in a real-time and cloud-based monitor. These monitors can facilitate the clinical validation of the developed algorithms. The core of the real-time monitor includes different algorithms translated into a low-level programming language. On one end, it connects to an EEG recording machine wirelessly, and on the other end, it communicates with the developed graphical user interface to monitor the signals and show the alarms. The cloud-based monitor has also different modules for uploading the EEG recordings, processing and storing the data on the server, and displaying the signals and alarms on a secure platform on the web. In the cloud-based system, the algorithms can run in their native languages which facilitates validation of newly developed algorithms in several centers. The details about these algorithms and monitors are explained in this thesis and their added value, limitations, and future potential are discussed.

Beknopte samenvatting

De dienst neonatale intensieve zorg is een gespecialiseerde eenheid voor de intensieve zorg voor ernstig zieke of premature baby's. Deze eenheid is uitgerust met incubators en verschillende apparatuur aan de bedzijde om de boorlingen gedurende enkele dagen of weken te monitoren en te verzorgen. In het afgelopen decennium werd het elektro-encefalogram (EEG) beschouwd als een belangrijk monitoring systeem op de afdeling neonatale intensieve zorgen. Verschillende soorten ritmes en patronen die zich in de EEG-signalen manifesteren, kunnen verband houden met de hersenontwikkeling, slaapstadia en verschillende hersenafwijkingen zoals epileptische aanvallen. Voor het detecteren van deze patronen, om ze te onderscheiden van artefacten en/of om afwijkingen te diagnosticeren, is speciale expertise vereist, die in vele centra niet permanent aanwezig is. Bovendien is deze analyse erg duur, arbeidsintensief en tijdrovend. Dit heeft tot gevolg dat enkel een beperkt aantal neonaten die dringende opvolging vereisen, ook op die manier gemonitord kan worden. De ontwikkeling van een hersenmonitor die is uitgerust met geautomatiseerde algoritmen en alarmen zal daarom de werklast en kosten verlagen, en de mogelijkheid bieden om meer neonaten gelijktijdig te monitoren.

Om een dergelijke neonatale hersenmonitor te ontwikkelen, richt dit proefschrift zich op twee doelstellingen: 1) ontwikkeling van nieuwe algoritmen, of verbetering van bestaande, voor automatische EEG-analyse, 2) praktische implementatie van de algoritmen voor klinisch gebruik. Voor de eerste doelstelling is het onderzoek gericht op het verbeteren van detectie van neonatale aanvallen, evenals de classificatie van neonatale slaapstadia. Hiertoe worden eerst de beschikbare opnamen gekarakteriseerd en de interbeoordelaarsbetrouwbaarheid van verschillende EEG experts onderzocht. Vervolgens wordt een meertraps-aanvalsdetector voorgesteld, die een uitbreiding is van een eerder ontwikkeld heuristisch algoritme. Deze meertraps-detector gebruikt een machinale leertechniek als een post-processor en een nieuwe verzameling kenmerken om aanvallen van artefacten te onderscheiden. Vervolgens wordt een adaptieve leermethode toegevoegd aan de detector om de betrouwbaarheid

en nauwkeurigheid van het algoritme te vergroten door het gebruik van een alarmsysteem. Verder worden twee diepe convolutionele neurale netwerken voorgesteld en toegepast om neonatale aanvallen te detecteren en slaapstadia te classificeren. Deze netwerken zijn in staat om de vereiste kenmerken automatisch te leren en kunnen daarom onbewerkte meerkanaals EEG-segmenten classificeren. Bovendien wordt een nieuwe manier voorgesteld voor het meten van de prestaties van algoritmen met een database met meerdere beoordelingen en verschillende graderingen van interbeoordeelaarsbetrouwbaarheid. In dit kader worden 6 veelgebruikte prestatiematen, 3 epoch-gebaseerde en 3 event-gebaseerde maten, uitgebreid met fuzzy logic en getest op de aanvalsdatabase die gelabeld werd door verschillende beoordeelaars. Deze resultaten worden vergeleken met de techniek van meerderheidsstemmen. Ten slotte wordt de aanvalsdatabase, geannoteerd door 4 onafhankelijke EEG-experts, ook aangewend voor de ontwikkeling van een leerplatform.

Voor de tweede doelstelling zullen verscheidene algoritmen, ontwikkeld door verschillende onderzoekers van onze groep, worden geïmplementeerd in een real-time en een cloud gebaseerde monitor. Deze monitors zullen de klinische validatie van de ontwikkelde algoritmen vergemakkelijken. De kern van de real-time monitor bevat verschillende algoritmen die zijn vertaald naar een low-level programmeertaal. Aan de ene kant verbindt het draadloos met een EEG-registratiemachine en aan de andere kant communiceert het met de ontwikkelde grafische gebruikersinterface om de signalen te monitoren en de alarmen te tonen. De cloud gebaseerde monitor bestaat ook uit verschillende modules voor het opladen van de EEG-opnamen, het verwerken en opslaan van de gegevens op de server en het weergeven van de signalen en alarmen op een beveiligd platform op het web. In het cloud gebaseerde systeem kunnen de algoritmen in hun eigen taal worden uitgevoerd, wat de validatie van nieuw ontwikkelde algoritmen in verschillende centra vergemakkelijkt. De details over deze algoritmen en monitors worden in deze thesis uiteengezet en hun toegevoegde waarde, beperkingen en toekomstige mogelijkheden worden besproken.

Nomenclature

Metrics

cm	centimeter
GB	gigabyte
Hz	Hertz
h	hour
min	minute
ms	millisecond
mV	millivolt
μV	microvolt
s	second

Abbreviations

AED	Anti-Epileptic Drug
aEEG	Amplitude-integrated EEG
ANN	Artificial Neural Network
AP	Action Potential
app	Application
AUC	Area Under the Curve
cEEG	Continuous Electroencephalogram
CFM	Cerebral Function Monitor
CNN	Convolutional Neural Network
CPU	Central Processing Unit
CSS	Cascading Style Sheets
DB	Database
DNN	Deep Neural Network
ECG	Electrocardiogram
EDF	European Data Format
EEG	Electroencephalogram
EER	Enhanced Entity-Relationship
EMCR	Erasmus Medical Center Rotterdam

EMG	Electromyogram
EOG	Electro-oculogram
ERS	Event-Related Synchronization
FAR	False Alarm Rate
FIR	Finite Impulse Response
GDR	Good Detection Rate
GPU	Graphics processing Unit
GUI	Graphical User Interface
HIE	Hypoxic-Ischemic Encephalopathy
HTML	Hypertext Markup Language
IBI	Inter-Burst Interval
LASSO	Least Absolute Shrinkage and Selection Operator
LDA	Linear Discriminant Analysis
LMS	Least Mean Squares
LS-SVM	Least Squares-Suport Vector Machine
LSTM	Long Short-Term Memory
MAF	Moving Average Filter
MLP	Multilayer Perceptron
MPC	Mean Phase Coherence
MRI	Magnetic Resonance Imaging
MSE	Mean Squared Error
NGDC	NeoGuard Data Collector
NGIS	NeoGuard Information System
NGLP	NeoGuard Learning Platform
NGSP	NeoGuard Scoring Platform
NICU	Neonatal Intensive Care Unit
NSD	Neonatal Seizure Detector
PMA	Post-Menstrual Age
PPV	Positive Predictive Value
RBF	Radial Basis Function
ReLU	Rectified Linear Unit
Resp	Respiration
RF	Random Forest
ROC	Receiver Operating Characteristic
SDA	Seizure Detection Algorithm
Sel	Selectivity
Sen	Sensitivity
Spe	Specificity
SQL	Structured Query Language
SSE	Sum of Squared Errors
SVM	Suport Vector Machine
TKE	Teager–Kaiser nonlinear Energy
USB	Universal Serial Bus
UZA	University Hospitals/Ziekenhuizen(Dutch) Antwerp
UZL	University Hospitals/Ziekenhuizen(Dutch) Leuven

Contents

Abstract i

Nomenclature v

Contents vii

List of Figures xv

List of Tables xxiii

Part I Introduction 1

1 General introduction 3

1.1 Problem Statement 3

1.2 NeoGuard consortium 4

1.3 Research objectives 6

1.4 Chapter overview 7

1.4.1 Part I, General introduction 7

1.4.2 Part II, Database and scoring 7

1.4.3 Part III, Algorithms 9

1.4.4 Part IV, Practical Implementation 10

1.4.5 Part V, Conclusion 10

2 Physiological background **11**

2.1 Human brain and electroencephalogram 12

2.2 Hypoxic-ischemic encephalopathy 14

2.3 Neonatal seizures 15

2.4 Neonatal Sleep-wake cycles 17

2.5 Artifacts in neonatal EEG 19

Part II Database and Scoring **25**

3 Database **27**

3.1 Introduction 28

3.2 NeoGuard database for term neonates 28

3.2.1 EMCR 32

3.2.2 UZL 32

3.3 NeoGuard database for preterm neonates 33

4 Inter-rater agreement of neonatal seizures **37**

4.1 Background 38

4.2 Data description 40

4.3 Methods 40

4.3.1 Seizure annotation 40

4.3.2 Seizure detection algorithm 40

4.3.3 NeoGuard Website 41

4.3.4 Scoring of the events 41

4.3.5 Measures of agreement 43

4.4 Results 44

4.4.1 Database 44
4.4.2 Majority vote and inter-rater agreement 44
4.4.3 Electrographic characteristics 46
4.5 Discussion 48
4.6 Conclusion 50

Part III Algorithms 53

5 Mathematical background 55
5.1 Support vector machines 56
5.1.1 Principle of SVM using hard margin 56
5.1.2 Soft margin 57
5.1.3 Kernel trick 58
5.2 Convolutional neural networks 59
5.3 Random Forest 61
5.4 Ridge regression 62
5.5 LASSO 63
5.6 Bootstrapping 64

6 Multi-stage neonatal seizure detection 67
6.1 Background 68
6.2 Data description 69
6.3 Proposed method 69
6.3.1 Stage I: Heuristic algorithm 70
6.3.2 Stage II: Data-driven post-processor 70
6.4 Results 77
6.4.1 Improvement of performance metrics 77

6.4.2 Analysis of event duration 77
6.4.3 Analysis of EEG experts' agreement 79
6.4.4 Feature Ranking 80
6.5 Discussion 81
6.6 Conclusion 83

7 Seizure detection using adaptive learning 85
7.1 Background 86
7.2 Data description 86
7.3 Methods 87
7.3.1 The multi-stage detector 87
7.3.2 Proposed method 87
7.4 Results and discussion 90
7.5 Conclusion 92

8 Seizure detection using convolutional neural networks 93
8.1 Background 94
8.2 Materials and methods 97
8.2.1 Data description 97
8.2.2 Segmenting 97
8.2.3 Heuristic method 98
8.2.4 Feature-based approaches 98
8.2.5 Proposed CNN-RF 99
8.3 Results 102
8.4 Discussion 107
8.5 Conclusion 113

9 Sleep stage classification using convolutional neural networks 115

9.1 Background 116
9.2 Materials and methods 118
9.2.1 Data description 118
9.2.2 Proposed method 118
9.2.3 Feature-based neonatal sleep stage classifier 120
9.2.4 Cluster-based Adaptive Sleep Staging (CLASS) 120
9.2.5 Computational Time 121
9.3 Results 121
9.3.1 Classification Performance 121
9.3.2 Computational Time 123
9.4 Discussion 123
9.5 Conclusion 125

10 Weighted performance metrics 127

10.1 Background 128
10.2 Data description 129
10.3 Methods 129
10.3.1 Classical metrics 129
10.3.2 Multi-label problems 132
10.3.3 Averaging (ideal reference) 133
10.3.4 Majority Voting 134
10.3.5 Proposed weighted metrics 135
10.3.6 Bootstrapping test 136
10.3.7 Tested Seizure Detectors 136
10.4 Results 137
10.4.1 Accuracy 137
10.4.2 Robustness 138

10.5 Discussion 139
10.6 Conclusion 143

Part IV Practical Implementations 145

11 NeoGuard information system 147

11.1 Background 148
11.1.1 Programming languages 148
11.1.2 The accessibility of the users 149
11.1.3 Database structure 149
11.2 NeoGuard Data Collector 152
11.2.1 Uploading centers 152
11.2.2 The GUI 152
11.3 NeoGuard Scoring platform 155
11.4 Conclusion 155

12 NeoGuard learning platform 157

12.1 Background 158
12.2 Data description 160
12.2.1 Recording data 160
12.2.2 Rescoring strategy 160
12.3 Method 161
12.3.1 Scoring labels 161
12.3.2 Level of agreement 162
12.3.3 Measuring raters' agreement 164
12.3.4 Storing the users' Info 165
12.3.5 Start page 165

12.3.6 Scoring page 165
12.4 Results 166
12.5 Discussion and Conclusion 169
12.6 Intellectual property 170

13 NeoGuard: A real-time monitor for NICUs 171
13.1 Introduction 172
13.2 Materials and methods 172
13.2.1 NeoGuard hardware 173
13.2.2 The core 173
13.2.3 Graphical user interface 176
13.3 Conclusion 177

14 NeoGuard: A cloud-based remote monitor for NICUs 179
14.1 Background 180
14.2 Design 181
14.3 Developed algorithms 182
14.4 Graphical user interface 183
14.5 Current centers and devices 183
14.6 Limitations 186
14.7 Conclusion 187

Part V Conclusion 189

15 Conclusion 191
15.1 Summary 191
15.1.1 Algorithms 191
15.1.2 Implementations 194

15.2 Future directions 195
15.2.1 Algorithms 195
15.2.2 Implementations 197

A Previously developed NeoGuard algorithms **199**

A.1 Heuristic seizure detector 199
A.1.1 Version 1 199
A.1.2 Version 2 201
A.2 Amplitude-integrated EEG 202
A.3 Inter-burst interval 205
A.3.1 IBI for preterm neonates based on line length 206
A.3.2 dynamic IBI for term neonates 207

Bibliography **209**

Curriculum vitae **231**

List of publications **233**

List of Figures

1.1 Official NeoGuard logo 5

1.2 Schematic overview of the chapters in this thesis. 8

2.1 Four lobes of the cerebral cortex of the brain [1] 12

2.2 locations of electrodes in 10-20 system. A) frontal view, B) occipital view , C) left temporal view, D) right temporal view, E) top view, and F) names of electrodes at their approximate positions in the top view [2] 13

2.3 A real example of an EEG segment having a seizure. The horizontal axis shows the time in seconds. The names of bi-polar channels are listed on the vertical axis. The distance between the channel baselines equal $176\mu V$. 14

2.4 Three real examples of neonatal seizure. a) a spike-train type seizure with high peak-to-peak amplitude mainly observed at C4. b) an oscillatory type seizure with peak-to-peak amplitude of $50\mu V$ observed at C3. c) a low amplitude seizure ($\leqslant 15\mu V$) mainly at Cz. 18

2.5 Changes of quiet sleep and active sleep with post-menstrual age (in weeks). 19

2.6 Example of an active sleep and a quiet sleep EEG segment at 32 and 38 PMA respectively. a) Active sleep with continuous activity and mixed frequencies associated with REM. b) Quiet sleep with discontinuous EEG tracing and temporal Theta activity. 20

2.7 ECG artifact mainly visible at O1 21

2.8 EMG artifact due to chin movement 21

2.9 EOG artifact mainly visible at F8 23

2.10 Respiratory artifact mainly visible at T3 and T5 23

2.11 Electrode artifact visible at O2 24

3.1 A definite mix-type seizure observed mainly at Cz. 30

3.2 An arrhythmic dubious seizure observed at C_3. The frequency of spikes varies from 0.25 to 0.5 Hz and the peak-peak amplitude is lower than 20 µV. This dubious seizure is associated with very severe abnormality of EEG background (grade 8). 31

3.3 Neonatal EEG montages. A: full 10-20 system of electrode placement using 17 electrodes, B) a restricted 10-20 system using 13 electrodes, C) a restricted 10-20 system using 9 electrodes. The arrows in these montage maps define the bipolar channels (e.g. $C_Z - C_3$) and their lengths do not represent any meaningful value. 33

3.4 Example of a quiet sleep and non-quiet sleep EEG segment at 32 weeks and 2 days PMA. (a) Continuous tracing during active sleep. Delta brushes in temporal and occipital regions, irregular breathing pattern. (b) Discontinuous tracing during quiet sleep. IBI shorter than 15 sec, temporal theta activity and occipital delta brushes, more regular breathing pattern. 35

4.1 Examples of the 3 different event categories. A) This event was scored as 'definite seizure' by all raters. B) This event was scored as 'dubious' by 3 raters and as 'non-seizure' by 1 rater. C) This event was scored as 'non-seizure' by all raters. 42

4.2 Database overview: the number of all the events, classified as seizures and non-seizures with a given median duration based on majority voting. 45

4.3 Majority voting for all the events, seizures and non-seizures. Poor agreement (Poor): 2 raters agreed, moderate agreement (Mod.): 3 raters agreed, high agreement (High): 4 raters agreed. 45

4.4 Inter-rater agreement among all raters and for all the events, presented with Fleiss' Kappa (k). P: primary rater. SR 1-3: secondary raters. 46

4.5 Effect of seizure/non-seizure duration on majority voting. Poor agreement (Poor): 2 raters agreed, moderate agreement (Mod.): 3 raters agreed, high agreement (High): 4 raters agreed. 47

5.1 Histogram of the means of the bootstrapped samples 57

5.2 Architecture of a convolutional neural network. This network is used in chapter 9 for automatic neonatal sleep stage classification. 60

5.3 Histogram of the means of the bootstrapped samples 65

6.1 Diagram of the multi-stage detector. The input of the post-processor is the detected segments of the heuristic model 70

6.2 The behavior of the amplitude, frequency, and morphology evolution for a detected seizure. (a) Original detected EEG segment lasting for 75 seconds. The vertical red line shows the detected center of the segment where TKE is maximum. (b) Smoothed signal of rectified EEG amplitude (black line), detected local peaks of the smoothed amplitude (red dots), and the fitted line over the peaks (blue dashed line). The slope of this line determines the Amplitude Evolution. (c) The mean power frequency of 2s epochs of the EEG (red dots) and the fitted line (blue dashed line). The slope of this line determines the Frequency Evolution. (d) The normalized cross-correlation between the EEG and the template (black line), detected local peaks of the cross-correlation signal (red dots), and the fitted line (blue dashed line). The slope of this line determines the Morphology Evolution. In the traces (b-d), the reader's left side shows the 'onset' subsection and the right side shows the 'end' subsection. 73

6.3 Two detected segments by the heuristic algorithm, (a) is a falsely detected ECG artifact (MPC: 0.21) and (b) is a truly detected rhythmic seizure (MPC: 0.04). 75

6.4 Diagram of the decision making layer 77

6.5 ROC curves of variation of the Good Detection Rate (%) against False Alarm Rate (h^{-1}) for the four datasets. The marked points indicate the performance at different sensitivity thresholds: 0 (filled circle), 0.1 (square), and 0.3 (rectangle) 78

6.6 Histogram of the true and false detections for the original heuristic method (dashed bars) and the proposed multi-stage method (grey bars) as function of duration 79

6.7 Histogram of the definite seizures and artifacts for the original heuristic method (dashed bars) and the proposed multi-stage method (grey bars) as a function of rater's agreement 80

6.8 Performance of the proposed multi-stage classifier tested on DB2-4 compared to the reported performance of other methods. The grey circles show the previously reported performance of the heuristic method tested on different databases. The dark circles show the performance of the heuristic method and its extensions when it is applied on DB2-4. Furthermore, the light diamond shows the GDR (with one false alarm per hour) reported in [3] while the dark diamond shows the GDR of its re-implementation tested on DB2-4. 84

7.1 Diagram of the proposed technique. The TH used in the post-processor is adaptively re-turned by the proposed 3rd stage based on the current feedback, elapsed time from that last feedback, detected channel, and the old value of TH. 89

7.2 The ROC curve of the seizure detector with fixed threshold (dashed line) as well as the proposed adaptive threshold (continuous line) for DB2 (top) and DB3 (bottom). 91

7.3 Histogram of the detected seizures (top), the number of false alarms (middle), and the PPVs (bottom), for the original method with fixed threshold (light gray) and the proposed adaptive learning method (dark gray). In the upper chart, the white stacked bars show the number of missed seizures. 91

8.1 Schematic overview of the proposed CNN-RF method. 100

8.2 Schematic structure of the proposed CNN method. 101

8.3 Summary of the automatically extracted features (f1-20) by the CNN from the test dataset. For each feature, the first and second boxes are corresponding to the non-seizure and seizure segments. The features starting with star (*) correspond to the features removed after the pruning process. All features are plotted after being normalized between 0 and 1. 104

8.4 The AUCs of different classifiers when the features are extracted by the CNN (a) and Alg. 3 (b). 105

8.5 The ROC curves of the heuristic, feature-based, and proposed CNN-RF methods for the test data; (A) for all neonates in the test dataset and (B) after excluding 7 neonates which did not have appropriate training patterns in the training dataset. . . . 105

8.6 Qualitative example of a seizure segment and outputs for some layers. A) A seizure segment with 20 bipolar channels. The x-axis is time in seconds. B) The output of Conv layers 7 and 13, as well as Pooling layers 17 and 18, and the final output of the CNN after the classification layers. The red-highlighted boxes correspond to the filters removed by the pruning process. . . . 107

8.7 Qualitative example of a non-seizure segment and outputs for some layers. A) A non-seizure segment with 20 bipolar channels. The x-axis is time in seconds. B) The output of Conv layers 7 and 13, as well as Pooling layers 17 and 18, and the final output of the CNN after the classification layers. The red-highlighted boxes correspond to the filters removed by the pruning process. 108

8.8 The execution time is shorter for the CNN-RF when compared to the heuristic and feature-based methods. The time is measured in seconds for each segment (90 s, 20 channels). 108

9.1 Architecture of the convolutional neural network. 119

9.2 The ROC curves for the proposed CNN sleep stage classifier. The light grey ROC curves show the performance of the classifier for each recording. The black dashed and full line represent the mean ROC and median ROC curve respectively. 122

9.3 The histogram of the training data segments is displayed in light grey (left y axis). The squares show the AUC for each recording from the test set (right y axis). 122

9.4 The average recall time for 2 h multi-channel EEG segments for each of the three algorithms. The total recall time (left) and the time for classification (right) are shown. 124

10.1 An imaginary example of an event and epoch-based assessment. A) The ground truth including 3 scored seizures in 3 hours. B) The output of an automated detector when each line segment corresponds to an epoch. C) Event-based segments checking the overlap of the ground truth with the detections (I) and vice versa (II). D) Epoch-based segments and the number of involved epochs. 130

10.2 An agreed seizure starting from second 10 and lasting for 70 seconds. This event was primarily scored as 'seizure' by the primary neurophysiologist with rhythmic oscillation and sharp spikes (mixed) at C_4 and T_4 (peak to peak amplitude of spikes is 200 µV). All secondary raters also rescored it as 'seizure'. . . 132

10.3 A moderately agreed dubious seizure starting from second 10 and lasting for 20 seconds. This event was primarily scored as 'dubious seizure' by the primary neurophysiologist with arrhythmic oscillation and sharp spikes (mixed) at $T_6 - O_2$ (peak to peak amplitude of spikes is 40-50 µV). It was rescored as 'Seizure', 'Dubious', and 'Dubious' by the secondary raters. . . 133

10.4 Disagreed seizure starting from second 10 and lasting for 25 seconds. This event was primarily scored as 'seizure' by the primary neurophysiologist with rhythmic oscillation and sharp spikes (mixed) at C_4 (peak to peak amplitude of spikes is 10-20 µV). It was then rescored as 'Seizure', 'Dubious', and 'Non-seizure' by the secondary raters. 134

10.5 Probability density function of GDR, FAR, PPV, SEN, SPE, and SEL when the detector is A) the heuristic algorithm proposed in [4], B) the improved heuristic algorithm proposed in [5], and C) the multi-stage classifier proposed in [6]. In all traces, continuous, dotted, and dashed lines are respectively corresponding to the PDFs of the reference, classical metrics using majority voting, and proposed weighted metrics. 140

11.1 The EER diagram of the designed SQL model. 151

11.2 Welcome page of the NGIS 154

11.3 Events as displayed on the NeoGuard Website for scoring. Screenshot of image 3, a detailed view of the 20 - 40 seconds of the event. 156

12.1 Schematic diagram of the multi-scoring scenario. R1-3 denote the 3 secondary raters and ASD is the automated seizure detector. 161

12.2 A definite artifact that was scored as non-seizure by the primary neurophysiologist and all secondary raters. 162

12.3 A definite seizure, starting from second 10 and lasting for 19 seconds over the right temporal region. This event was scored as 'seizure' by the primary neurophysiologist and all secondary raters. 163

12.4 A dubious seizure starting from second 10 and lasting for 15 seconds, with arrhythmic oscillations, without clear evolution in morphology, at Cz electrode, with an amplitude of $25\mu V$. This event was scored as 'dubious seizure' by the primary neurophysiologist. It was then rescored as (Dsz, Dsz, Sz) by the secondary raters. 163

12.5 A screenshot of the NGLP start page 166

12.6 The view of the upper part of the scoring page including the scoring buttons, and images 1-3. In each image, the EEG and polygraphic signals are shown. 167

12.7 The view of the lower part of the scoring page including the images 4-5 and the Print button. In each image the EEG and polygraphic signals are shown. 168

12.8 A screenshot for the result of a user who scored 10 events . . . 169

13.1 The main modules of the NeoGuard real-time monitor 173

13.2 The prototype of the NeoGuard hardware 174

13.3 The main units of the NeoGuard core 175

13.4 The main screen of the NeoGuard Monitor 177

14.1 The main modules of the NeoGuard cloud-based system 181

14.2 The GUI of the data uploader 184

14.3 The GUI of the data visualizer 185

14.4 The histogram of detected IBIs over the last hour of recording 186

15.1 A simple collaboration diagram of the methods used in this thesis. The solid and dashed lines similarly show the connections. The white blocks located in the "Older Algorithms" group correspond to the methods developed by other PhD researchers and used in this study mainly in the NeoGuard monitors. 192

A.1 Schematic overview of the heuristic method. The upper and lower lines show the spike-train type and oscillatory type seizure detectors respectively. 200

A.2 Example of a definite seizure detected by the spike-train detector. The gray highlighted spikes represent those detected by the algorithm. 201

A.3 Example of a definite seizure detected by the oscillatory detector. The rectangles represent the oscillatory seizure activities that were detected by the algorithm. 201

A.4 Example of a definite seizure detected by both spike-train and oscillatory detectors (mixed type). The highlighted spikes and solid rectangles represent respectively the spike-train and oscillatory patterns that were detected by the algorithm. 202

A.5 Illustration of the new features of the autocorrelation function used in version 2 [7]. 203

A.6 The bod plot of the asymmetrical filter 204

A.7 Example of an estimated aEEG. The patterns with high amplitude correspond to electroencephalographic seizures. . . . 205

A.8 Example of 65 seconds multi-channel EEG recording from a preterm infant and automatically detected IBI. The upper traces are the raw EEG and the lower one is the line length. The diamonds and the horizontal gray line respectively correspond to the burst segments and the threshold [8]. 206

A.9 Example of three detected dynamic IBIs in a multi-channel EEG. The blue, pink, and red rectangles respectively correspond to the low, medium, and high amplitude clusters. The thick solid rectangles show the detected IBIs after majority voting. The red and yellow arrows show the IBIs marked by an expert EEG reader [9]. 208

List of Tables

3.1 NeoGuard seizure database from term neonates 29

6.1 Selected signal features 76

6.2 The difference in performance between the heuristic and proposed methods 78

6.3 Ranked feature sets using ridge regression and MI 80

6.4 Performance of the original heuristic algorithm and its extensions compared to the proposed method 83

8.1 Extracted features used in Alg. 1 - 3 99

8.2 Layers of the designed network before pruning 103

8.3 Comparison of the performance metrics for the CNN and Heuristic methods 106

9.1 Layers of the designed network before pruning 119

9.2 The average classification performance for the proposed CNN, the CLASS and feature-based approach without (NP) and with (PP) post-processing. The area under the mean ROC curve (AUC), average sensitivity (Sen) at specificity (Spe) of 80% and 90%, and kappa (Spe = 90%) are set out. 123

10.1 Similarity of reference with proposed metrics and classical metrics using majority voting 138

10.2 Difference between the 1st and 2nd scoring trials 139

11.1 The different levels of accessibility rights 149

A.1 Asymmetrical filter characteristics 204

Part I

Introduction

Chapter 1

General introduction

1.1 Problem Statement

After delivery, sick newborn babies, which need intensive medical attention, will be admitted to neonatal intensive care units (NICUs). This need of intensive attention is mainly due to maternal factors (e.g. bleeding, age, diabetes), delivery factors (e.g. birth asphyxia, nuchal cord), and/or baby factors (e.g. age, weight, infection, seizure). Critically ill newborns, such as those with seizure, with severe asphyxia, and very premature neonates, need extra attention and special monitoring in the NICUs. Based on statistics, the prevalence of clinical seizures in newborn babies is estimated to be 2 to 3 per 1000 live births. In the developed countries, for every 1000 live births, 1 to 6 neonates with perinatal asphyxia have been reported. And on a local scale, in Flanders, 7% of neonates (among 70,000) are being prematurely born every year. In order to protect and treat these babies, continuous high-quality monitoring, especially from their brains, are needed.

In most NICUs, dozens of physiological vital signals including the electrocardiogram (ECG), heart rate, blood pressure, blood oxygen saturation, and amplitude-integrated electroencephalogram (aEEG), are often continuously being monitored. Among these, the aEEG, which is a compressed, rectified version of the EEG, is the only signal that represents the brain activity. Cerebral function monitor (CFM^{TM}) is the best known and widely used monitor for plotting the aEEG using 2 to 4 EEG channels. The CFM generally displays about 5 hours of a recording on one page and is utilized in many NICUs because of its ease of use and support. However, brief-duration, low-amplitude, or

focal/regional activities may be invisible on an aEEG signal. Furthermore, some types of artifacts can easily be misinterpreted, especially by non-experts.

Therefore, the multi-channel continuous EEG monitoring (cEEG) is the ideal method for monitoring the brain activities and detecting seizures. Generally, visual inspection of multi-channel cEEG along with video by an expert clinical neurophysiologist is considered to be the gold standard for diagnosing neonatal seizures. However, EEG interpretation is more expensive and labor-intensive as compared to the CFM. Additionally, reading EEG needs special knowledge and training which may not be available around the clock in many NICUs, even in big centers. Hence, a cEEG monitor equipped with some supportive tools , like seizure detection, is needed in order to alleviate the workload of clinicians, decrease the expenses of monitoring, and importantly improve the quality of monitoring and subsequently the treatment.

1.2 NeoGuard consortium

As mentioned, there has been a huge need for developing an automated monitor for the NICUs. To this end, studies have been pursued to automate the required algorithms mainly in the recent decade. In our research group, STADIUS-Biomed[1], the basic algorithm for automated neonatal seizure detection was developed in 2008. Then, an international consortium, namely NeoGuard (figure 1.1), was formed around this algorithm in 2010 with the goal of making a prototype automated neonatal brain monitor. Other algorithms have also been added by several researchers working in this consortium over the past years. The NeoGuard partners and collaborators are listed below.

- University Hospitals Leuven, department of neonatology and department of development and regeneration, Neonatal intensive care unit. Prof. G. Naulaers, A. Dereymaeker, K. Jansen, J. Vervisch, and L. Thewissen. The project was guided and coordinated by this center. Additionally, parts of the data used in this project were recorded in this center and annotated by its expert EEG interpreters.
- University of KU Leuven, department of electrical engineering (ESAT), division of dynamical systems, signal processing and data analytics (STADIUS). Prof. S. Van Huffel, A.H. Ansari, V. Matic, N. Koolen, O. De Wel, and A. Caicedo. The main body of the algorithms and the core of the NeoGuard prototype were developed in this division.

[1] Group of biomedical signal processing, division of STADIUS, center of dynamical systems, signal processing and data analytics, Department of Electrical Engineering (ESAT), KU Leuven, Belgium



Figure 1.1: Official NeoGuard logo

- University of KU Leuven, department of electrical engineering (ESAT), division of microelectronics and sensors (MICAS). Prof. R. Puers, H. De Clercq, and P. Pelgrims. The NeoGuard EEG recorder, the hardware, was developed in this division.
- Erasmus Medical Center and University of Rotterdam, department of neonatology. Prof. P. Govaert, L. De Wispelaere. J. Dudink, and R. Swarte. Parts of the data used in this project were recorded in this center.
- Erasmus Medical Center and University of Rotterdam, department of clinical neurophysiology. J.P. Cherian and G. Visser. Parts of the data used in this project were annotated by the expert clinical neurophysiologists of this center.
- Ghent University, department of telecommunications and information processing (TELIN). Prof. W. Philips, prof. E. Vansteenkiste, I. Despotovic, D. Babin, D. Van Haerenborgh. The graphical user interface of the NeoGuard prototype was developed in this center.

Furthermore, Prof. M. De Vos (University of Oxford, Oxford, UK) and Dr. C. Dielman (ZNA Koningin Paola Kinderziekenhuis, Antwerp, Belgium) were collaborating in this project. Besides the aforementioned specific responsibilities of each center, they were participating in regular meetings and had important and active roles in guiding, brainstorming, and supporting to make a proper user-friendly prototype.

1.3 Research objectives

The studies and implementations that will be explained in this thesis were part of the mentioned NeoGuard project that were carried out by this PhD candidate. Therefore, similarly to the NeoGuard project, the main goal of this thesis is developing a neonatal brain monitor equipped with automated algorithms and alarming systems for the NICUs. To this end, 6 challenges are considered in this thesis as follows:

- *Challenge 1:* in order to develop and validate the algorithms, a database is needed to be collected and annotated from different centers from two countries. Therefore, a secure web-based platform for collecting and annotating the data is needed to be provided. This platform is explained in chapter 11. The platform is used to collect parts of the EEG database, which is described in chapter 3, and is also used to re-annotate seizures by several expert clinical EEG readers. Analysis of this multi-rated database is described and discussed in chapter 4. Furthermore, a learning platform is made using this multi-rated database for public trainees and tutors, which is explained in chapter 12 and is publicly available on the project website (https://neoguard.net).

- *Challenge 2:* the mentioned basic neonatal seizure detector needs to be improved to have a higher accuracy and lower false alarm rate. To this end, 3 new algorithms using support vector machines, adaptive learning, and convolutional neural networks are proposed and validated. They are explained respectively in chapter 6, 7, and 8.

- *Challenge 3:* a new algorithm is needed for neonatal sleep stage classification. This method is developed by a deep convolutional neural network and can achieve the performance of two state-of-the-art algorithms. This is described in chapter 9.

- *Challenge 4:* In the presence of several annotations in our multi-rated seizure database and various degrees of inter-rater disagreement, a new framework is needed to be made in order to measure the overall performance. This framework is described in chapter 10.

- *Challenge 5:* in order to make the NeoGuard prototype, the back-end core should be developed which includes the main automated algorithms. To do so, the developed automated algorithms, which were mainly implemented in MATLAB, need to be translated and implemented in a low-level, multi-thread programming language, particularly C++, and be equipped with a user-friendly graphical user interface. It is considered in chapter 13.

- *Challenge 6:* in order to validate the newly developed algorithms (mainly in MATLAB) in different centers, a hardware-independent platform was needed. To this end, a cloud-based monitor was made and is explained in chapter 14.

In the next section, all chapters of this thesis are listed and briefly explained.

1.4 Chapter overview

This thesis is organized in 5 parts and 15 chapters. Figure 1.2 displays the schematic overview of these parts and chapters. In this figure, in the third part, Algorithms, the considered problems and the used methods are also shown. In the following paragraphs, each chapter is briefly explained.

1.4.1 Part I, General introduction

chapter 1, Introduction:
Current chapter introducing the thesis

chapter 2, Physiological background:
In this chapter, general physiological and clinical background concerning the human brain, electroencephalography, asphyxia, neonatal seizures, neonatal sleep stages, and artifacts are given.

1.4.2 Part II, Database and scoring

chapter 3, Database:
In this chapter, our databases including 1) term neonatal datasets recorded in two centers and used for seizure detection and 2) preterm dataset recorded in one center and used for sleep stage classification are explained.

chapter 4, Inter-rater agreement of neonatal seizures:
In this chapter, our multi-rated seizure database is investigated. To this end, 1919 events extracted from 280 h of EEG recordings from 71 neonates are reviewed by 4 independent raters. Next, the inter-rater agreement is calculated and its variability in identifying rhythmic ictal/non-ictal events is considered.

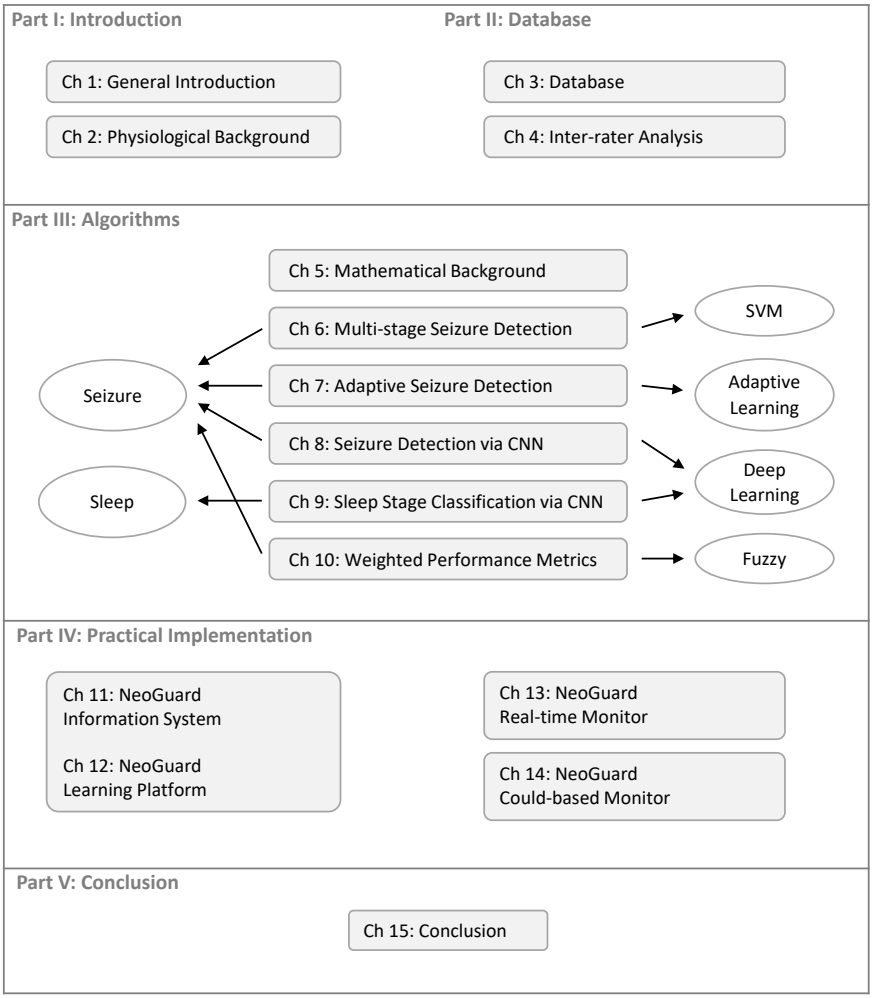


Figure 1.2: Schematic overview of the chapters in this thesis.

1.4.3 Part III, Algorithms

chapter 5, Mathematical background:
In this chapter, some general mathematical and algorithmic explanations concerning support vector machines, convolutional neural networks, random forests, regression models, and bootstrap are given.

chapter 6, Multi-stage neonatal seizure detection:
In this chapter, a new multi-stage seizure detector, which can improve the performance of our basic previously developed heuristic algorithm, is explained. In this method, after identifying the most seizure-relevant segments by the heuristic algorithm, a data-driven post-processor using a novel set of features is applied. The results show a significant improvement especially for false alarm rate per hour.

chapter 7, Seizure detection using adaptive learning:
In this chapter, a new stage is proposed to be added to the multi-stage classifier explained in the previous chapter, in order to use feedback of the clinical neurophysiologist. The idea is to adaptively re-tune a threshold of the second stage to improve the performance of detection of brief-lasting seizures. As a result, the false alarm rate significantly decreases.

chapter 8, Seizure detection using convolutional neural networks:
In this chapter, a convolutional neural network (CNN) is designed and proposed for neonatal seizure detection. This method does not need any hand-crafted feature engineering. It takes a raw multi-channel EEG segment and assigns a label (seizure/non-seizure) to the segment. It is shown that although it cannot outperform the heuristic and multi-stage methods, it can surpass the other state-of-the-art feature-based algorithms.

chapter 9, Sleep stage classification using convolutional neural networks:
In this chapter, a CNN is also designed and proposed for neonatal sleep stage classification. It is shown that this method can achieve the state-of-the-art performance with some extra advantages.

chapter 10, Weighted performance metrics:
In this chapter, a new framework is proposed for classification of 2-class problems when the labels are not clearly defined. Although this framework is mainly developed and tested for our multi-rated seizure database in the presence of dubious seizures and high inter-rater disagreement, it can be applicable for all 2-class problems like multi-scored sleep-stage classification.

1.4.4 Part IV, Practical Implementation

chapter 11, NeoGuard information system:
In this chapter, the developed NeoGuard information system (NGIS) which is a web-based platform for collecting EEG recordings and scoring EEG segments is explained.

chapter 12, NeoGuard learning platform:
Although the NeoGuard learning platform is part of the NGIS, which is explained in the previous chapter, because of its importance, it is described separately here. This chapter introduces and explains the NeoGuard public learning platform that can be used by trainees, tutors, and expert EEG readers who are interested to test their knowledge and learn from neonatal EEG-polygraphic segments scored by several expert EEG readers.

chapter 13, NeoGuard - A real-time monitor for NICUs:
In this chapter, the NeoGuard prototype, which is a real-time bedside monitor is explained. This prototype is composed of hardware and software (back/front-end) developed in three centers as mentioned in section 1.2. The back-end was implemented by this PhD candidate.

chapter 14, NeoGuard - A cloud-based remote monitor for NICUs:
In this chapter, the NeoGuard cloud-based monitor is described. This monitor has two key features: 1) it uses the MATLAB engine on the server-side, and, therefore, can easily run newly developed algorithms without the need of translation to a secondary language; and 2) the server-side codes (algorithms) and GUIs are designed to be able to work independently from the EEG recorder. It is hoped that this system is used in the future in clinical practice in several NICUs with different EEG machines to validate new algorithms.

1.4.5 Part V, Conclusion

chapter 15, Conclusion:
In this chapter, the thesis is summarized, the general conclusions are drawn, the main advantages and disadvantages of each method are listed, and finally the potential of the developed methods as well as future work is pointed out.

Chapter 2

Physiological background

This chapter provides a brief overview of the physiological and clinical background and principles treated in this thesis. It begins with physiology and anatomy of the human brain and basics of the electroencephalogram. Then, it introduces the hypoxic-ischemic encephalopathy as the main cause of morbidity and mortality in newborn babies. Next, neonatal seizures and their different types, as well as several related terminologies, are explained. Neonatal sleep-wake cycles and their general characteristics are introduced as well. Finally, the important artifacts contaminating the electroencephalogram in the neonatal intensive care units are reviewed.

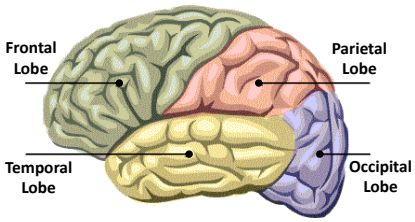


Figure 2.1: Four lobes of the cerebral cortex of the brain [1]

2.1 Human brain and electroencephalogram

The human brain is built up of approximately 10^{12} linked neurons which are forming a large neural network. The brain consists of the cerebrum, the cerebellum, and the brainstem. The outer and larger region of the cerebrum is the cerebral cortex and has an important role in memory, attention, cognition, language, etc. The cerebral cortex is anatomically divided into 4 areas: frontal, parietal, occipital, and temporal lobes (figure 2.1). In the cerebral cortex, as well as the whole central nervous system, the information is transmitted via moving action potentials (APs). Since APs are generated by the movement of some positive and negative ions, e.g. Na^{+1}, K^{+1}, Ca^{+2}, Cl^{-1}, across the cell membrane, changes of the electric or magnetic fields in the cortex can be associated with the activation of the corresponding neurons. Therefore, analyzing these electromagnetic fields and their changes through time can be used to investigate some behavior of the cortical neurons.

Electroencephalography (EEG) is a (non)invasive method for recording electrical activity generated by pyramidal cells on top of the human (or animal) cortex. For recording the EEG, some electrodes are placed above (noninvasive approach) or beneath (invasive approach) the scalp in order to measure the changes of the electric fields. Since the size of electrodes are much bigger than a single neuron, the averaged electrical activity generated by thousands of cells located beneath or close to the electrode are acquired by each electrode. When these cells activate synchronously, the electric fields are summed up and therefore the amplitude of the averaged acquired signal increases, and vice versa. Monitoring these fluctuations over time can reveal some special patterns corresponding to normal or abnormal brain activities [10–12].

Technically, the EEG is represented in a graph of voltage (vertical axis) versus time (horizontal axis). The voltage is obtained as the difference of the potential of two electrodes located on/beneath the scalp. The electrode positioning typically follows the 10-20 international system (figure 2.2). In EEG machines,

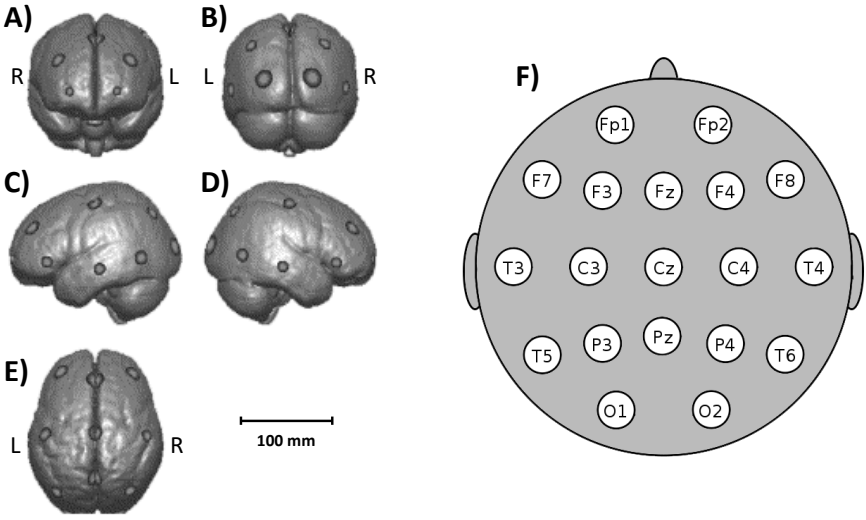


Figure 2.2: locations of electrodes in 10-20 system. A) frontal view, B) occipital view , C) left temporal view, D) right temporal view, E) top view, and F) names of electrodes at their approximate positions in the top view [2]

one electrode is usually defined as reference and the voltage of all electrodes are compared with it (uni-polar). However, in some applications, for processing or monitoring the EEG, the differences between pairs of specific electrodes are used as bi-polar channels. The names and orders of these electrodes are usually defined in a montage map which differs from one application to another [13]. For instance, in this thesis, in the sleep stage classification uni-polar EEG is used while in the seizure detection methods bi-polar EEG is used. Figure 2.3 shows a real example of a recorded EEG with 12 bi-polar channels. The rhythmic and high-amplitude pattern visible in some channels corresponds to an epileptic seizure which started from second 5 and lasted for about 70 seconds.

In order to characterize the EEG signals, analyzing the frequency distribution is important. Different oscillations and EEG phenomena, which are linked to underlying mechanisms, can be described by their frequency content. The EEG is commonly broken down into 4 sub-bands as follows [14–22]

- Delta activity (0.5-4 Hz): this pattern is a low-frequency and usually high-amplitude wave and is associated with slow-wave sleep when the declarative memory is probably being consolidated.

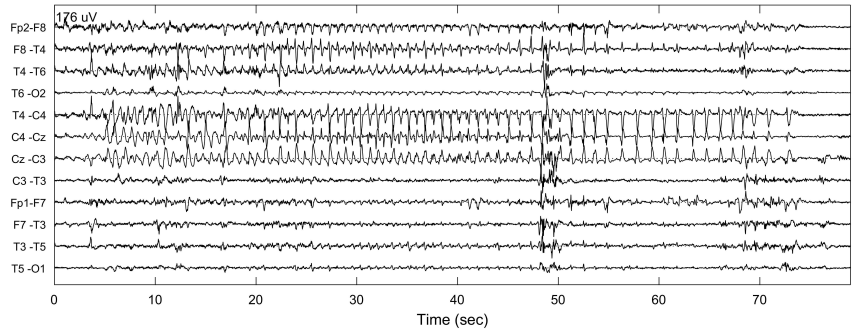


Figure 2.3: A real example of an EEG segment having a seizure. The horizontal axis shows the time in seconds. The names of bi-polar channels are listed on the vertical axis. The distance between the channel baselines equal $176\mu V$.

- Theta activity (4-8 Hz): this type of brain waves is generated by either the cortex or hippocampus with amplitudes of more than 20 μV. Several lines of evidence showed the roles of the Theta waves in working memory. Theta activity is mainly present during active sleep with Rapid Eye Movements (REM), meditative, or drowsy states. It has also been reported that it is observable during various types of voluntary, preparatory, orienting, or exploratory locomotor activities.
- Alpha activity (8-13 Hz): this wave has relatively a high amplitude of 20-100 μV and is usually the most dominant wave in EEG. Alpha activity typically appears with eyes closed, and lessens with eyes opening. Several studies have postulated that the higher cognitive performance is associated with the higher power of Alpha and lower power of Theta frequency bands.
- Beta activity (>13 Hz): Beta is the common waking rhythm of the human brain with amplitude mainly below $30\mu V$. It is usually associated with active attention, active thinking, solving complex problems, or concentration on the outside world. In some studies, the waves with frequency content above 30 Hz are called fast Beta waves, or Gamma waves. These infrequent waves have usually very low amplitude and are associated with event-related synchronization (ERS) of the brain.

2.2 Hypoxic-ischemic encephalopathy

Different etiologies may affect the EEG patterns and their characteristics (e.g. changing the frequency and morphology of seizures or background abnormalities).

Hypoxic-ischemic encephalopathy (HIE), or asphyxia, is one of the most important ones that can potentially cause morbidity and mortality in the perinatal periods and usually associated with severe neurological problems in survivors including cerebral palsy, epilepsy, sensory and cognitive problems. Hypoxia is a condition in which a cell is deprived of sufficient supply of oxygen, ischemia denotes an insufficient supply of blood to the cell, and encephalopathy refers to brain disorders. The prevalence of HIE is 1 - 6 infants in 1000 live term neonates. The problems during pregnancy, labor, delivery, and even post-delivery can cause HIE [23–27].

In detail, there are three major types of cell death: necrosis, apoptosis, and autophagy. Necrosis occurs when the plasma membrane of the cell is damaged, and therefore, the intracellular fluid is rapidly released into the extracellular space. It is usually caused by factors from outside of the cell, such as infection and toxins [28]. Apoptosis is a programmed cell death which means the cell intentionally suicides for a reason [28, 29]. Autophagy is also a natural and regulated process by which a cell can degrade and recycle surplus proteins and damaged organelles [29, 30]. Unlike necrosis, which results from cellular injury and usually happens suddenly, apoptosis and autophagy are normally regulated, controlled, and gradual mechanisms [28, 30–32]. The interval between the start of degeneration and the complete cell death is called *window of intervention*.

In cases with very severe HIE, it has been shown that the brain cell degeneration is dominantly necrosis. Nevertheless, in cases with milder encephalopathy, it is mainly apoptosis (and autophagy or hybrid apoptosis-necrosis). Although apoptosis is required for normal brain development to refine the pathways and connections between cells, HIE accelerates the cell death and consequently causes brain injury [33]. It is believed that during the window of intervention, interrupting the degeneration by protective strategies, like hypothermia, can help save the brain cells from damaging degeneration. [23, 31, 33–35].

2.3 Neonatal seizures

The incidence of neonates with seizure has been reported 1.5 to 5.5 per 1000 living births [36] and is even higher (9 - 11%) in very low-birthweight, premature neonates [37]. The actual prevalence is likely to be underestimated as neonatal seizures are often subtle or subclinical, and clinical observation alone is known to be unreliable in their diagnosis [37]. Since neonatal seizures are one of the most important signs of acute brain dysfunction [38], they need immediate medical attention [39]. It has been reported that the main etiology of symptomatic neonatal seizures is HIE (40 - 45 %) with a very early onset and various seizure

types [36, 40]. There are also other etiologies that may cause neonatal seizure, such as metabolic disorder, genetic malformation, and brain infection, which are less frequently linked to epileptic seizures. Since the main studies in this thesis are focused on neonatal seizures, it is important to define some terminologies in this context:

- *Epilepsy* refers to a group of neurological syndromes in which their common major feature is suffering from recurrent and unprovoked seizures [41].
- *Seizures* refer to abnormal, paroxysmal, excessive, synchronous electrical discharges from a cluster of neurons in the brain [27]. In adults or children, seizures are classified as epileptic or non-epileptic respectively caused by neurological and physiological/psychological conditions [42]. However, most neonatal seizures are acute, provoked and show very different manifestations compared to the usually recurrent, unprovoked epileptic or non-epileptic seizures in adults or older children [27].
- *Convulsions* refer to the abnormal movements of body and can be epileptic or non-epileptic [36].
- *Ictal*, in this context, means the interval between the onset and offset of a seizure. Subsequently, *pre-ictal*, a short period before the start of a seizure, *post-ictal*, a short period after the end of a seizure, and *inter-ictal* or *non-ictal*, the interval between two seizures, can be defined.
- *Seizure burden* equals the cumulative duration of seizures in a time-frame. For instance, a seizure burden can be 15 minutes in a frame of 1 hour (25%). The total seizure burden usually means the total seizure duration in the whole recording.
- *Status Epilepticus (ES)* usually expresses the 'maximum expression of epilepsy' [43] and refers to an epileptic seizure prolonged at least for 5 minutes or to several recurrent seizures occurred in 5 minutes (high seizure burden) so that the patient does not return to the normal neurological baseline [44]. However, this interval has been denoted differently in the literature ranging from 5 minutes to 1 hour depending on the age of the patient and type of seizures. [43–45]. There are also some conceptual definitions in which the specific time-frame is not explicitly determined [43, 46].

More details about the neonatal seizures are given in the following sentences based on [27, 36, 38]. Typically, a seizure can be detected by either observing clinical symptoms (such as lip smacking, sucking, chewing, and blinking), called *clinical seizure*, or detecting specific patterns presenting in EEG, called

electroencephalographic/electrographic seizure. A seizure without (or with subtle) clinical symptoms is called *subclinical* seizure. Furthermore, a seizure with both clinical and electroencephalographic manifestations is referred to as *electro-clinical* seizure. Despite being clinical or electroencephalographic, neonatal seizures are divided into four major types as follows:

- *Clonic*, which refers to the seizures with rhythmic jerking.
- *Tonic*, in which the tone or stiffness of the muscles increases gradually or suddenly. The speed of movement depends on the quickness and duration of the seizure.
- *Myoclonic*, which is a seizure with a brief and sudden muscle contraction with no rhythmic behavior.
- *Subtle*, which refers to the seizures that are not clearly clonic, tonic, or myoclonic, without (or with subtle) external motor manifestation. The main approach to diagnosing these seizures is visual analysis of EEG. Since some of these seizures might not be electroencephalographic, the EEG is usually inspected along with video. In neonates, most seizures are subtle and, therefore, clinical observation alone is an unreliable diagnosis. This is the reason why EEG monitoring is highly needed in the NICUs.

In the NICUs, the current gold standard for detecting seizures is visual analysis of multi-channel continuously recorded EEG (cEEG) along with video by an expert clinical neurophysiologist [23, 47–54].

2.4 Neonatal Sleep-wake cycles

Term and preterm (premature) infants sleep for long periods of time in a day. Generally, an infant is in one of the following three states at each moment:

- awake,
- active sleep (AS), also called rapid eye movement sleep (REM), see figure 2.6.a,
- quiet sleep (QS), see figure 2.6.b.

The duration and proportion of these states depend on the age of the infant. Changes between these states are not detectable when the post-menstrual age is lower than 27 weeks, where post-menstrual age (PMA) means the time elapsed

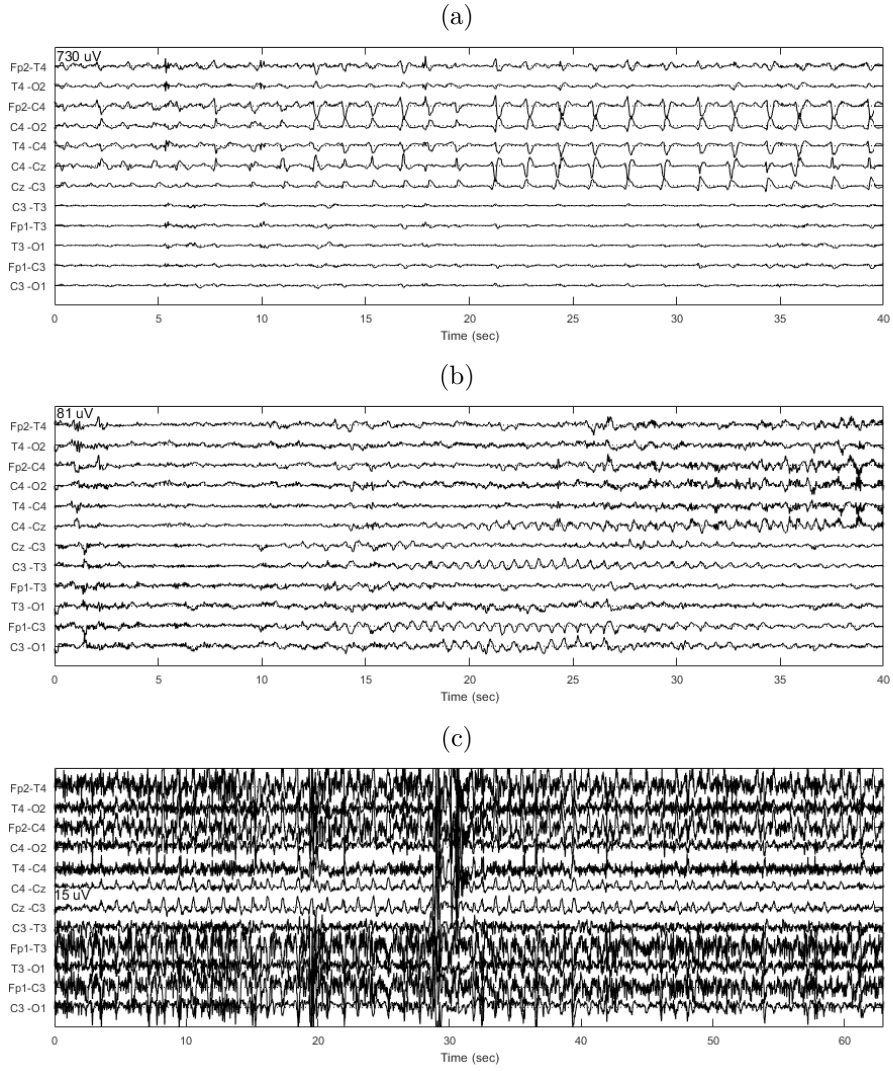


Figure 2.4: Three real examples of neonatal seizure. a) a spike-train type seizure with high peak-to-peak amplitude mainly observed at C4. b) an oscillatory type seizure with peak-to-peak amplitude of $50\mu V$ observed at C3. c) a low amplitude seizure ($\leqslant 15\mu V$) mainly at Cz.

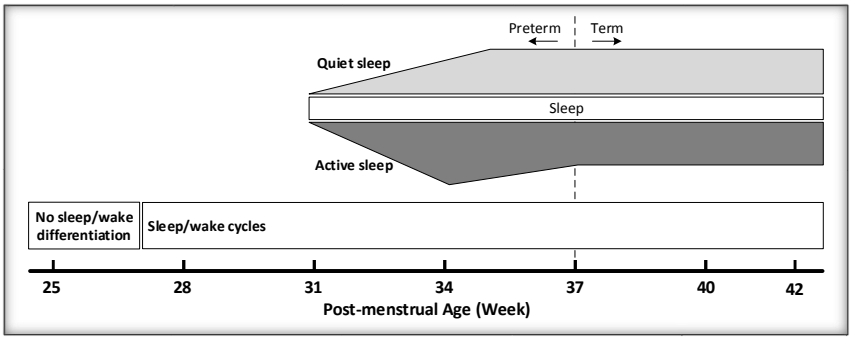


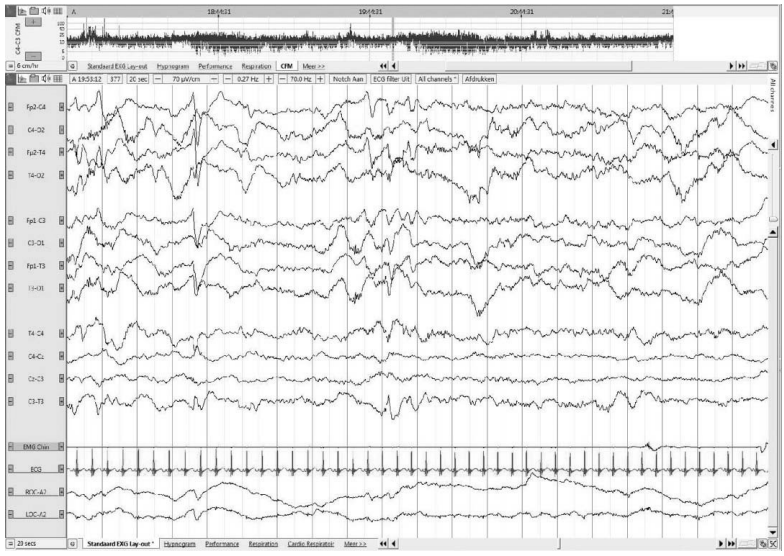
Figure 2.5: Changes of quiet sleep and active sleep with post-menstrual age (in weeks).

between the first day of the last menstrual period of the mother and the time of EEG recording. However, only after 31 weeks, the cycles between states are well-established [55]. Figure 2.5 illustrates the changes of QS and AS with post-menstrual age. Usually preterm sleep stage classification via EEG is limited to separating QS from non-QS (AS + awake) [56–59]. It is evident that the morphology of the EEG patterns for AS and awake are indistinguishable, except for the movement artifacts, which are usually more frequent during the awake period. However, polygraphic signals, like the electromyogram (EMG), may reveal the differences [55]. Figure 2.6 illustrates two segments of AS and QS recorder at PMA 32 and 38 respectively.

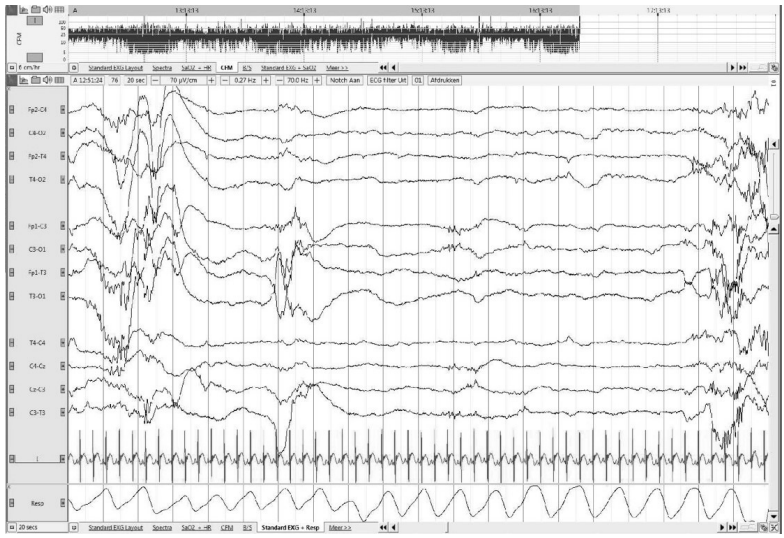
2.5 Artifacts in neonatal EEG

One of the main challenges of EEG analysis is the presence of artifacts in the EEG signal. The source of these artifacts can be biological or non-biological. Some of the artifacts contaminating the EEG are listed below [7, 10, 60–63].

- Cardiac artifact: two main types of artifacts are generated by the heart: I) ECG artifact (electrical) and II) blood vessel pulsation (mechanical) [10]. In the former one, the electric field generated by the heart, which is also the origin of the ECG signal, can be seen in the EEG [10], usually more clearly in the left side of the brain [10, 64] (see figure 2.7). The latter one, on the other hand, is the mechanical movement of one/some electrode cap(s) caused by pulsations of the vessels beneath the electrode(s) and, therefore,



(a)



(b)

Figure 2.6: Example of an active sleep and a quiet sleep EEG segment at 32 and 38 PMA respectively. a) Active sleep with continuous activity and mixed frequencies associated with REM. b) Quiet sleep with discontinuous EEG tracing and temporal Theta activity.

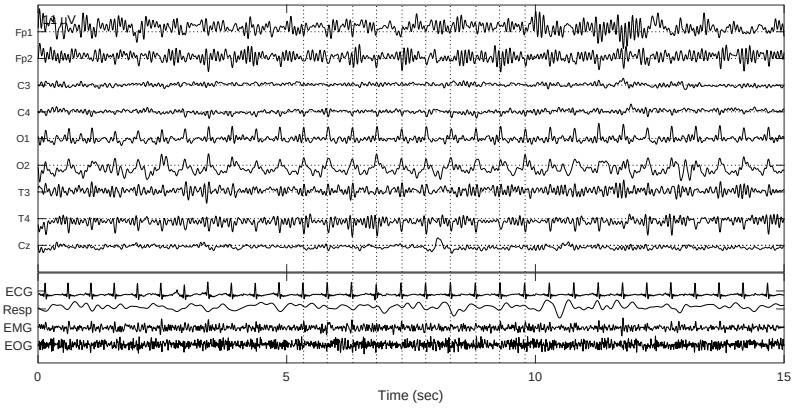


Figure 2.7: ECG artifact mainly visible at O1

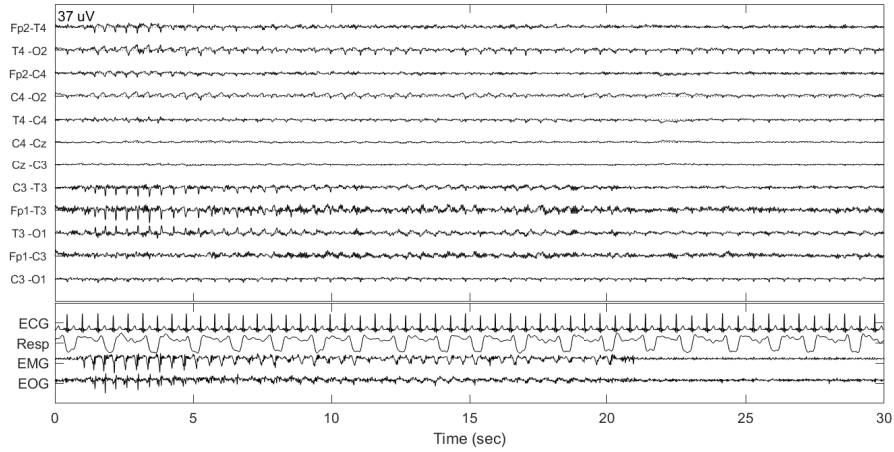


Figure 2.8: EMG artifact due to chin movement

it is usually classified as an electrode artifact. It usually manifests itself in the EEG signal as periodic slow waves with regular intervals and has about 200 ms delay with respect to the peaks of the ECG artifact, and is mainly measured over the frontal and temporal regions [10].

- EMG artifact: electromyographic (EMG) artifact is a kind of muscle activity-related artifact, in which the EEG signal is contaminated with EMG patterns. Although a wide range of frequency components have been reported for the EMG artifacts, in most cases, those of the artifact are higher (> 40 Hz) than those of the common interesting patterns in the EEG (< 40 Hz) [61] (see figure 2.8).

- Glossokinetic artifact: it is another kind of muscle artifact which refers to change of electric fields in the brain when the tongue is moving (like chewing or swallowing). It is mostly observed in the frontal area and resembles vertical eye movement or frontal delta activity [65].

- Ocular artifact: this type of artifact is generated by movement of the eyeballs or blinking. Each eyeball acts as a dipole pointing from retina (negative) to cornea (positive). A fixed eyeball does not yield an artifact. However, when the eyeballs move, depending on the direction and speed of the movement, an artifact may appear. In general, if these artifacts exist, those related to the vertical and horizontal movements might be maximally seen on $Fp_{1,2}$ and $F_{7,8}$ respectively [65] (see figure 2.9).

- Respiratory artifact: this mechanical, low-frequency artifact results from the movement of the bed, pillow, electrodes, electrode wires, or an unstable recording board, due to the breathing of the neonate. In some cases, repositioning the neonate can solve it. Nevertheless, if it is observed in some channels after recording, the slow-wave patterns in the EEG synchronized with polygraphic signals, particularly the abdomen respiratory signal, can be detected or eliminated via signal processing algorithms [7, 27, 66] (see figure 2.10).

- Electrode artifact: it can be defined as all kinds of artifacts related to the recording electrodes, e.g. respiration, cardiac vessel pulsation, sweating, movement. However, it mainly refers to the non-biological discharges, also known as electrode pop, and is completely different from cerebral potentials. It manifests and may repeat itself in a very restricted field (usually limited to a single electrode). This artifact can appear due to poorly applied electrodes, broken electrode wires, drying of electrode gels, or changes in the scalp leads [63] (see figure 2.11).

- Environmental artifact: it refers to all kinds of artifacts generated by sources and equipment in the recording room. This is mainly non-

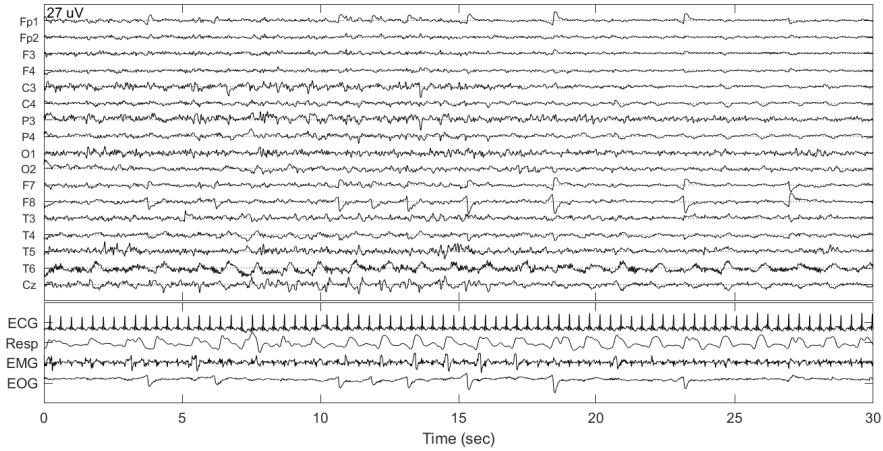


Figure 2.9: EOG artifact mainly visible at F8

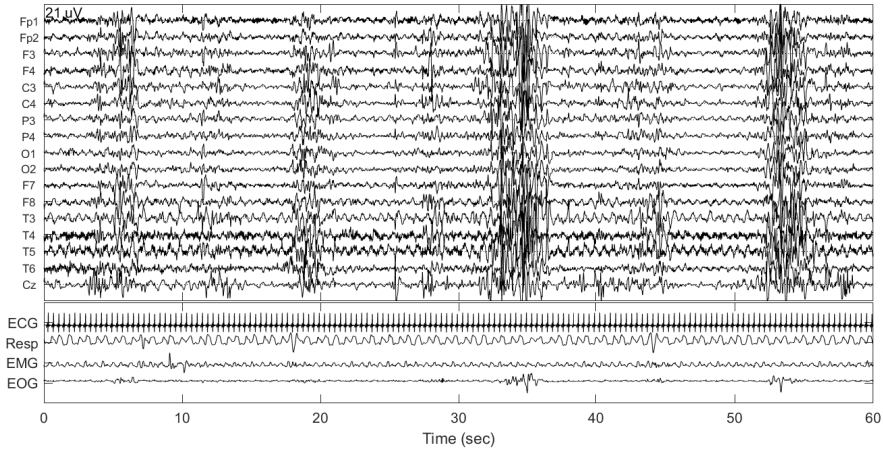


Figure 2.10: Respiratory artifact mainly visible at T3 and T5

biological and results from interfering of electric fields from the power-line, mechanical ventilators, incubator, cell phones, etc. [63].

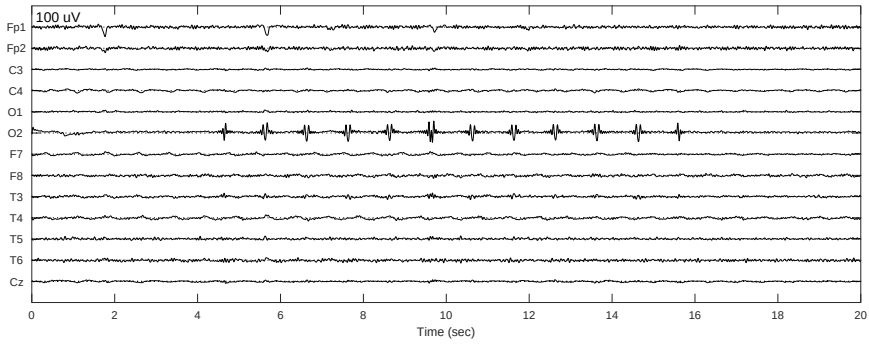


Figure 2.11: Electrode artifact visible at O2

Part II

Database and Scoring

Chapter 3

Database

In this chapter, all data used in this thesis to develop, train, and validate the algorithms are described. The data are partitioned into two databases. The first one is seizure database which was recorded from full-term neonates in two different medical centers and annotated by two expert clinical neurophysiologists. Part of this database was additionally re-scored by 3 independent expert EEG readers which is studied in more detail in chapter 4. The second database was recorded from premature neonates in one medical center and the sleep stages were annotated by two expert EEG interpreters.

3.1 Introduction

One of the most important materials for developing methods and applications is a proper database including a large number of cases with accurate labels defining the ground truth. In this study, two databases were used:

- NeoGuard database for term neonates
- NeoGuard database for preterm neonates

The first database includes the EEG-Polygraphic recordings from term neonates recorded in the NICU of the Erasmus University Medical Center Rotterdam (EMCR) and the NICU of the University Hospitals Leuven (UZL). Part/all of these recordings are used in Chapters 4, 6, 7, 8, 10, 12, and 14. Additional more specific information of the used data is provided in each chapter as 'Data description'.

The second database includes the recordings of only preterm neonates. These recordings were acquired only in the NICU of UZL. This database is used to develop a sleep stage classifier in chapter 9 and an inter-burst interval (IBI) estimator implemented in chapter 14, which was previously developed by Koolen et al. [8].

3.2 NeoGuard database for term neonates

The EEG recordings in this database were obtained from 81 neonates recorded in EMCR and UZL. This database is partitioned into 4 datasets according to the centres and durations of the used EEG recordings, called DB 1-4. The 48 neonates in DB1-DB3 had seizures and were enrolled from the neonatal intensive care unit (NICU) of the Erasmus University Medical Center Rotterdam (EMCR). The 23 neonates in DB4 were enrolled from the NICU of University Hospitals Leuven. They presented with seizures in the neonatal period and had continuous EEG monitoring during the standard clinical care. DB1 and DB2 include the full recordings (27 h per neonate on average) while DB3 and DB4 includes parts of the recordings selected by the expert EEG annotators when at least one seizure was observed. In some of the chapters, instead of using the whole recordings of DB1 and DB2, some selected windows of the DB1 and DB2 are selected by the expert clinical neurophysiologists in which at least one seizure was observed. The window length varied from 4 h to 8 h for different neonates. These so-called DB1-Small and DB2-Small datasets include also 10 additional recordings from 10 neonates (with severe

Table 3.1: NeoGuard seizure database from term neonates

Nama	Center	Number of neonates	Duration (h)	Number of seizures	Total seizure burden(h)	Average seizure duration
DB1	EMCR	17	461	1398	16.6	43 s
DB2	EMCR	18	489	1758	29.5	60 s
DB3	EMCR	13	27	217	4.0	66 s
DB4	UZL	23	46	120	5.6	168 s
- DB1-Small	EMCR	17 + 5	136 + 52	481	8.7	96 s
- DB2-Small	EMCR	18 + 5	72 + 20	532	7.0	69 s

to moderately severe HIE) in which no seizure was observed during the EEG recording. Table 3.1 lists some of the general characteristics of these datasets. In the last two rows, the plus sign, '+', separates the values corresponding to the neonates with seizure from those without seizure. Note that the 17 and 18 notates used in DB1-Small and DB2-Small are the same as DB1 and DB2 respectively. The polygraphic signals were also recorded along with the EEG during the recordings and include electrocardiogram (ECG), electro-oculogram (EOG), chin or limb surface electromyogram (EMG), and abdominal respiratory movement signal (Resp). All recordings used in this thesis were filtered in the software between 1-20 Hz using a finite impulse response (FIR) filter of length 1 second. The recordings were anonymized in the corresponding hospitals. The studies performed in the NeoGuard project were conducted with the approval from the Ethical Review Boards in both hospitals.

In all studies explained in this thesis, there is no pre-selection of data and no EEG recordings were excluded due to low-quality of EEG recordings, artifact contaminations, or expression of dubious seizures. Definite seizures were defined as paroxysmal EEG patterns with a change from ongoing background activity with repetitive spike-trains, oscillations or a mixture there of, with clear-cut onset and offset, lasting for at least 10 seconds. Figure 3.1 illustrates a definite mixed-type seizure lasted for about 1 minute. In this figure, the polygraphic signals are also displayed. The dubious seizures, on the other hand, are paroxysmal EEG events lasting for at least 10 seconds, composed of arrhythmic mixed oscillations or sharp waves of low amplitude ($<30\,\mu V$) with irregular variation in amplitude, frequency and morphology (without well-defined evolution) [67]. Figure 3.2 illustrates an arrhythmic dubious seizure with a low frequency and amplitude, with ill-defined onset and offset. In practice, the clinicians in the NICUs do not initiate treatment with anti-epileptic drugs (AED) when a dubious seizure is detected, unless this pattern frequently repeats itself or is accompanied by definite seizure patterns.

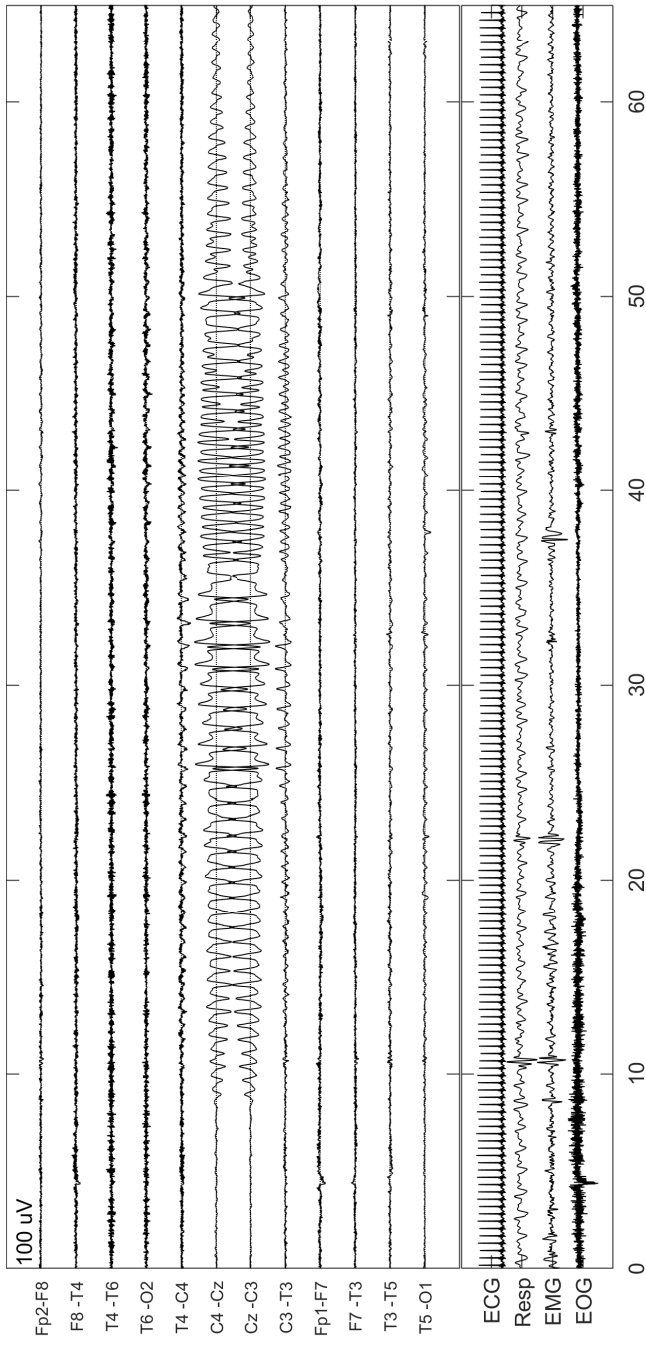


Figure 3.1: A definite mix-type seizure observed mainly at Cz.

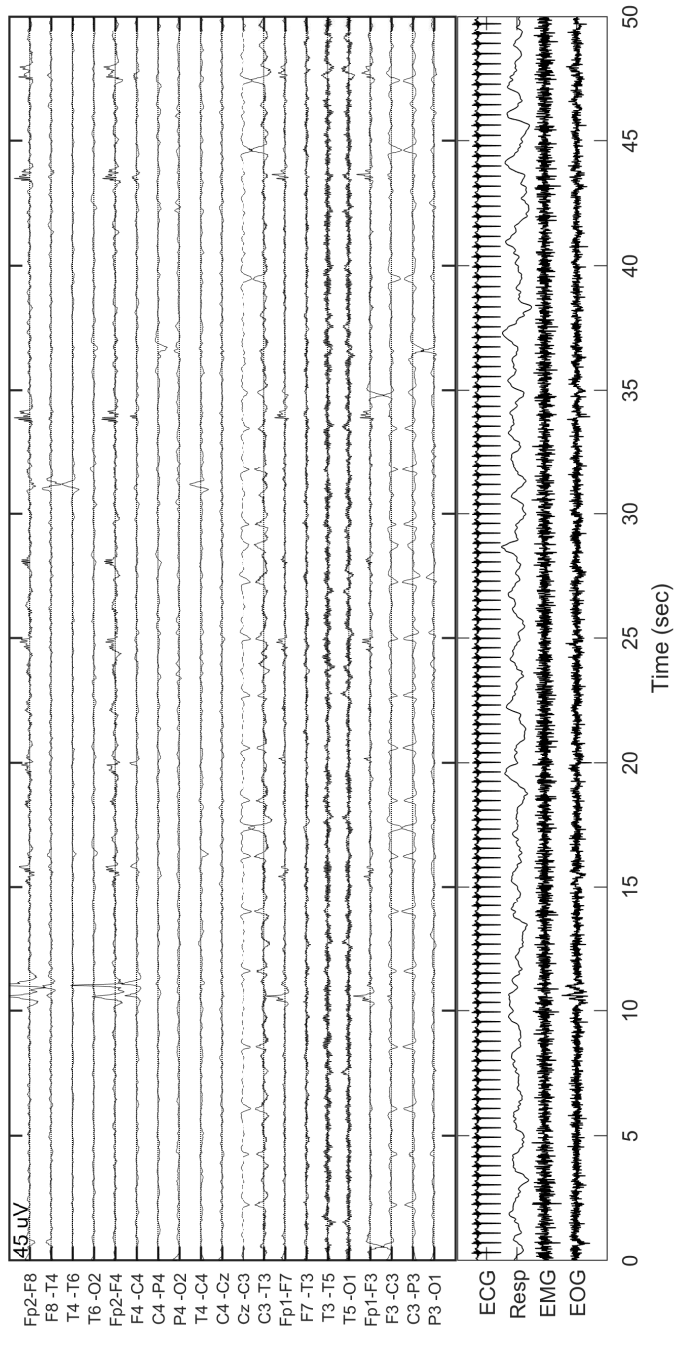


Figure 3.2: An arrhythmic dubious seizure observed at C_3. The frequency of spikes varies from 0.25 to 0.5 Hz and the peak-peak amplitude is lower than 20 µV. This dubious seizure is associated with very severe abnormality of EEG background (grade 8).

In the following subsections, more specific details concerning the recordings of each medical center are given.

3.2.1 EMCR

EEG-polygraphic data of 58 term neonates (DB1-3 + 10 neonates with no seizure in DB1,2-Small) were recorded in this center between 2003 and 2012. All neonates were term (with gestational age ≥ 36), admitted to the NICU with presumed postasphyxial HIE, and underwent cEEG and MRI. The inclusion criteria for asphyxia were either a five minute Apgar score below 6 or arterial pH of the umbilical cord blood less than 7.1 and clinical encephalopathy according to Sarnat score. Among these, 48 neonates had seizures. Treatment with anti-epileptic drugs was initiated by a protocol as explained in [55] whereas either electrographic or electroclinical seizures were detected. The neonates with heart or brain malformation were excluded and no other preselection was made. The other 10 neonates without seizures had also HIE with moderate to severe encephalopathy.

For 45 neonates, a full 10-20 International System using 17 electrodes, figure 3.3 (A), was used. In the recordings from 10 neonates, the electrodes F_{3-4} and P_{3-4} were not available and a restricted system using 13 electrodes, figure 3.3 (B), was used. For the remaining 3 neonates, electrodes F_{7-8} and T_{5-6}, in addition to F_{3-4} and P_{3-4}, were not recorded, hence a restricted system using 9 electrodes, figure 3.3 (C), was used [55]. All data were recorded by a NervusTM monitor (Taugagreining hf, Reykjavik, Iceland) or a Brain RT EEG system (OSG BVBA Rumst, Belgium). The sampling frequency was 256 Hz and the initial filtering was between 0.3 and 70 Hz.

3.2.2 UZL

From this center, EEG data from 23 term neonates (DB4), who were admitted to the NICU and had seizures in the neonatal period, recorded between 2013 and 2015 were selected. Among these, 6 neonates had HIE, 5 had metabolic problems, 5 had stroke, 2 had genetic malformations, and the remaining 5 had other etiologies. The neonates with heart malformations were excluded and no further preselection was performed. The EEG-polygraphic signals were recorded by a Brain RT EEG system (OSG BVBA Rumst, Belgium), using 9 EEG electrodes, Fp_{1-2}, T_{4-5}, C_{4-5}, Cz, O_{1-2} placed according to the restricted 10-20 International System. Twelve bipolar channels were then made from these available electrodes [55], figure 3.3 (C). The polygraphic signals were the same as for the EMCR data, mentioned above. The data were recorded at 250 Hz

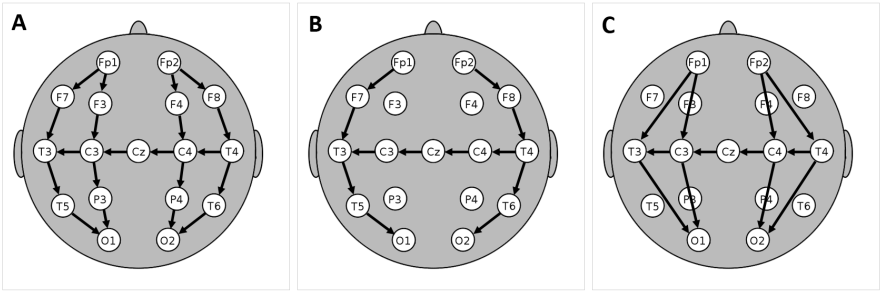


Figure 3.3: Neonatal EEG montages. A: full 10-20 system of electrode placement using 17 electrodes, B) a restricted 10-20 system using 13 electrodes, C) a restricted 10-20 system using 9 electrodes. The arrows in these montage maps define the bipolar channels (e.g. $C_Z - C_3$) and their lengths do not represent any meaningful value.

initially and then resampled to 256 Hz. Two-hour EEG recordings from each neonate in which at least one seizure has been observed were selected and scored by an experienced clinical neurophysiologist.

3.3 NeoGuard database for preterm neonates

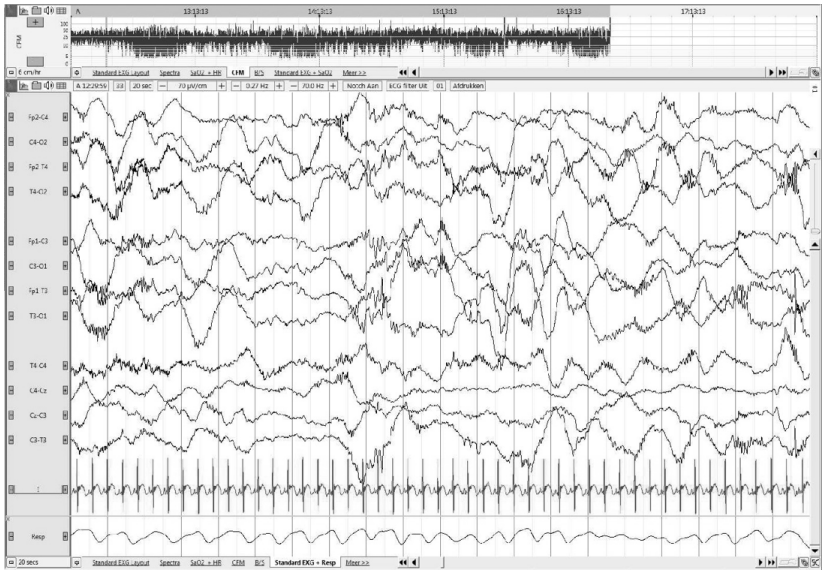
The data used in this database were collected between 2012 and 2014 at UZL. The EEG signals were recorded after approval by the Ethics Committee of University Hospitals Leuven and informed parental consent were obtained. Serial EEG recordings collected from 26 preterm infants born before 32 weeks of gestation were used in this database. The EEG of each baby is recorded at least twice during their stay in the NICU, resulting in 97 multi-channel EEG recordings between 27 and 42 weeks post-menstrual age (PMA). All neonates enrolled in the study had a normal neurodevelopmental outcome score at 9 and 24 months corrected age (Bayley Scales of Infant Development-II, mental and motor score > 85). Moreover, none of the subjects were under sedative or anti-epileptic medication during the EEG registration or had severe cerebral lesions [1].

The EEG was recorded using 9 electrodes: Fp_{1-2}, C_{3-4}, T_{3-4}, O_{1-2} and reference electrode Cz, according to the modified international 10-20 system [55].

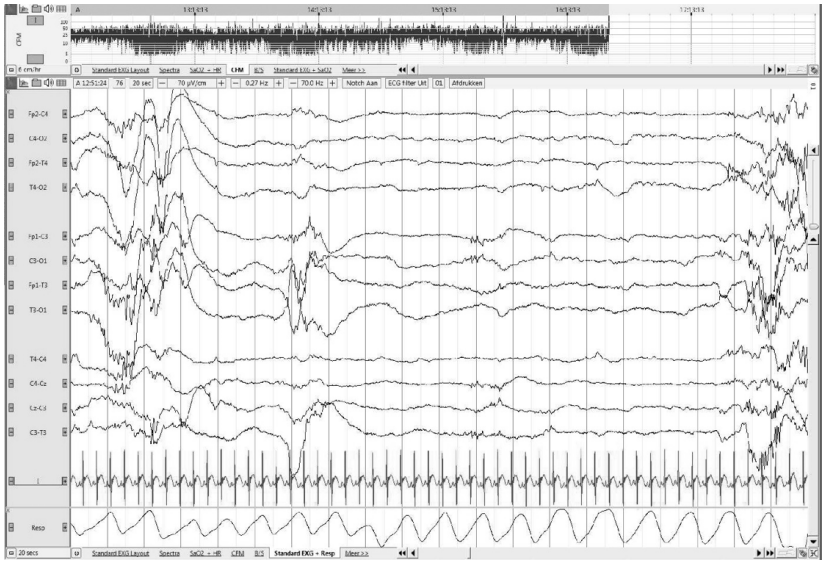
[1] normal cerebral ultrasonography or intraventricular haemorrhage grade $\leqslant$ II, no periventricular leukomalacia or ventricular dilatation exceeded the 97th percentile (p97) according to Levene [68]

The monopolar EEG set up was used and the reference electrode Cz was discarded during the analysis phase. A Brain RT EEG system (OSG bvba, Rumst, Belgium) was employed to acquire the data. The data were initially filtered between 0.3 and 70 Hz and sampled at 250 Hz.

The quiet sleep segments were annotated by 2 independent expert clinicians upon consensus. Wakefulness and other sleep stages (active sleep and indeterminate sleep) are merged and labelled as non-quiet sleep. An example of a quiet sleep and non-quiet sleep EEG segment is shown in figure 3.4.



(a)



(b)

Figure 3.4: Example of a quiet sleep and non-quiet sleep EEG segment at 32 weeks and 2 days PMA. (a) Continuous tracing during active sleep. Delta brushes in temporal and occipital regions, irregular breathing pattern. (b) Discontinuous tracing during quiet sleep. IBI shorter than 15 sec, temporal theta activity and occipital delta brushes, more regular breathing pattern.

Chapter 4

Inter-rater agreement of neonatal seizures

In this chapter, the seizure database explained in chapter 3 and the NeoGuard Information System explained in chapter 11 are used to provide and analyze a multi-rated EEG database. To this end, a consensus decision is made based on 'majority voting' and inter-rater agreement is estimated using Fleiss' Kappa. Then, the influences of different factors on the agreement are determined. This study is published in modified form in [69]. It is a joint work with Dr. Anneleen Dereymaeker. The contribution of this PhD candidate relies on designing and developing the framework, as well as the technical and statistical analysis.

4.1 Background

In recent years, there is an increasing focus on neuromonitoring of neonates at high risk for cerebral dysfunction. Neonatal encephalopathy and seizures are the most common and distinctive signs of underlying brain injury and warrant neuromonitoring. There is emerging evidence from experimental research that prolonged or recurrent seizures add injury to the developing brain [70–73]. Clinical studies supporting these findings are still scarce. However, few studies have reported that seizures worsen brain damage in neonates with hypoxic-ischemic encephalopathy (HIE) [74, 75], a higher seizure burden is correlated with adverse outcome [73, 76–79] and that neonatal seizures are independently associated with long-term neurological impairment [79]. These observations suggest that early identification and treatment might diminish morbidity, however there is a pressing need for well-designed prospective studies to support these observations.

Diagnosing neonatal seizures remains a challenge due to their unique characteristics [40]. First, bedside recognition of seizures is not accurate as the clinical manifestations of neonatal seizures are often subtle and highly variable [80–82]. Second, critically ill neonates often manifest only electrographic seizures [82], and anti-epileptic drugs (AED) may suppress clinical seizures while ictal EEG discharges persist (electroclinical dissociation) [83]. Continuous video-electroencephalography (cEEG) is the gold standard for accurate diagnosis of both electroclinical and electrographic seizures [84, 85] and prolonged EEG monitoring is mandatory to evaluate the efficacy of their treatment with AED [86]. The implementation of cEEG monitoring is resource intensive and requires expertise not available in all neonatal units. Automated quantitative analysis of cEEG, including background activity and neonatal seizure detection algorithms (NSDs), will improve clinical applicability. In the last decade, NSDs have been developed by our group [4,6,7] and others [3,87,88] with the ultimate goal to aid the clinician in the field, however the results vary widely with different performance outputs reported. The challenge of reproducing published results might be due to the nature of visual scoring criteria used by the clinician or due to the nature of the data of a particular database [6, 67, 89–92]. First, visual detection of seizures in cEEG can only be accepted as gold standard when inter-rater variability is low. The NSDs are designed based on the current EEG definition of a neonatal seizure [93, 94]. The algorithms at their best will match the decision made by the clinician. If there is no good agreement on the label of a 'true' seizure, then the performance of the algorithm will be affected [90]. Inter-rater agreement for assessing electrographic seizures varies widely, with a Kappa value of 0.4 to 0.93 in different studies [67, 92, 95–99]. High agreement is achieved when scoring typical seizure patterns with well-defined morphology [83]

whereas the agreement may differ significantly in more dubious patterns [67]. Moreover, the inherent ambiguity of visual classification might be influenced by the composition of the databases used in the different studies, and cause differences in output performances. Both electrographic characteristics as well as technical factors can result in the construction of databases with highly different levels of arduousness. Features such as total seizure number, duration, burden, morphology and rhythmicity of the electrographic seizure pattern in relation to the overall background activity and the use of AED [96], might affect seizure pattern recognition, both visually and technically [67] [92]. Finally, assessment of inter-rater agreement is problematic in itself and there is a lack of comparable measures to assess inter-rater agreement. Different approaches presented in literature are Kappa coefficients or Intraclass Correlation Coefficients (ICC), which can be used on 'event-based'(comparisons of annotation of an event by event basis) [67, 96, 97], and 'time-based' or 'epoch-based' metrics (comparisons of annotation on a second by second basis) [98] or identification of 'any' seizure within a recording [99]. Time-based metrics will reflect the amount of seizures (seizure burden) but are labour intensive to assess on long duration EEG monitoring, whereas the event-based metrics provide the percentage of seizures that will be detected [90]. Both measures are significantly influenced by the exact research query (e.g. event-based, recording-based) and the prevalence of findings within a sample, which can cause methodological issues for its interpretation and comparison [100, 101]. This underscores the necessity to define more accurately agreement estimates based on well-classified datasets which contain clinically relevant events. Only this will allow assessment of reliability of visual EEG interpretation and comparison of NSDs with the same performance measure as used for visual annotation.

The aim of the present chapter was twofold: the first aim was to provide an objective and comprehensive assessment of inter-rater agreement of neonatal rhythmic patterns and define events as 'ictal' or seizures based on majority voting. For this, we designed an online platform based on a large seizure database of multichannel EEGs, recorded at two university hospitals (NeoGuard database). A previously published heuristic NSD [4], which is explained in section A.1.1, was applied to detect suspected seizure events and 4 raters reviewed the events. The second aim was to determine the influences of different factors on inter-rater agreement. In-depth analysis of relevant seizure characteristics and technical factors that affect the majority voting and therefore the recognition of seizures, was performed.

4.2 Data description

In this chapter, DB1-small, DB2-small, DB3, and DB4 are used. The 10 neonates without seizures are excluded from the first two datasets. The detailed descriptions about these datasets are given in section 3.2. Briefly, the EEG-polygraphic recordings from 71 neonates recorded in two medical centers, EMCR and ULZ, are used in this study. The EEG signals are filtered between 1-20 Hz and bipolar channels are made according to the montage maps explained in section 3.2.

4.3 Methods

4.3.1 Seizure annotation

The EEG recordings were first annotated by an independent, experienced clinician with extensive experience in neonatal cEEG, referred to as primary rater. An electrographic seizure was defined as ictal-appearing, electrographic discharges that showed a clear variation from background activity, displaying a repetitive, stereotyped EEG pattern of sinusoidal oscillations or sharp waves, or a mixture of both, for at least 10 s, with evolution in amplitude and frequency over time [93, 94]. Status epilepticus was defined as recurrent seizure activity lasting $\geq$50% of the time for a minimal EEG record duration of 1hour [102]. Seizures were annotated by the primary raters with the beginning and end of each seizure collectively over all channels. Discharges were classified as 'dubious seizures' if the rater was unsure about ictal appearing electrographic discharges, such as: (a) runs of $>$ 10 s sharp waves/oscillations or a mixture of both, occurring arrhythmically (with marked variability in the interval and morphology between individual complexes in a seizure discharge for the major part of its duration) or (b) rhythmic discharges of shorter duration ("brief rhythmic discharges"), or periodically occurring complexes consisting predominantly of sharp waves with no clear-cut evolution [67]. Non-seizure/artifacts included typical artifacts (such as pulsation, respiration, ECG, tremor) which mimicked the stereotyped rhythmic seizure patterns that are often seen in neonates. The raters had no access to video or clinical information.

4.3.2 Seizure detection algorithm

The previously developed heuristic NSD that is explained in Appendix A.1 (Version 1) was applied on the EEG recordings. Briefly, this algorithm consists

of two detectors running in parallel: (a) a spike-train detector that detects high-energy segments of the EEG and analyses the correlation between them and (b) oscillatory seizure detector that detects increase in low-frequency activity (1–8 Hz) with high autocorrelation. The algorithm provides an output of event-based detection. Both events of the NSD detections and the primary raters (missed seizures = annotated by the primary raters but not-detected by the NSD and false detections = detections by the NSD but not labelled as seizures by the primary rater) formed the database used in this study.

4.3.3 NeoGuard Website

The implementation of continuous EEG-monitoring in neonates requires a 24h-surveillance and a certain expertise which is not yet available or applicable in every NICU. To solve this debacle, there is a growing interest to use automated detection algorithms to analyse the EEG background and detect ictal patterns. This study is a first approach to apply and evaluate an NSD in the same way as it will be used in the clinical setting. The algorithm will 'alert' the clinician to review a certain event as ictal or artifact. In this context, we choose not to review the full EEG epoch, but only the events as detected by the NSD or the primary rater. At present, we opted to use two types of events in the database. Our reason to do so is twofold: first, the so-called 'ground truth' is still ambiguous. Although there are clear, theoretical definitions and guidelines for seizures, in practice, EEG readers are faced with numerous events which are more dubious, whether or not, true or falsely detected by the NSD. Second, the NSD theoretically still can miss seizure events compared to the 'visual gold standard'.

The technical explanation of the website is provided in section 11.3 (chapter 11).

4.3.4 Scoring of the events

Three secondary raters (SR 1-3), blinded to the primary scores, reviewed independently all events (all NSD detections + missed seizures) as 1. Definite seizure, 2. Dubious seizure, 3. Non-seizure/artifact. EEG reviewing was restricted to the selected events. Examples of the 3 different event categories are presented in figure 4.1.

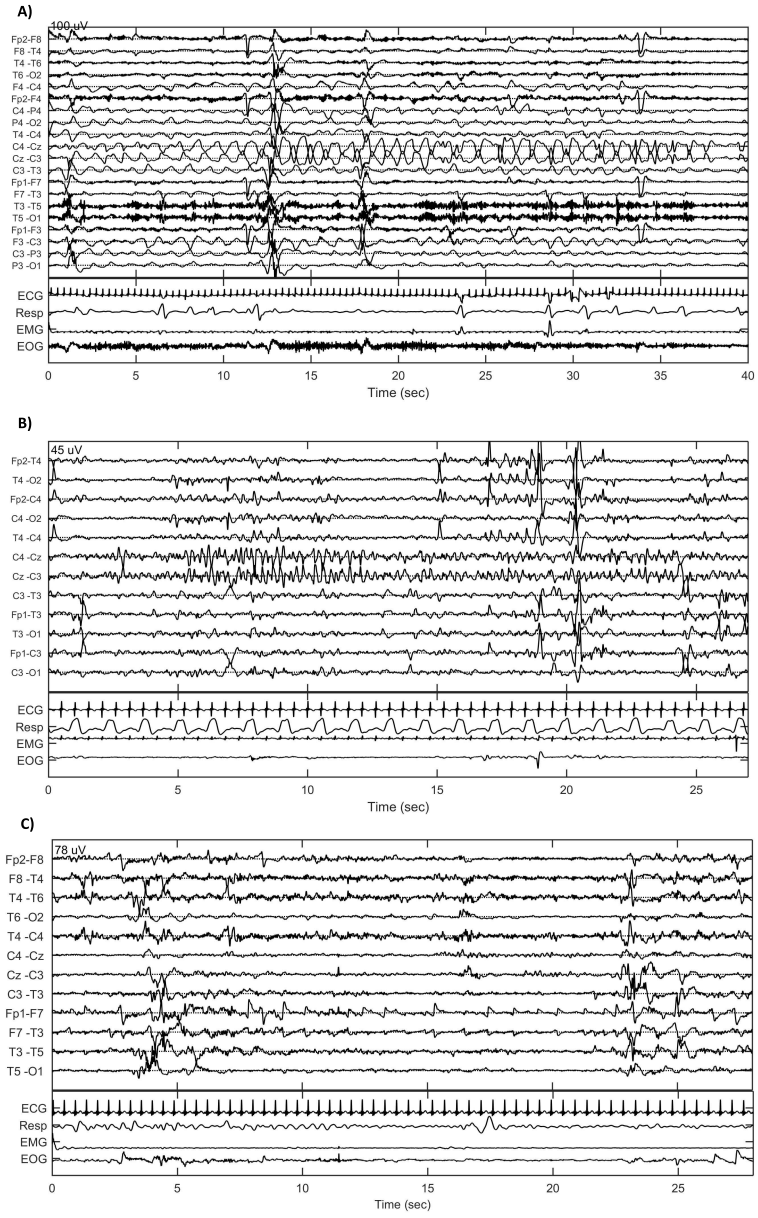


Figure 4.1: Examples of the 3 different event categories. A) This event was scored as 'definite seizure' by all raters. B) This event was scored as 'dubious' by 3 raters and as 'non-seizure' by 1 rater. C) This event was scored as 'non-seizure' by all raters.

4.3.5 Measures of agreement

Majority voting

For the purpose of this study, a "majority vote" was used to describe how well raters agreed on an event. A binary classification (seizure/non-seizure) was assigned to each event according to the label of the 4 raters (1 primary and 3 secondary raters) and a majority vote score derived. Dubious events were classified in the non-seizure group, since those events, although paroxysmal, do not fit in the operational definition of seizures and their clinical significance is controversial. All events were classified into 'poorly agreed' (2 raters agreed or 50%), 'moderately agreed' (3 raters agreed or 75%) and 'highly agreed' (all 4 raters agreed or 100%), and the percentage of the majority voting according to those three categories was calculated. In case of equal votes, the following tie-breakers are used (Seizure: S, Dubious: D, Non-seizure: N): 1) S, S, N, N $\Rightarrow$ S, 2) D, D, S, S $\Rightarrow$ S, 3) D, D, N, N $\Rightarrow$ N. Starting from the hypothesis that the accuracy of the "ground truth" is obscure, the votes of the primary and secondary raters were equally weighted.

Inter-rater agreement

Agreement was assessed using Fleiss' Kappa coefficients for categorical variables. The concept of our study only allowed assessment of event-based measures. Fleiss' Kappa is applicable for any number of raters [103]. Kappa coefficient is a measure of agreement for categorical data while controlling for agreement by chance. Kappa values often range from 0 (inter-rater agreement does not differ from chance) to +1 (total agreement). However, if the observed agreement is less than chance agreement (disagreement), Kappa could be negative. The level of agreement measured by Kappa was classified as follows: 0-0.20 slight agreement; 0.21-0.40 fair agreement; 0.41-0.60 moderate agreement: 0.61-0.80 substantial agreement; 0.81-1.00 almost perfect agreement. Here, Kappa was measured for a three class problem (seizure, dubious, non-seizure) and additionally for a two class problem (seizure/non-seizure), like it was done for the majority voting since Kappa tends to decrease with the number of categories [104].

Electrographic characteristics

Factors that may affect seizure recognition and inter-rater agreement were examined. In-depth analysis of database composition and electrographic characteristics such as duration, peak to peak amplitude and power ($\mu V^2/s$)

were assessed to evaluate the effect of these characteristics on ictal pattern recognition, automatically calculated based on the event-based categorization by all raters, after final labelling. Electrographic characteristics were compared between seizures and dubious classified events. Additionally, the inter-rater agreement for the classification of seizure intensity was also assessed using Fleiss' Kappa. Single seizure events were accumulated and quantified within any hour of the clipped recording. According to this, Kappa values for 1-hour recording with 10 min versus 10-20 min and >20 min seizure activity were compared.

Agreement was also assessed in the context of the presumed seizure etiology in DB4. Data of seizure characteristics were log-transformed (due to the very large range of values noted) to guarantee the symmetry of the data (more uniform data spread) and allow further statistical testing [59]. Wilcoxon rank-sum test was used to investigate the effect of seizure features on seizure recognition by comparing the differences in electrographic characteristics between seizures and dubious events. Data were analyzed using Matlab^{TM} (The Mathworks, Natick, MA, USA). All tests were 2-sided and a p-value <0.05 was considered significant.

4.4 Results

4.4.1 Database

1919 Events in 280 h of multi-channel EEG from 71 neonates with seizures, were rated by 4 independent observers. 66% of events in this database were less than 60 s, and 45% less than 30 s (figure 4.2). The most common diagnosis was HIE (DB1-3: 48 patients with HIE, 3/48 treated with therapeutic hypothermia). The presumed etiology of seizures for the 23 neonates in DB4 was: HIE (n=6), stroke (n=5), intracranial haemorrhage (n=2), metabolic (n=5), genetic (n=2), meningitis (n=1), viral encephalitis (n=1) and malformation of cortical development (n=1). All neonates received treatment with at least one AED.

4.4.2 Majority vote and inter-rater agreement

A binary classification (seizure/non-seizure) was assigned to the events according to the annotations of the 4 raters and a majority vote score derived. From a total of 1919 events, 44% events were classified with high, 36% with moderate, and 20% with poor agreement (figure 4.3).

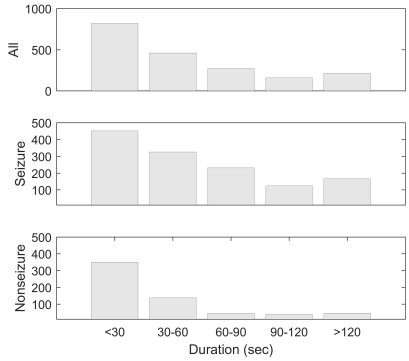


Figure 4.2: Database overview: the number of all the events, classified as seizures and non-seizures with a given median duration based on majority voting.

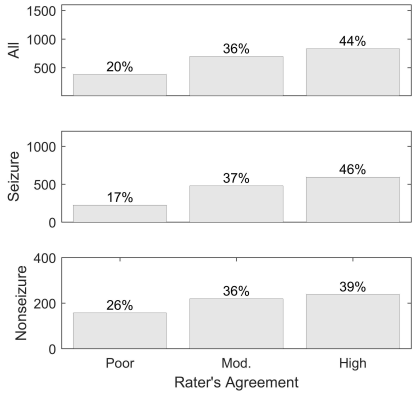


Figure 4.3: Majority voting for all the events, seizures and non-seizures. Poor agreement (Poor): 2 raters agreed, moderate agreement (Mod.): 3 raters agreed, high agreement (High): 4 raters agreed.

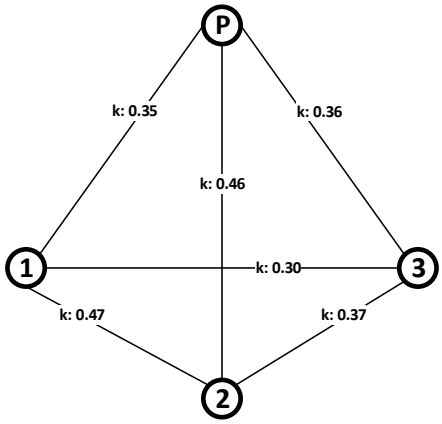


Figure 4.4: Inter-rater agreement among all raters and for all the events, presented with Fleiss' Kappa (k). P: primary rater. SR 1-3: secondary raters.

Inter-rater agreement among all raters and for all the events (for a 3 class problem: seizures/dubious/non-seizures), resulted in a total Fleiss' Kappa value of 0.39 (±0.007). Kappa values were comparable for DB1-Small, DB3 and DB4 whereas Kappa was slightly lower for DB2-Small: DB1-Small: 0.39 (±0.012), DB2-Small 0.29 (±0.013), DB3 0.42 (±0.018), DB4 0.42 (±0.020). Average Kappa values for each rater varied between 0.34 (±0.031)-0.43 (±0.045) (figure 4.4).

Fleiss' Kappa was also measured for a binary outcome (seizure/non-seizure) (n=1231 events). The Fleiss' Kappa improved to 0.62 (±0.012), post hoc, when all segments which were scored 'dubious' by at least one rater (=36% of all events) were removed, which improved substantially the agreement.

4.4.3 Electrographic characteristics

Based on majority voting, 68% of the events were labelled as seizures. In 83% of them, at least 3 raters were convinced about electrographic seizures and in 46% (n=600), all of them agreed (figure 4.3). The most common seizure duration in this dataset was below 30 s (43% of all seizure events). Raters agreed the best for events of 60-120 s. The seizure duration with the highest level of agreement was 90-120 s (63%) (figure 4.5). Higher levels of disagreement were noted in brief seizures (<30 s) and long-duration seizures (>120 s) (poor agreement: respectively 21% and 23% versus 7% in seizures with duration of 60-90 s). In

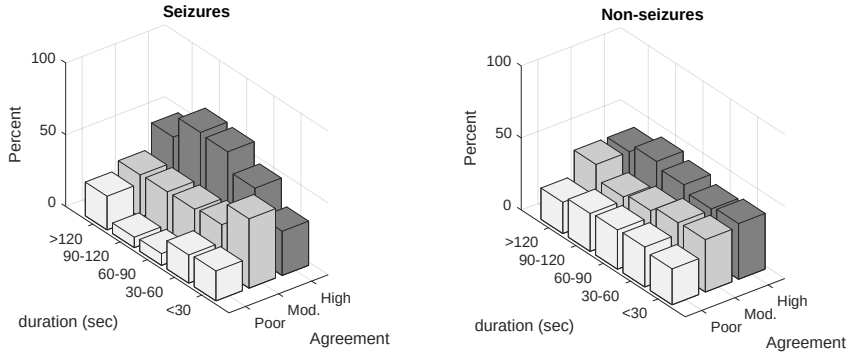


Figure 4.5: Effect of seizure/non-seizure duration on majority voting. Poor agreement (Poor): 2 raters agreed, moderate agreement (Mod.): 3 raters agreed, high agreement (High): 4 raters agreed.

non-seizures, events of 60-90 s and 90-120 s were recognised with the highest certainty (figure 4.5).

There was a significant difference in the electrographic characteristics of the seizures with moderate and high agreement versus dubious events, with seizures having higher median duration ($p<0.001$), higher median power ($p<0.001$) and higher median peak to peak amplitudes ($p< 0.01$).

Assessment of seizure burden (quantity) resulted in Kappa values of 0.42 ($\pm$0.013) for <10 min, 0.39 ($\pm$0.012) for 10-20 min and dropped to 0.26 ($\pm$0.013) for >20 min maximum seizure burden per hour.

Inter-rater agreement was assessed according to the etiology of seizures in DB4. Only a limited number of patients (n=23) were available, however a total of 251 events were included. Three groups were compared: group 1 (HIE: 99 events), group 2 (stroke + intracranial haemorrhage: 66 events), group 3 (metabolic, genetic, meningitis, encephalitis and malformations of cortical development: 89 events). Fleiss' Kappa for group 1 was 0.42 ($\pm$0.03), group 2 0.48 ($\pm$0.04), but Kappa decreased notably in group 3 to 0.34 ($\pm$0.03). Subgroup analyses revealed that events in group 3 were of shorter duration (median 24 s, IQR: 17-49) whereas the median duration of all events was 35 s ($\pm$0.017). More events were classified as dubious in group 3 compared to group 1 and 2. Fleiss' Kappa for group 1 with HIE (0.42 $\pm$ 0.020), was comparable to Kappa of DB1-Small, DB2-Small, and DB3, respectively 0.39 (($\pm$0.012) and 0.42 ($\pm$0.018), with a median duration of events of 45-47.5 s. Total Kappa of DB2-Small with HIE was noticeably lower: 0.29 ($\pm$0.013). The duration of events was also briefer in DB2-Small (median 30 s, IQR: 18-60).

4.5 Discussion

This is the first study using majority voting to define neonatal electrographic seizures based on a comprehensive assessment of a large online database with rhythmic EEG patterns. Without any assumption of the 'ground truth' of a true 'electrographic seizure' pattern, we opted to use a hybrid collection of rhythmic events, as marked by a human rater and a seizure detection algorithm, and equally weighted the votes of primary and secondary raters. With this approach, we aimed to tease out the different factors that have influenced inter-rater agreement and ictal pattern recognition.

The design of our study focused on an in-depth analysis on event-by-event basis, which is how a clinician would score and treat seizures detected on cEEG. Our current results suggest that there is a wide variability in identifying some rhythmic ictal and non-ictal EEG features, and that only the most robust ictal patterns are consistently agreed upon since only 600 of 1919 events were labelled as 'definite' seizures by all 4 raters. As a consequence, inter-rater agreement is fair at best, with Fleiss' Kappa values around 0.4 using 3 categories, increasing to 0.6 when applying Kappa for a binary outcome. This allows a more direct comparison with other study results, however different metrics and outputs used for the assessment of inter-rater agreement still complicate exact evaluation.

Time-based measures for seizure occurrence may yield different, possibly higher estimates since it takes into account also the longer epochs of negative agreement [96]. In particular, when using long duration monitoring, these epochs could be more straightforward to interpret. Moreover, Kappa will also be influenced by the prevalence of seizures. Few studies that have compared Fleiss' Kappa for seizure occurrence in neonates, based on event-based measures in reasonably sized patient cohorts, have reported Kappa values in the same range for two raters [67]. Abend et al. [97] found a Kappa of 0.4 in older infants, assessed by three experienced raters. In contrast, more recently, Wusthoff et al. [99] reported a much higher agreement (Fleiss' Kappa of 0.93) for 3 experienced raters, however the binary outcome used was the presence or absence of 'any' seizure within a 3h-recording. When focusing more on seizure quantity categories (based on the number of events within a recording), which more accurately reflect an event-by-event evaluation, the Kappa dropped remarkably with increasing number of seizures (respectively k=0.72 and k=0.35). Furthermore, 'brief rhythmic discharges' were scored separately and induced much higher disagreement (k=0.19), whereas these more 'dubious' events were simultaneously evaluated in our analysis. Shah et al. [96] reported a much higher agreement (Cohens) Kappa of 0.84 for two raters, however only 7 neonates with seizures were included. Stevenson et al. [98] reported high levels of agreement (k=0.83) for the identification of 'any' seizures based on a more conventional review of

long duration multi-channel EEG of 70 neonates (35 neonates with seizures, 35 neonates without seizures). Their results are certainly encouraging and impressive. However, besides the use of different metrics, their database has around 55% of seizures between 30-120 s and only 16% of seizures <30 s, whereas our database contains 45% events of shorter duration <30 s. They have shown that the reliability of seizure recognition was low for seizures with a duration less than 30 s. The results of our majority voting support these findings and confirm that greater certainty could be achieved when the ictal pattern was long enough and displayed a clear-cut evolution. Moreover, we and others have shown that automated detection of brief seizures is also challenging [6, 88, 92]. Correlation of human raters for seizure recognition has also identified that longer duration seizures with ambiguous offset and associated with high seizure burden, mixed up seizure annotation and resulted in lower agreement [67, 105], which is further depicted in our results of inter-rater agreement and seizure quantity. Quantitative analysis of electrographic features of the events in our database indicates that dubious events can be discerned from definite seizures, based on their shorter duration, lower amplitude, and power, which confirms that these events are genuinely different.

Due to the retrospective character of our study, we have a bias towards birth asphyxia as main etiology, however this is still the most common etiology of neonatal seizures and thus a reflection of clinical practice [106]. Interestingly, subgroup analysis based on etiologies, showed a notable decrease in Kappa in the group with metabolic and genetic causes compared to HIE and seizures due to a focal lesion. Those seizures were of shorter duration, and expressed more variation in seizure morphology and spatial distribution in individual patients, which might have complicated seizure annotation. The drop in Kappa value for DB2-Small (HIE) can be explained by the briefer duration of seizures (<30 s) and very suppressed, low amplitude EEG, as described by Cherian et al. [67].

Some limitations of this study should be recognised. We are well aware that this study may have underestimated the agreement. First, the more 'artificial' way of reviewing the EEG may have complicated seizure annotation, since it does not take into account the accumulation of seizure evidence within a recording. This may also have introduced bias, being unclear to the rater whether brief rhythmic discharges represented an inter-ictal pattern or the start or end of a longer ictal pattern, or accumulated over time. However, since events of 60-120 s are classified with the highest agreement, we assume that this effect is probably small, since this only affects events with longer duration (>120 s). Improvements to the NeoGuard website can be easily made by displaying the whole event in detail, with successive 20 s epochs, to better discriminate the evolution in rhythmicity and amplitude at the beginning and end, to permit reviewers to gain better insight in the background activity.

Second, teasing out the effect of "experience" is possibly the most subjective and complex task of this and any similar study. The variation in Kappa values among the different raters is only small, suggesting no striking outlier. While efforts have been made by experts in the field in the last few years to provide clearly worded theoretical guidelines to improve neonatal EEG interpretation [94] [85], studies to bridge the gap between 'theoretical knowledge' and applying those definitions in clinical practice, and helping to generalize the results of individual research, have been rather limited. By combining different raters from 3 different centres and the application of a multi-centre, hybrid database, we wanted to create a replicable scenario. Since 'level of experience' or the 'ground truth' (of visually scored seizure) will always be 'subjective', it can only be compared when all participants are examined with exactly the same questions. This will eventually lead to the identification of reproducible EEG features (both ictal/inter-ictal) that are relevant both for clinical practice as well as for research. We are willing to share this web-based system with other researchers as well as with interested clinicians to demystify this black box. In future, we will extend this interesting query to premature EEG recordings, where various rhythmic patterns are even more common.

Third, this database includes critically ill neonates who have been treated with AED and sedative medications at the time of recording. These drugs may have suppressed EEG background (amplitude and continuity) [107], influenced seizure morphology and rhythmicity and affected seizure recognition [67], which was not taken into consideration in the visual assessment, nor was this accounted for in the quantitative analyses. Extra clinical information as well access to video-EEG segments may improve characterisation of these events in future applications of the website. This information can assist EEG readers to make a seizure diagnosis, interpret dubious patterns or recognise the cause of seizure-like artifacts. However, only this kind of scoring methods mimics the actual working mechanism of an NSD and allows direct comparison with NSD performance. Based on the objective assessment of the 'core characteristics' of visual patterns in the neonatal EEG, research into the pathophysiological basis at cellular, synaptic and network levels, can be conducted. This will help to elucidate the effect of modulators such as severity and nature of underlying brain injury and effect of treatment on neurodevelopmental outcome.

4.6 Conclusion

Interpretation of neonatal seizures remains challenging, however it is imperative to know the accuracy of visual classification to further improve our knowledge. In-depth analysis of database composition and electrographic characteristics,

revealed important factors that influenced seizure recognition. The results suggest that there is a wide variability in identifying rhythmic ictal and non-ictal EEG features, and that only the most robust ictal patterns are consistently agreed upon. The use of well-described databases and input of different EEG experts will help to improve neonatal EEG interpretation and define uniform, general accepted seizure definitions which can be reproduced by EEG readers as well as used in automated algorithms.

Part III

Algorithms

Chapter 5

Mathematical background

This chapter introduces the required mathematical background and machine learning techniques that are used in this thesis to develop new algorithms. First, three types of supervised pattern classifiers including support vector machines, deep convolutional neural networks, and random forests are introduced. Then, two regression methods including ridge regression and 'least absolute shrinkage and selection operator', which are used in this thesis mainly for feature selection and feature ranking, are explained. Finally, the bootstrap statistical test, which is used in chapter 10 for analyzing the proposed performance metrics, is described.

5.1 Support vector machines

Support Vector Machine (SVM) is a (non)linear supervised learning method using kernel tricks made for 2-class classification problems originally proposed in [108] and later completed in [109] and [110]. The main idea of this technique is maximizing the margin between border hyper-planes of the two classes.

5.1.1 Principle of SVM using hard margin

Firstly, assume C_{-1} and C_{+1} are two linearly separable classes. For each class, we define

$$\begin{cases} w^T x_i + b \geq +1 & if \quad x_i \in C_{+1} \\ w^T x_i + b \leq -1 & if \quad x_i \in C_{-1} \end{cases} \tag{5.1}$$

where w and b are the parameters of the border planes and i ($i = 1, \ldots, Q$) is the data point index. In this formula, the equalities express the border plane of the classes where all data points of the corresponding class are on one side of (or on) the plane. If the class label (y_i) is defined as -1 and +1 for the data belonging to C_{-1} and C_{+1} respectively, (5.1) can be rewritten as

$$(w^T x_i + b) y_i \geq +1 \tag{5.2}$$

As mentioned above, the idea of SVM is maximizing the margin which is the distance between these two border planes. Therefore, the goal is

$$\begin{aligned} &\arg \max_w \left\{ \tfrac{|(b-1)-(b+1)|}{\|w\|_2} \right\} \\ &subject\ to \quad (w^T x_i + b) y_i \geq +1 \end{aligned} \tag{5.3}$$

where $\|.\|_2$ denotes L2 norm (Euclidean norm). It is equal to

$$\begin{aligned} &\arg \min_w \left\{ \tfrac{1}{2} w^T w \right\} \\ &subject\ to \quad (w^T x_i + b) y_i \geq +1 \end{aligned} \tag{5.4}$$

One can use the Lagrange multiplier technique to find the dual problem as

$$\begin{aligned} &L(w, b, \alpha) = \tfrac{1}{2} w^T w - \sum_{i=1}^{Q} \alpha_i [y_i (w^T x_i + b) - 1] \\ &subject\ to \quad \alpha_i \geq 0 \end{aligned} \tag{5.5}$$

where Q is the number of data points and α_i are the Lagrange multipliers. This is a quadratic optimization problem subject to linear constraints which has a unique minimum. In the dual problem, as an important consequence of the

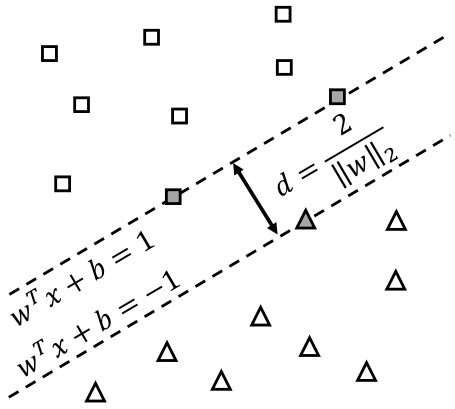


Figure 5.1: Histogram of the means of the bootstrapped samples

Kuhn-Tucker necessary optimality conditions [111], α_i are only nonzero for the data points locating on the class borders. In other words, only the data points closest to the other class play a role in forming the decision surface. These data points are called support vectors (SV).

5.1.2 Soft margin

Now, instead of using a hard margin, the formula is modified so that small errors are tolerated in order to increase the margin. This so-called soft margin technique improves the generalization power of the SVM. To this end, (5.2) is changed to

$$(w^T x_i + b) y_i \geq 1 - \xi_i \tag{5.6}$$

where ξ_i is a slack variable to relax the constraint. This modification leads to a bigger margin at the expense of some miss-classification. Clearly, the stack variables should also be entered to the minimization process in order to control the miss-classification error. Thus, (5.4) will be changed to

$$\begin{aligned} &\arg \min_{w} \left\{ \tfrac{1}{2} w^T w + cF(\textstyle\sum {\xi_i}^{\sigma}) \right\} \\ &\textit{subject to} \quad (w^T x_i + b) y_i \geq 1 - \xi_i, \quad \xi_i \geq 0 \end{aligned} \tag{5.7}$$

where σ is a positive variable, $F(u)$ is an arbitrary, convex, monotonic function with $F(0) = 0$, and c is the regularization parameter. The larger the c, the narrower (harder) the margin. For simplification, usually $\sigma = 1$ and $F(u)$ equals

an identity function, as we used in this thesis. In the primal problem, it results in

$$\begin{aligned} &\arg \min_{w,b,\xi} \left\{ \tfrac{1}{2} w^T w + c \sum \xi_i \right\} \\ &subject\ to \quad (w^T x_i + b) y_i \geq 1 - \xi_i, \quad \xi_i \geq 0 \end{aligned} \tag{5.8}$$

And the corresponding dual problem is as follows:

$$\begin{aligned} &arg \min_{\alpha} \left\{ \sum_{i=1}^{Q} \alpha_i - \tfrac{1}{2} \sum_{i,j=1}^{Q} y_i y_j x_i^T x_j \alpha_i \alpha_j \right\} \\ &subject\ to \quad \sum_{i=1}^{Q} \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq c \end{aligned} \tag{5.9}$$

This is still a quadratic optimization problem with a unique minimum.

5.1.3 Kernel trick

A simple method to use the aforementioned linear SVM in a nonlinear classification problem is firstly use a nonlinear function, Φ, to map the primary space, x_i, to the secondary space, $\Phi(x_i)$ and then find a linear classifier in the second space by the method described above. The problem arises in computing the $\Phi(x_i)^T \Phi(x_j)$ presented in the dual problem (5.9) (x_i is replaced with $\Phi(x_i)$) because the dimension of $\Phi(x_i)$ can be much higher than the primary space. Nonetheless, $\Phi(x_i)^T \Phi(x_i)$ is an inner product which is a kind of similarity metric. With the aid of Hilbert-Schmidt's and Mercer's theories [112, 113], this similarity can be approximated by different kernels using different nonlinear mappings so that

$$K(U, V) = \Phi(U).\Phi(V) \tag{5.10}$$

where K is the kernel function and ' . ' denotes the inner product. Now, in the optimization process, the inner product is replaced by this kernel function. As a result, (5.9) is changed to

$$\begin{aligned} &arg \min_{\alpha} \left\{ \sum_{i=1}^{Q} \alpha_i - \tfrac{1}{2} \sum_{i,j=1}^{Q} y_i y_j K(x_i, x_j) \alpha_i \alpha_j \right\} \\ &subject\ to \quad \sum_{i=1}^{Q} \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq c \end{aligned} \tag{5.11}$$

Thus, instead of calculating the inner product in the secondary space, the kernel is applied in the primary space. Note that in this case, the nonlinear functions, Φ, and subsequently the secondary space are not even needed to be

computed. This is one of the key characteristics of the SVM method. Below, some commonly used kernels are listed.

linear:	$K(U,V) = U^T V$
polynomial:	$K(U,V) = (U^T V + \tau)^P$
radial basis function (RBF):	$K(U,V) = \exp\left(\frac{-\|U-V\|_2^2}{2\sigma^2}\right)$
hyperbolic tangent (MLP):	$K(U,V) = \tanh(aU^T V - b)$

Note that the parameters (e.g. τ, P, σ, a, b) of the kernel function as well as the regularization parameter, c, are the hyper-parameters and should be defined before solving (5.11). In order to optimize/tune the hyper-parameters, trial and error, grid search, or cross validation can be applied [114–117].

5.2 Convolutional neural networks

Convolutional neural networks (CNNs), also known as ConvNet, are classified as a special type of feed-forward artificial neural networks (ANNs). These networks are biologically inspired by the processing of the mammalian visual cortex in which each neuron is stimulated by only a restricted region of the visual field, known as the receptive field [118–120]. CNNs usually have a deeper structure compared to ANNs (e.g. multilayer perceptrons, radial basis function networks), because they need fewer parameters for the same number of hidden layers. This parameter reduction, which is achieved by 3 key features of CNN: local receptive fields, parameter sharing and pooling, makes the computation more efficient and helps preventing overfitting. Using local receptive fields means that each neuron is only connected to a limited region of the input volume, in contrast with fully connected neurons. As a result, the network is able to learn local spatial/temporal correlations from the input. Parameter sharing means that all neurons located in one layer have the same parameters. This allows the network to detect a shifted target pattern, which is known as the shift invariant property. The last characteristic that makes the CNN different from other ANNs is pooling, which is a nonlinear down-sampler. Pooling progressively reduces the size of the data representation and improves translational invariance.

In general, a CNN consists of multiple stacked layers from 3 different types: convolutional layer (Conv), nonlinear layer, and pooling layer (see figure 5.2). Note that the input of each layer can be a 1, 2, or 3 dimensional volume (Tensor).

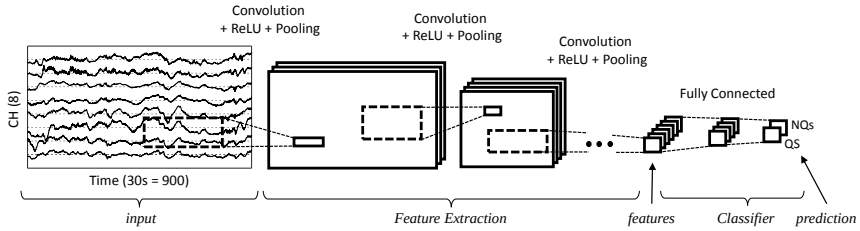


Figure 5.2: Architecture of a convolutional neural network. This network is used in chapter 9 for automatic neonatal sleep stage classification.

Conv layer: This layer, which is the main block of CNN, is composed of a bank of finite impulse response filters (FIR) (also called kernels) that operate on the input as follows:

$$O(i,j,k) = \sum_{p=1}^{P} \sum_{n=1}^{N} \sum_{m=1}^{M} f_k(m,n,p) I(i-m, j-n, p) \tag{5.12}$$

where $I(i,j,p)$ is the input of the Conv layer, where (i,j,p) represent the dimensionality of the input data, and $f_k(m,n,p)$ are the coefficients of the kth filter which consists of $M \times N \times P$ coefficients, where M and N represent the size of the filters, and P represents the number of filters in the previous layer. $O(i,j,k)$ is the output of Conv layer, resulted from the convolution operator of the filter f_k and the input I through the 1st and 2nd modes. The filter coefficients, f_k, are the only unknown parameters of the CNN which should be found in the training process by a backpropagation method. However, the size of filters (M,N), known as receptive field, as well as the number of filters of each Conv layer should be predefined in the design process (hyper-parameters). The output size of the Conv layers in the first and second modes resulting from the convolution operator equals the size of the input subtracted by the length of the filter plus one. The output size of the 3rd mode is equal to the number of filters in that layer. For instance, in the proposed CNN method for the neonatal seizure detection problem in chapter 8, see table 8.2, the first Conv layer is composed of 5 filters each one of them of size 1×5. The size of the output of this layer in the 2nd mode is $2696 (= 2700 - 5 + 1)$.

Comparing the FIR filters of the Conv layer with common neurons in ANN shows that each filter is like a layer of simple linear neurons with two important characteristics: 1) The weights of all neurons located in the layer are shared between the neurons. 2) Neurons only connect to a limited number of inputs

with overlap. Applying these two characteristics on a layer of simple neurons converts the layer to the mentioned convolutional FIR filter.

Pooling layer: The main aim of pooling layers is reducing the number of outputs of the Conv layers by a nonlinear sub-sampling function in local regions. In practice, taking the maximum and averaging are the two most common operations being used in pooling layers. The stride of pooling should be predefined as a hyper-parameter. The output volume size of this layer in each mode equals

$$\lfloor \frac{S_{input} - S_{filter}}{stride} \rfloor + 1 \tag{5.13}$$

where S_x represents the size of x and $\lfloor\ \rfloor$ is the floor function. For instance, in such example, see table 8.2, the pooling in the 8th layer is a 1×3 max pooling with stride 2 which decreases the size of the 2nd mode from 667 to 333. Note that pooling layers have no trainable parameters.

Nonlinear layer: This is a nonlinear unit which increases the nonlinearity and power of the network. The most commonly used function in CNNs is rectified linear unit (ReLU) which is defined as follows:

$$O(x) = \max(0, x) \tag{5.14}$$

where x is the input value and $O(x)$ is the output. In other words, this function is a half-wave rectifier which replaces the negative values of the Conv layer output with zero. It has no effect on the size of data. In addition, in the very last layers of CNN, which are performing the classification task, where the 1st and 2nd modes are completely aggregated by pooling, the Sigmoid unit is also suitable which is computed by

$$O(x) = \frac{1}{1 + e^{-x}}. \tag{5.15}$$

For instance, in the example, table 8.2, the ReLU and Sigmoid units were used in the 12th and 22nd layers respectively.

5.3 Random Forest

In general, there are different types of ensemble learning methods based on decision trees, such as random subspace method, ensemble bagged trees, and boosted trees. In this thesis, ensemble bagged trees are explained and addressed as RF. Generally, an RF is composed of tens or hundreds of bagged decision trees, also known as weak learners. In order to train an RF, first, $\sqrt{N}$ features,

where N is the total number of features, are randomly chosen. Second, a new set of Q data points is created from the Q available training segments using random selection with replacement, this procedure is normally called bagging or bootstrapping. Then, a decision tree was trained using the Q segments and $\sqrt{N}$ features based on "classification and regression tree" analysis, namely CART, proposed in [121]. Briefly, first, all possible binary splits of each feature are performed and the Gini's diversity index (GDI) of the tree after each split is calculated. Second, the split which has maximized the GDI is selected and the two consequence child leaves are formed. The procedure recursively repeats for each leaf until one of the following stopping conditions is reached: 1) when the tree depth equals a predefined maximum depth (Max_{Depth}), 2) when the number of segments in a leaf is smaller than a predefined threshold (Min_{Leaf}), or 3) when a node purely includes segments of one class. In recall (test) mode, the outputs of the RF are the probabilities corresponding to the classes averaged from all outputs of decision trees [121–123].

5.4 Ridge regression

In simple multiple linear regression models, the least mean squares algorithm (LMS) is often used to optimize the coefficients. In this algorithm, the mean of squared errors is minimized as follows:

$$\min_{\beta} \left\{ \frac{1}{N} \|y - X\beta\|_2^2 \right\} \tag{5.16}$$

where N is the number of observations, y is the vector of the observed outputs, X is the matrix of the observed inputs, called regressors, and β is the vector of the coefficients. This minimization results in

$$\widehat{\beta} = (X^T X)^{-1} X^T y \tag{5.17}$$

where $\widehat{\beta}$ is the vector of optimal coefficients. A problem arises when collinearity between the regressors (columns of X) leads $X^T X$ to be singular. In practice, even approximate collinearity will result in a small determinant of $X^T X$. Thus, its inverted matrix includes very big numbers and is very sensitive to the noise presented in the observed inputs, X.

Nevertheless, using Tikhonov regularization, in addition to the least squared error, the square of the coefficients are also added to the minimization problem in order to avoid having very big coefficients as follows:

$$\min_{\beta} \left\{ \frac{1}{N} \|y - X\beta\|_2^2 + \|\Gamma\beta\|_2^2 \right\} \tag{5.18}$$

where Γ (Tikhonov matrix) includes the weights of the coefficients which should be properly defined. In order to simplify it, in ridge regression, this matrix is chosen as λI where I is the identity matrix and λ is the scalar ridge factor. Therefore, (5.18) can be rewritten as follows:

$$\min_{\beta} \left\{ \frac{1}{N} \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2 \right\} \tag{5.19}$$

Adding this extra term to the regression formula will change the solution from (5.17) to what follows below.

$$\widehat{\beta} = (X^T X + \lambda I)^{-1} X^T y \tag{5.20}$$

Therefore, a small value of λ prevents $(X^T X + \lambda I)$ to be singular and consequently decreases the variance of estimation. However, this extra term increases the bias and therefore decreases the accuracy. Thus, a trade-off between decrease of variance and increase of bias should be found. It can be performed by trial and error in simple problems or cross validation in complex problems (explained in the next subsection) [124–126].

5.5 LASSO

Least absolute shrinkage and selection operator (LASSO) was proposed by Tibshirani [127] in order to reduce the number of regressors in a linear regression problem. In this method, the second squared L2 norm used in (5.19) is replaced by a L1 norm (absolute norm) as follows:

$$\min_{\beta} \left\{ \frac{1}{N} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\} \tag{5.21}$$

The key point of this replacement is the fact that using absolute norm can result in a sparse vector of coefficients. In other words, it can enforce some coefficients (β_j) to become exactly 0 in the optimization process. Therefore, it can be used as a feature selector so that the features having zero coefficent can be eliminated [117, 127, 128].

In order to select features, the λ in (5.21) is varying from a very small number (e.g. 10^{-3}) to a big number (e.g. 10^{1}) and subsequently different vectors of β has been calculated. The greater the λ, the sparser the β vector, the more features we eliminate. Furthermore, for each value of λ, and subsequently each set of selected features, the MSE of the regression can be calculated. Then, the

proper set of features can be chosen when a sufficient number of features are eliminated while the MSE is still satisfyingly low.

However, this kind of feature selection has a high dependency on the used data (X). In order to overcome it and increase the robustness of the feature selection process, LASSO is usually used along with cross-validation. In this case, the data is split to k folds and for each fold the LASSO is applied. Then, the largest λ with minimum overall MSE (or with minimum MSE plus 1 standard deviation), is chosen and the features with zero coefficients are removed [129, 130].

LASSO is usually being used as a feature selection technique in different applications [128, 131–134]. Note that if the feature vectors are orthonormal ($X^T X = I$), the features with zero weights can be interpreted as uninformative features having no linear correlation with the outputs. However, in general case, in the presence of inter-feature correlation, it is not correct such that a feature with a high correlation with the outputs can have zero weight. In addition, because of the sparse outputs of LASSO, many features may have zero weights and therefore LASSO is not a good method for feature ranking or interpreting.

5.6 Bootstrapping

Bootstrapping is a powerful analysis technique for measuring the characteristics of a variable from real observations. This method is useful in complicated situations when the characteristics of the real model generating the data are unknown. The main idea of bootstrap is mimicking the real generator behavior by repetitive selections of 'n-out of-m' samples from the real observations with replacement. In many problems, n (the number of selection) equals m (the number of observation). After bootstrapping, one vector with m observations will be converted to N vectors of n selected observations, where N is the number of repetitions. For each vector, the target variable (e.g. mean) is computed. It results in N computed variables in total from the N bootstrapped vectors. Then, the mean, median, variance, PDF, or other required statistical characteristics of the target variable are calculable [135–137].

To illustrate, assume a statistical sample is $\{1, 2, 3, 4, 5\}$ and the goal is estimation of the mean by a bootstrap test. In what follows below, 10 bootstrapped samples and their means are listed:

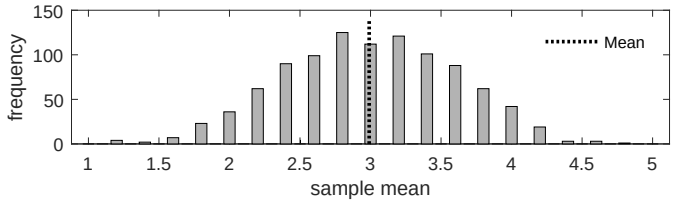


Figure 5.3: Histogram of the means of the bootstrapped samples

5, 1, 4, 4, 3 (3.4)	5, 5, 1, 1, 2 (2.8)
1, 5, 5, 2, 4 (3.4)	5, 3, 5, 1, 4 (3.6)
4, 5, 4, 1, 1 (3.0)	1, 1, 4, 5, 3 (2.8)
2, 3, 4, 4, 3 (3.2)	3, 5, 2, 2, 4 (3.2)
5, 4, 4, 5, 4 (4.4)	5, 5, 1, 1, 4 (3.2)

while the numbers in the parentheses represent the mean. In order to remove the effect of random selection, 1000 of these bootstrapped samples were selected and the mean of each was calculated. The final estimated mean and standard deviation of the sample means equal 3.0 ± 0.6. Figure 5.3 shows the histogram of the bootstrapped sample means. The mean of the statistical model generated the observed samples ($\{1, 2, 3, 4, 5\}$) has estimated to 3 by bootstrap method (equal to the sample mean $(1 + 2 + 3 + 4 + 5)/5$). However, because of the few number of observations (only 5), the bootstrap shows a high deviation for this mean (0.6) indicating this mean is not highly reliable. Note that the calculated mean by bootstrapping may be slightly different (has bias) from the sample mean which is not the case here.

Chapter 6

Multi-stage neonatal seizure detection

This chapter introduces the first algorithm developed in this thesis for neonatal seizure detection. This method is a multi-stage seizure detector consisting of a previously developed heuristic algorithm, which is mimicking a human EEG reader, and a machine learning algorithm. In the first stage, the heuristic algorithm detects some candidates as potential seizures. In the second stage, the main characteristics of the potential seizures are extracted by 5 proposed sets of features including synchronization, evolution, retention, segment, and signal features. Next, the features of those potential seizures are evaluated by a supervised classifier and then falsely detected candidates, which are mainly artifacts, are detected and removed. The advantages and disadvantages of this algorithm is discussed in this chapter. This study has been published in modified form in [6] and is extended in chapter 7.

6.1 Background

As mentioned in chapter 2, seizures are a common and distinctive sign of serious brain dysfunction in neonates [38]. The majority of neonatal seizures have an acute symptomatic basis and one of the most important causes is hypoxic ischemic encephalopathy (HIE) [67, 138]. Clinical presentation can be highly variable and manifestations of neonatal seizures can be subtle, absent or resemble normal behavior. It is known that after treating with anticonvulsants, clinical seizures will change in subclinical seizures [139, 140] . Hence, clinical observation alone is ill-suited for their identification and monitoring [47, 80, 141]. Monitoring of the electroencephalogram (EEG) along with video is the gold standard for diagnosing and monitoring neonatal seizures [47]. However, most clinicians in NICUs opt to use amplitude integrated EEG [aEEG or cerebral function monitoring (CFM^{TM})] instead, because of the ease of use and minimal need for support from EEG technology and clinical neurophysiology services [47, 142]. Since single channel aEEG often misses short, low-amplitude, or focal seizures [47,143], reliable automated neonatal seizure detection using continuous multi-channel EEG monitoring using 13 to 21 scalp electrodes has the potential to help clinical decision making in the NICUs and alleviate significantly the workload of the EEG interpreters.

In the literature, a few heuristic and model-based algorithms have been proposed to detect neonatal seizures. They typically consist of if-then rules and thresholds and are called heuristic because they are derived from mimicking the experts and their strategies used to detect neonatal seizures. Autocorrelation techniques [144], rhythmic discharge detection [145], model-based EEG parameterization [146], modeling and complexity analysis [147], wave-sequence analysis [87], pseudo-periodicity analysis [148] and atomic decomposition [149] are some of the best known methods. Furthermore, an automated neonatal seizure detector mimicking a neonatal seizure expert was proposed in our group by [4] and was refined in [5]. In addition, artifact removal using different blind source separation techniques has been added to the detector to improve the performance [7]. Additionally, the performance of this method has been validated on an extensive dataset of asphyxiated neonates in the NICU of the Erasmus University Medical Center (EMC) Rotterdam [67]. The total good detection rate and positive predictive value (PPV) of this method, primarily reported to be 62% and 74% respectively, improved to 84% and 90% after removing four specific patients and some dubious seizures [67]. In this chapter, this method is referred to as "heuristic" algorithm.

On the other hand, machine learning approaches have also been applied to train data-driven classifiers for this problem. The following methods have been considered: time-frequency based analysis and multi-layer perceptrons

(MLPs) [150], quantitative features and a linear discriminant classifier [39], support vector machine (SVM) based classifier [151], Bayesian classifier via Gaussian mixture models [152], adaptive multi-channel information fusion [153], SVM classifier and Kalman filter [154], and trend template analysis with SVM classifier tested on fetal lambs [155].

In addition, multi-stage classification composed of heuristic rules supplemented by a data-driven classifier was applied in [156] and [88]. In the former, a heuristic algorithm is used for artifact removal and EEG segmentation. Afterwards, the features are extracted from the segments and MLPs are applied as a classifier to identify the seizures. In the latter, conversely, MLPs and a clustering technique are used to detect and cluster seizures (stage I, II) and then a heuristic model is applied to remove artifacts (stage III).

In this chapter, we describe a method for improving the heuristic neonatal seizure detector, a so-called multi-stage classifier, as explained in [6, 91]. In the first stage, the heuristic algorithm mimicking an expert EEG reader detects the potential seizures. Then, in the second stage, a data-driven post-processor identifies the main characteristics of the detected segments such as evolution of spikes, synchronization of EEG and polygraphic signals, and other time-frequency domain features, in order to remove the falsely detected segments. An extensive test on three independent datasets exhibits the improved false alarm rate (FAR) in comparison to the original heuristic algorithm and its extensions.

6.2 Data description

In this chapter, DB1-DB4 are used. The detailed descriptions about these datasets are given in section 3.2. Briefly, the EEG-polygraphic recordings from 71 neonates recorded in two medical centers, EMCR and ULZ, are used in this study. The EEG signals are filtered between 1-20 Hz and bipolar channels are made according to the montage maps explained in section 3.2. DB1 is used to develop and train the proposed algorithm while DB2-DB4 are used to validate.

6.3 Proposed method

The proposed method comprises two main stages: a heuristic classifier and a data-driven post-processor. In the first stage, the heuristic algorithm detects both the spike-train and oscillatory seizure type segments. At this stage, some artifacts or non-seizure segments having similar oscillatory or spike-train patterns (such as ECG artifacts) are incorrectly detected as seizure. In the second stage,

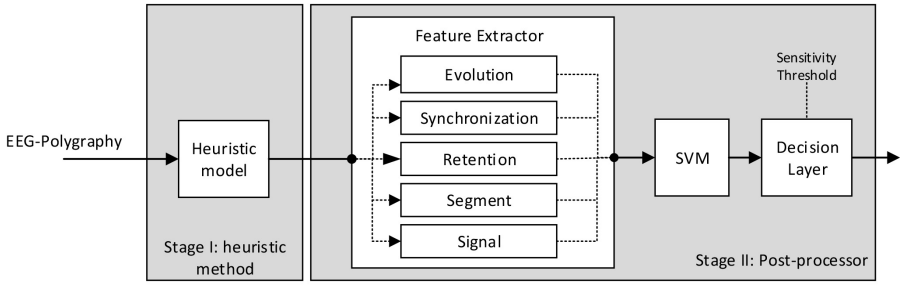


Figure 6.1: Diagram of the multi-stage detector. The input of the post-processor is the detected segments of the heuristic model

a pre-trained data-driven post-processor reanalyzes the heuristic detections using some informative features and a support vector machine. Figure 6.1 shows the schematic overview of the stages. The following subsections describe each step in detail.

6.3.1 Stage I: Heuristic algorithm

The heuristic algorithm is described in appendix A.1. Briefly, it is composed of two parallel procedures; detection of spike-train and oscillatory patterns. Despite the fact that the 2nd version of this algorithm is the improved version, in the current multi-stage method, the basic algorithm, version 1, without the improvements was used for the first stage. The main reason is that the 1st version detects more seizures and false alarms and the proposed post-processor (the second stage) is strong enough for removing such false alarms. Consequently, more seizures are detected at the end and it results in a higher sensitivity when compared to the 2nd version. In the Discussion, the performance of the multi-stage algorithm is compared with both versions.

6.3.2 Stage II: Data-driven post-processor

The detected segments of the heuristic algorithm are the inputs of the second stage. The following steps describe the procedure.

1. All heuristic detections are split into 8s epochs with 50% overlap.
2. Five sets of features (50 features in total) are extracted for each epoch (Feature extraction).

3. A pre-trained SVM classifier assigns a class membership probability to each epoch (Classification).
4. A decision layer aggregates the probabilities of all channels and classifies the detections into falsely and truly detected segments (Decision Making Layer).
5. The segments classified as falsely detected segments are removed and the performance is measured.

Feature extraction

Five sets of features (50 features in total) are extracted from each segment detected as seizure by the heuristic algorithm in the first stage. They include Evolution (12), Synchronization (4), Retention (5), Segment (5), and Signal features (24). The first four sets are extracted for the entire segment and repeated similarly in all epochs of that segment and the last set is extracted for each epoch separately.

Evolution features: The evolution of seizure patterns is one of the most characteristic features of neonatal seizure patterns as determined by visual analysis and is very useful in distinguishing it from rhythmic artifacts. In this method, three types of evolution have been taken into account: amplitude, frequency, and morphology. Because of the fact that the onset and the end of seizures have usually different (or opposite) evolutions, the segment is split into two subsections ('onset' and 'end') by determining the center of the segment. Then, evolutionary features of each subsection are extracted separately.

- Detecting the Center:
  - TKE of EEG signal is measured.
  - TKE is smoothed by a central linear moving average filter (MAF) (3s window).
  - The center is located where the filtered energy is maximum. This temporal point represents the time instant at which the overall EEG activity is maximum (figure 6.2.a).
- Evolution of Amplitude:
  - The absolute value of the EEG signal is filtered by a MAF (0.5s window).
  - The local peaks of the filtered signal are determined.

  - A line is fitted through the peaks using robust linear regression [157].
  - The number of peaks is used as a feature, so-called validity of regression.
  - The slope of the fitted line is also used as evolution of amplitude (figure 6.2.b).

- Evolution of Frequency:
  - The current subsection is split into 2s epochs with 75% overlap.
  - The power spectral density (PSD) is computed using the fast Fourier transform (FFT).
  - The mean normalized frequency (center of gravity of the PSD) is measured for each epoch.
  - A line is fitted through the mean frequencies of all epochs using robust linear regression [157].
  - The number of epochs is used as validity of regression.
  - The slope of the fitted line is used as evolution of frequency (figure 6.2.c).

- Evolution of Morphology:
  - The epoch that started 250 ms before the center of the segment and lasting for 500 ms is selected as template.
  - The normalized cross-correlation is measured between the template and the current subsection.
  - The local peaks of the cross-correlation coefficients are identified.
  - A line is fitted through the peaks using robust linear regression [157].
  - The number of peaks is used as validity of regression.
  - The slope of the fitted line is used as evolution of morphology (figure 6.2.d).

As a result, for each of the two subsections ('onset' and 'end'), 3 evolutionary slopes and 3 numbers of validity of regression are extracted. In total, every segment has 12 evolutionary features.

Synchronization features: One of the causes of false detections of automated seizure detectors are the physiological artifacts contaminating the EEG signals such as ECG spikes, respiratory artifacts, tremor artifacts, eye movements, and blood vessel pulsations. In practice, clinical neurophysiologists often employ the recorded polygraphic signals to distinguish these kinds of artifacts during

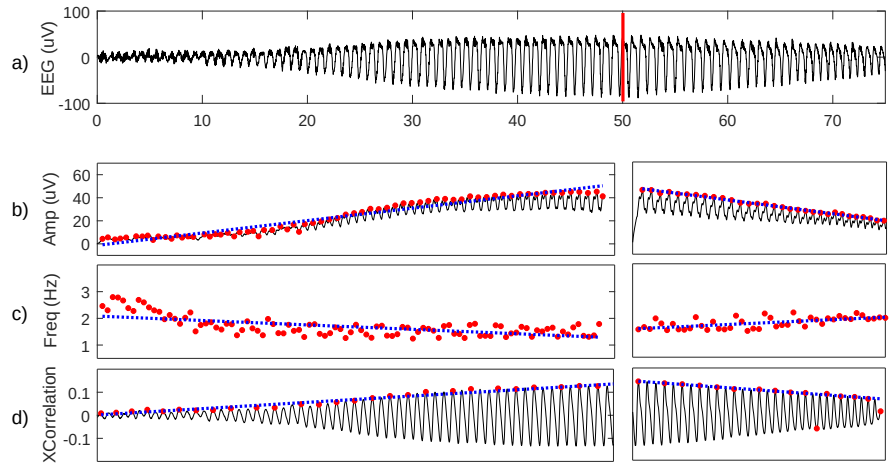


Figure 6.2: The behavior of the amplitude, frequency, and morphology evolution for a detected seizure. (a) Original detected EEG segment lasting for 75 seconds. The vertical red line shows the detected center of the segment where TKE is maximum. (b) Smoothed signal of rectified EEG amplitude (black line), detected local peaks of the smoothed amplitude (red dots), and the fitted line over the peaks (blue dashed line). The slope of this line determines the Amplitude Evolution. (c) The mean power frequency of 2s epochs of the EEG (red dots) and the fitted line (blue dashed line). The slope of this line determines the Frequency Evolution. (d) The normalized cross-correlation between the EEG and the template (black line), detected local peaks of the cross-correlation signal (red dots), and the fitted line (blue dashed line). The slope of this line determines the Morphology Evolution. In the traces (b-d), the reader's left side shows the 'onset' subsection and the right side shows the 'end' subsection.

the visual assessments. In other words, they look for synchronization between the EEG spikes and polygraphic events (such as the QRS complex in the ECG signal). In order to quantify such synchronization, Mean Phase Coherence (MPC) of angular distributions is employed and calculated by:

$$MPC_{(ab)} = \left(\left[\frac{1}{N} \sum_j \sin \phi_{ab}(j\Delta t) \right]^2 + \left[\frac{1}{N} \sum_j \cos \phi_{ab}(j\Delta t) \right]^2 \right)^{\frac{1}{2}} \qquad (6.1)$$

where N is the number of samples, $j \in [0, N-1]$, and $\phi_{ab}(t)$ is the phase difference as:

$$\phi_{ab}(t) = \phi_a(t) - \phi_b(t) = \arctan \frac{\widehat{s_a}(t)s_b(t) - s_a(t)\widehat{s_b}(t)}{\widehat{s_a}(t)\widehat{s_b}(t) + s_a(t)s_b(t)} \tag{6.2}$$

with $s_x(t)$ and $\widehat{s_x}(t)$ denoting the signal and its Hilbert transform. The latter one is calculated by:

$$\widehat{s_x}(t) = -i\mathscr{F}^{-1}\{\mathscr{F}\{s_x(t)\}sign(\Omega)\} \tag{6.3}$$

where $\mathscr{F}\{.\}$ and $\mathscr{F}^{-1}\{.\}$ are the Fourier transform and its inverse respectively [158]. The main advantage of using MPC is that the highly discrepant amplitudes of polygraphic signals and EEG signals would have no effect on the synchronization measurement. Moreover, MPC is a robust measurement against noise for phase synchronization of biomedical time series [158]. To illustrate the difference of MPC between an artifact and an actual seizure, figure 6.3 plots a falsely detected ECG artifact (a, MPC: 0.21) and a truly detected rhythmic seizure (b, MPC: 0.04). In the current work, the MPC of EEG segment and the polygraphic signals ECG, EMG, EOG, and Resp. are measured and used as four synchronization features [91].

Retention features: In clinical practice, the physician interpreting the ongoing cEEG signal would initially appraise a number of seizure-like patterns and would consider start of AED treatment only if the patterns are repeated over time. Furthermore, seizure burden and the spatial concentration can affect the visual detection and subsequent decision to initiate treatment. In order to take this fact into account, five features are extracted from one-hour signals before the onset of every detected segment, called Retention features. They consist of (I) the number of seizures detected by the heuristic algorithm in the same channel and (II) in all channels per hour and in addition, (III) the seizure burden detected in the same channel and (IV) in all channels per hour. In order to solve the lack of data in the beginning of recordings or after missing data, a fifth feature is defined. It is 1 if at least one-hour data before the segment is available. Otherwise, it equals the duration (in hours) of the available data (so it is always between 0 and 1). In other words, the fifth feature notifies the classifier how reliable the other four Retention features are.

Segment features include 5 features extracted from the detected segment: 1) the length of the segment, 2) the number of channels expressing this seizure, 3) the type of the detection (1: spike-train, 2: oscillatory, or 3: mixed), and 4,5) (x, y) coordinates of the detected channel using a simplified two-dimensional grid on the 10-20 system of electrode placement map [91].

Signal features comprise 24 features extracted directly from the EEG epochs (8 s). These features are selected from the features used in [39] and [3] by a

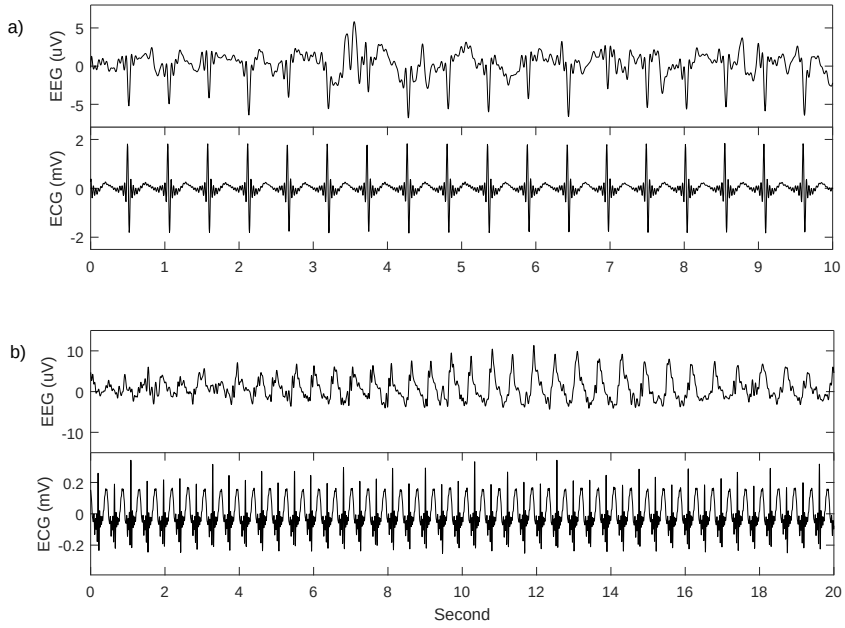


Figure 6.3: Two detected segments by the heuristic algorithm, (a) is a falsely detected ECG artifact (MPC: 0.21) and (b) is a truly detected rhythmic seizure (MPC: 0.04).

Lasso feature selection technique [159] and listed in table 6.1. The formulas of the features are exhaustively described in [39].

Classification

In order to classify the detected segments, a support vector machine (SVM) with radial basis kernel function (RBF) is used. For training the SVM, 10000 segments where randomly selected from about 280000 available training segments. For optimizing the hyper-parameters, leave-'one patient'-out (LOPO) cross validation was applied on the training dataset and the best set of the hyper-parameters were selected. Basically, the output of the SVM for each epoch is a binary value 0,1 corresponding to the class labels false detection, true detection. However, for measuring the membership probability, a scaling method was proposed in [160] and improved in [161]. The latter one is used here to assign a probability of being a true detection to every epoch. Consequently, a

Table 6.1: Selected signal features

Type	Feature
Time Domain	• The total number of maxima and minima • Root mean square amplitude • Hjorth Complexity • Auto regression modelling error (order: 9) • Skewness • Kurtosis
Frequency Domain	• 6 spectral powers in sub-bands (1-3 Hz, 3-5 Hz, 4-6 Hz, 6-8 Hz, 9-11 Hz, 10-12 Hz) • 6 normalized spectral powers in sub-bands (1-3 Hz, 2-4 Hz, 3-5 Hz, 4-6 Hz, 6-8 Hz, 7-9 Hz) • Peak frequency • Spectral edge frequency (80%) • The wavelet energy of (1-2 Hz)
Information Theory	• Shannon entropy • Singular value decomposition entropy • Spectral entropy

detected segment which was split into 8 s epochs (with 50% overlap) is mapped by the SVM and the scaling method onto a vector of probabilities. Then, the decision making layer processes this probability vector.

Decision Making Layer

The task of this layer is aggregating the probabilities of epochs and channels of each detected segment. First, for aggregating the epochs of every detected channel, the vector of probabilities is compared to a threshold, called 'sensitivity threshold', and majority voting is applied to assign one label (0: 'false detection' or 1: 'true detection') to that channel. Second, for aggregating the channels, a logical "OR" operator is applied to the labels from the channels to define the truly detected segments. It means that if a detected segment has at least one channel labeled as 'True detection' (after the majority voting), it remains in the detection list. Otherwise, the segment is marked as false detection and is removed (figure 6.4).

In order to choose the sensitivity threshold, first it varies from 0 (where no seizure is removed) to 1 (where all seizures are removed) and forms the Receiver Operating Characteristic (ROC) curve. Then, the elbow of the ROC showing the trade-off between the good detection rate and false alarm rate of the training data (DB1) measured by LOPO is selected.

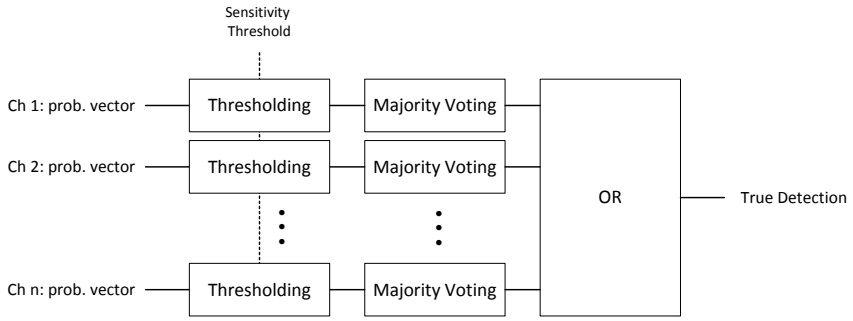


Figure 6.4: Diagram of the decision making layer

6.4 Results

6.4.1 Improvement of performance metrics

Figure 6.5 shows the ROC curves for DB1-4. The curve of DB1 (training dataset) results from LOPO cross validation. When the sensitivity threshold (TH) is 0, none of the detections are removed by the second stage. Therefore, the dark circles (where TH=0) are demonstrating the performance of the heuristic method (the first stage). Increasing the TH results in removing more detections and subsequently decreases the GDR and the FAR. Two elbows of the curve of the training dataset (DB1) expressing the trade-offs between decrease of the GDR and FAR have been approximately chosen at 0.1 and 0.3. In this figure, the GDR and FAR of those THs for each dataset are defined by the squares and triangles respectively. Furthermore, table 6.2 compares the event-based and epoch-based performance metrics of the heuristic and the proposed methods on the test datasets at these thresholds. Additionally, the averaged performance metrics measured on all patients of the test DBs are reported. It shows that the proposed post-processor (when TH=0.3) is able to decrease FAR by 64% whereas the GDR decreases 7%.

6.4.2 Analysis of event duration

Figure 6.6 shows the histograms of the total true and false detections of the test datasets as a function of segment duration. Comparison of the first bars of the upper and lower histograms (< 30 s) to the other bars reveals that detecting very short seizures lasting less than 30 seconds is the most challenging task of

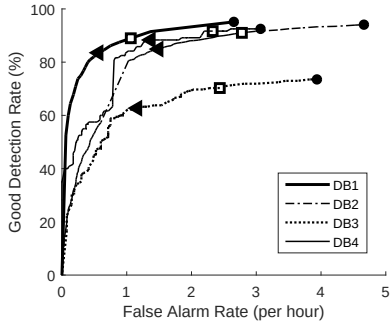


Figure 6.5: ROC curves of variation of the Good Detection Rate (%) against False Alarm Rate (h^{-1}) for the four datasets. The marked points indicate the performance at different sensitivity thresholds: 0 (filled circle), 0.1 (square), and 0.3 (rectangle)

Table 6.2: The difference in performance between the heuristic and proposed methods

Method		DB2	DB3	DB4	Average
Heuristic	GDR %	94	74	93	88
	FAR (h^{-1})	4.66	3.94	3.07	3.81
	Sensitivity %	92	61	80	79
	Specificity %	78	92	92	88
	kappa	0.28	0.54	0.63	0.49
Proposed TH = 0.1	GDR %	91	70	92	86
	FAR (h^{-1})	2.78	2.44	2.33	2.50
	Sensitivity %	91	60	80	79
	Specificity %	80	93	93	89
	kappa	0.30	0.56	0.64	0.51
Proposed TH = 0.3	GDR %	85	63	88	81
	FAR (h^{-1})	1.49	1.16	1.37	1.36
	Sensitivity %	89	58	79	77
	Specificity %	82	95	93	90
	kappa	0.31	0.57	0.66	0.52

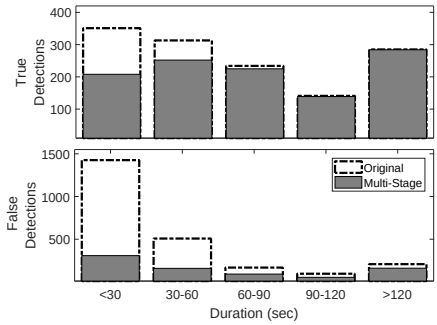


Figure 6.6: Histogram of the true and false detections for the original heuristic method (dashed bars) and the proposed multi-stage method (grey bars) as function of duration

the seizure detector since 26% of true and 60% of false detections are shorter than 30 seconds. However, the post-processor removes 40% and 78% of such short true and false detections respectively and results in an increase of the PPV from 20% to 40% (for detections < 30 s). However, it is not yet very satisfactory. The detections lasting more than one minute are almost not falsely removed by the post-processor. The analysis also shows that the duration of the seizures need to be taken into account when studies regarding seizure detection are evaluated.

6.4.3 Analysis of EEG experts' agreement

All detections of the DB3 and DB4 were rescored by four secondary independent EEG experts and labeled as 'Definite seizure', 'Definite artifact', or 'Dubious'. Subsequently, for each detection, 5 labels are available (One primary and four secondary). Then, the majority voting is applied to assign the final label and the percentage of the majority vote is calculated. As a result, all the detections are classified into 'poorly agreed' ($<$60% of votes), 'moderately agreed' (60% to 80%), and 'highly agreed' ($\geq$80%). Figure 6.7 shows the histogram of the definite seizures and the definite artifacts as a function of the agreement classes. The post-processor (TH = 0.3) removes only 4% of highly agreed seizures, while 66% of highly agreed non-seizure detections are removed. It increases the PPV of this class from 57% to 79% which means 22% more reliable detections of well-defined events.

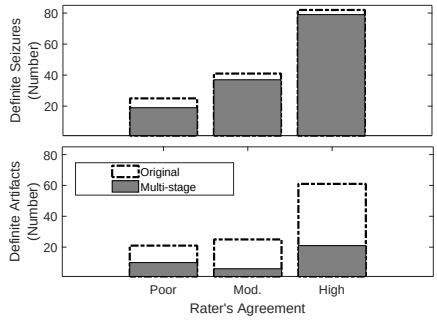


Figure 6.7: Histogram of the definite seizures and artifacts for the original heuristic method (dashed bars) and the proposed multi-stage method (grey bars) as a function of rater's agreement

Table 6.3: Ranked feature sets using ridge regression and MI

Rank	Ridge (Standardized Coefficients)	MI (MI value)
#1	Segment (9.00)	Retention (0.20)
#2	Evolution (4.98)	Segment (0.14)
#3	Signal (0.02)	Evolution (0.14)
#4	MPC (0.02)	Signal (0.05)
#5	Retention (0.01)	MPC (0.03)

6.4.4 Feature Ranking

In order to rank the feature sets, two methods have been used to assign an importance score to each feature. The first method is fitting a linear ridge regression and using the absolute value of the standardized coefficients of the feature vectors [127]. The second one is measuring mutual information (MI) between the features and the seizure labels [159]. Table 6.3 sorts the feature sets based on the averaged coefficients and mutual information values. The order of the features is similar in both methods, except for the Retention features. The reason may be high nonlinear correlation between the Retention features and the labels which is not measurable by the linear ridge method.

6.5 Discussion

In typical pattern recognition problems such as neonatal seizure detection, the ground truth must be initially identified in order to train and test the method. In neonatal seizure detection, due to the lack of other biological markers for neonatal seizures, only visual analysis of expert clinical neurophysiologists is used and considered the 'gold standard'. Therefore, automated seizure detection (ASD) needs to be considered as an extension of the visual analysis. It can thus function as a tool for supporting clinical decision making in a busy NICU. The final decision regarding treatment initiation is best left to the team of treating physician and clinical neurophysiologist.

However, visual analysis is imperfect science considering the rather modest inter-rater agreement especially in the presence of dubious seizures. In other words, high agreement between human raters is obtained in only very typical/classical seizure patterns with well-defined morphology, reasonable amplitude and duration. Thus, if an automated seizure detector is developed solely to identify such well-defined patterns, it would mean not acknowledging the rich variety encountered in seizure patterns for practical applications. In order to build a realistic decision support tool, different types of clear and dubious patterns should be used to develop, train, and validate the algorithms. Besides, a multi-rater analysis of the automatically detected seizures can measure the overall satisfaction of different experts if they want to use that system in their NICUs. Without automated methods, sufficiently powered large multi-center cEEG studies cannot be done. There is a pressing need for such studies to address gaps in present knowledge, such as whether aggressive treatment of subtle and subclinical neonatal seizures with antiepileptic medication will improve the outcome of neonates.

In this work, different patients recorded in two centers and having different etiologies were used to train and test the proposed method and compared with the original one. Using the clear and dubious seizures at the same time provides a challenge for the methods, as would be expected in real-life scenarios. The multi-rater analysis showed the added-value of the proposed method for detecting the well-defined events.

During visual neonatal seizure detection by clinical neurophysiologists while providing clinical service to a NICU, in addition to the raw EEG data from a specific time frame, different types of information are used to identify the seizure, such as spatial distribution, corresponding compressed EEG trend signal such as the aEEG signal, the overall background activity and status of previous or subsequent seizures, evolution of patterns, etc. In addition, the EEG signal is usually being observed in frames of different duration depending

on the background activity, type of seizure, morphology, EEG amplitude and other characteristics that the expert has learned over years. In ASD, in order to have a similar approach and extract some of this extra information, the length of the processing window should be variable. For instance, a short processing window (a few seconds) is required to extract fast-rate features like frequency information, while a longer window (about a minute) is needed for considering the evolution of patterns, or very long window (about hours) should be used to take the features of the previous seizures into account.

In this method, unlike most retrospectively proposed automatic seizure detectors having a (relatively) fixed-size window, different window lengths varying from 8 seconds to 1 hour are applied to extract the 'core characteristics' of seizures including: the features of the EEG signal at a specific time (signal features), the spatial information (segment features), the evolution of patterns (evolution features), the correlation of EEG and polygraphic signals (synchronization features), the seizure burden and repetition detected in the previous one hour (Retention features). As a result, the FAR drops by 64% while the GDR decreases by 7% when the TH is 0.3, or the FAR decreases by 32% with only 2% reduction of the GDR when TH equals 0.1.

However, it should be taken into consideration that the described 'core characteristics' of seizures in this chapter are purely based on visual patterns and not based on any pathophysiological basis for these patterns. Without more basic research into underlying seizure mechanisms at cellular, synaptic and network levels, the processes underlying these characteristics cannot be elucidated. Other modifying factors such as severity and nature of underlying brain injury and treatment with antiepileptic medications also need to be taken into account.

Table 6.4 lists the performance of the heuristic algorithm and its extensions compared to the proposed multi-stage classifier. It shows that, the refined version of the heuristic algorithm [5] results in a GDR of 67% with 1.7 FAR per hour on the test DBs on average per patient. However, by keeping the same FAR (corresponding to TH=0.23) the multi-stage method produces a GDR of 82%. Besides, if the blind source separation (BSS) technique [7] is applied on the EEG signal in order to pre-process and remove the polygraphic-related artifacts, the FAR decreases to 1.2 (h^{-1}) with a GDR of 66%. However, the multi-stage method (without pre-processing) has a 14% higher GDR for the same FAR (where TH = 0.33). Additionally, because the multi-stage method has less time complexity (in recall) compared to BSS, it post-processed the test DBs about 3 times faster than the pre-processor which is an important advantage for real-time implementation.

Figure 6.8 shows the performance of the proposed method compared to the other

Table 6.4: Performance of the original heuristic algorithm and its extensions compared to the proposed method

Method	DB2		DB3		DB4		Average	
	GDR	FAR	GDR	FAR	GDR	FAR	GDR	FAR
Deburchgraeve [4] (Ver. 1) = Multi-stage (TH=0)	94%	4.6	74%	3.9	93%	3.1	**88%**	**3.8**
Deburchgraeve [5] (Ver. 2)	76%	2.6	49%	2.1	72%	0.9	67%	1.7
Multi-stage (TH=0.23)	87%	1.7	65%	1.6	88%	1.8	**82%**	**1.7**
De Vos [7]	69%	1.4	49%	0.8	73%	1.5	66%	1.2
Multi-stage (TH=0.33)	84%	1.4	62%	1.0	88%	1.3	**80%**	**1.2**

reported ones. In this figure, the grey circles show the reported performance of the heuristic algorithms and its extensions in the original studies. However, the dark circles are corresponding to the grey ones when those algorithms are again applied on the test DBs used in this study. Furthermore, the white diamond shows the performance of the algorithm that was proposed in [3] while the dark diamond displays its performance when the algorithm is re-implemented in our group and tested in DB2-4. The considerable difference between the performance of the dark and light symbols demonstrates that comparing different algorithms while using different databases results in a very inaccurate judgment. There are many database-related factors that may lead to such inaccuracy like the difference in the number of patients, the density (prevalence) of seizures, the length of recordings, the distribution of seizure durations, and the grade of the background activities [67, 89, 148]. The most important reason that makes our test DBs more challenging is that the dubious seizures were not removed in our datasets. Generally, detecting the dubious seizures which are usually shorter and of a very low-amplitude is a difficult task and needs a fine seizure-sensitive detector which would produce many false alarms in return. Various seizure-related factors such as severe brain injury as well as aggressive treatment with AEDs such as midazolam and lidocaine might influence seizure morphology and rhythmicity and lead to more dubious patterns [67]. The automated seizure detection algorithm would help tackle this kind of problems.

6.6 Conclusion

A major improvement of a previously developed automated neonatal seizure detector was achieved by combining a machine learning technique with the heuristic algorithm. In such a multi-stage detector, first a simplified visual seizure detection approach of an EEG expert is formulated by a few heuristic if-then rules and thresholds. Then, the core characteristics of seizure patterns are extracted by five sets of features including I) "evolution features" for

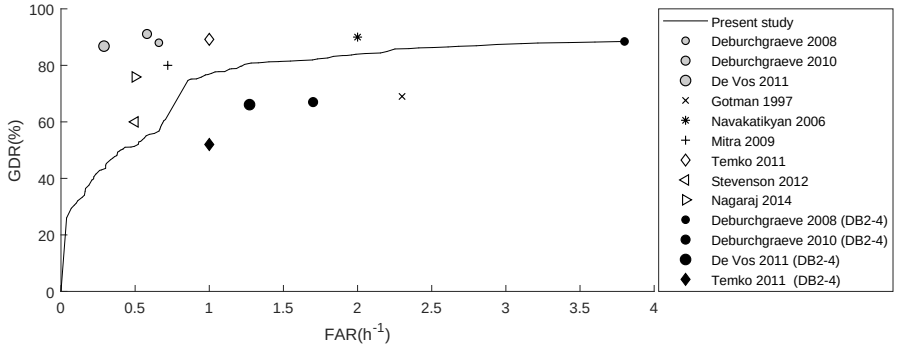


Figure 6.8: Performance of the proposed multi-stage classifier tested on DB2-4 compared to the reported performance of other methods. The grey circles show the previously reported performance of the heuristic method tested on different databases. The dark circles show the performance of the heuristic method and its extensions when it is applied on DB2-4. Furthermore, the light diamond shows the GDR (with one false alarm per hour) reported in [3] while the dark diamond shows the GDR of its re-implementation tested on DB2-4.

quantifying the evolution of amplitude, frequency and morphology of detections, II) "synchronization features" for extracting the correlation of detected spikes of the EEG with those of the polygraphic signals, III) "Retention features" for measuring the seizure burden and repetition occurring before the detections, IV) "Segment features" for using general information of the detections such as spatial spread or type of seizure, and V) "signal features" for extracting the common signal processing features. Next, the features are fed into an SVM in order to distinguish between the truly and falsely detected segments and minimize classification error. The extracted features are based either on the underlying behavior of the EEG seizure patterns such as evolution, retention, synchronization etc., or their mathematical characteristics such as kurtosis, Hjorth complexity etc. The core features of neonatal seizure patterns extracted in this study through a data driven approach are suitable for researchers developing automated seizure detection methods for further testing on larger databases. We have shown that a combination of the heuristic and data-driven approaches yields a significantly improved performance.

Chapter 7

Seizure detection using adaptive learning

This chapter introduces an extension of the multi-stage neonatal seizure detector explained in the previous chapter, 6. In this method, we propose to add a third stage to the detector in order to use feedback of the clinical neurophysiologist and adaptively re-tune a threshold of the second stage to improve the performance of the detections, particularly of the brief-lasting seizures. This study has been published in modified form in [162].

7.1 Background

As mentioned, occurrence of neonatal seizures usually denotes an underlying serious neurological dysfunction. As the majority of these seizures have a subtle or no clinical correlate, their timely detection requires continuous electroencephalographic monitoring (cEEG). However, due to lack of expert support, there is a significant unmet gap in implementing cEEG in neonatal intensive care units (NICUs) [27, 38, 47]. Furthermore, occurrence of seizures may potentially aggravate brain injury, as has been shown in laboratory animal studies, although it is still debatable in human newborns [163–167]. Brain monitoring is usually performed in NICUs by using a compressed single channel EEG trend, known as amplitude integrated EEG (aEEG), because of its ease of use and interpretation. However, this has many limitations. Visual interpretation of multi-channel cEEG by an experienced clinical neurophysiologist is the ideal method for detecting seizures. However, such expertise is often not available around the clock. Moreover, this method is expensive and labor-intensive [47, 67]. Therefore, an accurate automated neonatal seizure detector can be a valuable supportive tool for NICUs and it can potentially improve the quality and decrease the expenses of continuous monitoring of brain function.

In recent years, many publications have described different methods of automated neonatal seizure detection, varying from heuristic methods [4, 7, 147–149], data-driven classifiers [3,150,154,155], and their combinations [6,88,91,156]. One of the greatest challenges for these methods is the detection of brief-lasting seizures [3, 92, 95, 149] because typical characteristics of seizures, such as evolutionary patterns, are less clear-cut and at the same time they may be misinterpreted as artifacts [6]. Therefore, improved detection of these can boost the performance of an automated detection algorithm. The main goal of the method proposed in this chapter is improving the overall performance of a previously developed multi-stage neonatal seizure detector particularly by improving the performance of the short seizure detections. To this end, after alarming for each detected seizure, the feedback from the expert is collected and used immediately to enhance the detector for subsequent detections. Thus the detector is constantly tuned for different channels and patients.

7.2 Data description

In this chapter, DB1-DB3 are used. The detailed descriptions about these datasets are given in section 3.2. Briefly, the EEG-polygraphic recordings from 48 neonates, all with HIE, recorded in one medical centers, EMCR, are used in

this study. The EEG signals are filtered between 1-20 Hz and bipolar channels are made according to the montage maps explained in section 3.2.

To develop and train the method, DB1 including 17 neonates each of which has 27 hours on average is used. For testing, we used 31 different neonates with the same measurement protocol from the same center. In total, the database includes 461 hours of training and 516 hours of test data. The mean and median seizure duration of the test sets, including 1975 seizures, are respectively 67 and 40 seconds.

7.3 Methods

7.3.1 The multi-stage detector

The multi-stage algorithm is described in chapter 6 in detail. Briefly, in the first stage, a heuristic algorithm mimicking a human expert is applied on the EEG signals in order to define some potential seizures. Afterwards, in the second stage, the EEG-polygraphic signal of each potential seizure is split into shorter epochs (8 s, 50% overlap) and 50 features from each epoch are extracted. Then, a pre-trained support vector machine (SVM) is applied on the extracted features and a probability is assigned to each epoch. Finally, all probabilities of each potential seizure are compared with a predefined threshold (TH) and aggregated by a majority voting technique. In this way, some potential seizures primarily detected by the heuristic algorithm are removed as falsely detected seizures by the post-processor in the 2nd stage. As a result, the false alarm rate (FAR) significantly diminishes by 64% while the good detection rate (GDR) decreases only by 7%, when the TH is fixed to 0.3 based on used training data.

7.3.2 Proposed method

The main idea of this proposed method is that in order to improve the performance of detecting short seizures, the sensitivity of the detector should vary according to the three following facts: 1) different patients may have different grades of background activity abnormality. As it is shown in [27], there is a relation between the background activity and characteristics of seizures, such as duration, repetition, rhythmicity, and evolution. The neonates with severe brain dysfunction have usually shorter and more repeated seizures. Therefore it is expected that the sensitivity of the detector should change from patient to patient. 2) Many short seizures are either focal, i.e., when a seizure activity appears only on one EEG electrode, or regional, i.e., when two or more

adjacent electrodes display the seizure activity. Hence, only a limited number of channels need to be investigated for detecting seizures. In that case, using other irrelevant channels may inadvertently increase the false alarm rate. Thus, it is expected that the detector operates differently on different channels. 3) In some cases, especially for neonates with more severe brain injury and very short and repeatedly occurring seizures, the repetitiveness may disappear after treatment with anti-epileptic drugs (AEDs). Accordingly, it is expected that when a few seizures with robust characteristics are detected, the method becomes tuned to detect new seizures in that channel, even if these may not express all the characteristics of the previous seizures. Conversely, when there are no well-defined seizures or if some artifacts are wrongly detected in a channel, the detector needs to become less sensitive to the detection of dubious seizure-like patterns for that specific channel. To take these facts into account, the threshold (TH) of the second stage of the mentioned multi-stage seizure detector will be adaptively re-tuned in a third stage using the output of the second stage and feedback of an expert user. Every false/true detection will increase/decrease the TH of that channel and other channels in the neighborhood. This change applied on the TH will gradually be forgotten unless other similar feedback is regularly being received. The penalizing/rewarding rate as well as the forgetting factor should be optimized based on the training dataset. Figure 1 shows a simple block diagram of the proposed multi-stage classifier. As it is clear in this diagram, after receiving feedback for each detected seizure by the post-processor, the new value of TH is recalculated based on the received feedback, the elapsed time from the last feedback, the channel in which the current seizure is detected, and the old value of the TH. The following steps thoroughly describe the proposed scenario when a potential seizure is detected on channel ch^* in the first stage:

1. The potential seizure is split into 8s epochs with 50% overlap.
2. The features proposed in in the previous chapter are extracted for each epoch.
3. A pre-trained SVM (with probabilistic output) is applied on the features and a probability is assigned to each epoch.
4. The thresholds of all available channels are primarily updated by (7.1), derived from the transient and steady-state responses of resistor-capacitor circuits with an initial condition.

$$TH_{ch} \leftarrow TH^0 + (TH_{ch} - TH^0)e^{-\frac{\Delta t}{\tau}} \tag{7.1}$$

TH_{ch} is the sensitivity threshold of the updating channel. TH^0 is a predefined steady-state sensitivity threshold, Δt denotes the elapsed time

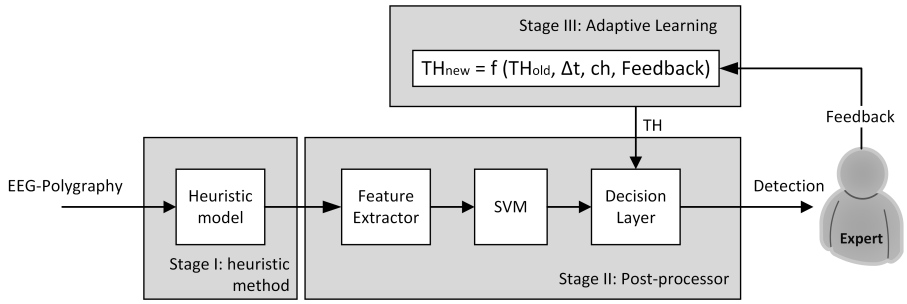


Figure 7.1: Diagram of the proposed technique. The TH used in the post-processor is adaptively re-turned by the proposed 3rd stage based on the current feedback, elapsed time from that last feedback, detected channel, and the old value of TH.

(per second) from the previous update, and τ is a keeping factor (inverse of the forgetting factor) which shows how long a feedback will affect the threshold (the feedback will be used in step 6).

5. Then, all assigned probabilities of the current potential seizure are compared with the new TH_{ch^*}. If a majority of epochs has probabilities smaller than TH_{ch^*}, the potential seizure is labeled as false detection and removed. Otherwise, it is labeled as seizure and the detector alarms.

6. If a new seizure is detected, the real label (y: seizure, dubious, non-seizure) of that seizure scored by an expert is used to update the thresholds of all available channels by

$$TH_{ch} \leftarrow TH_{ch} - \frac{\gamma}{d(ch, ch^*) + 1} R(y) \tag{7.2}$$

where γ is a pre-defined learning rate, R is a rewarding function defined by (7.3), and $d(ch, ch^*)$ is the graph-based distance between the updating channel (ch) and the detected channel ch^* on the montage graph.

$$R(y) = \begin{cases} +1 : y = Seizure \\ 0 : y = Dubious \\ -1 : y = Nonseizure \end{cases} \tag{7.3}$$

As a consequence, if a seizure is detected on a channel and the expert user believes that the seizure has truly been detected, the threshold of that channel

will decrease by γ, and subsequently, the thresholds of its neighbor channels with one common electrode decrease by $\gamma/2$, and likewise for other channels. Conversely, if the user scores the detected seizure as false alarm, the threshold increases by γ. This change ($\pm\frac{\gamma}{n}$) added to the threshold exponentially decays in time based on (7.1) so that if no alarm is found or no feedback received for 3τ seconds, the threshold will become almost equal to TH^0. If the seizure is scored as dubious, it has no effect on the threshold.

7.4 Results and discussion

In this work, the used learning rate (γ) and keeping factor (τ) were optimized by a leave-'one patient'-out (LOPO) cross-validation method applied on the training dataset and equal to 0.1 and 10 minutes respectively.

Figure 7.2 shows the performance of the proposed method as well as the original method with fixed threshold for the test DBs. As is clear in this figure, the overall area under the ROC increased by the proposed method. In the previous chapter, it was shown that TH=0.3 is a good trade-off between the GDR and the FAR based on the training data. The same threshold has been used here as TH^0 (squares and circles in the figure). As a result, the overall FAR (for DB2&3) decreased by 35% while the GDR did not decrease. At this point, the PPV also increased by 5%.

Figure 7.3 displays the histogram of the detected seizures, the number of false alarms, and the PPVs as a function of seizure duration for both methods and both DBs together. Almost all seizures longer than 1 min were detected by both methods (upper chart), while the short seizures are still not entirely detected. Nevertheless, the number of false detections of very short seizures (< 30 s) decreased by 50% which is the main contribution of this work. The PPV of very short seizures also increased from 41% to 59% by the proposed method which shows higher reliability of the alarms.

The other advantage of this method is that since the feedback of the expert user is continuously used to re-tune the threshold, the method can tailor to the needs of the user. In other words, if an expert user prefers to detect more seizures including the dubious and definite ones or to detect only definite seizures, the method can adaptively learn these preferences. Consequently, it is possible that this method would have higher overall performance in a multi-rater scored database compared to other kinds of fixed detectors. But, this needs to be tested in a future work.

The main drawback of this technique is the need for collecting feedback from the

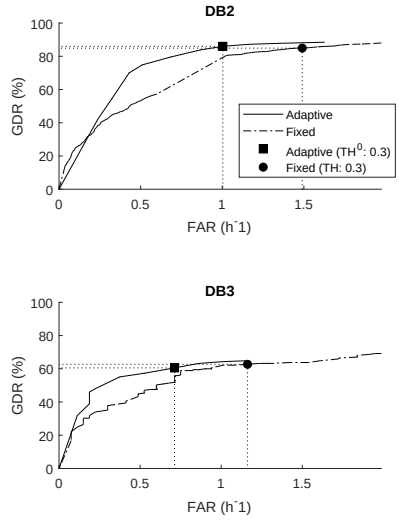


Figure 7.2: The ROC curve of the seizure detector with fixed threshold (dashed line) as well as the proposed adaptive threshold (continuous line) for DB2 (top) and DB3 (bottom).

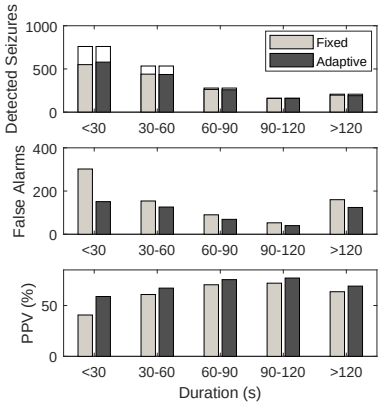


Figure 7.3: Histogram of the detected seizures (top), the number of false alarms (middle), and the PPVs (bottom), for the original method with fixed threshold (light gray) and the proposed adaptive learning method (dark gray). In the upper chart, the white stacked bars show the number of missed seizures.

expert users in a real NICU environment. A major problem is the limited EEG expertise among front-line carers in the NICUs, the nurses and neonatologists. These primary caregivers generally call a neurologist or clinical neurophysiologist for consultation when they see repetitive patterns that frequently occur or are considering initiating treatment with anti-seizure medications. In this research it is presupposed that after each alarm, the feedback of an expert EEG reader is available. In future research, the practical limitations, such as the limited knowledge of primary caregivers and a delay between an alarm and its feedback for each patient can be addressed.

7.5 Conclusion

In this chapter, an adaptive learning method was proposed to decrease the false alarm rate and subsequently increase the positive predictive value of a previously developed multi-stage neonatal seizure detector particularly for very short seizures and false alarms. This method effectively improves the performance such that the FAR decreases by 35% and PPV increases by 5% on average, while the good detection rate remains unchanged. Moreover, for very short events, which are very challenging problems in automated neonatal seizure detection, the FAR decreases by 50% and PPV increases by 18%.

Chapter 8

Seizure detection using convolutional neural networks

In this chapter, we introduce a new algorithm based on deep convolutional neural networks and random forest to automatically optimize feature selection and classification. Unlike previous algorithms, this method does not need any hand-engineered feature extraction. The input of the proposed classifier is raw multi-channel EEG and the output is the class label: seizure/non-seizure. By training this network, the required features are optimized, while fitting a nonlinear classifier on the features. After training the network with the training EEG recordings, the end layers of the network performing the classification are replaced with a random forest classifier in order to improve the performance. The architecture of the network and its results, as well as the advantages and disadvantages are discussed in this chapter. This study has been published in modified form in [168].

8.1 Background

As mentioned in the previous chapters, neonatal seizures usually indicate serious neurological dysfunction, and could potentially worsen underlying brain injury [38, 47]. The majority of neonatal seizures are acute symptomatic events, unlike the unprovoked epileptic seizures observed in older children and adults [27, 38]. These seizures may have absent or subtle clinical manifestations, which may resemble normal behavior, such as lip smacking, sucking, chewing, and blinking. This makes neonatal seizure detection very difficult and inaccurate if it solely relies upon clinical observation [37, 80, 169]. It has been shown that the most accurate method for their detection is visual interpretation of continuous multi-channel EEG along with video by an expert clinical neurophysiologist [47]. However, such interpretation is extremely labor-intensive, time-consuming, and importantly, needs special expertise which is not available around the clock in many neonatal intensive care units (NICUs). A reliable and accurate automated neonatal seizure detector using multi-channel continuous EEG can be a very helpful supportive tool, particularly for the NICUs.

In the literature, there are several model-based methods for the automatic detection of neonatal seizures, usually developed based on heuristic if-then rules and some thresholds and parameters. Lie et al. computed the periodicity score using autocorrelation techniques and used 3 if-then rules and 5 thresholds to detect seizures [144]. Gotman et al. proposed a rhythmic discharge detector using 3 parallel methods in order to detect rhythmic discharges, multiple spikes, and very slow rhythmic seizures. This algorithm used 10 different thresholds in total [145]. Celka and Colditz used a singular spectrum analysis and compared the 'optimum required model order' with a threshold to detect seizures [147]. Navakatikyan et al. applied a wave-sequence analysis and used about 9 thresholds to detect seizures [87]. Furthermore, a heuristic algorithm mimicking a human EEG reader was developed in our group, NeoGuard, and published in [4], improved in [7], and validated in [67]. In this method, spike train type and oscillatory type seizures are detected by two parallel algorithms. The common feature of the aforementioned methods is that the thresholds and parameters were found empirically, usually by trial and error, and therefore they might not be optimized.

Other research groups focused on the development of data-driven methods for this task. Among them, the following algorithms have been considered: Hassanpour et al. used a singular value decomposition and 'successive spike interval analysis' in order to extract respectively the low and high frequency features. Then, the features were fed into two separate artificial feed-forward neural networks, each of which has two-hidden layers [150]. In [39], twenty one features, including frequency domain, time domain, and entropy based

features, were extracted from 2s epochs and then used into a classifier based on linear discriminant analysis. In [152], a Gaussian mixture model (GMM) was applied on 55 extracted features from 8s epochs. Temko et al. and Ahmed et al. employed the same set of features with a support vector machines (SVM) classifier, using a radial basis function (RBF) and a 'Gaussian dynamic time warping' kernel [170], [3]. In [149], a dictionary was created using an atomic decomposition technique applied on the training data. The complexity of the atoms were then measured and aggregated to define seizures. In [155], five features were extracted and an SVM classifier was used to detect newborn lamb seizures.

In addition, in some proposed methods, hybrid models combining data-driven methods, heuristic rules, and an empirically found set of parameters/thresholds have been used. Aarabi et al. used some if-then rules with predefined thresholds to remove artifacts, and then applied a multi-layer perceptron (MLP) with two hidden layers to detect seizures [171]. In [88], an MLP and a clustering method were used to detect and cluster seizures. Then, three rules with some predefined thresholds remove the artifacts and decrease the number of false detections. Lastly, as explained in chapter 6, Ansari et al. developed a multi-stage classifier composed of a heuristic method to detect potential seizures and a data-driven post-processor to remove artifacts. In the post-processor, different sets of features were introduced and extracted and an SVM was then used to classify the detected potential seizures [6]. In all the previously mentioned data-driven approaches, the parameters of the classifiers were optimized by machine learning techniques. However, their performance was determined by the quality of the chosen features and finding appropriate features was a big challenge, which is typically performed by trial and error. This problem can be solved by using deep neural networks (DNNs).

In general, DNNs refer to artificial neural networks with several hidden layers [172]. Unlike the shallow (not deep) artificial neural networks used for seizure detection [173–177], the deep networks do not need any hand-designed feature extraction unit. Different types of DNNs exist, e.g. convolutional neural network (CNN) [178], deep belief network [179], stacked denoising auto-encoder [180], long short-term memory (LSTM) [181], tensor deep stacking network [182], etc. Among dozens of different DNNs, the convolutional neural network, CNN or ConvNet, has generated good results in image and speech processing applications [178, 183–187]. Recently, these networks have also been found useful in EEG analysis: Cecotti and Graser used a CNN embedded with a Fourier transform to classify steady-state visual evoked potential activities [188]; Furthermore, they developed some CNN methods in [189] for detecting P300 responses in a brain computer interface. Mirowski et al. applied a previously developed CNN for document recognition, called LeNet, on seizure prediction data. They

showed a significantly better prediction rate for the CNN method compared with an SVM-based and logistic regression approaches [190]. Recently, Acharya et al. proposed a new CNN method for automated seizure detection for adult patients. They designed a 13-layer network to process a single-channel EEG and achieved 95% sensitivity and 90% specificity [191].

The main goal of this chapter is introducing a seizure detection algorithm using CNN, specifically for neonatal seizures, which takes a segment of raw-EEG data and then labels it as seizure or non-seizure. Unlike the previously proposed methods in this field, this method needs neither pre-processing of the EEG data nor any hand-engineered feature extraction process. In other words, this method is automatically extracting the best required features and classifying each segment of raw data based on those features. Once the CNN is trained, it can be merged with other classifiers, like LSTM [192, 193], random forest (RF) [194], SVM [195, 196], etc., to improve its performance.

In this chapter, the proposed CNN is merged with an RF to detect neonatal seizures from 90s multi-channel EEG segments. This method is compared with our previously developed heuristic approach, as well as with three feature-based data-driven algorithms (Alg. 1-3). The main objective of this chapter is to introduce a CNN-based algorithm and compare it with hand-designed feature-based and heuristic methods with no complex pre/post-processing steps. Dozens of pre/post-processing algorithms exist for improving the performance of neonatal seizure detectors proposed in the literature. For instance, De Vos et al. used blind source separation techniques for removing artifacts as a pre-processor [7]. Temko et al. applied adaptive background modeling to adaptively change the latent variable with respect to the background activity as a post-processor [197]. A data-driven post-processor was proposed in chapter 6 to find evolutionary patterns of seizures to distinguish between the real seizures and 'polygraphic signals'-related artifacts (e.g. ECG artifacts). Furthermore, an adaptive learning technique to apply the experts' feedback to tune the latent variable was proposed in chapter 7. These algorithms can also be applied on the outputs of the methods considered in this chapter in order to improve their performance. However, this is outside the scope of this study.

The next sections of this chapter are organized as follows: the database and the used methods including a heuristic, 3 feature-based, and the proposed CNN methods are explained in section 8.2. The results of the methods are reported and compared in section 8.3. Discussion is given in section 8.4 and conclusions are drawn in section 8.5.

8.2 Materials and methods

8.2.1 Data description

In this chapter, DB1-Small, DB2-Small, and DB3 are used. The 10 neonates without seizures are excluded from the first two datasets. The detailed descriptions about these datasets are given in section 3.2. Briefly, the EEG-polygraphic recordings from 48 neonates, all with HIE, recorded in one medical centers, EMCR, are used in this study. The EEG signals are filtered between 1-20 Hz and bipolar channels are made according to the montage maps explained in section 3.2. In this study, DB1-small, in addition to 9 random neonates from DB2-Small and DB3, totaling 26 neonates, are used to train the algorithms, and the remaining 22 neonates to test. The next subsection describes that how these recordings are segmented and used for training, validating, and testing the algorithms.

8.2.2 Segmenting

In order to train the different classifiers, 4344 segments (50% seizure and 50% non-seizure) have been selected manually from the 26 neonates. Note that some segments can correspond to different parts of the same seizure or even can have overlap with each other. Each of these segments includes 90 seconds of EEG data from all available bipolar channels. In order to have a sufficient number of training data, the data was segmented using consecutively overlapping segments with overlap of 2, 4, and 6 seconds to the left and the right, obtaining a total of more than 30,000 segments to be used for training. The training data were split into training (75%) and validation sets (25%) to stop the back-propagation algorithm. For the test dataset, the whole recordings of the 22 remaining neonates were split to 90s segments with 60 seconds overlap. This length is considered long enough to extract the dynamics and evolutionary characteristics of brief-lasting seizures ($<$ 1 min), which are reported as the most difficult seizures for automatic detection [3, 6, 92, 149, 162]. None of the testing neonates or segments has been used in the training process. All data-driven methods, which will be explained in this section, used this training and test dataset. In contrast, the heuristic algorithm did not need training data. However, part of this training dataset was used by Deburchgraeve et al. to tune the parameters and thresholds. To guarantee full independence with the test set, the EEG data from neonates previously used for developing the heuristic method were excluded in any test performed with the classifiers considered in this chapter.

8.2.3 Heuristic method

In this method, we used a previously developed heuristic model that mimics a human EEG reader to compare with the proposed CNN method. This algorithm is described in appendix A.1.2 (Version 2).

8.2.4 Feature-based approaches

In order to compare the classic feature-based approaches to the proposed CNN-based method, which needs no predefined features, three feature-based algorithms are proposed (Alg. 1-3). All the algorithms used the same feature set including 28 features from the frequency domain, 23 from the time domain, and 4 from information theory (in total 55 features), which are listed in table 8.1. These features have been proposed originally in [39, 88, 145, 198] and used in different methods for neonatal seizure detection [3, 91, 92, 152, 170, 197]. More information about the computation and meaning of these features can be found in the references. The classifier used in these algorithms is a bagged random forest. However, the splitting of the EEG or the aggregating of the channels are different depending on the algorithms.

Alg. 1: Fifty five features from each channel (each 90 s) are calculated and concatenated to make a vector with 1100 (=55 features * 20 channels) features in total. Then, these features are fed into the classifier. The probabilistic output of the classifier is compared with a threshold to define the label. Since the number of channels should be constant to result in a fixed input size, a zero vector is used for the unavailable bipolar channels.

Alg. 2: In this algorithm, the features are extracted and classified for each channel separately. Therefore, the input of the classifier is 55 features extracted from each channel separately. Then, the probabilistic output of the classifier for each channel is compared with a threshold. If at least the output of one channel is greater than the threshold, the whole segment is considered as seizure ('OR' operator on channels).

Alg. 3: In this algorithm, first, the EEG data of each channel was split into 8 s epochs with 50% overlap. Second, 55 features of the epochs are extracted and the classifier is applied. Third, the probabilistic outputs of all epochs in each channel are smoothed with a moving average filter (N=15) and compared with a threshold. If at least one epoch of a channel contains a seizure, the whole segment is considered as seizure ('OR' operator on epochs and channels). The general idea of this method was derived from the method proposed by Temko et al. in [3]. However, there are two differences between them: 1) instead of using

Table 8.1: Extracted features used in Alg. 1 - 3

Type	Features
Frequency Domain	• Wavelet energy • Total power • Peak frequency • Spectral edge frequency (80%, 90%, 95%) • Power (0 - 2 Hz, 1 - 3 Hz, ..., 10 - 12 Hz) • Normalized power (0 - 2 Hz, 1 - 3 Hz, ..., 10 - 12 Hz)
Time Domain	• Curve length • Root mean square amplitude • Hjorth parameters • Zero crossing (raw, Δ, $\Delta\Delta$) • Variance (Δ, $\Delta\Delta$) • Skewness • Kurtosis • Nonlinear energy • Number of maxima and minima • Autoregressive modeling error (n: 1 - 9)
Information Theory	• Shannon entropy • Singular value decomposition entropy • Spectral entropy • Fisher information

an SVM classifier, an RF is used in order to have a fair comparison with the proposed method which uses an RF. For our dataset, the RF method results in a higher performance than the SVM classifier as is reported in the result section. 2) The collar method used by Temko et al. for correcting the onset and offset of detections is not used here. As mentioned before, the previously proposed pre/post-processing methods (e.g. collar) can be similarly applied on the CNN or feature-based approaches to improve the performance.

In Alg. 1-3, the mentioned threshold is varied from 0 to 1 in order to construct the Receiver Operating Characteristic (ROC) curve. Since in Alg. 2 and 3 the classifier is respectively applied on channels and epochs of a segment separately, in order to improve the training, the exact moment and the channels representing seizures were pre-marked in each segment of the training dataset by the method developer.

8.2.5 Proposed CNN-RF

We propose the use of a CNN for the automatic detection of epileptic seizures. As mentioned before, the main advantage of the proposed method is that there is no need to select any features manually. In other words, the classifier takes the raw multi-channel EEG data and automatically optimizes the features and classifier at the same time. In order to improve the classification performance, when the CNN was trained, the classifying end layers were removed and replaced by an RF. In this method, an RF classifier was selected since it performs better than other classifiers. In order to test this, a bootstrap test was applied on the training data. Firstly, the training data was split into 75% training and 25% validating subsets. Then, four classifiers including LDA, two SVMs (with

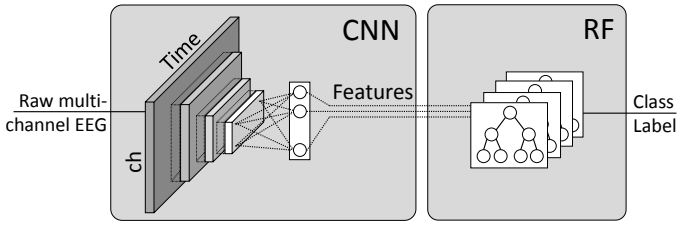


Figure 8.1: Schematic overview of the proposed CNN-RF method.

linear and RBF kernels), and an RF were trained. A 10-fold cross-validation and grid search was used to optimize the hyper-parameters of the SVM-RBF and RF methods. Next, the AUC of each classifier on the validation set was calculated. The splitting has been repeated 100 times. As presented in the result section, the RF has a significantly better performance than others. As a result, in the final model, the CNN is considered as an automatic feature extractor and the RF is the classifier (figure 8.1). The following subsections describe this approach in detail.

Structure of the proposed CNN

A generic overview of the CNN method is given in section 5.2. In this method, the input of the proposed CNN is firstly filtered between 0.5 and 15 Hz. In order to decrease the complexity of the CNN network, the EEG is down-sampled to 30 Hz. As a result, each segment of EEG which is 20 channels by 90 seconds is converted to an image with size of 20 * 2700 where the 1st and 2nd modes correspond to channel and time. For the neonates having a fewer number of channels, zero vectors are used to make a homogeneous input image with the same size. These images are the inputs of the proposed CNN.

The structure of CNN is formed by 23 layers which are listed in table 8.2. In this table, the dimensions of the filters, as well as the number of Conv filters, stride of pooling layers, and the output size of each layer are shown. The first 15 layers are composed of 5 blocks of (Conv + max pool + ReLU) in order to extract the features related to seizure patterns. The beginning blocks extract very local abstractions, like the slope of lines, whereas the deeper blocks extract more global ones, like the spike and oscillation patterns. In layer 16 and 17, a maximum and an averaging pooling layer aggregate all time samples. It means that each output of the 17th layer represents the abstraction of the whole 90 s of the corresponding channel. Due to the fact that in our database, the recording

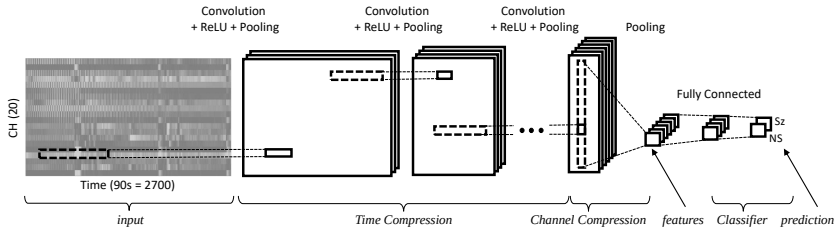


Figure 8.2: Schematic structure of the proposed CNN method.

of different neonates included a different number of electrodes and subsequently different bipolar channels all Conv and pooling layers in these first 17 layers are (1 * N) operators. It means that they filter and aggregate only the time information, and has no effect on channel (spatial) information. In other words, in these layers, the process is performed on each channel separately. Next, in layer 18, a max pooling is performed on all 20 channels in order to aggregate the features of different channels. The used maximum operator, which can be considered as an 'OR' operator in fuzzy logic, for the channel aggregation is supported by the fact that in a clinical neurophysiologists' point of view, if one channel of EEG represents a seizure pattern, the whole segment is marked as seizure. As a result, the output of the 18th layer are 20 numbers (features) representing the characteristics of all channels and the whole 90 seconds. Then, the remaining 5 fully connected layers including 2 hidden Conv layers, 2 sigmoid nonlinear units, and finally a loss function for computing the classification error are performing the classification task. This structure, as well as the hyper-parameters were chosen by trial and error. When the network is trained, these end 5 layers are removed and replaced with an RF. figure 8.2 schematically shows the designed layers, features, and the fully connected classifier.

The CNN implementation was performed by the MATLAB toolbox MatConvNet [199]. For training the CNN, the weights of the Conv layers were initiated with normally distributed random numbers generated by $N(0, \sigma^2)$, where the standard deviation, σ, equals 0.2 and 0.1 for layers 1-18 and 19-23 respectively. All bias weights of the Conv layers were initiated by zero. The learning rates were varying from 0.3 to 0.003 with respect to the layer and epoch numbers. The learning batch size was 20 segments.

Pruning and tuning

As explained previously, a ReLU replaces negative outputs of a Conv layer with zero. Thus, if the output of a ReLU is always zero, for all seizure and non-seizure segments of the training dataset, it means that the corresponding filter of the Conv layer located before that ReLU is always producing negative outputs and, therefore, has no influence on the final output of the network. In order to rank the effectiveness of the filters and prune them with respect to the aforementioned fact in the presence of outliers, the 99% percentile, p_{99}, of the outputs of each filter is calculated through the training dataset. If the p_{99} is negative, that filter, as well as all corresponding parameters of the Conv layer of the next layer, is removed. For positive but small values of p_{99}, the procedure is continued till the validation error increases. The total number of parameters of the CNN reduced by 58% after pruning (from 7600 to 3300), which increases the generalization power of the network as well as the training and recall speed. When the selected filters and parameters are removed, the network was re-tuned by the training data for a few extra epochs.

Using random forest

The generic structure of random forests are given in section 5.3. In the current CNN-RF method, an ensemble bagged trees classifier with 100 weak learners has been applied and is addressed as RF. When the CNN is trained, the last five classifying layers were replaced with an RF classifier. Therefore, the first 18 layers of the CNN act as an automatic feature extractor and the RF is a classifier.

In the proposed method, Max_{Depth} and Min_{Leaf} were respectively equal to (Q-1) and 1, which means that if the 3rd stopping condition is not reached, the tree can be as deep as possible, having one leaf for each bootstrapped training segment. As it is mentioned, 100 of these decision trees are trained over bootstrapped training segments with randomly selected features. For each test segment, the obtained seizure probability (p_s) is compared with a threshold to score the segment as seizure or non-seizure. This threshold is varied from 0 to 1 in order to construct the ROC curve.

8.3 Results

Figure 8.3 shows box-plots of all extracted features by the CNN method. The features which were removed by the pruning process are marked by a star (*)

Table 8.2: Layers of the designed network before pruning

		Layer Info	Output Size		
		Input:	(20,	2700,	1)
Feature Extraction	1	Conv (1, 5) × 5	(20,	2696,	5)
	2	MPool (1, 3), s: 2	(20,	1347,	5)
	3	ReLU	(20,	1347,	5)
	4	Conv (1, 5) × 8	(20,	1343,	8)
	5	MPool (1, 3), s: 2	(20,	671,	8)
	6	ReLU	(20,	671,	8)
	7	Conv (1, 5) × 10	(20,	667,	10)
	8	MPool (1, 3), s: 2	(20,	333,	10)
	9	ReLU	(20,	333,	10)
	10	Conv (1, 5) × 15	(20,	329,	15)
	11	MPool (1, 3), s: 2	(20,	164,	15)
	12	ReLU	(20,	164,	15)
	13	Conv (1, 20) × 20	(20,	145,	20)
	14	MPool (1, 10), s: 5	(20,	28,	20)
	15	ReLU	(20,	28,	20)
	16	MPool (1, 5), s: 3	(20,	8,	20)
	17	APool (1, 8), s: 1	(20,	1,	20)
	18	MPool (20, 1), s: 1	(1,	1,	20)
Classifier	19	Conv (1, 1) × 5	(1,	1,	5)
	20	Sigmoid	(1,	1,	5)
	21	Conv (1, 1) × 2	(1,	1,	2)
	22	Sigmoid	(1,	1,	2)
	23	Loss	(1,	1,	1)
Total Number of Parameters:			7600		

Conv: convolutional layer, the information is given in the following format (dimension in channel, number of coefficients in time) × number of filters
MPool: Pooling by Maximum operator, the information is given in the following format (dimensions in channel, number of coefficients in time) s: stride
APool: Pooling by Average operator
ReLU: Rectified linear unit
Loss: loss function

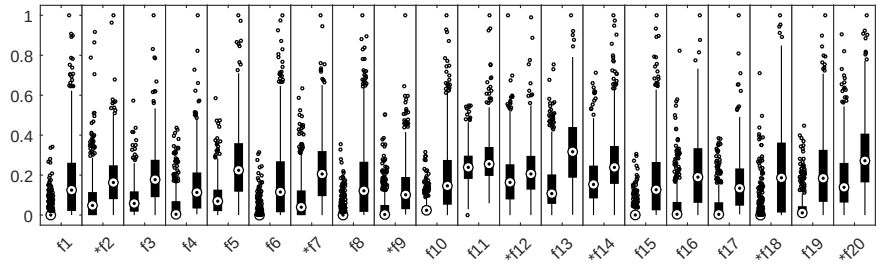


Figure 8.3: Summary of the automatically extracted features (f1-20) by the CNN from the test dataset. For each feature, the first and second boxes are corresponding to the non-seizure and seizure segments. The features starting with star (*) correspond to the features removed after the pruning process. All features are plotted after being normalized between 0 and 1.

character. For each feature, the first and second boxes are corresponding to the non-seizure and seizure segments respectively. For each feature, the first and second boxes are corresponding to the non-seizure and seizure segments respectively. In each plot, the filled black boxes show the 1st and 3rd quartiles ($Q_{1,3}$) and the thin lines display the Whisker range from $(Q_1 - 1.5 \times IQR)$ to $(Q_3 + 1.5 \times IQR)$, where IQR is $(Q_3 - Q_1)$. The small circles in the plots show the outliers and big circles with a dot in the center show the median values. All features are plotted after being normalized between 0 and 1.

As mentioned, a bootstrap test was applied to compare different classifiers including LDA, SVM (linear and RBF kernels), and RF. Next, the AUC of each classifier was calculated when the features extracted by the CNN and Alg. 3 are used (figure 8.4). The Mann–Whitney–Wilcoxon statistical test applied on the results from RF versus SVM-RBF (* and ** in the figure) shows that the RF classifier has a significantly higher performance.

Figure 8.5 shows the ROC curves on the test dataset for the heuristic, feature-based algorithms (Alg. 1-3), and the proposed CNN after pruning and connecting to the RF. The left-side curve, A, is the result of applying the proposed classifier to all test neonates (22 neonates), while the right-side one, B, shows the ROC when 7 neonates who expressed a completely different seizure pattern having no training patterns in the training dataset were excluded from the training set. Results from the complete CNN, without pruning, are similar to the displayed CNN. The AUC of the CNN-RF method is also 8% higher than the pure CNN with the fully connected network (83% vs. 75%).

Furthermore, in table 8.3, the epoch-based sensitivity, specificity, and kappa, as

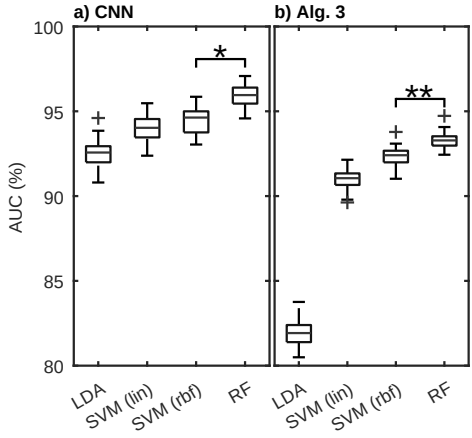


Figure 8.4: The AUCs of different classifiers when the features are extracted by the CNN (a) and Alg. 3 (b).

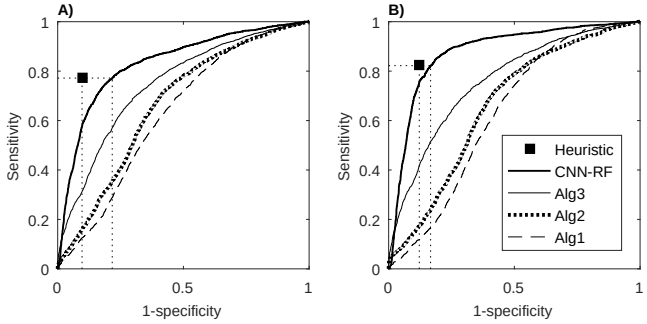


Figure 8.5: The ROC curves of the heuristic, feature-based, and proposed CNN-RF methods for the test data; (A) for all neonates in the test dataset and (B) after excluding 7 neonates which did not have appropriate training patterns in the training dataset.

Table 8.3: Comparison of the performance metrics for the CNN and Heuristic methods

Metric	Total Database		After Exclusion of 7 Neonates	
	Heuristic	**CNN**	**Heuristic**	**CNN**
Sensitivity %	77	77	82	82
Specificity %	90	78	88	84
Kappa	0.68	0.53	0.60	0.66
GDR %	77	77	78	78
FAR (h^{-1})	0.63	0.90	0.77	0.73

well as event-based good detection rate (GDR) and false alarm rate per hour (FAR) are reported. These performance metrics are thoroughly explained in chapter 10.3.1. In this table, in order to make the comparison simpler, the threshold of the CNN-RF was chosen where the sensitivity of CNN-RF and heuristic methods are equal (the horizontal dashed-lines in figure 8.5). As is clear from the table, after excluding the 7 neonates, the specificity of CNN-RF is 5% less while the averaged false alarm rate per hour is 0.04 better than those of the heuristic method.

Figures 8.6 and 8.7 show two qualitative examples of a seizure and a non-seizure segment respectively. In these figures, the outputs of the 7th and 8th Conv layers, as well as the outputs of the 17th and 18th pooling layers are shown. The red-highlighted images are corresponding to the filters which were removed by the pruning process. Each image of layers 7, 13, and 17 displays an output which has 20 channels (y-axis), whereas the output of the layer 18, after pooling the channels, has only one value for all channels. Furthermore, as explained, the resolution of time (samples in x-axis in these figures) decreases after each pooling layer so that layers 17 and 18 have only one value in time. Therefore, the output of layer 18 includes 20 values (20 before the pruning, and 13 after pruning) each of which is a 1×1 (channel $\times$ time) number. These values are considered as the automatically extracted features. As is clear in the seizure example, figure 8.6, the seizure occurred in the right (almost bottom) side of the segment. The outputs of the Layer #7 show some variations in this area which is different from the left side of the segment. These differences are more pronounced in the outputs of the layer #13 where almost all filters were activated for the seizure area. For the layer #17, when the time information is completely compressed, the channels with maximum activation of some of the filters are clearly distinguishable. In contrast, for the non-seizure segment displayed in figure 8.7, there is no clear activation in the layers #7 and #13, and consequently no distinct channels in layer #17.

Figure 8.8 shows the recall computational time (in test phase) calculated for

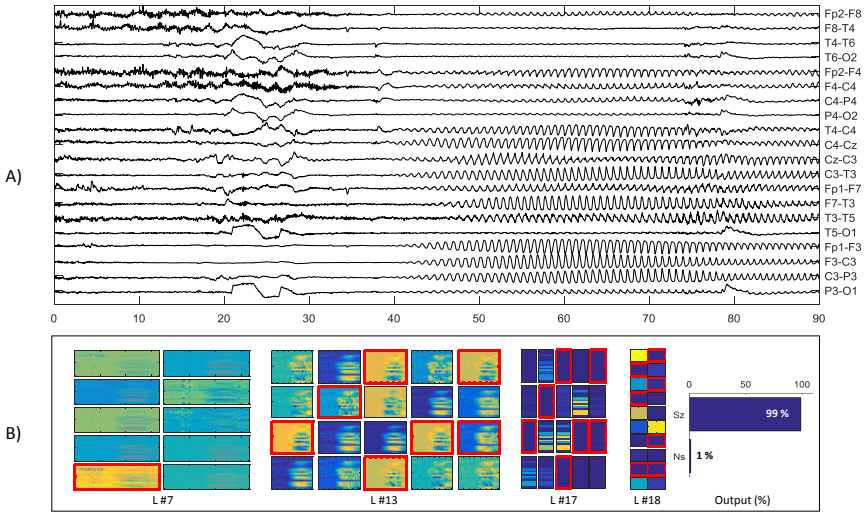


Figure 8.6: Qualitative example of a seizure segment and outputs for some layers. A) A seizure segment with 20 bipolar channels. The x-axis is time in seconds. B) The output of Conv layers 7 and 13, as well as Pooling layers 17 and 18, and the final output of the CNN after the classification layers. The red-highlighted boxes correspond to the filters removed by the pruning process.

the heuristic, feature-based (Alg. 1-3), and proposed CNN-RF methods. The time is shown in seconds per each segment in a logarithmic scale. In order to have a correct comparison and overcome variable CPU loading, the methods ran chronologically for each segment and the elapsed time for each method and each segment was stored. Then, the median, Q_1 and Q_3 of the whole segments were calculated and shown in a box-plot. The time was measured from the moment of loading data to when the label was defined including pre-processing, feature extraction, and classification time. The algorithms ran in Matlab^{TM} platform, v. 9.1.0 (2016b) (The MathWorks, Natick, MA, USA), and on a server computer, Intel(R) Xeon(R) CPU 2.20 GHz, with GNU/Linux operating system (Red Hat 4.8).

8.4 Discussion

In neonatal seizure detection using machine learning approaches, choosing a proper type of classifier plays an important role to have a good performance.

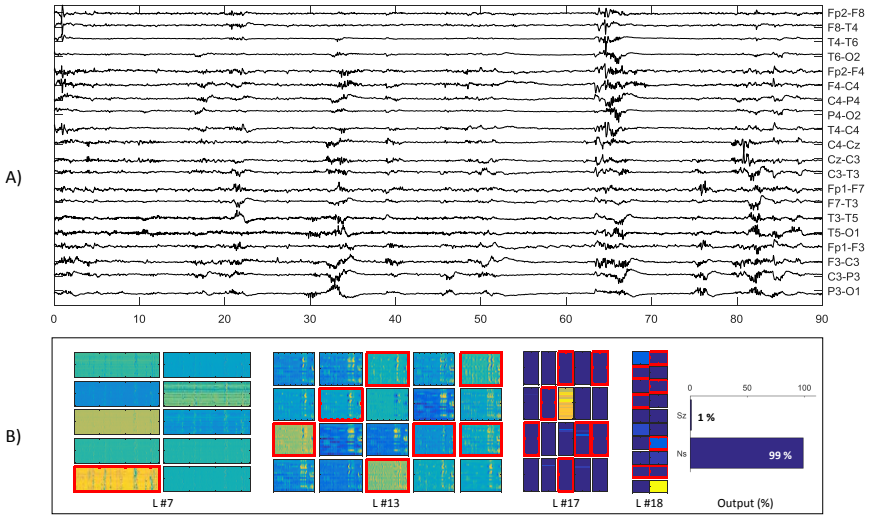


Figure 8.7: Qualitative example of a non-seizure segment and outputs for some layers. A) A non-seizure segment with 20 bipolar channels. The x-axis is time in seconds. B) The output of Conv layers 7 and 13, as well as Pooling layers 17 and 18, and the final output of the CNN after the classification layers. The red-highlighted boxes correspond to the filters removed by the pruning process.

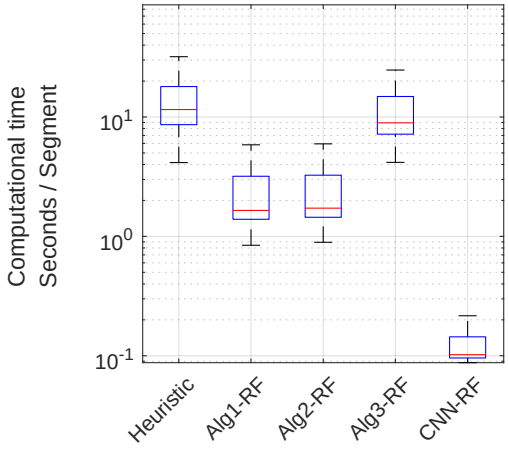


Figure 8.8: The execution time is shorter for the CNN-RF when compared to the heuristic and feature-based methods. The time is measured in seconds for each segment (90 s, 20 channels).

Moreover, finding appropriate features which are discriminative and informative is another challenge and has a big influence on the performance. In neonatal seizure detection, like other classification problems, some researchers have discovered and proposed new features to enhance the performance, while others improved the classification strategies. By using a deep convolutional neural network, both the features as well as the classifiers, are optimized simultaneously. In this chapter, a CNN with 18 layers was designed in order to automatically extract the required features from raw multi-channel EEG segment and a random forest was used to classify them. It is important to note that the final layers of the CNN method were also able to classify the segment based on the features extracted in the previous layers. However, they were replaced with a random forest, after training the network, in order to have a higher performance and a smoother ROC curve.

In the classic feature-based seizure detectors, like Alg. 1-3, the features are hand-engineered and usually found by trial and error. It is possible that some information which is important for classification is fully or partly missed in the selected features. However, in the proposed method, since the feature extraction is optimized based on the training data, the maximum information, related to the classification, is potentially able to be extracted which improves the performance of the classification (compare CNN-RF and Alg. 1-3 in figure 8.5). As shown in figure 8.3, the features extracted by the CNN are discriminative to separate seizure from non-seizure segments. This resulted from the optimization of feature learning process in the CNN, which is considered as the main advantage of the proposed method.

On average, the proposed CNN-based method performs better than the tested methods relying on features. However, as plotted in figure 8.5.A, the proposed method has lower performance than the heuristic method due to the limited number of training neonatal EEG data sets. If a larger scored database was available for training, this could improve the performance of the method in a patient-independent framework. When lacking training data, it seems likely that heuristic methods, which are not directly trained based on the training data, will perform better. This is because of the fact that for developing the heuristic method mimicking an expert human EEG scorer, the knowledge of the neurophysiologist which has been collected over years from hundreds of neonates was used. It seems likely that to have a fair comparison between the heuristic method and data-driven techniques, adequate training data should be provided for the data-driven methods. If such a large training database is collected in future years, better methods could be trained. These methods have some important advantages, like re-train-ability, more accuracy, and more flexibility. In the test dataset, 7 neonates with moderate to severe hypoxic brain injury were seen to have very unique seizure patterns. Similar patterns

were not available in the training dataset. It is unlikely that a data-driven classifier will detect a pattern which was not presented in the training data. As it is shown in table 8.3 and figure 8.5.B, if these 7 neonates are excluded from the test dataset, the performance of the proposed method is similar to the heuristic method and better than other data-driven methods. It is evident that by using a larger training dataset in the future, the performance of the trainable classifiers will be improved, while the heuristic 'un-trainable' methods will be unchanged. When the latent variable of the CNN method was fixed to provide the performance reported in table 8.3, 7% and 10% of seizure segments were detected only by the CNN-RF method and only by the heuristic method respectively. This shows that even with the current lack of training data, the proposed CNN-RF method is a complementary method which can detect 7% of the seizures which were missed by the heuristic approach.

Since neonatal seizures are usually very focal or regional, which means that only one or a few channels display seizure patterns, retrospectively developed algorithms were designed as a single-channel detector so that they are applied on each EEG channel separately. If a seizure is detected in at least one channel, the segment is classified as seizure ('OR'). This idea is also applied in the proposed CNN method. As was explained, all filters in the Conv layers, as well as pooling layers, developed for extracting features, are $1 \times N$ operators. It means that they only affect the time mode and have no operation on the channel mode. However, it is likely that there is some spatial information, at least among adjacent electrodes, which show how seizures spread through channels in regional seizures, involving contiguous brain regions. Furthermore, each electrode is usually used more than once in different montage maps. For instance, in the full montage map, C4 is used 4 times in F4-C4, C4-P4, T4-C4, and C4-Cz. Hence, if a seizure occurs in a brain area close to the electrode C4, these 4 bipolar channels should display it. This spatial connection of channels can be very useful even for very focal seizures, and it can be a distinctive characteristic for distinguishing seizures from some artifacts. The CNN structure is easily able to extract this information by increasing the dimension of filters in the channel mode, as it is working in diverse image processing applications. However, since some electrodes were not recorded or available in our training database for some neonates and different bipolar montage maps were subsequently constructed, this information was not extractable. In case of datasets with homogeneous recordings, the filters can become 2-dimensional and it is expected that the performance will be significantly improved. This is also true for Alg. 1 which concatenates the features of different channels.

As previously explained, in Alg. 2, the method is applied on each channel individually and an OR operator is then used to aggregate the channels. Since most neonatal seizures are regional/focal, the channel(s) representing seizure

patterns should be predefined by the developer or an Expert EEG reader for each training segment in order to train the method with correct training data. This is a very labor-intensive task especially in big datasets. This problem is exacerbated for Alg. 3 where not only the true channels should be predefined, but also the proper epochs of those channels should be marked. Nevertheless, one of the main characteristics of the CNN is a shift-invariant property which means shifting the target pattern (like seizure pattern in this problem) through the 1st or 2nd modes (time or channels here) has no effect on the output. This characteristic results from the parameter sharing of the neurons in the convolutional layers, as well as from the maximum operation in the pooling layers. Therefore, for the proposed CNN method, it is not important in which channel or at what time the seizure activity is manifested. The CNN can automatically find the related region, so-called region-of-interest. To illustrate this, in the example of figure 8.6, where the seizure emerges in the almost bottom-right of the segment, the output of Layer 13 shows that the CNN neurons were successfully activated in the seizure area, compared to figure 8.7 for the non-seizure segment.

Furthermore, in neonates, most seizures are rhythmic, with evolution of amplitude. For these seizures, the exact moment of onset or offset of seizures are sometimes not very clear due to the fact that the seizure patterns (oscillatory or spike-train) start with a very low amplitude and gradually increase over time (and vice versa for the offset). It is evident that even expert EEG readers may not agree with each other about the exact time and duration of seizures in this case [69]. Furthermore, in focal/regional seizures, the channels which are not close to the center of the seizure might show some low amplitude seizure patterns, increasing the uncertainty for the classifier. Due to the mentioned shift-invariant property, the CNN method can overcome this inherent fuzzy onset and offset of seizures, as well as representing channels, so that it does not need to know when and where exactly the location of the seizure is within the segment.

In addition, retrospective studies of neonatal seizure detection have reported varying level of agreement between expert EEG readers, with kappa coefficients ranging between 0.4 and 0.93 [67,69,96–99]. It shows that different experts, and consequently different centers, use empirically different definitions of seizure and the gold standard is not yet clear-cut. One advantage of data-driven methods, like the proposed one, compared to fixed heuristic ones, which may work very well in one center, is that the network can be retrained or re-tuned in different centers in order to tailor to their needs. Moreover, an advantage of artificial neural networks, like the proposed CNN, is that different training segments can have different weights in the backpropagation training. Therefore, the segments upon which the experts have higher agreement can have more influence in the

cost function during training than segments with a larger uncertainty. This weighted training technique can increase the overall satisfaction of the experts from the final outputs when the method is being used in different centers. Although the labels of only one expert reader were used for this proposed method, a multi-score analysis can be performed in the future.

Finally, one of the most important advantages of the proposed CNN-RF method is that it is made by simple FIR filters, maximum, and averaging operators. Consequently, the recall time is much faster than other tested heuristic and feature-based methods, see figure 8.8. This method is about 115, 89, and 17 times faster than the heuristic, Alg. 3 and Alg. (1, 2) respectively in recall. The computational time is very important for the real-time implementation. Although each of them can individually work in real-time and needs less than 90 seconds to process a 90s segment, the real-time ratio (= average computational time needed for a segment / 90 s) can be important when the final product acquires, filters, down-samples, stores, monitors the EEG, sends the data to a cloud system, and performs many other possible tasks and processes. Faster processing often results from less operational calculations which means lower energy requirements and it is very important in portable/wearable devices with limited source of energy.

However, the proposed approach has some disadvantages: first, this method, and in general all deep networks, are very time-consuming in the training process. Thus, optimizing the hyper-parameters or improving the design of the network are much harder than in a simple feature-based data-driven technique. Second, the suggested network, like other DNNs, needs a large amount of data to be trained. If sufficient training data are not available, it is very likely to over-fit due to the high number of layers and parameters. Third, compared to heuristic methods, the process is not transparent and it is not immediately evident why a certain segment is classified as seizure or non-seizure. Finally, compared to regular feature-based methods, the extracted features are just some numbers resulting from filtering, pooling, and rectifying of the EEG, which make it difficult to provide a tangible interpretation.

Several limitations of this study need to be acknowledged. First, the designed network and its structural parameters, including the number of layers, the length of filters, the strides, etc., have been chosen by trial and error and consequently they are not guaranteed to be optimal. Second, the scored seizures used for training and testing were labeled by only one expert clinical neurophysiologist. Third, the data used in this study were recorded in one center only. In order to have a more generalizable comparison, the methods should be tested on an extensive and multi-rated multi-center database.

8.5 Conclusion

A novel neonatal seizure detector using convolutional neural networks and random forest was introduced in this chapter. The main advantage of the proposed method is that it does not require hand-engineered feature extraction process, but it automatically extracts the required features and optimizes them based on the training data. We show that this proposed method outperforms the tested feature-based approaches. Compared to the previously developed heuristic detector, the proposed method is not yet superior because of the limited number of training neonates. However, it seems possible that by having more training data in future, it can reach the performance of the heuristic method as well. At last, it was also shown that the proposed method is remarkably faster than other tested algorithms, which is very important for real-time applications. However, further studies need to be carried out in order to validate this algorithm in a multi-center and multi-scored database. Furthermore, it seems that using sequence learners like hidden Markov models or LSTM instead of the used RF classifier can enhance the performance.

Chapter 9

Sleep stage classification using convolutional neural networks

In this chapter, we design and implement a novel deep convolutional neural network to detect quiet sleep in premature infants. As mentioned in the previous chapter, this type of artificial feed-forward neural networks is able to learn the required features automatically and does not need any hand-crafted features. The structure and properties of this network are described. Then, the performance of the proposed method is compared with two state-of-the-art algorithms. Finally, the advantages and disadvantages are discussed. The modified version of this study has been submitted for publication. This study is a joint work with Ir. Ofelie De Wel. My contribution relies on designing and implementing the architecture of the convolutional neural network proposed in this study.

9.1 Background

Preterm infants are born during a critical period of rapid growth and development of the brain and central nervous system. In these newborns, the fast neural growth that normally occurs in the womb during the third trimester of gestation is interrupted and has to take place in the neonatal intensive care unit (NICU). Therefore, a primary concern of NICUs is to support optimal brain development of these vulnerable infants, who are at increased risk of neurodevelopmental disorders [200, 201]. Continuous electroencephalography (EEG) monitoring is considered a valuable noninvasive tool to assess and track brain maturation of newborns in the NICU. Sleep wake cycling (SWC) undergoes fast changes and is one of the main hallmarks of neurodevelopment in preterm infants. Since sleep state organization reflects the level of functional brain development, a first step towards automated analysis of cerebral function is to identify the different sleep stages [202, 203].
Moreover, evidence suggests that sleep is of utmost importance in the development and maturation of neural pathways. As a consequence, the neonate's sleep should be protected and promoted during their stay in the NICU [204, 205]. Automated algorithms for real-time neonatal sleep state identification could optimise the scheduling of neonatal interventions and care procedures in the NICU. Up to now, the gold standard for the assessment of sleep EEG is visual analysis by an expert clinician. However, visual interpretation of neonatal EEG is a tedious and time consuming task for the clinician and requires expertise that might not always be available in the NICU [203]. This can be overcome by developing automated algorithms to discriminate neonatal vigilance states.

A growing body of literature has investigated algorithms for neonatal EEG sleep scoring, which typically rely on the difference in continuity [203, 206] and frequency content [59] between neonatal sleep states. Palmu et al. [207] compared the proportional duration of bursts, called Spontaneous Activity Transients (SAT%), in the EEG between neonatal sleep states. They confirmed the existence of genuine vigilance states in early preterm infants and found that the fluctuation of the SAT% signal corresponds to sleep stage cycling. Stevenson et al. [208] continued this work and derived quantitative measures of brain activity cycling from the frequency domain representation of the SAT% time series. Besides analysis of the SAT% signal, EEG time profiles have been used for automated sleep stage detection as well [56, 209, 210]. The EEG time profiles are constructed by adaptive segmentation of the EEG and subsequent clustering of features extracted from each segment. Thresholding of the processed time profile, which shows the class membership of the EEG segments over time, will then lead to labels of the sleep states. This concept of EEG time profiles

for sleep stage modelling was explored by Barlow et al. [211], extended by Krajca et al. [209, 210] and further improved by Dereymaeker et al. [56]. In addition to the approaches described above, a number of studies have developed a neonatal sleep stage classifier based on a set of discriminative EEG features in combination with a machine learning classifier [57, 206, 212, 213].

While a wide range of features have been used to develop neonatal sleep stage classifiers, there is still no consensus on the optimal combination of features. Moreover, those hand-crafted features are based on prior human knowledge about EEG sleep patterns, hence only a limited set of features have been examined. In general, the process of selecting the optimal feature set and classifier is a challenging task.

This chapter will address these problems by adopting a deep convolutional neural network (CNN) to detect quiet sleep in preterm infants. Deep neural networks learn the features directly from the data, which eliminates the need for manual feature extraction. CNNs have been extensively used in the field of image and speech processing [178, 183–187]. Only recently, deep neural networks have been attracting a lot of interest for biomedical applications (e.g. EEG seizure detection [191], medical image analysis [214], detection of myocardial infarction [215], etc.). CNNs have been applied for sleep stage scoring in adults [216–218], however to the best of our knowledge this is the first study adopting CNNs for neonatal sleep stage classification.

The aim of this chapter is to design and train a deep convolutional neural network to identify sleep stages in preterm infants. The performance of this novel classifier will be compared to two reference algorithms described in the literature: (1) a support vector machine (SVM) classifier using a set of spectral features [59] and (2) Cluster-based Adaptive Sleep Staging (CLASS) developed by Dereymaeker et al. [56].

The remaining part of the chapter proceeds as follows. We start by describing the database that has been used to develop, train and test the classifier. We will then go on to a thorough description of the architecture of the network developed for sleep scoring. The pipeline of the two reference algorithms described in the literature will be briefly explained as well. Next, the performance of the proposed CNN and two reference algorithms will be reported and compared. At last, the advantages and disadvantages of the proposed method and future directions will be discussed.

9.2 Materials and methods

9.2.1 Data description

For this chapter the database recorded from preterm infants are used. The generic explanation about this database are given is section 3.3. This database, which recorded from 26 preterm infants, were split into a training set and a test set. The training set containing 54 recordings from 13 patients (about 32000 EEG segments) was used to develop and train the convolutional neural network. The total duration of training data is equal to 269 hours, with 69 h of quiet sleep, and 200 h of non-quiet sleep. The test set consisting of 43 recordings from 13 independent patients (about 27000 EEG segments) was used to assess the classification performance. The test set consists of 223 h of EEG with 53 h during quiet sleep and 170 h during non-quiet sleep. This inter-patient data splitting avoids that the model is tested on data from a patient that is also used to train the model. In this way, we expect that patient specific characteristics can not bias the classification performance.

9.2.2 Proposed method

Pre-processing

Prior to using the EEG data as input for the CNN, some pre-processing steps are performed. First of all, the EEG is band-pass filtered between 1 and 15 Hz. The filtered EEG is then down-sampled to 30 Hz in order to reduce the complexity of the network. The training data is normalized so that the mean and standard deviation of each EEG channel across the whole data set is equal to zero and one respectively. The parameters from the training data are then used to normalize the test data. At last, the EEG is segmented into windows of length 30 s, resulting in EEG segments of size 8 (channels) $\times$ 900 (samples).

CNN method

These pre-processed EEG segments are then fed into a 18-layer CNN, whose architecture is shown in figure 9.1 and table 9.1. The CNN is designed and implemented in Matlab using the MatConvNet toolbox [199]. The first 14 layers perform feature extraction and the last 4 layers including the fully connected layers perform classification. The mask (kernel) size, stride, and the number of feature maps used in the conv and/or pooling layers are also mentioned in the table.

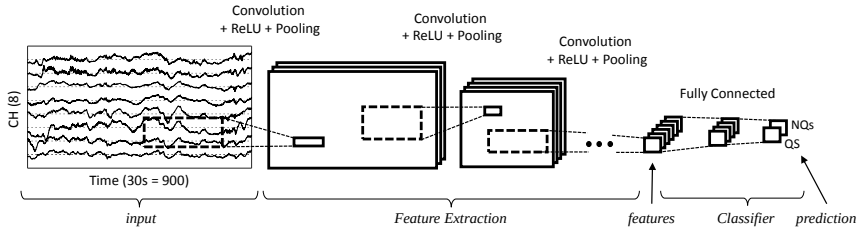


Figure 9.1: Architecture of the convolutional neural network.

Table 9.1: Layers of the designed network before pruning

		Type	Mask Size ∼ Stride
Feature Extraction	1	Conv	$(1, 10) \times 3$
	2	ReLU	-
	3	Conv	$(3, 1) \times 3$
	4	ReLU	-
	5	Mpool	$(2, 3) \sim (2, 2)$
	6	Conv	$(1, 5) \times 5$
	7	ReLU	-
	8	Conv	$(3, 1) \times 5$
	9	ReLU	-
	10	Mpool	$(1, 5) \sim (1, 3)$
	11	Conv	$(1, 5) \times 7$
	12	Mpool	$(1, 6) \sim (1, 4)$
	13	ReLU	-
	14	Conv	$(1, 37) \times 10$
FC	15	Sigmoid	-
	16	Conv	$(1, 1) \times 2$
	17	Sigmoid	-
	18	ReLU	-
Total Number of Parameters:			3000

Conv: Convolutional layer
MPool: Pooling by Maximum operator
ReLU: Rectified linear unit
Loss: Loss function
FC: Fully connected classifier

For training the CNN using the backpropagation algorithm, the learning rate changes from 10^{-2} to 10^{-4} dependent on the number of training epochs and the number of layers. The back-propagation batch size was set equal to 20 EEG segments and the maximum number of training epochs to 1000. However, after training, epoch number 602 was selected based on the validation error.

Post-processing

According to the literature, a sleep stage should last at least three consecutive minutes or three out of four consecutive minutes [219]. Relying on this fact, a moving average filter with length 6 (6 segments of 30 s = 3 min) is used as a post-processing step.

The proposed method will be compared to two methods from the literature, which will be briefly explained in the next subsections.

9.2.3 Feature-based neonatal sleep stage classifier

This neonatal sleep stage classifier is based on a set of spectral features described by Piryatinskaa et al. [59] which are fed into a support vector machine (SVM) classifier.

After band-pass filtering between 1 and 20 Hz and segmenting the multi-channel EEG in epochs of 30 s, 9 spectral features are extracted from each EEG segment: (1-4) relative power in 4 EEG frequency bands (δ: 0.5 - 4 Hz, θ: 4 - 8 Hz, α: 8 - 12 Hz, β: 12 - 15 Hz), (5-6) spectral edge frequency (75% and 90%), (7) spectral moment, (8) spectral entropy and (9) amplitude entropy. Each multi-channel EEG segment is then characterized by a total of 72 (9×8 channels) features and is marked as quiet sleep or non-quiet sleep according to the clinical label. The hyper-parameters of the SVM with radial basis function (RBF) kernel are set using 5-fold cross validation on the training data. This SVM classifier is then trained using the training set and the performance is measured on the test set.

As in the proposed CNN, a moving average filter with length 6 is used to remove the transient discontinuities.

9.2.4 Cluster-based Adaptive Sleep Staging (CLASS)

This algorithm is based on the relatively higher discontinuity in quiet sleep compared to non-quiet sleep and has been developed by Dereymaeker et al. [203]. Briefly, the pipeline of the algorithm is as follows. First, a pre-processing step consisting of a band-pass filter from 1 to 40 Hz and a notch filter at 50 Hz is performed. Artifact subspace reconstruction is then adopted to reject remaining artifacts, which could otherwise be wrongly detected as EEG discontinuities. After artifact removal, the cleaned EEG is down-sampled with a factor three

(sampling frequency = 83 Hz) to reduce the computational time during the subsequent steps.

To deal with the non-stationarity of the EEG signal, the recordings are adaptively segmented resulting in quasi-stationary segments with variable length. Features extracted from these segments are then used to cluster segments with similar characteristics. A cluster time profile is obtained by representing each sample by its cluster label. Once the cluster time profile is extracted for each EEG channel, some processing steps are performed to obtain a single smooth curve which can be thresholded to detect quiet sleep segments. In the final stage of the algorithm, quiet sleep detections shorter than three minutes are removed [219].

9.2.5 Computational Time

In order to compare the recall time of the 3 algorithms, sleep stage classification of the complete test set in blocks of 2 h of EEG is performed with each of the algorithms. This experiment was run on a workstation with an Intel(R) Core(TM) i7 3.6 GHz processor and 16 GB RAM, implemented in MATLAB R2016a software (The MathWorks, Natick, MA, USA). The computational time of the different algorithm are measured and compared.

9.3 Results

9.3.1 Classification Performance

The classification performance of the proposed CNN is illustrated in figure 9.2, which presents the ROC curves for each of the recordings separately, and the mean and median ROC curve. The area under the mean and median ROC curve are equal to 92% and 97% respectively. The histogram of the training data segments and the AUC for each recording of the test set as function of post-menstrual age are plotted in figure 9.3. It can be seen from the histogram that training data is available at all ages. From the AUC in function of PMA, it is apparent that most of the test recordings are measured between 30 and 38 weeks of PMA. Moreover, the figure shows that the algorithm performs well over a wide range of PMA. Table 9.2 compares the AUC, as well as the sensitivity and kappa at specificity equal to 80% or 90% across the 3 methods. It can be seen from the table that the proposed CNN has similar performance compared to the CLASS and feature-based algorithm. In addition, it is clear that the post-processing step significantly improves the performance of the sleep stage classifier.

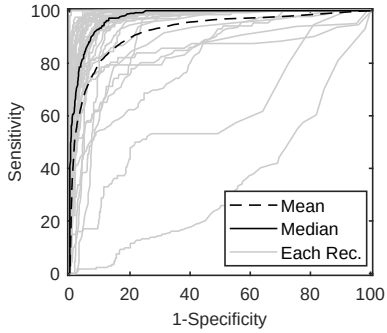


Figure 9.2: The ROC curves for the proposed CNN sleep stage classifier. The light grey ROC curves show the performance of the classifier for each recording. The black dashed and full line represent the mean ROC and median ROC curve respectively.

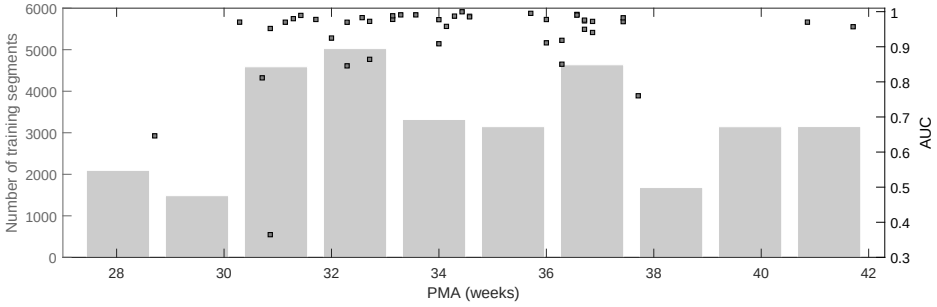


Figure 9.3: The histogram of the training data segments is displayed in light grey (left y axis). The squares show the AUC for each recording from the test set (right y axis).

Table 9.2: The average classification performance for the proposed CNN, the CLASS and feature-based approach without (NP) and with (PP) post-processing. The area under the mean ROC curve (AUC), average sensitivity (Sen) at specificity (Spe) of 80% and 90%, and kappa (Spe = 90%) are set out.

	AUC	Sen (Spe-80%)	kappa (Spe-80%)	Sen (Spe-90%)	kappa (Spe-90%)
CNN	0.92	0.90	0.62	0.80	0.69
CLASS	0.92	0.91	0.59	0.84	0.69
Feat-based NP	0.83	0.76	0.52	0.63	0.57
Feat-based PP	0.93	0.89	0.63	0.81	0.68

9.3.2 Computational Time

The bar graph in figure 9.4 illustrates the mean recall time for 2 h EEG segments. The dark blue box represents the average total time for loading, pre-processing and classifying 2 h EEG data. The light blue box shows the time required for classification. In case of CLASS, the classification consists of performing CLASS, starting from the artifact subspace reconstruction up to defining the quiet sleep periods. For the feature-based approach, the computational time for feature extraction, classification using the SVM and the post-processing are merged in the classification time. Finally, the time for classification and post-processing are combined and marked as classification for the CNN.

From this chart, it can be seen that overall the CNN (7 ± 0.65 s) is 11 times faster than CLASS (80 ± 2 s) and slightly faster than the feature-based (9 ± 0.80 s) approach. The most remarkable difference is the classification time, which is significantly lower for the CNN (0.11 ± 0.03 s) compared to CLASS (67 ± 1.62 s) and the feature-based classification (1.9 ± 0.07 s).

9.4 Discussion

The aim of the current study was to implement a convolutional neural network for neonatal sleep stage classification. By adopting convolutional neural networks which are able to learn relevant features automatically from the raw EEG data, the difficult and time-consuming process of selecting the proper feature set, which requires domain knowledge from experts, can be avoided. Moreover, the CNN optimizes the feature extraction and classification simultaneously. This

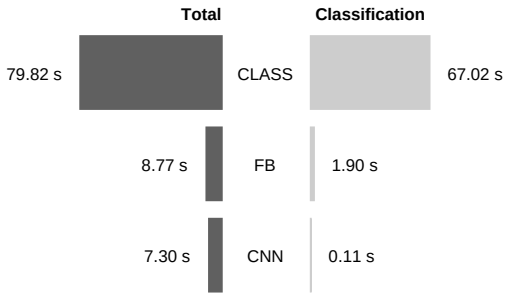


Figure 9.4: The average recall time for 2 h multi-channel EEG segments for each of the three algorithms. The total recall time (left) and the time for classification (right) are shown.

novel data driven approach successfully identified quiet sleep in preterm infants with an average and median AUC of 92% and 98% respectively.

While this performance is comparable to the CLASS algorithm developed by Dereymaeker et al. [203], the CNN has some advantages compared to the cluster time profiles approach. First of all, except for band-pass filtering the proposed method does not use any advanced artifact removal, whereas the CLASS algorithm uses artifact subspace reconstruction as additional artifact rejection. Secondly, the proposed CNN can be retrained if additional data collection is performed while this is not straightforward for the CLASS algorithm. As sleep patterns will be better represented in a larger training database, the classification performance will improve. Although the proposed CNN uses only 10 features as input for the fully connected classification stage, it reaches the same performance as the feature-based approach which is based on 72 features [59]. To conclude, it can be seen in figure 9.4 that the presented CNN is considerably faster compared to the current state-of-the-art.

One of the key challenges of neonatal sleep stage classification is to deal with the fast alterations in the EEG patterns during sleep ontogenesis. It is apparent from figure 9.3 that the network is able to train the age specific EEG characteristics, resulting in a good performance over a wide PMA range (30 - 38 weeks PMA). Due to the small number of recordings below 30 weeks and beyond 38 weeks PMA, we can not make any statements about the performance in those age ranges. Both in figure 9.2 and 9.3, 2 outliers can be observed.These recordings also had a low AUC in the CLASS and feature-based approach. The outlier

at 28 weeks and 5 days (AUC equal to 65%) is probably due to an electrode impedance problem, while the low performance of the recording at 30 weeks and 6 days (AUC equal to 36%) is likely caused by an intravenous infusion motor artifact. Moreover, both recordings were affected by movement artifacts.

In this study, the overall AUC is computed as the area under the mean or median ROC curve. As explained above, this mean ROC curve is obtained by averaging the sensitivity and specificity at an incremental threshold between zero and one, which is called threshold averaging [220]. It is important to note that this procedure of first computing the average ROC and then the AUC is not the same as computing the AUC for each recording separately and then taking the average. Computing the mean of individual AUCs is called vertical averaging, which corresponds to taking the average of the true positive rates at a fixed false positive rate [220]. In this application we opted for threshold averaging rather than vertical averaging, because implementation of the algorithm in a brain monitor for use in clinical practice requires the choice of one fixed threshold.

The present study has relevance for clinical practice. First of all, the proposed CNN has a lower computational time compared to the reference algorithms, hence it is more suitable for real time sleep stage monitoring. Real-time sleep scoring is important in the NICU as it can aid in making the neonatal care more patient-driven and avoid disruption of the infant's sleep. In addition to implementation of the algorithm in a brain function monitor to promote neonatal sleep in the NICU, this algorithm is of great interest in maturation studies as we have to account for the cyclic nature of EEG during brain maturation analysis. Furthermore, the sleep EEG undergoes fast maturational changes during early brain development.

The main limitation lies in the fact that the training phase is computationally expensive and requires a large database. However, training time is secondary since the network only has to be trained once. After training, the recall time is the important parameter, which is much smaller for CNN compared to the existing algorithms. A weakness of this study was the paucity of EEG recordings below 30 weeks and beyond 38 weeks post-menstrual age. Another drawback of CNNs is the lack of interpretability of the features. Further work needs to be done to establish which features are extracted by the network.

9.5 Conclusion

In this chapter, we designed and implemented a deep convolutional neural network which automatically extracts optimal features to discriminate quiet sleep from non-quiet sleep in preterm infants. The proposed network achieved

the state-of-the-art performance over a wide range of PMA without using domain-specific knowledge. Furthermore, the presented sleep stage classifier has a low computational time, which makes the CNN based sleep stage classification suitable for real-time sleep scoring in the NICU.

Chapter 10

Weighted performance metrics

In this chapter, a new framework for measuring the performance of algorithms when the database is scored by different raters is proposed. As shown in chapter 4, there is a wide variability in identifying rhythmic ictal and non-ictal EEG features. Therefore, when several readers score seizures independently, disagreement can be high and, therefore, commonly used metrics such as good detection rate and false alarm rate derived from data scored by multiple raters have their limitations. In this chapter, instead of defining the labels by consensus or majority voting, six popular metrics are modified such that they can take different scores into account. The result is compared with the majority voting technique and is confirmed using a bootstrapping test. This study has been published in modified form in [221].

10.1 Background

Detection of seizures has a high priority in the neonatal intensive care unit (NICU) because of high susceptibility of newborn brain to seizure expression [38]. Occurrence of neonatal seizures usually indicates a serious underlying neurological dysfunction. Unlike epileptic seizures in adults, most neonatal seizures have subtle clinical manifestations or are subclinical and are not readily recognized by clinical observation alone [80, 141]. Although monitoring of the electroencephalogram (EEG) accompanied by video is considered to be the gold standard in diagnosing neonatal seizures [47], reading EEG signals is labor-intensive, expensive, and importantly needs special expertise which may not be available around the clock in some NICUs. Therefore, a reliable, automated neonatal seizure detector is extremely useful and is required in the NICUs. However, measuring the reliability of such methods is not always straightforward, particularly when use of anti-epileptic drugs (AEDs) suppresses the EEG signal and modifies typical seizure patterns. Moreover, the varying degree of expertise among raters and the nature of the database, frequency of occurrence of seizures, presence of artifacts, etc., affect the visual scoring and to account for these in the performance assessment is not easy.

In recent years, there has been an increasing amount of literature on automated neonatal seizure detection using different techniques of pattern recognition and classification [3, 5–7, 148, 149, 154, 222]. However, very little research has been devoted to different approaches of measuring the performance of seizure detectors. Sensitivity, specificity, and selectivity were used to measure the performance in [144], one of the first studies in this field. False alarm rate per hour, which is more tangible than specificity for clinicians, is applied by Gotman et al. [145]. Navakatikyan clarified the difference between event-based and time-based metrics and explained different approaches for performance measurement [87]. The mean false detection duration was proposed by Temko et al. [89]. Mathieson et al. used Cohen's kappa between expert's scores and the output of an automated seizure detector [92]. In these reviewed studies, the ground truth was clearly defined by an expert or by consensus of different experts. However, Smit et al. employed 3 experts and showed that there is only moderate agreement between the raters, with a mean inter-observer agreement of 0.57. Subsequently, the final labels were reached by consensus of the raters. In that study, instead of considering two labels, 'seizure' and 'no-seizure', which is very typical, a third label called 'possible seizure, was used [95]. Cherian et al. also reported a fair agreement between used raters (kappa = 0.4). They called the third label 'dubious seizure' and showed a remarkable difference between the results when keeping or removing the dubious seizures in the test database [67]. This difference was also reported by Ansari et al. for several

methods [6]. Since the only possible way to measure the reliability of a seizure detector is comparing its outputs with the labels scored by an expert EEG reader, disagreement of the experts leads to uncertainty of the ground truth and makes the measurement labeling ambiguous. Although building a consensus solves this problem, it is not always possible in practice and is not the best solution, because of possible effect of social group dynamics among the expert EEG readers [90]. Therefore, there is a need to introduce a new set of metrics taking the dubious seizures as well as agreement of the raters into account. To this end, in this study, some important typically used metrics for performance measurement of neonatal seizure detection are extended.

10.2 Data description

In this chapter, DB1-small, DB2-small, DB3, and DB4 are used. The detailed descriptions about these datasets are given in section 3.2. Briefly, the EEG-polygraphic recordings from 81 neonates (353 h) recorded in two medical centers, EMCR and ULZ, are used in this study. The EEG signals are filtered between 1-20 Hz and bipolar channels are made according to the montage maps explained in section 3.2. The exact start-time and duration of each seizure were marked by the expert. Then, the automated neonatal seizure detector described in [4] was applied on the data and the false alarms were defined. All seizures scored by the clinical neurophysiologist, as well as the false alarms of the detector, were shuffled and relabeled by 3 independent EEG readers. As a result, 4 labels, 1 from the primary and 3 from the secondary raters, were assigned to each event, where event refers to a labeled seizure or a false/true detection. Note that the secondary raters could not change the start time or duration of the events. They only scored each event as "definite seizure", "dubious seizure", or "definite non-seizure".

10.3 Methods

10.3.1 Classical metrics

In order to measure the performance of a seizure detector, different metrics are used in the literature. In general, these metrics can be classified into two groups: event-based and epoch-based (time-based) metrics. In the former, each event equally influences the performance metrics regardless of its duration. Therefore, a very short detected seizure has the same effect on the performance as a very long one. However, in the latter, each event affects the performance metrics

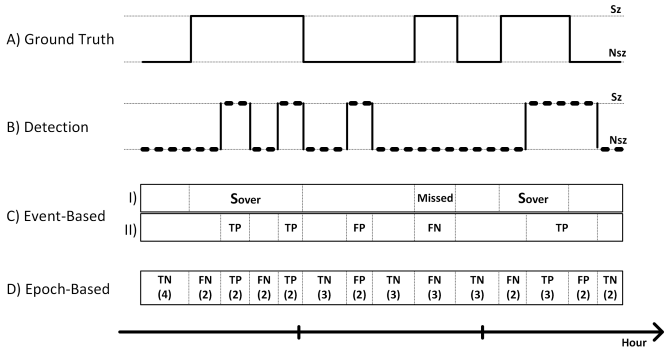


Figure 10.1: An imaginary example of an event and epoch-based assessment. A) The ground truth including 3 scored seizures in 3 hours. B) The output of an automated detector when each line segment corresponds to an epoch. C) Event-based segments checking the overlap of the ground truth with the detections (I) and vice versa (II). D) Epoch-based segments and the number of involved epochs.

proportionally to its duration. Subsequently, if the epoch-based metrics are used to optimize a seizure detector, many short seizures may be sacrificed for a very long one, which may not be satisfactory in some applications. Accordingly, different researchers may use one of these groups, based on their need and problem definition. In the following, the three most typical metrics of each group (Event/Epoch-based) are described. Furthermore, to illustrate the terms and definitions, figure 10.1 shows an imaginary example of 3 scored seizures as well as 4 automated detections over 3 hours.

Good Detection Rate (event-based): Good detection rate (GDR), also known as event-based sensitivity, expresses the overall percentage of the detected seizures and is calculated by

$$GDR = \frac{S_{over}}{S_{tot}} \times 100 \tag{10.1}$$

where S_{over} is the number of labeled seizures having overlap, even very tiny, with at least one of the detections and S_{tot} is the total number of labeled seizures. For instance, in figures 1, 2 out of 3 seizures are truly detected and the GDR equals 67%.

False alarm rate (event-based): False alarm rate (FAR) demonstrates the average number of false alarms that one can expect to see every hour and is defined as

follows:

$$FAR = \frac{FP^{Event}}{T_{total}} \tag{10.2}$$

where FP^{Event} is the number of falsely detected events and T_{total} is the total duration of the recordings per hour. In figure 10.1, one false alarm in 3 hours are falsely detected and the FAR is 0.3 h^{-1}.

Positive Predictive Value (event-based): Positive predictive value (PPV) shows how reliable automated detections are and is measured using (10.3).

$$PPV = \frac{TP^{Event}}{TP^{Event} + FP^{Event}} \times 100 \tag{10.3}$$

where TP^{Event} shows the number of truly detected events which means the number of events having overlap with at least one of the labeled seizures. In figure 10.1, PPV equals 75% (3 out of 4). Note that TP^{Event} can be different from S_{over} used in GDR. The former shows the number of detections having overlap with labeled seizures, whereas the latter is the number of labeled seizures having overlap with the detections. In figure 10.1, TP^{Event} and S_{over} equal 3 and 2 respectively.

Sensitivity (epoch-based): Sensitivity, also called true positive rate, corresponds to GDR while the number of epochs is counted instead of the number of events. It is defined as

$$SEN = \frac{TP^{Epoch}}{TP^{Epoch} + FN^{Epoch}} \times 100 \tag{10.4}$$

where TP^{Epoch} and TN^{Epoch} are the number of truly detected (True Positive) and falsely not-detected (False Negative) epochs respectively. In figure 10.1, sensitivity is 44% (7 out of 16 epochs).

Specificity (epoch-based): Specificity, also known as true negative rate, measures the percentage of non-seizure epochs which are truly not identified by the seizure detector as follows:

$$SPE = \frac{TN^{Epoch}}{TN^{Epoch} + FP^{Epoch}} \times 100 \tag{10.5}$$

where TN^{Epoch} shows the number of truly not-detected epochs (True Negative) and FP^{Epoch} denotes the number of falsely detected epochs (False Positive). In figure 10.1, specificity is 79% (15 out of 19).

Selectivity (epoch-based): Selectivity, also called epoch-based PPV or precision, shows the reliability of the detections based on their duration and is measured by (10.6). In figure 10.1, selectivity equals 64% (7 out of 11).

$$SEL = \frac{TP^{Epoch}}{TP^{Epoch} + FP^{Epoch}} \times 100 \tag{10.6}$$

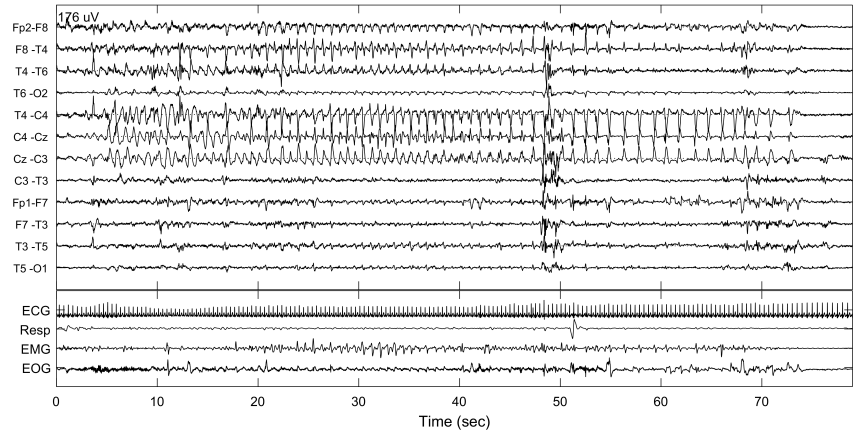


Figure 10.2: An agreed seizure starting from second 10 and lasting for 70 seconds. This event was primarily scored as 'seizure' by the primary neurophysiologist with rhythmic oscillation and sharp spikes (mixed) at C_4 and T_4 (peak to peak amplitude of spikes is 200 µV). All secondary raters also rescored it as 'seizure'.

10.3.2 Multi-label problems

The aforementioned metrics are applicable when the ground truth is clearly predefined. In other words, the events should be clearly labeled as seizure or non-seizure by an expert beforehand. However, in neonatal seizure detection, it is possible that some repetitive patterns suspicious for seizures, called 'dubious seizures', appear without fulfilling all characteristics that define a seizure particularly after treatment with high dose of anti-epileptic drugs (AEDs) or in neonates with very severe encephalopathy. These seizures usually have suppressed amplitude, and lack of typical seizure evolution patterns. Consequently, it is possible that different experts have different opinions, resulting in several labels for each event. To illustrate, figure 10.2 shows a seizure having clear evolution of frequency and morphology on C_4 and T_4. This seizure was scored as 'seizure' not only by the primary neurophysiologist, but also by all secondary raters. Figure 10.3 displays another event that lasted for 20 s and scored by the primary and 2 of secondary raters as 'dubious seizure', while one of the raters believed that it is a 'seizure'. Finally, figure 10.4 shows an event that was primarily score as 'seizure', but the secondary raters completely disagreed with each other and scored it as 'seizure', 'dubious', and 'non-seizure'. In such cases, there are different ways to measure the performance, as explained in the following subsections.

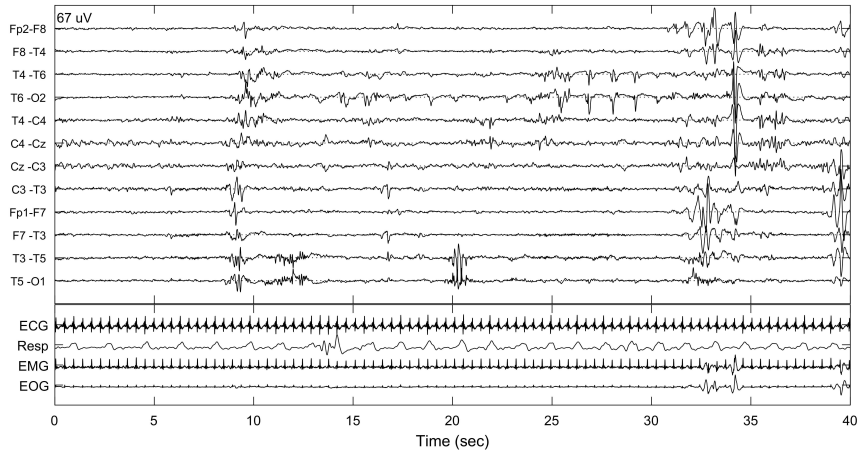


Figure 10.3: A moderately agreed dubious seizure starting from second 10 and lasting for 20 seconds. This event was primarily scored as 'dubious seizure' by the primary neurophysiologist with arrhythmic oscillation and sharp spikes (mixed) at $T_6 - O_2$ (peak to peak amplitude of spikes is 40-50 µV). It was rescored as 'Seizure', 'Dubious', and 'Dubious' by the secondary raters.

10.3.3 Averaging (ideal reference)

For measuring the performance of an algorithm in a multi-scored database, the simplest approach is measuring each performance metric multiple times based on each expert's labels independently and then average them. For example, the output of the algorithm is compared with each rater's scores and sensitivity is measured for each rater. Then, the average is reported as total sensitivity. Given the assumption that all used raters have enough experience and so their labels are reliable, this averaged metric can be the best approach since it exactly measures the average satisfaction if each of these experts uses such an algorithm in clinical practice. In this study, the averaged metrics are used as 'ideal reference' and other approaches are compared with them. However, in practice, recording EEG data and scoring seizures are very time-consuming and big datasets have gradually been made over years. Therefore, it is possible that several experts label different parts of the database. Similarly, multi-center databases may also have different scorers. It means that such averaged performance is not practically available and there is a need to introduce new performance metrics for multi-labeled databases.

Note that for measuring the performance based on each rater's scorings, the

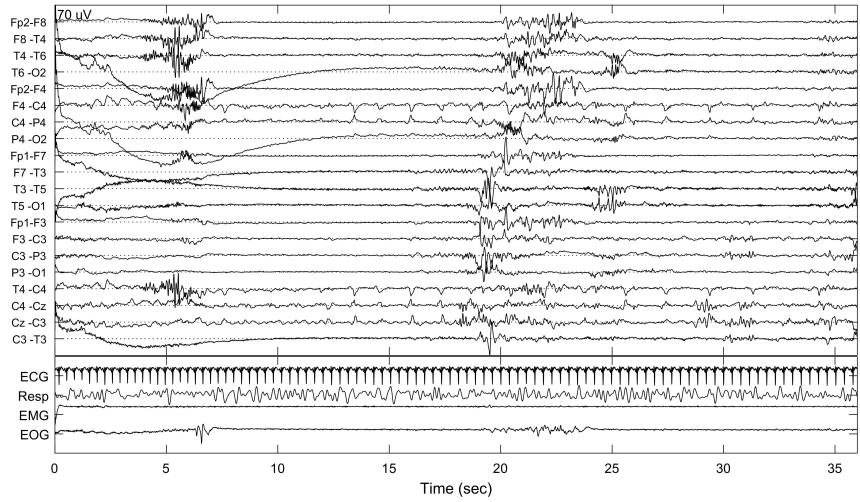


Figure 10.4: Disagreed seizure starting from second 10 and lasting for 25 seconds. This event was primarily scored as 'seizure' by the primary neurophysiologist with rhythmic oscillation and sharp spikes (mixed) at C_4 (peak to peak amplitude of spikes is 10-20 µV). It was then rescored as 'Seizure', 'Dubious', and 'Non-seizure' by the secondary raters.

dubious events should be skipped. It means that if a segment is scored as dubious, neither it is counted as false alarm if a classifier detects it, nor it is calculated as missed seizure if a classifier cannot detect it [6, 67]

10.3.4 Majority Voting

In this approach, the most frequent label is defined as the aggregated label for each event individually. In this way, unlike the averaging method, different events may primarily have different numbers of labels. The main issue of this method is that the minority votes are not taken into account in any case. For instance, a seizure with 100% of votes have the same importance as a seizure with (50%+1) votes. However, it seems evident that seizures with a high degree of agreement are more important than the controversial ones and are expected to have more effect on the performance measurement.

10.3.5 Proposed weighted metrics

The key aspect is taking the minority votes into the measurement in order to quantify the overall satisfaction of the raters. Since there is not yet a clear "gold standard" for the interpretation of the EEG signals for defining seizures in the literature, using Fuzzy logic may help compensate for the inherent uncertainty of such problem. Assuming the events are labeled as 1) definite seizure (Sz), 2) dubious seizure (Dsz), and 3) definite non-seizure (Nsz), three corresponding fuzzy sets are derived and subsequently each event has three membership values: μ_S, μ_D, and μ_N showing the membership of the event to each set. These values are defined based on the percentage of their corresponding labels. Thus, $\mu_S + \mu_D + \mu_N = 1$. For instance, μ_S, μ_D, and μ_N of an event which is labeled as [Sz, Sz, Sz, Dsz, Nsz] by 5 raters are respectively 0.6, 0.2, and 0.2. Since it is rational to say that detecting dubious events is neither true nor false, the defined μ_D is not used from here on and does not appear in any formula. The mentioned classical metrics are extended as follows:

Weighted good detection rate (WGDR) is measured by

$$WGDR = \frac{\sum_{n=1}^{N_{Scored}} \mu_S(n) o_{Scored}^{Event}(n)}{\sum_{n=1}^{N_{Scored}} \mu_S(n)} \times 100 \tag{10.7}$$

where N_{Scored} is the total number of labeled events and $O_{Scored}^{Event}(n)$ equals 1 if there is overlap between the nth labeled event and at least one of the automated detections, and otherwise, it equals 0.

Weighted false alarm rate (WFAR) is calculated by

$$WFAR = \frac{\sum_{n=1}^{N_{Detection}} \mu_N(n) o_{Det}^{Event}(n)}{T_{total}} \tag{10.8}$$

where $N_{Detection}$ is the total number of detected events and $O_{Det}^{Event}(n)$ equals 1 if there is overlap between the nth detected event and at least one of the labeled events, and otherwise, it is 0.

Weighted positive predictive value (WPPV) is defined as

$$WPPV = \frac{\sum_{n=1}^{N_{Detection}} \mu_S(n) o_{Det}^{Event}(n)}{\sum_{n=1}^{N_{Detection}} \left(\mu_S(n) + \mu_N(n)\right) o_{Det}^{Event}(n)} \times 100 \tag{10.9}$$

In the same way, *weighted sensitivity (WSEN)*, *weighted specificity (WSPE)*, and *weighted selectivity (WSEL)* are defined by the (10.10)–(10.12).

$$WSEN = \frac{\sum_{p=1}^{P} \mu_S(p) o^{Epoch}(p)}{\sum_{p=1}^{P} \mu_S(p)} \times 100 \tag{10.10}$$

$$WSPE = \frac{\sum_{p=1}^{P} \mu_N(p) \left(1 - o^{Epoch}(p)\right)}{\sum_{p=1}^{P} \mu_N(p)} \times 100 \tag{10.11}$$

$$WSEN = \frac{\sum_{p=1}^{P} \mu_S(p) o^{Epoch}(p)}{\sum_{p=1}^{P} \left(\mu_S(p) + \mu_N(p)\right) o^{Epoch}(p)} \times 100 \tag{10.12}$$

P is the total number of epochs and $O^{Epoch}(p)$ equals 1 if the p_{th} epoch is detected by the automated classifier, and otherwise, it equals 0.

Note that if $\mu_{N,S} \in \{0, 1\}$, which means only one rater scored the data or there is 100% agreement of the raters, the weighted and classical metrics are exactly equal. For instance, in the formula of *WSEL*, $(\mu_S(p)O^{Epoch}(p))$ equals 0 for all values of p except when the p_th epoch is detected by the algorithm and labeled as seizure by the rater. Therefore, its summation over all p equals the total number of true positives, the numerator of *SEL*, in (10.4). Thus, these proposed metrics can be seen as a generalization of the classical metrics for multi-rater approaches.

10.3.6 Bootstrapping test

In this chapter, in order to compare the accuracy and robustness of the classical metrics and proposed weighted metrics, a bootstrapping test was applied. The mathematical explanations of this technique has been provided in section 5.6. Since the effect of the change in the raters' scorings is to be considered in this chapter, the bootstrap is applied on the raters' opinions. It means that for each event, 4 labels were selected with replacement from the 4 existing labels. This selection is performed for the whole database and all metrics are computed using the bootstrapped labels and the fixed classifier output. This selection and computation is repeated 10,000 times resulting in 10,000 values for each metric. Then, a PDF is estimated for each metric by a kernel smoothing technique explained in [223] and the results are plotted and discussed.

10.3.7 Tested Seizure Detectors

In order to reduce the effect of the classifier on the comparison of the considered measurement metrics, three different previously published neonatal seizure detectors including a heuristic classifier [4], an improved version of the heuristic classifier [5, 7] , and a multi-stage detector (see chapter 6 or [6]) were used.

These algorithms are described respectively in Appendix A.1 version 1, version 2, and chapter 6.

Note that all parameters or variables of these three used algorithms (such as trained-classifiers, the hyper-parameters, and final thresholds) were exactly defined based on their original algorithms and were not retrained or re-tuned for the test explained in this chapter. Furthermore, they were not subjected to any changes during the performance measurement for different raters or different metrics.

10.4 Results

10.4.1 Accuracy

As it is mentioned, the 'ideal reference' of each metric is defined by averaging that metric over the raters, which shows the averaged raters' opinions. The accuracy of a substitute metric is corresponding to proximity of such metric to its ideal reference. Therefore, in order to compare the accuracy of the classical metrics using majority voting with the proposed metrics, first, all epoch/event-based metrics were measured based on each rater's labels individually. In this case, since only one rater is used for each metric, the classical and proposed metrics are equal. Then, the average of each metric was measured and used as an 'ideal reference'.

Second, the classical metrics using majority voting and proposed metrics are computed based on all raters' labels and compared with the ideal reference. The results are listed in table 10.1 where the boldface items of the last two columns represent the values which are closer to the ideal reference (are more accurate). Note that for measuring the epoch-based metrics, the epochs were selected as 1s windows with no overlap. However, this length does not play a major role in the performance measurement.

Note that for measuring the epoch-based metrics, the epochs were selected as 1s windows with no overlap. However, this length does not play a major role in the performance measurement.

The results shown in table 10.1 indicate that the proposed metrics produce results which are, in general, more similar to the ideal reference (12 out of 18). Because of the fact that the ideal reference is not always practically accessible, as it was previously mentioned, the proposed metrics can be a more accurate substitute for the average performance assessment than classical metrics using majority voting.

Table 10.1: Similarity of reference with proposed metrics and classical metrics using majority voting

Detector	Metric		Ideal Ref. [b]	Classical [c]	Proposed [d]
Heuristic [4] (Version 1)	Event	[W]GDR (%)	92.1	95.0	**91.6**
		[W]FAR (h^{-1})	3.17	3.14	**3.17**
		[W]PPV (%)	53.7	59.2	**53.9**
	Epoch	[W]SEN (%)	77.5	**78.1**	74.3
		[W]SPE (%)	89.7	**90.5**	88.1
		[W]SEL (%)	50.4	58.7	**52.7**
Heuristic [5] (Version 2)	Event	[W]GDR (%)	73.0	**77.8**	67.1
		[W]FAR (h^{-1})	1.60	1.30	**1.60**
		[W]PPV (%)	62.2	70.5	**63.5**
	Epoch	[W]SEN (%)	55.2	**56.5**	48.6
		[W]SPE (%)	97.0	97.7	**96.9**
		[W]SEL (%)	63.5	74.8	**65.2**
Multi-stage Detector [a] [6]	Event	[W]GDR (%)	82.2	89.6	**82.2**
		[W]FAR (h^{-1})	1.47	1.31	**1.47**
		[W]PPV (%)	62.4	71.5	**64.8**
	Epoch	[W]SEN (%)	73.2	**76.0**	69.9
		[W]SPE (%)	91.1	**92.1**	89.6
		[W]SEL (%)	55.8	65.7	**58.2**

[a] The multi-stage classifier when its defined threshold (TH) equals 0.3
[b] Averaged metrics over all raters
[c] Classical metrics using Majority Voting technique
[d] Proposed weighted metrics

10.4.2 Robustness

In order to compare the robustness of the two approaches for the assessment of the performance of the classifiers, majority voting and the proposed one, a realistic experiment and a bootstrapping test were designed. For the former, the secondary raters were asked to rescore 65% of the events for the second time. The interval between the scoring times was 1 year on average. Given that there are many dubious seizures in the used database, some events were labeled differently in the second trial (17% of the scorings, 847 out of 4980). For instance, the disagreed event presented in figure 10.4 was labeled as seizure by the 3rd secondary rater in the second trial, while it had been labeled as non-seizure primarily by this rater in the first trial. It is expected that a robust metric tolerates this change. To reveal the robustness of the metrics, the performance was computed using the new labels and it was compared with the results obtained with the initial labels. Table 10.2 lists the differences between the two trials for each metric, separately. The boldface values indicate the items with smaller change. As it is shown, in 16 cases out of 18, the proposed metrics have smaller difference, which indicates that the proposed metrics are more robust to small change in the labels.

Table 10.2: Difference between the 1st and 2nd scoring trials

Detector	Metric		Δ Classical	Δ Proposed
Heuristic [4] (Version 1)	Event	[W]GDR (%)	**0.47**	0.84
		[W]FAR (h^{-1})	**0.22**	0.25
		[W]PPV (%)	6.10	**1.79**
	Epoch	[W]SEN (%)	2.84	**0.30**
		[W]SPE (%)	2.41	**0.40**
		[W]SEL (%)	5.33	**1.88**
Heuristic [5] (Version 2)	Event	[W]GDR (%)	2.33	**0.11**
		[W]FAR (h^{-1})	0.65	**0.31**
		[W]PPV (%)	3.59	**0.95**
	Epoch	[W]SEN (%)	5.82	**0.58**
		[W]SPE (%)	1.49	**0.26**
		[W]SEL (%)	5.00	**0.78**
Multi-stage Detector [6]	Event	[W]GDR (%)	1.45	**0.20**
		[W]FAR (h^{-1})	0.19	**0.14**
		[W]PPV (%)	5.02	**3.04**
	Epoch	[W]SEN (%)	3.35	**0.04**
		[W]SPE (%)	2.39	**0.31**
		[W]SEL (%)	5.94	**2.81**

In practice, it is not possible to repeat the rescoring procedure by the same expert EEG readers for many times, not only because it is time-consuming and expensive, but also because of the memorizing effect. Therefore, a bootstrapping test is designed, as described in the previous section, and used to repeat the rescoring for statistical comparison. As a result, figure 10.5 shows the PDFs of all epoch/event-based metrics for the three algorithms separately. In this figure, it is shown that the PDFs of the proposed weighted metrics (dashed line) are generally more similar to the reference PDFs (continuous line) than the majority voting technique (dotted line). Furthermore, for most PDFs, the standard deviations of the proposed metrics are smaller than those of the majority voting, which confirm their higher robustness.

10.5 Discussion

The development of a classifier for neonatal seizure detection with adequate performance is undeniably needed in NICUs. However, measuring the performance of such a system is problematic since the gold standard for the interpretation of the EEG signals is not yet clearly defined. In other words, several expert clinical neurophysiologists may score a database differently, which results in different performances for one algorithm. This problem is exacerbated when dubious seizures are present in a database containing recordings from neonates with severe neurological problems, such as severe HIE. The best and

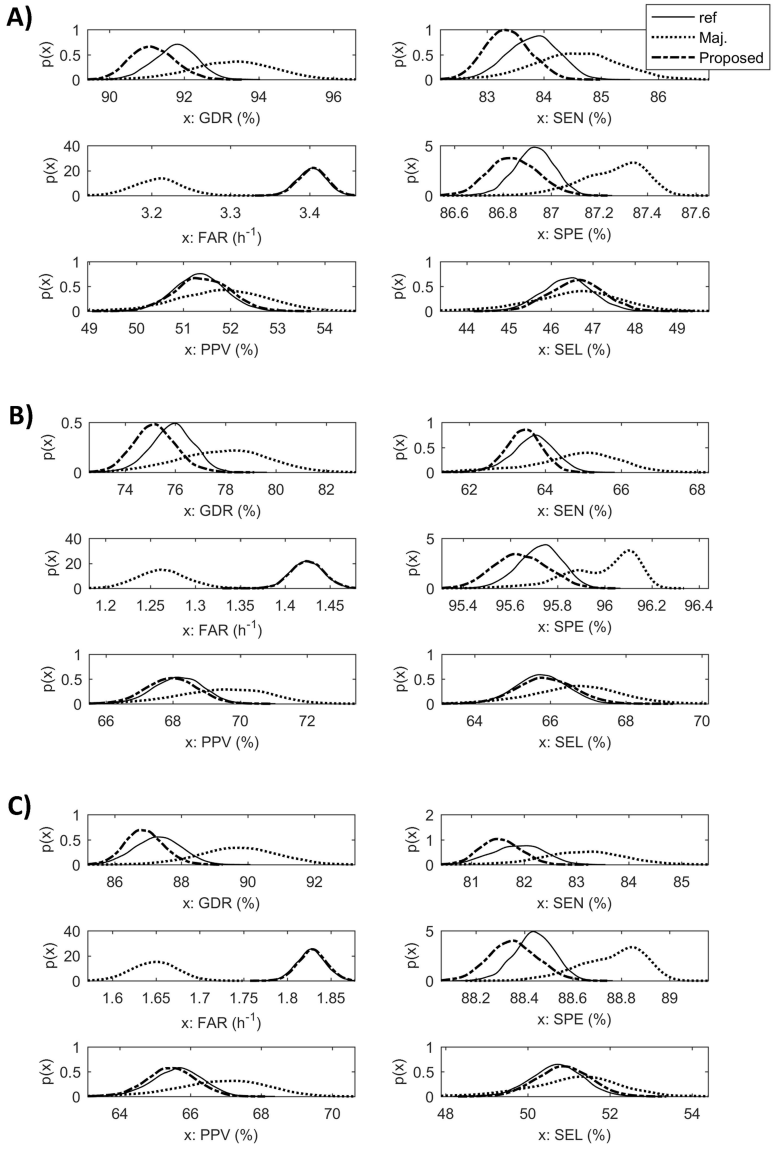


Figure 10.5: Probability density function of GDR, FAR, PPV, SEN, SPE, and SEL when the detector is A) the heuristic algorithm proposed in [4], B) the improved heuristic algorithm proposed in [5], and C) the multi-stage classifier proposed in [6]. In all traces, continuous, dotted, and dashed lines are respectively corresponding to the PDFs of the reference, classical metrics using majority voting, and proposed weighted metrics.

simplest solution for such ambiguity of ground truth is averaging the performance over all the expert raters. This averaged performance is the ideal reference since it expresses the averaged satisfaction of the raters if they use that classifier individually at their centers. However, in practice, the scores of all raters are not always available for all events, especially in large and multi-center datasets. For instance, the used database in this study was collected over 10 years in two centers and there is no guarantee that all present experts who scored it are available in the future. (Note that in this study, all events were scored by the primary and secondary raters to be able to compare the performance of the proposed metrics with the ideal ones, although it is not required for the proposed metrics as mentioned.) Furthermore, for training some seizure detection algorithms, the label of each event should be explicitly defined. In that case, a common approach is using majority voting to aggregate different labels of each event. However, since majority voting does not use the minority votes, it does not measure the overall performance and seems to be different from the defined ideal reference. Therefore, a new framework taking all raters' opinions into account was needed for measuring the performance of neonatal seizure detectors. To this end, in this chapter, a new set of performance measurement metrics was proposed and compared with the majority voting approach.

The main feature of the proposed weighted metrics is the fact that it takes minority votes into account besides the majority votes. The events with larger agreement in the votes have higher influence on the proposed metrics while in majority voting there is no difference between two events having 100% and (50%+1) of votes. Therefore, the overall performance measured by the newly proposed metrics is closer to the ideal reference than majority voting, as it has been practically confirmed by the results provided in this manuscript. In addition, the dubious seizures may be scored differently when a rater scores them several times. Moreover, these dubious seizures are usually controversial and have less agreement. Hence, it is probable that whenever a dubious seizure is rescored, the majority vote changes which results in a decrease of the robustness of the majority voting technique. Nevertheless, since minority votes also play a role in the proposed metrics, changing a label has a small effect on the metrics, thereby increasing the robustness besides the accuracy.

The other advantage of the proposed metrics, compared to the majority voting approach, is that when the raters have different levels of expertise, they can influence the metrics differently. To this end, each rater can have a specific weight which corresponds to his/her experience, knowledge, agreement with others, etc. This weight can be easily used in the calculation of the μ_x using weighted averaging. For instance, the membership value of being seizure (μ_S) of an event scored as (Sz, Sz, Nsz) by 3 raters having (1, 0.75, 0.5) weights is 0.78 ($= (1 + 0.75)/(1 + 0.75 + 0.5)$). As a result, the labels of more experienced or

preferred raters have more impact on all the modified performance measurement metrics than those of the other raters. In this chapter, for the ideal reference, it has been assumed that the raters have the same experience, which rarely happens in real practice. In that case, the mentioned weighted averaging technique can also be used for the ideal metrics. Note that for this solution, the levels of expertise should be quantified which is not always straightforward.

Moreover, it seems logical that detecting the seizures which are agreed by different experts is more important than controversial events. Hence, if we can employ the scores of raters in developing or training the classifiers in order to prefer the highly agreed events to the poorly ones, it may increase the overall performance. Another important finding of the proposed technique is that each event has three membership values expressing their agreement. If the training method accepts weights for the training data, e.g. weighted least squares support vector machines [224], the μ_S which corresponds to the agreement of being seizure is a good option to be used as the weight. Furthermore, if the classifier supports rewarding and penalizing strategies, like reinforcement learning, μ_{Ns} can be used as a penalty to correct the model.

In addition, since the most alarming systems in hospitals only supported on/off inputs, it was previously expected that the output of seizure detectors should be either seizure or non-seizure, Sz/Nsz. On the contrary, new generations of alarming systems, such as colorful lighting systems, can be used in NICUs and according to the probability of detected seizures, change the color/luminance of the light. Although some published methods produced probability as output, like [3,6,149], they still applied thresholding combined with classical performance metrics. Hence, they do not measure the actual performance of their probability-based systems. Nevertheless, in the proposed metrics, the defined functions $O^x(y)$ showing the overlap between the events and detections can easily be replaced by the probability. If this is the case, the metrics take the probability of being a seizure into account and measure the overall performance of probability-based outputs.

However, a disadvantage of the proposed metrics compared to majority voting is that whenever a rater is not experienced enough, his/her opinion always influences the metrics as well as the experienced ones since in the proposed metrics the minority votes are always important. However, using majority voting can remove an outlier rater if present. Although using the aforementioned weights for the raters can partially solve this problem, finding the correct weights might be another challenge.

10.6 Conclusion

The purpose of this chapter was to determine a new framework for measuring the overall performance of automated neonatal seizure detectors in the presence of raters' disagreement. To this end, the typically used metrics were extended in order to take the agreement of raters on each scored/detected event into account and were tested on cEEG data from 81 neonates, recorded in 2 medical centers. A bootstrapping test was also applied on the data and supports the empirical results. The significant finding to emerge from these results is that the proposed metrics are more accurate and robust than majority voting because the minority votes also play a role in the metrics. The proposed metrics are generic and applicable for every pattern recognition 2-class problem in which the ground truth is differently defined by several experts. Furthermore, the new metrics automatically remove 'dubious'-scored events and allow having different raters for different parts of the database and for missing scores.

Part IV

Practical Implementations

Chapter 11

NeoGuard information system

In this chapter, the NeoGuard Information System is described. This online platform was used for collecting part of the data used in chapter 3 and the scorings used in chapters 4 and 10. The NeoGuard Learning Platform is also part of this online system. However, because of its importance, it is described separately in the next chapter.

11.1 Background

One of the most important materials for developing a supportive tool is an extensive database including diverse types of the goal patterns. In this PhD research, in order to develop a neonatal brain monitor, such a database should have been collected from different centers, scored by several expert raters working in different countries, and collected in a centralized storage. Then, this database can be used not only for developing and training the methods, but also for validating the performance. To this end, a secured web-based system, namely NeoGuard Information System (NGIS), has been developed with three main modules:

1. NeoGuard Data Collector (NGDC)
2. NeoGuard Scoring Platform (NGSP)
3. NeoGuard Learning Platform (NGLP)

In the next sections, the NGDC and NGSP, in addition to the general structure of the NGIS are described. The developed graphical user interfaces (GUI) and their usage are explained and finally the added value, as well as the limitations, are discussed. The NGLP is explained thoroughly in the next chapter, 12.

11.1.1 Programming languages

The NGIS, like other web-based platforms, consists of server-side and client-side scripts. For the former one, PHP [1] language has mainly been used to process the queries asked by the client-side scripts, to authenticate the users, connect to the database, save an uploaded file, etc. Furthermore, SQL [2] has been used to perform 'select', 'insert', 'update', and 'delete' commands in the database in order to make the queries or update the stored data. For the client-side scripts, HTML [3] (version 5) was used for the web elements. Javascript and CSS [4] were also utilized to make dynamic elements and design the style of pages respectively. Additionally, AJAX [5] technology was hired to create asynchronous queries.

[1] PHP: Hypertext Preprocessor
[2] SQL: Structured Query Language
[3] HTML: Hypertext Markup Language
[4] CSS: Cascading Style Sheets
[5] AJAX: Asynchronous JavaScript And XML

Table 11.1: The different levels of accessibility rights

		Ability					
		NGDC					NGSP
Level	Name	Manage users	Delete	Update	Add	View	Rate
1	Admin	✓	✓	✓	✓	✓	✓
2	Full		✓	✓	✓	✓	✓
3	Normal			✓	✓	✓	✓
4	Add-view				✓	✓	✓
5	Add only				✓		✓
6	View only					✓	
7	Rate only						✓

Besides, in order to generate the images of EEG segments used in NGSP and NGLP automatically, and to export the labels of the scoring system into MATLAB, MATLAB scripting and ODBC [6] drivers were used.

11.1.2 The accessibility of the users

The NGIS is a platform available for 1) clinicians recording data in hospitals by NGDC, 2) expert EEG interpreters annotating EEG segments by NGSP, and 3) public users using the learning platform by NGLP. Therefore, different levels of accessibility should be defined in the system. In this version of the NGIS, the public users have no specific account. Consequently, everyone who enters the website with no authenticated username and password is counted as public user with minimal accessibility rights and is only able to use the learning platform. For the other users, an account (with username and password) can be made by the website administrator(s) and the level of accessibility is defined. The provided levels and the user abilities are listed in table 11.1.

11.1.3 Database structure

For storing the data, two types of management systems were used: 1) common Unix folders, and 2) structured SQL database. The former one is used for storing the recording files uploaded by the users. These files are usually being saved in EDF [7] and have a large size which is not appropriate to be stored in the structured database. The latter one is for saving the specifications of the recordings, e.g. date of birth and Sarnat/Apgar scores, as well as the

[6] ODBC: Open Database Connectivity
[7] EDF: European Data Format

information of the users. The main advantage of the structured database is that the main commands including 'select', 'insert', 'delete', and 'update' are automatically processed by the database and therefore making queries is easier and faster.

In order to manage the structured database, MySQL database has been used. The enhanced entity-relationship (EER) diagram of the designed SQL tables is shown in figure 11.1. As is clear from the diagram, this database includes 9 tables as follows:

- tblUsers: to store the information about the users including encrypted username, encrypted password, level of accessibility, email, etc.
- tblInfo: to store the information about the uploaded recordings by the NGDC, including birth of date, center, case ID, etc.
- tblFiles: to store the system file-name of the uploaded recordings. This table keeps the link between the structured and unstructured databases.
- tblEvents: to store the information about the EEG segments which are being used in the scoring and learning platforms (NGSP and NGLP).
- tblRatings: to store the labels of the expert scorers made in NGSP.
- tblEventNotes: to store the comments and discussions that expert scorers may add for different events.
- tblTrainingProfile: to store the information of the public users of the NGLP.
- tblTrainingRatings: to store the labels entered by the public users of the NGLP.
- tblLog: to store all the logs generated in the NGIS.

Furthermore, different 'stored-procedures', views, and functions were made in MySQL to handle the requests of the website.

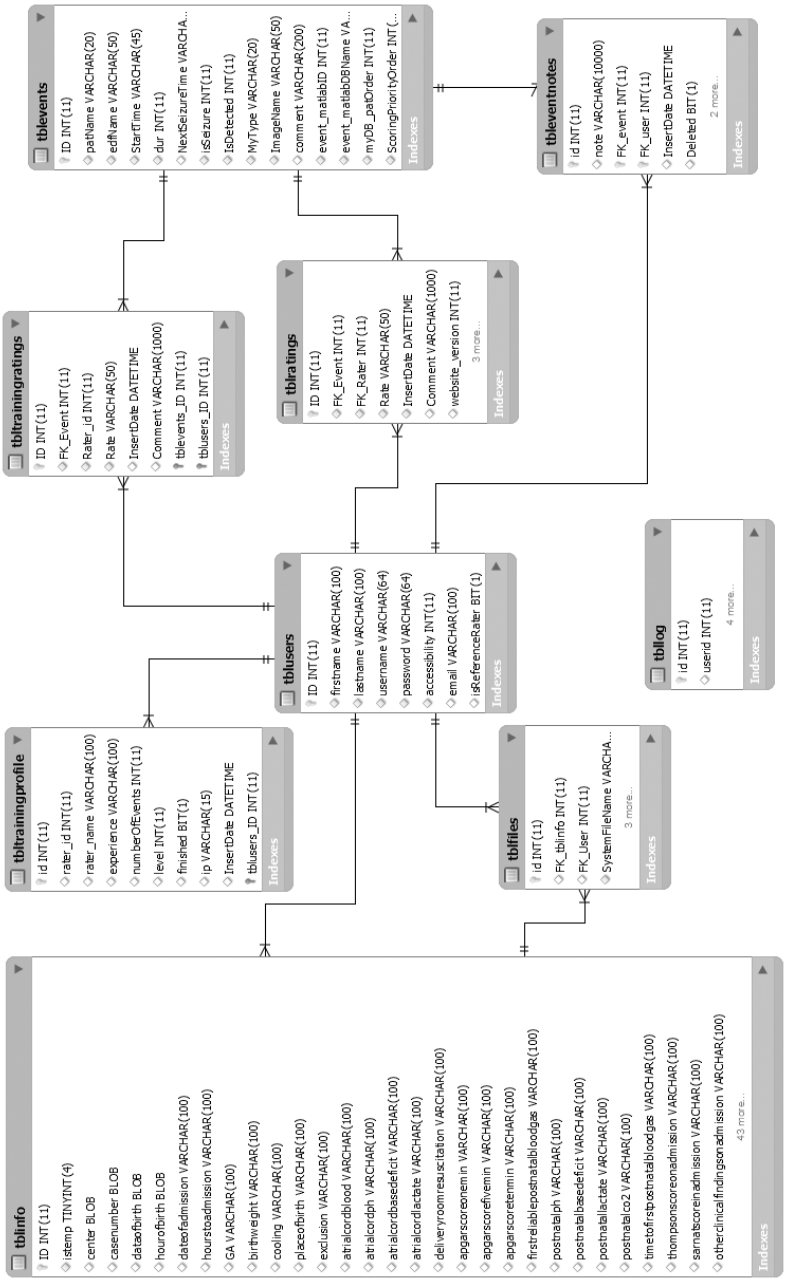


Figure 11.1: The EER diagram of the designed SQL model.

11.2 NeoGuard Data Collector

As mentioned, the NeoGuard Data Collector, NGDC, is a toolbox for clinical centers to upload their recordings to the NeoGuard centralized database. The main reason for developing this system was the fact that the use of common sharing platforms, such as Dropbox, Google drive, or WeTransfer, is not permitted for uploading such highly confidential data including personal information of neonates. The NGDC makes a secure connection to the server and also encrypts the confidential information. For higher security, the confidential data are being transferred and stored in encrypted format in the SQL database and are decoded on the client-side only when authenticated users want to read it.

Moreover, the users of NGDC are able to upload the recording files which are usually very big size (more than 1 GB). The client-side of the NGDC automatically splits the big-size files into smaller ones and uploads them sequentially using JavaScript. Then, on the server-side, the split files are merged to generate a copy of the original file. In this way, the users are able to upload big files of more than 100 GB to the NeoGuard server.

11.2.1 Uploading centers

So far, the NGDC has been used by three clinical centers including:

- The NICU of the University Hospitals Leuven, Belgium (UZL)
- The Sophia Children's Hospital (part of the Erasmus University Medical Center Rotterdam, The Netherlands) (EMCR)
- The NICU of ZNA Queen Paola Children's Hospital, Antwerp, Belgium (ZNA)

In future, other centers can also be easily added to the system by making new accounts and proper levels of accessibility by the administrator(s) of the NGIS.

11.2.2 The GUI

In addition to the general Welcome, Log-in, and Settings pages of the NGIS, the NGDC has 8 forms to insert and search the data as follows:

- Insert - General data: to add general information for the new neonates, such as center, date of birth, cause of asphyxia, result for exclusion, etc.
- Insert - The start: to add the arterial cord, Apgar scores, Sarnat scores and other clinical findings on admission.
- Insert - EEG: to add EEG information, such as the status of the CFM assessment and sleep/wake cycling on admission, CFM evolution, Thompson Scores, etc.
- Insert - Evolution: to add information concerning the evolution of cooling, such as cooling device, cooling side effect, rewarming events, etc.
- Insert - Imaging: to add imaging information, such as Ultrasound evolution, MR findings, etc.
- Insert - Discharge: to add discharge information, such as death type, date and hour of death, discharge exam, etc.
- Insert - Upload: to attach the required recordings and enter the file names.
- View: to observe the table listing all inserted recordings and their information. Additionally, some buttons for editing, deleting, updating, and downloading the data are available for users in case the proper access rights are granted.

In figure 11.2, the 'Welcome' page of the NGIS is shown. The user can enter the NGDC and NGSP by entering a registered username and password or can use the NGLP by clicking on "score for learning" link.

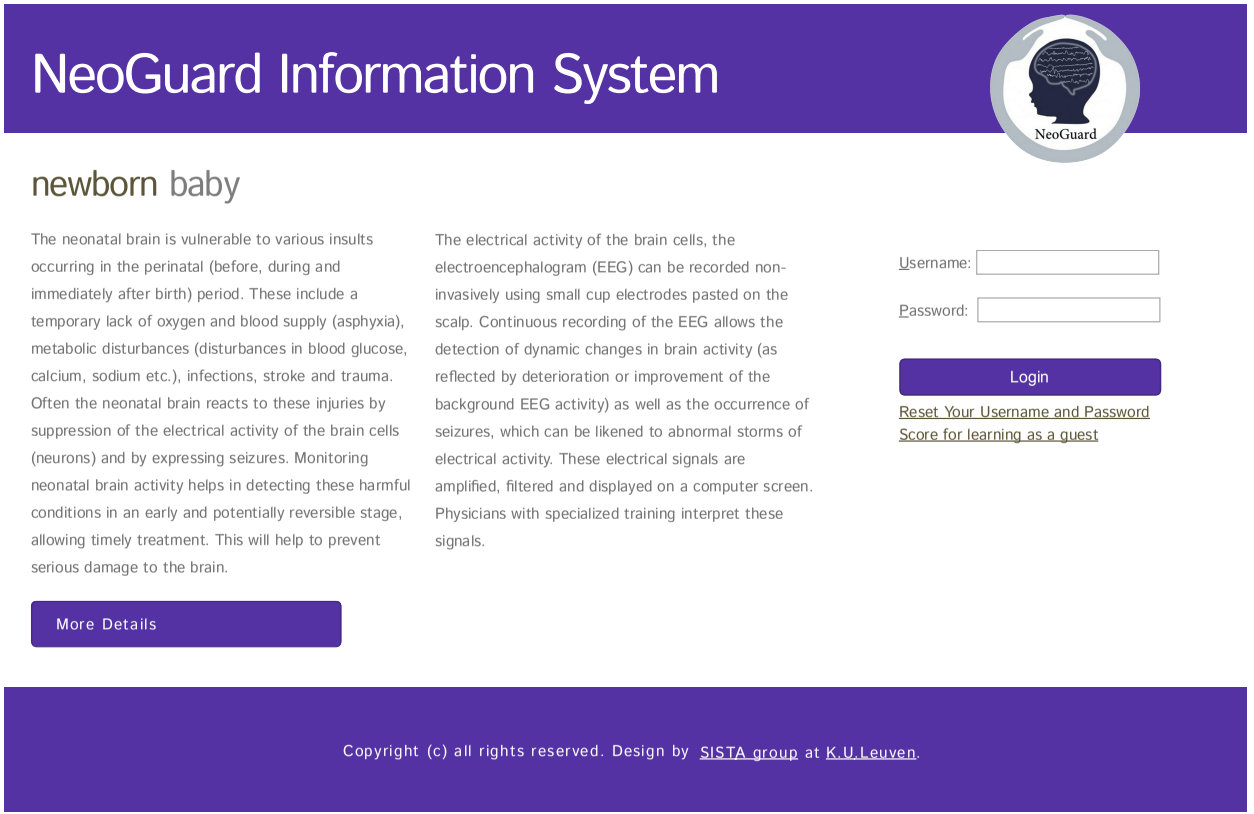


Figure 11.2: Welcome page of the NGIS

11.3 NeoGuard Scoring platform

NeoGuard scoring platform (NGSP) is a toolbox for expert EEG readers to investigate EEG segments and score them. This platform consists of two modules:

- MATLAB: This module has two important tasks: 1) loading EEG events and generating images for the website. These EEG events are selected based on the scored seizures by a primary neurophysiologist or are the outputs of a selected automated algorithm (see chapter 4). The images are generated with different sensitivities and different time frames. For instance, for the database used in chapter 4, it generates 1919 segments including seizures and artifacts from 71 term neonates. For each segment, 20 images (4 different sensitivities times 5 different time frames) were generated (see chapter 4). 2) exporting the scored segments from NGIS SQL database into the MATLAB database.
- Website: This is part of the NGIS and includes some web pages to show the segments and collect their labels from expert EEG readers. As mentioned in 11.1.2, the users who have access to the NGSP can open and score the segments by this platform. Figure 11.3 shows the main scoring page displaying an example event. The users can score the events by pressing one of the buttons located on top of the page as definite seizure, dubious seizure, and definite artifact.

11.4 Conclusion

In this chapter, we explained the general structure of the NeoGuard Information System. Furthermore, two modules of this system, namely the NeoGuard Data Collector for storing the recordings in a centralized database and the NeoGuard Scoring Platform for labeling EEG events by expert EEG readers have been described. These modules are used by the users who have a valid account in the system. The NeoGuard Learning Platform which is available for public users is described in the next chapter in detail.

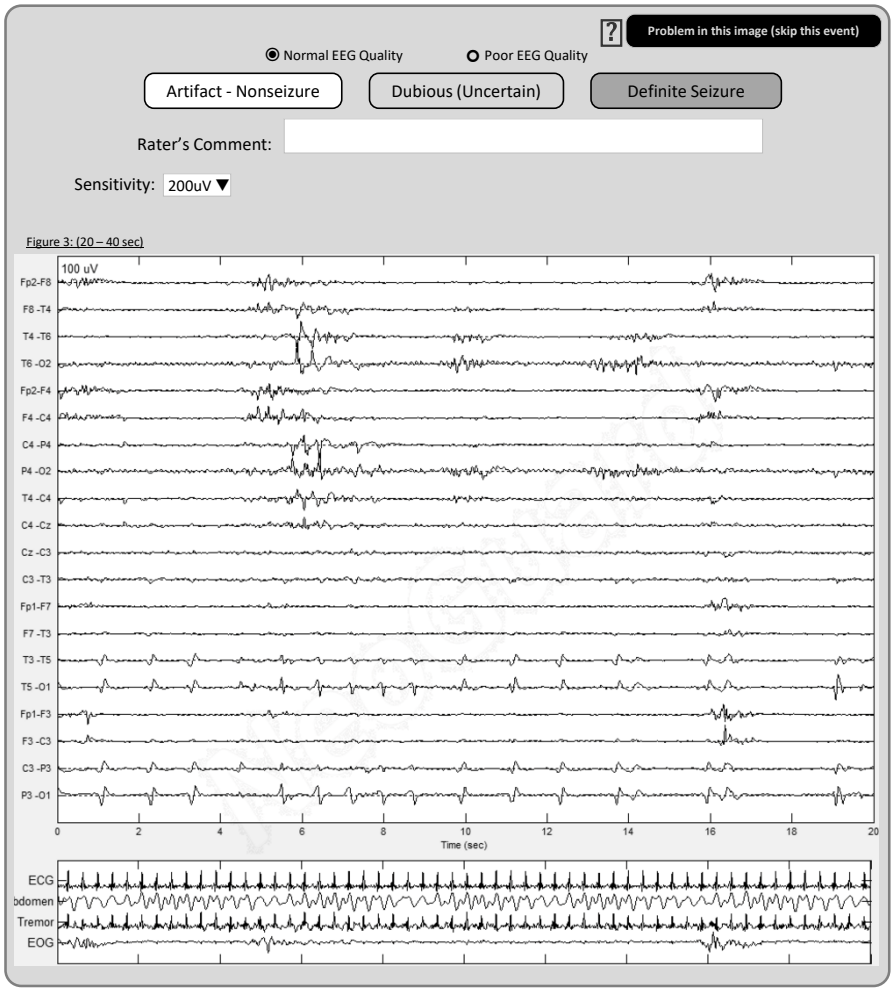


Figure 11.3: Events as displayed on the NeoGuard Website for scoring. Screenshot of image 3, a detailed view of the 20 - 40 seconds of the event.

Chapter 12

NeoGuard learning platform

In this chapter, we introduce and explain the NeoGuard public learning platform that can be used by trainees, tutors, and expert EEG readers who are interested to test their knowledge and learn from neonatal EEG-polygraphic segments scored by several expert EEG readers. For this platform, different clinically relevant segments recorded from term neonates, which are described in chapter 3, are used. A description of this publicly available platform, as well as its advantages and limitations, are given and discussed in this chapter.

12.1 Background

As mentioned in chapter 2, seizures commonly point to underlying brain injury in neonates and are generally associated with high mortality and morbidity. The prevalence of clinical seizures in newborn babies is estimated to be 2 to 3 per 1000 live births (0.2-0.3%), and is even higher (9-11%) in very low-birthweight or premature neonates [37,38,74]. The actual prevalence is likely to be underestimated as neonatal seizures are often subtle or subclinical, and clinical observation alone is known to be unreliable in their diagnosis [37]. Hypoxic-ischemic encephalopathy (HIE) is one of the most common and serious causes of neonatal seizures (30-53%). Other etiologies include metabolic disturbances, intracranial bleeding, infections, cerebral infarction as well as malformations. Neonatal seizures are brief lasting (few seconds to minutes) events of excessive and abnormal neuronal electrical activity and usually are regional or focal. Generalized tonic-clonic seizures are rarely seen in the first month of life [27,37]. Furthermore, the majority of neonatal seizures are clinically subtle or silent, so-called 'electrical only (electrographic) seizures', requiring the use of EEG monitoring for their accurate diagnosis and estimation of seizure burden. It is known that clinical observation alone results in marked underestimation of seizure burden, even by trained observers [37,80,141,225,226].

One of the best known and widely used EEG-based monitors is the cerebral function monitor (CFM^{TM}), which uses a compressed, rectified version of the EEG, called amplitude integrated EEG (aEEG) recorded using 2 to 4 EEG channels. The CFM generally displays about 5 hours of recording in one page and is used in many neonatal intensive care units (NICUs) because of its ease of use. However, it has been reported that aEEG often misses brief-duration (<30 s), low-amplitude, or focal/regional seizures. Furthermore, some types of artifacts can easily be misinterpreted as seizures in aEEG, especially by non-experts [47, 227]. Particularly in neonates with severe encephalopathy, treatment with high dose of anti-epileptic drugs (AEDs) may suppress EEG signals and result in brief-duration, low-amplitude seizures, which may be poorly detected by CFM.

Multi-channel Continuous EEG monitoring (cEEG) is the ideal method for detecting seizures. Generally, visual inspection of multi-channel cEEG complemented by video is considered to be the gold standard for diagnosing neonatal seizures, although their interpretation needs special knowledge and training. Moreover, cEEG is expensive and labor-intensive as compared to CFM [47,83,228,229]. One of the main challenges confronting the interpretation of EEG is the poor agreement of expert EEG readers for some types of seizures [67, 95], especially for 'dubious seizures'. Dubious seizures are low-amplitude, arrhythmic, paroxysmal EEG events lasting shorter or longer than

10 s, with no regular evolution of amplitude, frequency, and morphology. Such seizures commonly occur in neonates with severe encephalopathy or after treatment by high dose AEDs [67]. Seizure patterns with a higher disagreement among raters would also result in higher variability in treatment decisions. Furthermore, studies reporting automated neonatal seizure detection methods have highlighted the challenge of distinguishing brief-duration seizures from brief-lasting artifacts [3, 6, 149]. Lower inter-rater agreement between expert readers is also one of the hurdles for developing reliable automated seizure detection methods, when one realizes that the so-called "ground truth" is ambiguous.

It is likely that expert EEG readers working in one center have generally higher agreement compared to experts from other centers. The lack of a shared benchmark for neonatal seizure detection and scoring is a major issue. EEG trainers mostly use their local datasets for teaching trainees. This could contribute to the variability in pattern recognition and interpretation, especially when the characteristics of the patient populations differ. Creating a public platform for self-testing and comparing these tests with the opinion of different experts is essential for trainees and tutors. Such a platform can also be used by experts to share information, better define core characteristics of neonatal seizures, develop consensus and ultimately lead to improved standards of EEG interpretation, and impact quality of medical care and research.

As explained in the previous chapter, an online website, called NeoGuard Information System, has been developed by our research group and consists of three modules: 1) an EEG database, for collecting and storing EEG data from the participating teaching hospitals, 2) a scoring system, for presenting and scoring the EEG segments by our expert readers, and finally 3) the NeoGuard learning platform (NGLP), for sharing our experts' scores with trainees and other experts. The third module, NGLP, is described in this chapter. The main goal of the NGLP is to provide a platform teaching users how to annotate subclinical seizures from neonatal EEG-polygraphic recordings. This platform is open to researchers/clinicians with different levels of experience. The platform enables users to learn and compare their scores with the ones provided by our experts for EEG segments including definite seizures, dubious seizures, or seizure-like artifacts. In this chapter, the NGLP is introduced, its usage is described, and its added value, as well as limitations, are discussed.

12.2 Data description

12.2.1 Recording data

In this chapter, DB1-small, DB2-small, DB3, and DB4 are used. The 10 neonates without seizures are excluded from the first two datasets. The detailed descriptions about these datasets are given in section 3.2. Briefly, the EEG-polygraphic recordings from 71 neonates recorded in two medical centers, EMCR and ULZ, are used in this platform. The EEG signals are filtered between 1-20 Hz and bipolar channels are made according to the montage maps explained in section 3.2.

12.2.2 Rescoring strategy

The seizure segments were annotated in EMCR and UZL by two expert clinical neurophysiologists, one from each center. In total, 7% of recordings (25 h out of 353 h) were labeled as seizure. In order to construct the dataset of the training platform, some segments should be selected from the recordings and they should be re-scored by independent experts. If the segments are selected randomly from the recordings, it will lead to a very unbalanced dataset (7% seizure vs. 93% non-seizure). Instead, in order to have more clinically relevant segments, all annotated seizures have been selected, as well as the non-seizure annotations which were misclassified by an automated algorithm. Thus, the selected non-seizure segments usually include artifacts, dubious seizures that were not annotated as seizure according to the automatic detector. To this end, a previously developed heuristic neonatal seizure detector mimicking a human observer, developed in [4], was applied on the data. This algorithm is described in Appendix A.1 (Version 1). Then, the automatically detected segments were compared with the labeled seizures and, consequently, they were classified into two groups: falsely or truly detected segments. Next, the former group was merged and shuffled with the seizure group, forming the 'event pool'. Thus, the event pool includes all segments scored as seizure by the human experts, as well as all false alarms of a seizure detector (in total 1919 events). Finally, 3 independent expert EEG readers rescored all events and labeled them as 'definite seizure', 'dubious seizure', and 'definite artifact'. As a consequence, for each event, 4 scores are available, 1 provided by the primary and 3 provided by the secondary expert EEG readers. These labels are explained in the next section. Figure 12.1 schematically shows the mentioned scenario.

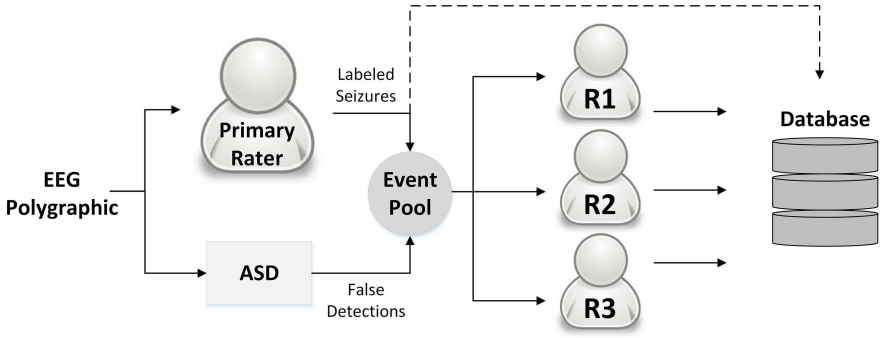


Figure 12.1: Schematic diagram of the multi-scoring scenario. R1-3 denote the 3 secondary raters and ASD is the automated seizure detector.

12.3 Method

As mentioned in the previous section, the goal of the NGLP is providing a platform for trainees and expert EEG readers to compare their knowledge and experience with experts of our group. To this end, a website was developed and is made publicly available. Anyone can open the website, choose the number of events that he/she wants to score, define some of the EEG review settings, and start to score. Each of these steps is explained in the following sub-sections.

12.3.1 Scoring labels

Each event can be scored as one of the following labels:

1) Definite Non-seizure/Artifact: our experts used this label when no pattern suspicious for seizure is present or when the EEG segment only expresses artifacts, such as those originating from the ECG, EOG, EMG, respiration, etc. Figure 12.2 shows a definite artifact which was scored as non-seizure by the primary clinical neurophysiologist and all secondary raters as well.

2) Definite seizure: our experts used this label when there is a clear pattern of rhythmic oscillations, repetitive sharp waves, or mixture of both on the EEG channels which lasted more than 10 seconds. This pattern should have clear evolution of frequency, amplitude, or morphology and it should look different from the ongoing background activity. Furthermore, the rater believed that clinicians will initiate or become highly encouraged to initiate treatment with anti-epileptic drugs (AED) because of this event. Figure 12.3 displays a definite

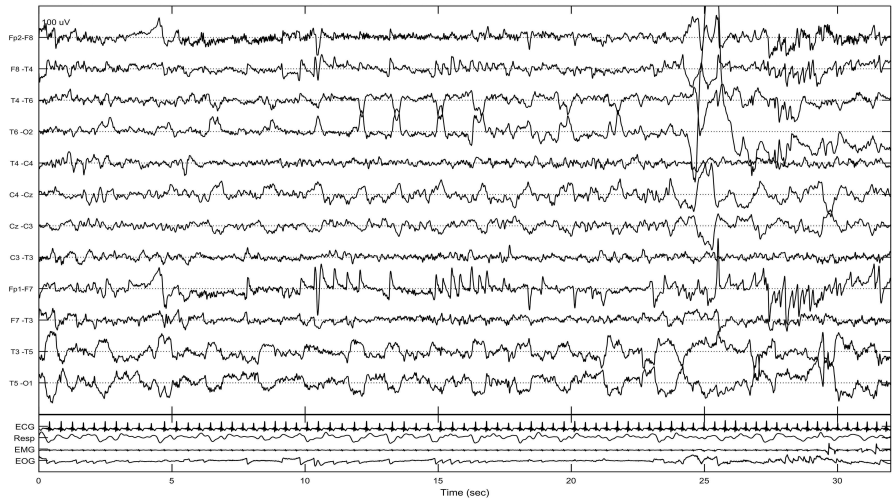


Figure 12.2: A definite artifact that was scored as non-seizure by the primary neurophysiologist and all secondary raters.

seizure lasting for 19 s which was scored as seizure by the primary and all secondary raters.

3) Dubious seizure: this label was used by our experts when there are dubious repetitive patterns in the EEG signal without sufficient characteristics of a definite seizure. This doubt usually occurs for very low-amplitude patterns with no clear evolution of amplitude, frequency, and morphology. They generally appeared in neonates with serious brain dysfunction and usually was associated with poor clinical outcome. For this type of events, the clinicians will become convinced and initiate the AED treatment only if the pattern repeats itself and leads to a high seizure burden. Therefore, in the raters' view, a single dubious pattern should be considered neither as a seizure nor as a non-seizure/artifact. Figure 12.4 shows a dubious seizure detected by the primary neurophysiologist occurring at Cz electrode for 15 seconds with arrhythmic oscillations and no clear evolution of amplitude, frequency, or morphology. This event was scored by the three secondary raters as dubious, dubious, and definite seizure.

12.3.2 Level of agreement

The events are partitioned into 3 classes based on the agreement of our expert raters. To this end, the most frequent scores of each event are counted and its

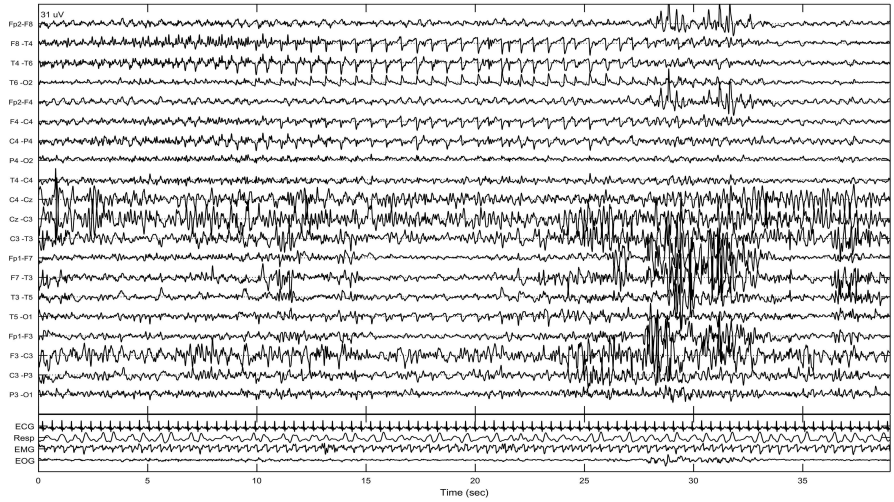


Figure 12.3: A definite seizure, starting from second 10 and lasting for 19 seconds over the right temporal region. This event was scored as 'seizure' by the primary neurophysiologist and all secondary raters.

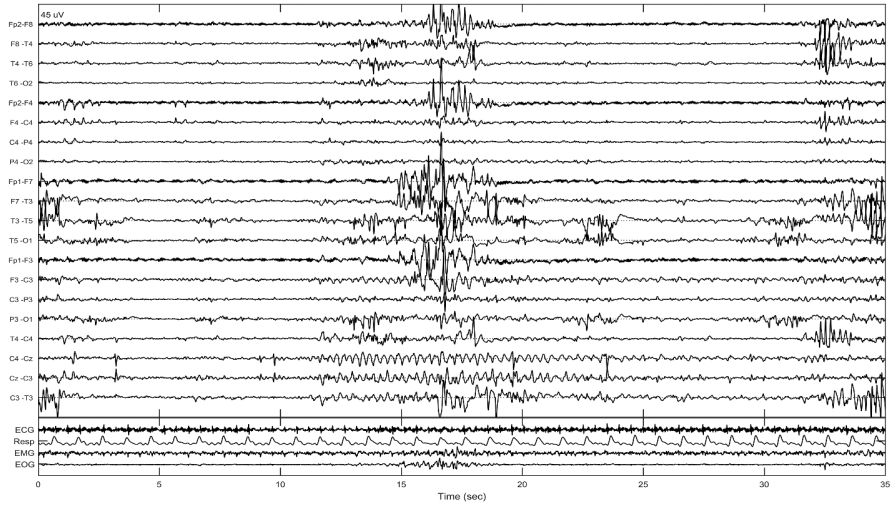


Figure 12.4: A dubious seizure starting from second 10 and lasting for 15 seconds, with arrhythmic oscillations, without clear evolution in morphology, at Cz electrode, with an amplitude of $25\mu V$. This event was scored as 'dubious seizure' by the primary neurophysiologist. It was then rescored as (Dsz, Dsz, Sz) by the secondary raters.

percentage, p%, defines the agreement of each event. Based on p, the 3 classes are defined as follows: poorly agreed ($p \leq 25\%$), moderately agreed ($25\% \leq p \leq 75\%$), and highly agreed ($p \geq 75\%$). A user wanting to start scoring the events can choose the event class, or choose the 'Random' option, which does not take the agreement into account.

12.3.3 Measuring raters' agreement

In order to score the user, the agreements between the user and our expert raters are measured by two inter-class correlation coefficients, which have commonly been used as inter-rater agreement metrics: I) joint probability of agreement, and II) Cohens' kappa. Both metrics are calculated between a couple of raters. The joint probability of agreement is computed as the number of events over which both raters agreed, divided by the total number of events. This metric does not take into account the fact that two raters may sometimes have similar opinions only by chance. The kappa is a corrected version of the joint probability of agreement and is calculated by

$$k = \frac{p_0 - p_e}{1 - p_e} \tag{12.1}$$

where p_0 is the joint probability of agreement and p_e is the hypothetical expected probability of chance agreement measured by the following equation.

$$p_e = \sum_{j=1}^{N} p_j^1 p_j^2 \tag{12.2}$$

where N is the total number of labels (=3 in this study) and p_j^r is the marginal probability of the rth rater for the jth class [103,230]. For example, assume that a rater scores 70%, 20%, and 10% of the events as 'Seizure', 'Dubious', and 'Non-seizure' respectively, whereas these values equal 60%, 10%, and 30% by the 2nd rater. In this case, p_e equals $(70\% \times 60\%) + (20\% \times 10\%) + (10\% \times 30\%) = 47\%$. It shows that for 47% of these imaginary events, the raters may agree only by chance. As it is clear in (12.1), this chance probability is removed from the joint probability in the kappa statistic. Both these metrics are calculated between the user and each of our expert raters and reported at the end of the scoring process.

12.3.4 Storing the users' Info

All scorings of expert or non-expert users are saved in the database of the NGLP. A user wanting to start the scoring process needs to agree that all entered inputs, including the user information (see figure 12.5), as well as the scorings, will be saved in our database and may be used for future research.

12.3.5 Start page

In order to start the scoring, the user should enter some initial information and settings. First, a CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) security code should be entered and validated (figure 12.5). CAPTCHA code is a worldwide Turing test to prevent automatic codes or malware from connecting to the website to make problems or steal the data. Second, the level of experience should be selected from 'Non-expert EEG Reader', 'Neurologist (Expert EEG Reader)', 'Neurophysiologist (Expert EEG Reader)', 'Pediatrician (Expert EEG Reader)', or 'Other (Expert EEG Reader)'. This information can help us to further refine the analysis of the current users' scorings. Third, the number of events that the user wants to score should be chosen (between 1 and 1000) and the level of agreement of upcoming events should be selected as explained. Users who would like to have us contact them may choose to enter their email or name. Finally, the user must allow our team to use the inputs for future analysis. Figure 12.5 shows a screenshot of the start page.

12.3.6 Scoring page

The user scores each event based on 5 images as follows: image 1 displays the EEG-polygraphic signals of the whole event in seconds. Images 2, 3 and 4, show the details of signals in three 20s-pages from 0-20 s, 20-40 s, and 40-60 s respectively. If the signal lasted more than 60 s, the detail images only show the beginning 60 s. On the other hand, if the signal lasted less than 20 s, images 3 and 4 will not appear (if the signal lasted less than 40 s, image 4 will not be available). Finally, image 5 gives a wider view of the event, starting 100 s before the start of the event and ending 100 s after the end of event. It can help the user to place the event in its context and help check whether it is an ictal event or part of a longer artifact. On top of the page, there are three buttons corresponding to the three labels: 'Artifact/Non-seizure', 'Definite seizure', and 'Dubious seizure'. Furthermore, the user can choose the sensitivity of the EEG signals between '20μV', '50μV', '100μV', and '150μV', except for

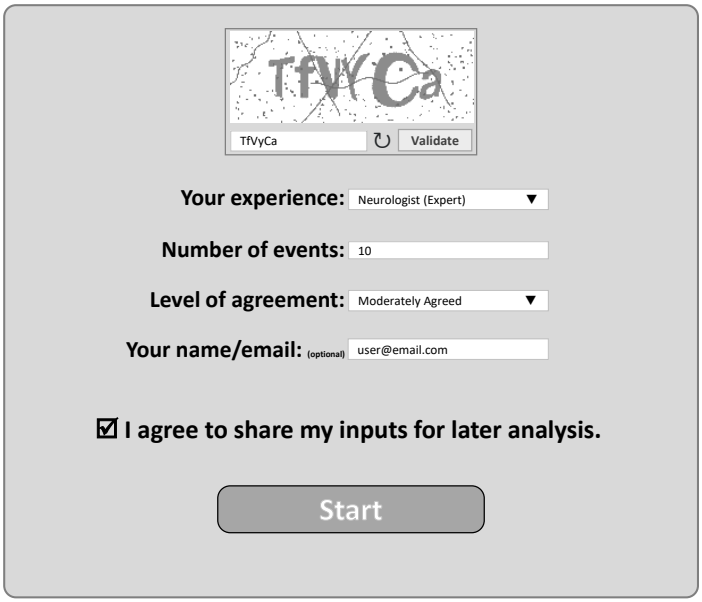


Figure 12.5: A screenshot of the NGLP start page

some specific events for which very small or very high sensitivity is needed. Some shortcuts are implemented for jumping between the images, scoring the events, and changing the sensitivity to speed up the scoring process. These are shown on this page. All images include a light 'NeoGuard' watermark in the background. Figures 12.6 and 12.7 are two screenshots of the scoring page showing respectively the upper and lower views of a scored seizure which lasted 50 seconds.

12.4 Results

When all events are scored by the user, the results page will be displayed with two tables. The first table shows all scorings in detail, with each row showing the user's choice of label and the corresponding scores of the experts. This allows the user to easily compare the scoring of each event. Furthermore, by clicking on the 'eye' icon in the first column of each row, one is able to review the images of the corresponding event. The second table displays the agreement metrics, Joint-probability and Kappa, between the user and each of our expert readers. If the number of scored events is less than 60, a warning will be shown

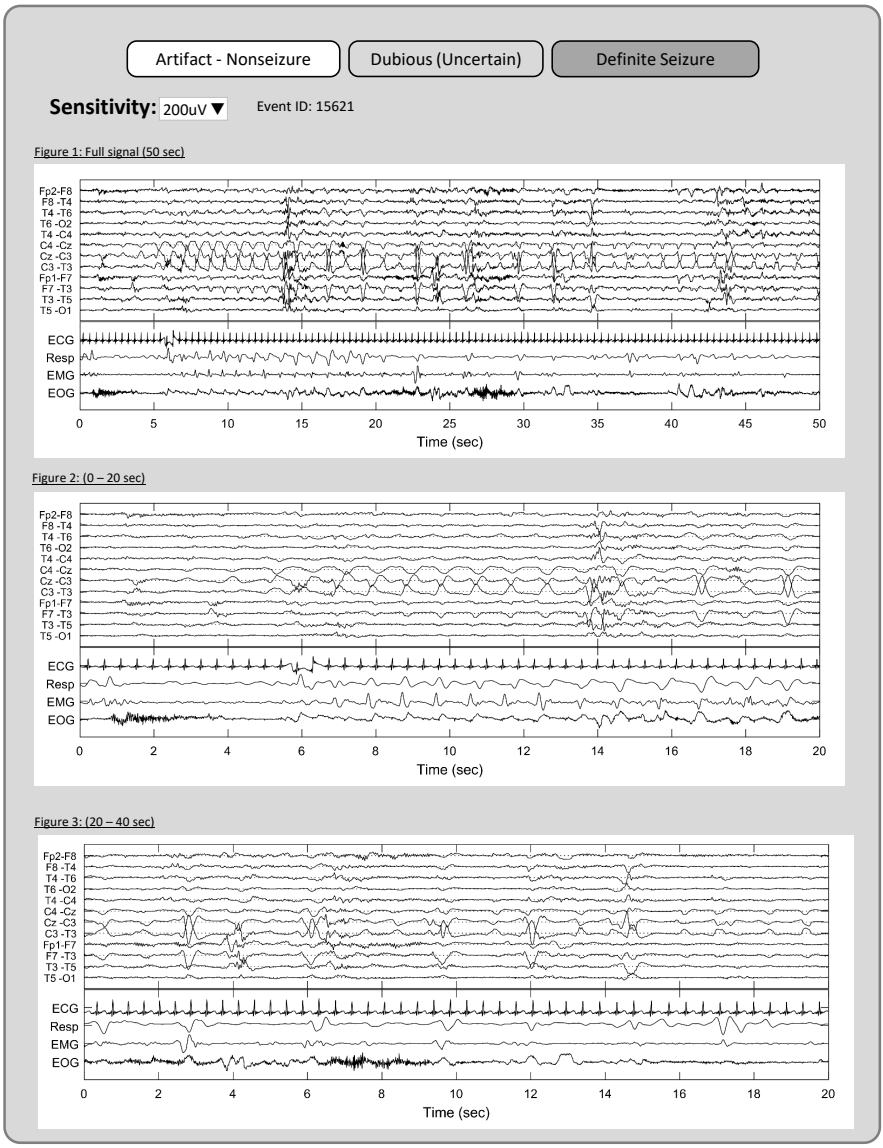


Figure 12.6: The view of the upper part of the scoring page including the scoring buttons, and images 1-3. In each image, the EEG and polygraphic signals are shown.

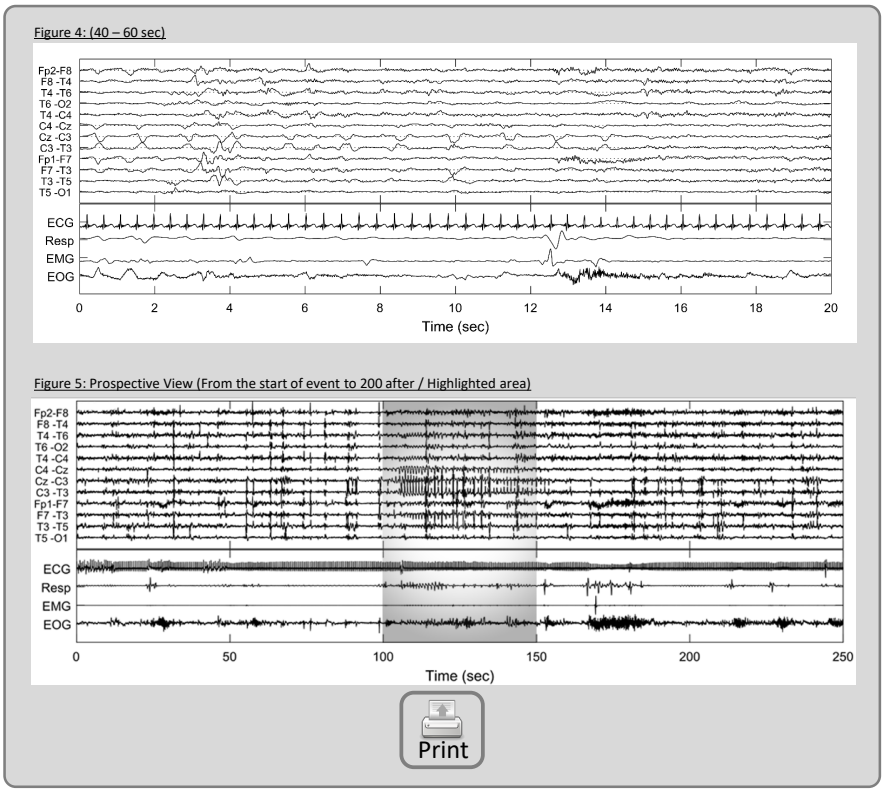


Figure 12.7: The view of the lower part of the scoring page including the images 4-5 and the Print button. In each image the EEG and polygraphic signals are shown.

to indicate that the agreement metrics may not be accurate enough due to an insufficient number of scored events. Figure 12.8 shows an example of a user who scored 10 events. In the first table, the green color shows the events with agreement between the expert readers and the user and the red ones shows events where they differ. The 'eye' icons which allow review of the corresponding EEG images, are located in the first column.



You have scored 10 events.

Your userid: 625136420

The details of your scoblack events

	Num	Event ID	Your Score	Rater #17	Rater #21	Rater #30	Rater #34
	1	1429	**Definite Non-Seizure**	Definite Non-Seizure	Definite Non-Seizure	Definite Non-Seizure	Definite Non-Seizure
	2	1629	**Definite Non-Seizure**	Definite Seizure	Definite Seizure	Definite Seizure	Definite Seizure
	3	219	**Dubious Seizure**	Definite Seizure	Definite Seizure	Definite Non-Seizure	Definite Seizure
	4	1336	**Definite Non-Seizure**	Definite Seizure	Definite Seizure	Definite Seizure	Definite Seizure
	5	1634	**Definite Non-Seizure**	Definite Seizure	Definite Seizure	Definite Seizure	Definite Seizure
	6	2309	**Definite Seizure**	Dubious Seizure	Definite Seizure	Definite Seizure	Dubious Seizure
	7	428	**Definite Non-Seizure**	Definite Non-Seizure	Definite Non-Seizure	Dubious Seizure	Definite Non-Seizure
	8	418	**Definite Seizure**	Definite Seizure	Definite Seizure	Definite Seizure	Definite Seizure
	9	2210	**Definite Seizure**	Dubious Seizure	Definite Seizure	Definite Seizure	Definite Seizure
	10	328	**Dubious Seizure**	Dubious Seizure	Dubious Seizure	Dubious Seizure	Definite Seizure

The agreement of the reference raters' scores with yours

Rater Id	Number of Events	Joint-Probability of Agreement (Rater - you)	Kappa (Rater - you)
17	10	55.00 %	0.08
21	10	65.00 %	0.34
30	10	60.00 %	0.21
34	10	55.00 %	0.01

* Warning: you scored only 10 events which is not statistically sufficient.
The agreement metrics may be not accurate!!

Print Restart

Figure 12.8: A screenshot for the result of a user who scored 10 events

12.5 Discussion and Conclusion

Several reports have shown that there is disagreement between expert EEG raters when they score neonatal seizures recorded on cEEG, particularly for lower amplitude patterns without very clear-cut evolution. This raises the issue that although there is a clear definition of neonatal seizure in theory, there is not yet a clear-cut gold standard for the interpretation of the EEG in practice, at least for the problematic presumed ictal patterns. Therefore, it seems essential for the EEG trainees who are learning to diagnose neonatal seizures to become familiar with the opinion of experts working in different centers. In this chapter, a novel learning platform was introduced and explained. This tool can be used by EEG trainees to improve their abilities and by tutors in their classes, workshops, or seminars, to demonstrate practically how to recognize seizures.

One of the advantages of this online system is that the expert readers, and events can be updated or increased easily in the future. Although, the NGLP currently includes the scorings of four expert EEG readers from our participating centers, we are keen to include more EEG samples and volunteer expert EEG readers from other centers to enrich the database and the NGLP. The NGLP is able to connect to our database and automatically add newly recorded segments. Subsequently, as soon as the secondary readers score them, they will be available in the system for the NGLP users.

The other advantages of this system is that as soon as a statistically sufficient number of trainees have used this platform, extra analyses are possible on the trainee' scorings. For instance, the following questions can be addressed: what is the correlation between the level of experiment and agreement of the user? What is the correlation of wrong/true scorings with the duration of events, etiologies, agreement of experts, types of seizures, etc.?

However, the NGLP also has some limitations. Firstly, the users cannot watch the video of the neonates. As it is explained above, the gold standard of detecting seizures is visual inspection of cEEG-video. However, sharing the videos of neonates on the web is clearly not possible because of privacy and ethical issues. Secondly, the users are constrained by the provided EEG settings and cannot change the montage of the EEG signals or cannot move the window over time. These limitations may make the scoring process artificial and more difficult when compared to the normal scoring practice of cEEG used by experts. In addition, due to these limitations, the agreement measures could be underestimated. Nonetheless, the NGLP is the first platform which publicly shares the scorings of several expert EEG readers from different centers on the web. It is hoped that some of the limitations mentioned are overcome in the future versions of the NGLP.

Furthermore, in future versions, the users will have a personal profile for reviewing the previous scorings, sharing their knowledge and opinions with other users and experts, as well as customizing the website settings. Also, some interactive training tools and educational content can be added to the website.

12.6 Intellectual property

The platform is publicly available at the NeoGuard website "www.neoguard.net". The users can freely copy, print, and distribute the images, events or the results of this platform, provided the author and original source are properly cited.

Chapter 13

NeoGuard: A real-time monitor for NICUs

In this chapter, the NeoGuard real-time monitor is described. This monitor was developed as the first prototype bedside monitor of the NeoGuard project. It includes different algorithms and processing units for automated EEG analysis which are explained and discussed in this chapter.

13.1 Introduction

In neonatal intensive care units (NICUs), there is an unmet need for around-the-clock interpretation of continuous EEG (cEEG) signals. However, due to the lack of continuous technical and clinical support, cEEG is not being implemented in many NICUs. One of the best possible solutions is developing a real-time EEG monitor that is equipped with accurate automated algorithms, which can help doctors, nurses, and care givers as a supportive tool. To this end, a prototype has been made and called NeoGuard real-time monitor to record and analyze multi-channel cEEG. In this system, different algorithms have been implemented to be tested in real practice including a seizure detector, inter-burst interval estimator, amplitude-integrated EEG (aEEG) estimator. These algorithms and the graphical user interface as well as its limitations are explained and discussed in this chapter .

13.2 Materials and methods

NeoGuard monitor is composed of three main modules (figure 13.1) which were developed in three centers:

1. the hardware, developed in KU Leuven, ESAT-MICAS[1]; it is an EEG machine including standard sockets to connect to common EEG electrodes, an electronic board to amplify and digitize the recordings, and a communication device to send EEG data to a computer.

2. the back-end core, developed in KU Leuven, ESAT-STADIUS[2] (by this PhD candidate); it is the main module which receives EEG and video signals, records and analyses them by different algorithms, and prepares the required data for the GUI. Furthermore, it processes the requests of the user sent by the GUI and provides the outputs or changes the corresponding settings.

3. the front-end GUI, developed in Ghent University, TELIN[3]; it includes the graphical user interface to monitor the EEG signals and video frames and display alarms and processed information.

[1]MICAS: Division of Microelectronics and Sensors
[2]STADIUS: Centre for Dynamical Systems, Signal Processing and Data Analytics
[3]TELIN: Department of Telecommunications and Information Processing

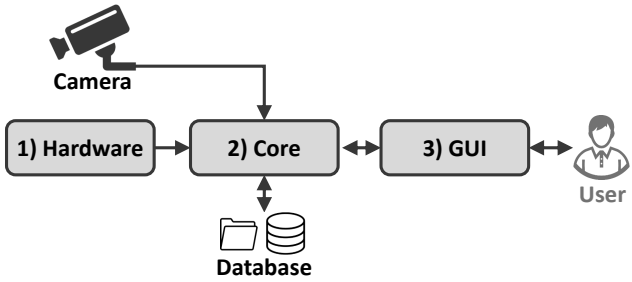


Figure 13.1: The main modules of the NeoGuard real-time monitor

13.2.1 NeoGuard hardware

The NeoGuard hardware includes a box and a USB [4] Bluetooth dongle (figure 13.2). The box has 18 standard sockets to connect to 14 EEG electrodes (including $F_{P1,2}$, $F_{3,4}$, $C_{3,4}$, $T_{3,4}$, $P_{3,4}$, $O_{1,2}$, C_Z, F_Z), 2 ECG electrodes, a reference electrode, and a bias electrode. An OpenBCI [5] EEG device (Cyton - 16-Channels) has been utilized in the core of this system to record, amplify, and digitize the EEG and ECG signals. The digitized data is then transmitted to the Bluetooth dongle and sent to the computer through USB port.

This hardware has 125 Hz sampling frequency and does not include a digital or analog filter. A rechargeable Lithium-ion battery has been used in the box that can supply the hardware to record more than 72 hours continuously.

13.2.2 The core

The core is the back-end of the monitor and is responsible for all processing functions. It is composed of 5 main layers (see figure 13.3) as follows:

- Hardware Layer: for connecting to the hardware, setting initial configurations, and receiving and buffering the recorded packages
- Video Layer: for connecting to the camera, setting initial configurations, receiving and buffering the video frames

[4] Universal Serial Bus
[5] Open Source Brain-Computer Interfaces

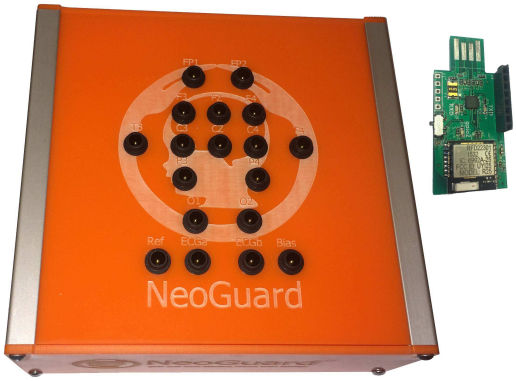


Figure 13.2: The prototype of the NeoGuard hardware

- Data Layer: for synchronizing the video frames and biomedical signals, decoding them for compressed saving, filtering the signals and removing the power-line noise and zero drifts.
- Algorithm Layer: for running the algorithms that are listed in the next subsection
- Communication Layer: for preparing and sending the data and alarms to the GUI. Furthermore, it processes the requests coming from the GUI and calls proper functions in the data and algorithm layers when needed.

Implemented algorithms

In this version of the NeoGuard monitor, 4 algorithms have been implemented as follows:

- Seizure detector: the improved version of the heuristic algorithm proposed in [5]. This algorithm is also briefly described in appendix A.1.2.
- aEEG estimator: the estimator of the amplitude-integrated EEG (aEEG) from the recorded raw EEG. This algorithm is described in appendix A.2.
- IBI calculator: the algorithm for calculating the duration and amplitude of inter-burst intervals (IBI) that was proposed in [9]. It is described in appendix A.3.2.

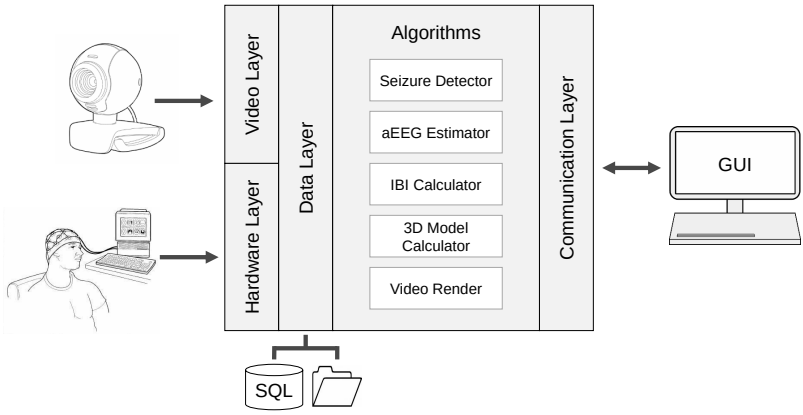


Figure 13.3: The main units of the NeoGuard core

- 3D model calculator: the algorithms for extracting the power of the α, β, δ, and θ frequency bands for each EEG channel in order to enable their display on the 3-dimensional brain model in the GUI.

Programming languages

The software of this monitor including the core and the GUIs has been implemented by C++ v.11 in Qt platform v.15. The algorithms were initially implemented and tested in MATLAB and then translated to C++. This system has been developed for Linux operating systems (tested on Ubuntu v.14.04). Since Qt is a cross-platform framework, it seems logical that the software is simply transferable to other operating systems, although it has not been yet performed.

Furthermore, in order to save the structured data, such as the neonates' information, the detected seizures and estimated amplitude/duration of IBIs, SQL scripting has been used in MySQL database.

13.2.3 Graphical user interface

Figure 13.4 shows the main screen of the NeoGuard monitor. As is clear from this figure, 6 blocks are visible in this form as follows:

- aEEG plot (left-top): It plots the estimated aEEG signal. The user is able to choose the proper channels of aEEG in the Settings form. As a common standard, every hour of aEEG signal is plotted in 6 cm. However, the user is able to zoom it in/out by pressing the corresponding buttons. He/she can also jump to the next or previous pages to observe different parts of a recording. The amplitude of the aEEG is also common to be semi-logarithmic, which means it is linear between 0 to 10 μV and is logarithmic for the amplitude above 10 μV. Furthermore, the detected seizures are marked in pink color on the aEEG plot (see figure 13.4).
- EEG plot (left-middle): It plots the mono/bi-polar EEG and ECG signals. The user can customize them in the Settings form.
- IBI plot (left-bottom): It plots the amplitude (in green) and duration (in blue) of the detected inter-burst intervals for every hour.
- Alarm (right-top): It shows the information about the latest detected seizure including the exact time of seizure onset and its duration. If no seizure has occurred or the user has clicked on the alarm (to reset it), the alarm will be in green color with no description.
- 3D model (right-middle): This is an artificial 3D brain model, which does not need MRI data. The EEG electrodes are marked on the brain. The color of the brain will be continuously changed locally based on the power of the EEG signal in the selected frequency band (α, β, δ, or θ).
- Video player (right-bottom): This is a video player which shows the synchronized video frames with the EEG recordings.

The user is also able to pause the monitor, when the recording is still continuing, and move the time cursor to an arbitrary moment. Furthermore, the user can start streaming from the arbitrary moment in offline mode or jump to the real-time stream when needed. In all these cases, the plots of the EEG, aEEG, IBI, and video frames are fully synchronized.

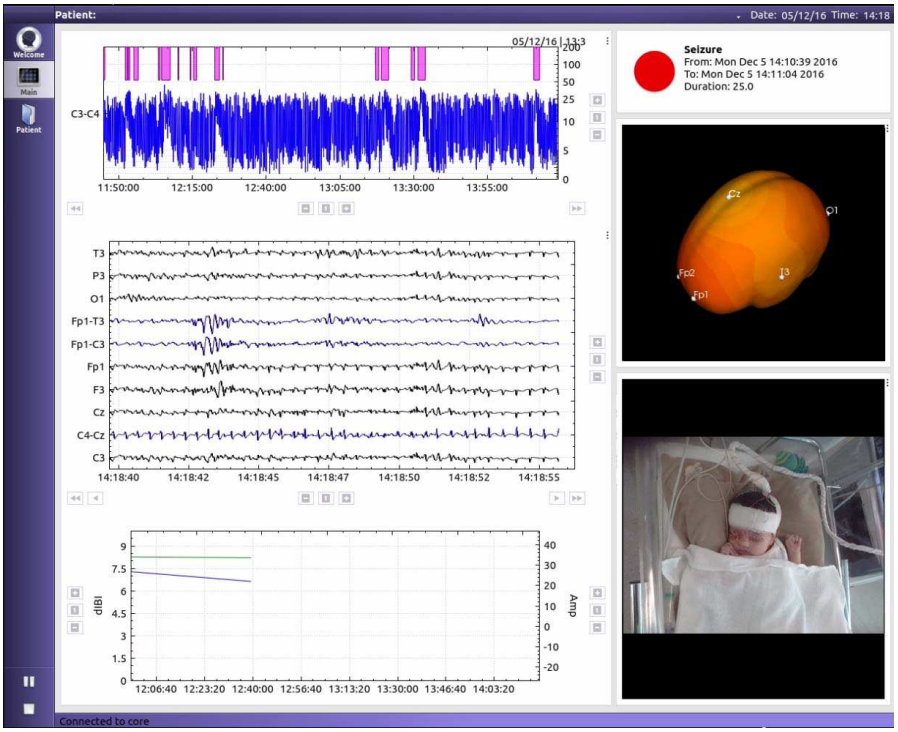


Figure 13.4: The main screen of the NeoGuard Monitor

13.3 Conclusion

In Neonatal Intensive Care Units (NICUs), using a real-time EEG monitor equipped with automated algorithms is extremely needed. Thus far, there is no any established monitor in the NICUs, although many automated algorithms for EEG analysis, such as seizure detection, sleep stage classification, inter-burst interval (IBI) classification, have been proposed in the literature. The main reason can be the fact that most of the proposed algorithms were developed and tested based on offline data under controllable conditions. However, in real practice, there are many unseen parameters which should also be taken into account. In this project, a prototype of a real-time EEG monitor has been developed and implemented. This monitor records the EEG, ECG, and video of a neonate and analyses the EEG signals in order to calculate the amplitude-integrated EEG, detect seizures, and estimate the duration and amplitude of the IBIs.

However, this prototype has still some limitations which should be solved in the final product:

- The hardware can only record continuously for about 2 hours because of temperature problems.
- Although the sound can be recorded along with the video frames, it cannot not be transmitted from the core layer to the graphical user interface due to the used TCP/IP [6].
- The users cannot change the filter settings while the recording is continuing.
- The artifact removal algorithms have not yet implemented in the system. Thus, the performance of the methods is not high in the presence of artifacts or when the quality of the signal is low.

[6]TCP/IP: Transmission Control Protocol/Internet Protocol

Chapter 14

NeoGuard: A cloud-based remote monitor for NICUs

This chapter introduces and describes the NeoGuard cloud-based monitor. This web-based platform is designed for practical and clinical validation of the developed algorithms without the need of language translations. In other words, the developed algorithms can be directly applied on recording data in their native languages (e.g. MATLAB, Python, C++, etc.). The advantages and limitations of this platform are also discussed in this chapter.

14.1 Background

As mentioned in the previous chapters, there is a growing need in neonatal intensive care units (NICUs) for an automated neonatal brain monitor. This monitor should be equipped with required supportive tools and alarms, such as seizure detector, aEEG display, inter-burst interval estimator, etc. Developing these algorithms requires some important steps including development/training, testing, and clinical validation. The latter one means that the algorithm should be implemented in a framework to be tested in a real environment for numerous patients. In some cases, like neonatal seizure detection, the algorithm must also be tested in several centers to guarantee that it can practically work with different EEG machines and in various conditions. However, this clinical validation encounters some difficulties. For instance,

- the algorithms should be translated from the original, scientific languages, like MATLAB or Python, into more advanced, low-level, multi-thread languages, like C++, C#, or JAVA, to be used in practice. This translation is usually very time-consuming.
- The algorithms should be able to import data from different EEG machines with different standards which is not always straight-forward.
- Since the algorithms might have several unexpected bugs when they run in a practical environment in different centers, fixing and debugging them is very difficult.

In order to overcome these difficulties, the NeoGuard cloud-based monitor is developed. The goal of this monitor is providing a framework for clinical validation of algorithms in the NICUS. This framework has three key features as follows:

1. It is a cloud-based system and only one version of the algorithms needs to be installed on a server. Therefore, in the debugging phase, when an algorithm is being corrected, all instances running in different centers are immediately updated.
2. Unlike the local desktop computers of clinical centers, the server used for this platform is fully controllable by the developer. Thus, in this remote system, the algorithms can run in their native languages, MATLAB, on the server. Therefore, there is no immediate need to translate the algorithms into a low-level language.
3. Except for the 'hardware layer' located in the 'data uploader' which is explained below, all other modules of the platform, from the data

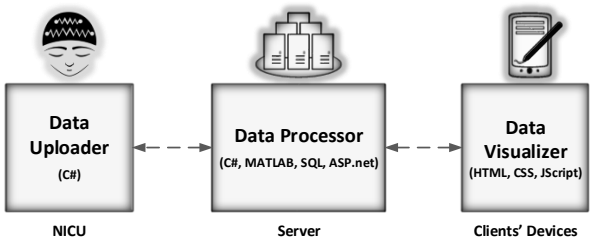


Figure 14.1: The main modules of the NeoGuard cloud-based system

transmitter to the algorithms to the data visualizer, are hardware-independent. It means that the developers of the algorithms can easily update, improve, and test their new algorithms in new centers, without the need to knowing the details of the EEG recording devises.

14.2 Design

This platform is composed of 3 modules as follows (see figure 14.1):

1. Data uploader: this module runs on a computer in the recording center. This computer should be connected to the EEG device (directly or via local network) and should also be connected to the web. The main tasks of this module are I) reading EEG from the device, II) converting the data to the European data format (EDF) and anonymize them, and III) uploading the EDFs to the server. This module was developed in C#.net.
2. Data processor: this module is the brain of the platform and responsible for collecting the data, running the algorithms, and providing the data and alarms for the GUI. It works on a virtual PC running on a Microsoft Azure cloud-based platform. This module includes 4 applications/layers collaborating together.
    - Request manager: as soon as the data is received from the data uploader, this layer starts the pipe-line. It manages the algorithms and their inputs. Then, it processes their outputs, saves them in a database and provides the information for the back-end visualizer. It was developed in C#.net.
    - Algorithm core: this external application (app) is a MATLAB compiled executable file which includes all required algorithms. This

app is started by the request manager when needed. The output of this app is several 'Comma Separated Values' (CSV) files. This module was wrapped into a C# app to be compatible with other .NET(dotnet)-based modules and applications.

- Database: the database and required functions have been implemented on Microsoft AZURE SQL Server via SQL scripting.
- Back-end visualizer: This is a server-side script which provides the data for the data visualizer in collaboration with the request manager. This module was developed in ASP.net.

3. Data visualizer: this module is the front-end (client side) of the platform. All processed information, signals, and alarms will be shown on the GUI of this module. It is developed for internet browsers (Chrome, Safari, FireFox, Internet Explorer, Opera) using HTML, java script, and CSS. Therefore, the users can easily see the signals on their laptops, tablets, or smart phones from everywhere.

14.3 Developed algorithms

On this platform, 6 algorithms have been compiled and being used so far as follows:

- aEEG: this algorithm was made based on the approach explained in [231] for estimating amplitude-integrated EEG (aEEG) from the recorded raw EEG signals. This algorithm is also described in appendix A.2.
- Improved heuristic seizure detector: this is a heuristic seizure detector proposed in [5]. This algorithm is described in appendix A.1.2.
- Multi-stage seizure detector: this is the algorithm explained in chapter 6 for automated neonatal seizure detection.
- IBI estimator for term babies: this algorithm was developed by Matic et al. [9] for calculating the duration and amplitude of the inter-burst intervals (IBI) for term newborn babies. This algorithm is described in appendix A.3.2.
- IBI estimator for preterm babies: this algorithm is for estimating the IBIs, particularly for preterm infants, using the 'line length' method developed by Koolen et al. [8]. This algorithm is described in appendix A.3.1.

- Sleep stage classifier: this algorithm is for classifying quiet sleep from non-quiet sleep using convolutional neural networks explained in chapter 9. Although this algorithm is already implemented in the server-side code, the data visualizer cannot display its outputs yet.

14.4 Graphical user interface

There are two important graphical user interfaces in this system:

- the Uploader form: this is the main form for the data uploader (figure 14.2). Using this form, the clinical users can choose an EEG device and the corresponding recording file, can select a case ID and start uploading. In order to access this GUI, the user should enter a valid username and password.
- the data visualizer: this is the main GUI of this system. After entering a valid username and password, the user can enter to this form. In this form, the aEEG signal, as well as IBI trend, list of detected seizures, and the histogram of IBIs will be plotted. In the current version the raw EEG is not shown. Figure 14.3 illustrates 3 hours EEG recordings of a term neonate opened in a Chrome browser. The upper trend shows the aEEG while the lower one displays the IBI trend. On the right panel, the detected seizures are listed. The user can adjust the resolution of the plots so that each hour of signal be plotted in 6 cm, as is common. Figure 14.4 shows the histogram of the detected IBIs from the last hour of recording. By pressing the arrow buttons above the form, the user can also see the histogram from the previous or next hours (if they are available). It has been shown that the evolution of such a histogram over time can be correlated with the development of a premature brain [8].

14.5 Current centers and devices

The NeoGuard cloud-based monitor has already been installed in UZ Leuven NICU and has been being tested on the local hospital network. In this center, it can read from EDF/EDF+ or from the recordings of the Brain RT EEG system (OSG BVBA Rumst, Belgium). To this end, an extra MATLAB-based converter was implemented and used in the data uploader. The second center in which the system is being installed is ZNA (Koningin Paola Kinderziekenhuis, Antwerp, Belgium). For this center, the system needs to connect to Medatec



Figure 14.2: The GUI of the data uploader

EEG devices (Medatec, Belgium) via socket programming as well as to Nihon Kohden EEG recorders (Nihon Kohden Corp., Tokyo, Japan) which is under construction. In the near future, it is hoped that this system is increasingly being used in other centers.

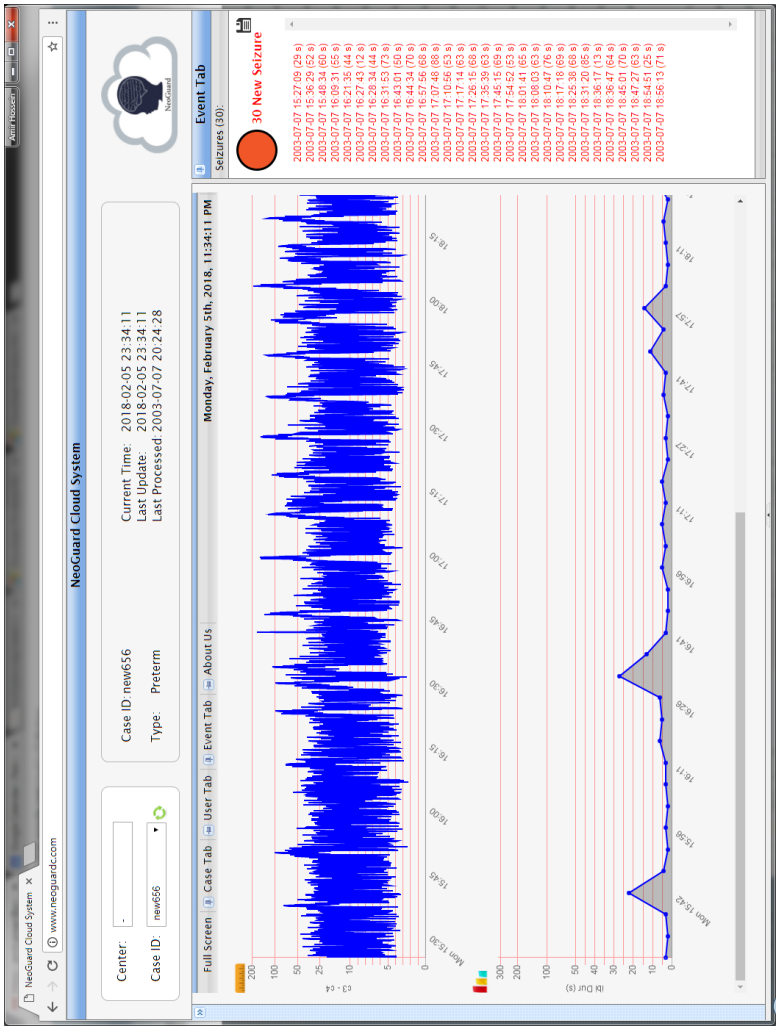


Figure 14.3: The GUI of the data visualizer

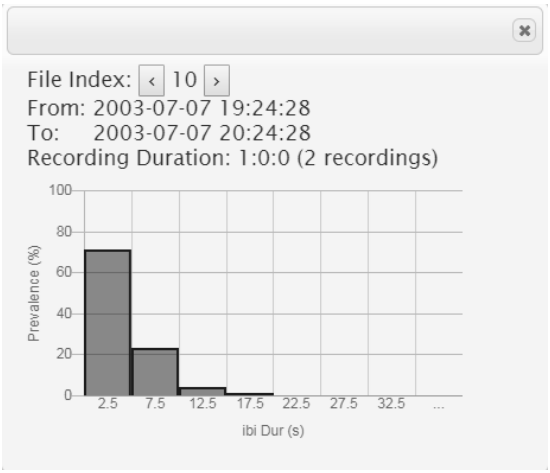


Figure 14.4: The histogram of detected IBIs over the last hour of recording

14.6 Limitations

This version of the monitor has 4 limitations:

1. Batch data: the current design of this monitor does not support real-time streaming. The EEG-polygraphic signals are collected in the data uploader for some minutes (varied between 5 min and 60 min depending on the configuration of the EEG recorder) and then will be sent to the server to be processed. The main reason is that for some EEG devices, like OSG, direct connection to the device is not allowed. Therefore, the system should wait till the recording file is split and finalized, and then it is ready to be read. Although some devices, like Medatec, do not have this limitation, this design has been created in order to support as many EEG devices as possible. In future versions, the system may support real-time streaming for some specific EEG devices as well.

2. Speed: since the algorithms run in their native high-level language, MATLAB, this system is considerably far slower than the NeoGuard real-time monitor which uses low-level C++ codes (see chapter 13). Although its speed is still high enough for online monitoring, which means it needs less than x minutes to process x minutes of EEG data, it may be too slow when hours of offline recorded data should be processed.

3. raw EEG plot: in this version, the raw EEG channels cannot be plotted on the GUI in the data visualizer. However, it is an essential item for

expert users when they want to validate the alarms and processed the trends. It should be added in the next version.

4. Sleep stage plot: although the sleep stage detector has already implemented in the back-end, the data visualizer cannot plot the output of the sleep stage classifier. It should also be provided in the next version.

14.7 Conclusion

In neonatal intensive care units (NICUs), a real-time brain monitor equipped with automated algorithms is extremely needed. In order to make such a monitor, newly developed or improved algorithms should be translated from their native programming languages into a proper low-level one. However, such a translation can be highly labor-intensive, time-consuming, and consequently expensive. Furthermore, debugging, updating, and controlling multiple instances of the algorithms in different languages are very delicate. Moreover, since there is not yet any established monitor with automated algorithms in the current NICUs, the real added value of these algorithms is not clear and there is no guarantee that they can be exactly tailored to the needs of clinicians in practice.

In this chapter, a new platform was introduced in order to hire new automated algorithms in real NICUs. The gold key is using a centralized and controllable cloud-based system that is able to run such algorithms in their native high-level programming languages. Therefore, there is no immediate need to translate them to a secondary low-level language. Furthermore, there is no need to control different versions of the algorithms running in different centers. Additionally, collecting the data, maintaining the systems, and updating them are much simpler. Note that this has been designed to validate the algorithms and is not intended to be the final monitor due to the relatively low speed of the high-level language, MATLAB.

Part V

Conclusion

Chapter 15

Conclusion

15.1 Summary

In this thesis, the development of an automated brain monitor, which is one of the major needs of neonatal intensive care units (NICUs) has been considered. To this end, several algorithms for neonatal seizure detection and sleep state classification have been proposed, the inter-rater agreement of expert EEG readers for annotating seizures has been investigated, and finally some practically implemented monitors, as well as collecting and training platforms, have been introduced and discussed. Figure 15.1 displays a diagram showing how these methods collaborating together. In this figure, the solid and dashed lines similarly show the connections between the methods. The algorithms showed in the "Older Algorithms" group are those developed by other PhD researchers in the NeoGuard project and used in this thesis mainly for the NeoGuard monitors. In the following paragraphs, these methods are summarized.

15.1.1 Algorithms

The contribution of this thesis from an algorithmic point of view consists of three novel seizure detectors, a performance framework, and a sleep state classifier. The seizure detectors and performance framework were tested on the recordings of term neonates with different types of encephalopathy, mainly HIE, recorded in two centers. The sleep study was performed on a group of preterm newborn babies with no encephalopathy recorded in one center.

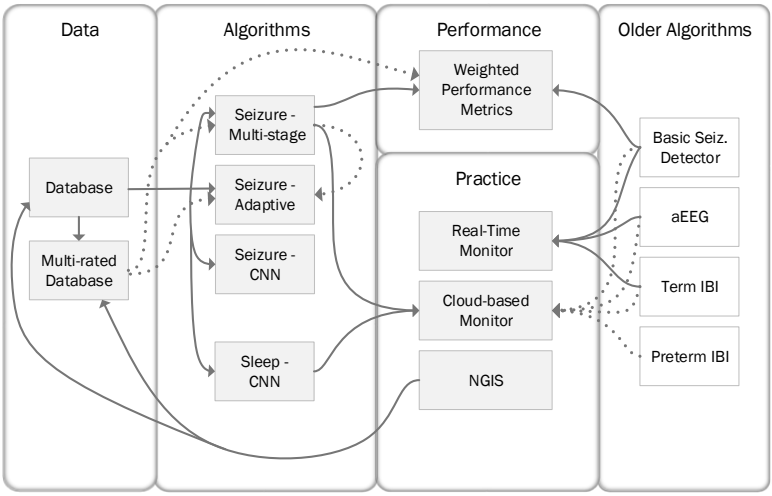


Figure 15.1: A simple collaboration diagram of the methods used in this thesis. The solid and dashed lines similarly show the connections. The white blocks located in the "Older Algorithms" group correspond to the methods developed by other PhD researchers and used in this study mainly in the NeoGuard monitors.

Seizure detection

The first algorithm that was proposed in this thesis, in chapter 6, was a multi-stage neonatal seizure detector (NSD). In this method, in the 1st stage, a previously developed heuristic algorithm mimicking a human EEG reader was applied on the EEG data in order to find potential seizures. Then, in the 2nd stage, a machine learning approach using SVM was used to re-analyze the potential seizures in order to filter out falsely detected segments, mainly artifacts. One can consider this algorithm as an extension of the heuristic NSD with the aid of a post-processing technique. In this post-processor, some novel features have been proposed including the evolution of amplitude/frequency/morphology and phase-synchronization between the EEG and polygraphic channels. It seems evident that these features also play an important role when an expert EEG reader wants to visually discriminate seizures from artifacts. It has been shown that this method is able to significantly drop the false alarm rate (FAR) of the heuristic method (by 34%) while its good detection rate (GDR) decreases by 2%. The structure of this post-processor is generic, improves our understanding of the core visually determined EEG features of neonatal seizures, and is applicable

to other NSDs.

In addition, in chapter 7, a 3rd stage has been proposed to be added to this multi-stage NSD in order to use feedback of the clinical neurophysiologists and adaptively re-tune a threshold of the 2nd stage. The main goal of this stage was decreasing the rate of short false alarms and subsequently increasing the positive predictive value (PPV) of short detections. As a result, the FAR of the short detections decreased by 50% and the PPV increased by 18% compared to the multi-stage NSD without the 3rd stage. At the same time, for all detections, the FAR decreased by 35% and PPV increased by 5% while the good detection rate remained unchanged.

Furthermore, a convolutional neural network (CNN) was designed and proposed in chapter 8 for detecting seizures. This method has a key advantage over the previously developed methods, such as the explained multi-stage NSDs, in that no hand-engineered feature selection is needed. It means that the NSD takes raw EEG data and automatically learns the required features while it fits a decision boundary to classify seizures from non-seizures. This method outperformed the considered feature-based approaches while it is not yet superior over the heuristic and multi-stage NSDs. This method is extremely fast in recall mode, which can be remarkably important when one wants to implement it in a real clinical monitor.

In addition, in chapter 4, it has been shown that the inter-rater agreement of expert EEG readers is not very high. In this study, after applying majority voting on the labels resulting from the annotations of 1919 clinically-relevant events by 4 independent raters, 44% of them were classified with high, 36% with moderate, and 20% with poor agreement, resulting in a Kappa value of 0.39. The events with lower level of agreement were mainly related to the brief-lasting seizures shorter than 30 s. It means that a wide variability in identifying rhythmic ictal and non-ictal EEG events exists and only the most robust ictal patterns are consistently agreed upon. Furthermore, it has also been shown that there is a significant difference in electrographic characteristics of seizures versus dubious events, with seizures having longer duration, higher power and amplitude. This information can be used for developing more reliable and generally satisfying algorithms.

In order to use the outputs of this inter-rater agreement analysis, a new framework for measuring the performance of NSDs has been defined in chapter 10. In this framework, 6 event-based and epoch-based performance metrics have been fuzzily extended. The new metrics can measure the performance when the seizures are annotated by one, two, or more experts. When the annotations from one single rater are used, the proposed performance metrics are exactly equal to the classic ones, which means the proposed metrics can be considered

as a generalized version of the classic ones. Furthermore, missing data and dubious seizures are supported in this framework. It has been shown that this framework has more accurate and robust output compared to the common majority voting approach. Although this framework was only described and tested for neonatal seizure detection, it seems evident that it can be applied to every 2-class problem when multiple labels are available for each instance.

Sleep stage classification
In order to classify the EEG segments into quiet sleep and non-quiet sleep, another CNN was designed and has been described in chapter 9. As mentioned, CNNs can simultaneously optimize the feature selection and classification automatically. The proposed CNN-based method achieved the state-of-the-art performance in a preterm database while it classifies much faster, as shown. This method has also been implemented in the back-end of NeoGuard cloud-based monitor and will be clinically validated soon.

15.1.2 Implementations

Collecting data and annotating them are two important steps for developing and validating new algorithms and subsequently developing new monitors. In this study, the NeoGuard Information System (NGIS) was one of the first developed platforms and has been being used from the beginning. This web-based platform is able to remotely collect large-size recordings from different centers. Part of the database used for seizure detection in this thesis was collected by this platform. Furthermore, it was used for gathering seizure labels from the secondary EEG raters working in several centers for inter-rater agreement analysis. Moreover, the collected annotations of the multi-rated database have been used in the NeoGuard Learning Platform (NGLP), which is now publicly available for trainees and tutors interested in EEG analysis (`www.neoguard.net`).

Furthermore, a real-time monitor has been implemented and described in chapter 13. This so-called NeoGuard Monitor includes the heuristic NSD, IBI detector, and aEEG estimator, as well as some inter-active graphs for plotting EEG, video, and a 3D brain model representing topographically the brain activities through time. Since this monitor was made before developing/improving the algorithms described in this thesis, it only used those that were developed previously by other researchers in the NeoGuard project. In order to implement newly developed methods, they should be translated into C++ language, which is very time consuming. Instead, another system, called NeoGuard cloud-based monitor, has been implemented to clinically validate these algorithms. If they are approved after this validation, they will be implemented in the real-time

monitor in the next step. The NeoGuard cloud-based monitor also helps to easily debug and maintain these algorithms during the clinical validation in the NICUs. Furthermore, the multi-center database which is being collected during this validation study on a cloud storage can be used for future improvement.

15.2 Future directions

In this study, some important steps towards developing a neonatal brain monitor have been taken. However, more work still needs to be done in order to practically transfer the suggested methods into the NICUs. To this end, different extensions and future directions can be identified for the methods explained in this thesis as considered below from both algorithmic and implementation points of view.

15.2.1 Algorithms

- In the database section, having more recordings will undoubtedly improve development and training of the algorithms. Furthermore, recordings from other centers and rated by several independent scorers are needed for a better clinical validation. This can be done with the aid of the NGIS and NeoGuard cloud-based monitor running in several centers.

- The post-processor proposed in the multi-stage seizure detector is semi-online. In other words, this method alarms when a seizure offset manifests in the EEG, is detected by the heuristic method in the 1st stage, and is accepted by the post-processor in the 2nd stage. It means that there is always a delay between the alarms and the seizure onset. Although this delay is practically negligible in brief-lasting seizures, it should be considered for long seizures. One practical solution can be splitting the long seizures, when their duration exceeds a given threshold, and running the post-processor multiple times. However, this threshold should be optimized and the performance of the method with this trick should be re-calculated.

- The adaptive learning method proposed for the 3rd stage of the multi-stage seizure detector has been tested in the presence of this assumption that the feedback from an expert EEG reader is always available after each alarm. However, it seems that the users of such a brain monitor will be 1) neurologists or clinical neurophysiologists with EEG expertise, and 2) primary care-givers working at the NICUs, such as nurses and neonatologists, with limited knowledge of EEG interpretation. Although

the second group usually consults the experts when they consider to initiate the treatment with anti-seizure drugs, it happens for recurrent alarms with enough seizure burden. Therefore, some alarms may not have any feedback from the expert users. In order to have a practical algorithm, this fact should be taken into consideration in developing as well as measuring the performance in this algorithm. This can be performed in the cloud-based monitor.

- Furthermore, in the adaptive learning algorithm, another option is selecting and replacing adaptively the support vectors presented in the 2nd stage, instead of changing the final threshold. It is expected that applying an advanced adaptive SVM or adaptive Least Squares-SVM (LS-SVM) can improve the performance.

- In the CNN-based algorithm for seizure detection, for training the network, a CPU has been used, and therefore, the training time was extremely high. Thus, the hyper-parameters were defined by trial and error with limited number of iterations. It is highly possible that those are not the optimal values. Nevertheless, using a GPU, instead of CPU, which is possible in the used parallelized toolbox (MatConvNet), can drop this time tremendously so that the hyper-parameters can be optimized as well. In addition, it seems logical that if the CNN method is trained with some more advanced techniques, such as drop-out and denoising mechanisms, the used random forest method is not needed anymore. Hence, it can make the algorithm faster with higher performance. Furthermore, this algorithm should be trained with more recordings from different centers and with different etiologies.

- For the CNN method proposed in the sleep stage classification, use of GPU, as mentioned above, for optimizing the hyper-parameters should be considered. Furthermore, it seems that using a sequence learner, specially the long short-term memory, instead of the used simple averaging, can increase the power of the network and consequently can improve the performance. Similarly, this method also should be validated with multi-center and multi-rated databases which will be done in the NeoGuard cloud-based system. Additionally, the performance of this method for sleep stage classification of term neonates should be measured. Furthermore, extending the network in order to use the polygraphic signals/video, in addition to the EEG, will improve the classification performance.

- In addition, for the developed seizure detectors, the degree of agreement of the experts collected in the inter-rater agreement analysis can be used. This degree can be represented as a weight for each seizure segment and subsequently influences the cost function in the training process. In this

way, the algorithms tend to detect the seizures upon which all raters agreed, compared to the dubious and disagreed seizures. Furthermore, in the chapter 4, it has been shown that there is a correlation between the agreement of the experts and some characteristics of the EEG segments. Therefore, a regression model can be fitted on these data to predict the level of agreement from the EEG characteristic. This model can be used besides a seizure detector in the brain monitor to predict the plausibility of alarming seizure.

15.2.2 Implementations

- In the NeoGuard Learning Platform (NGLP), more data from other centers scored by more experts should be used. The users should be able to shift the time frame in order to see pre-ictal or post-ictal periods. The filter and montage settings should be customizable. Some interactive learning materials should be added into the website. When a statistically sufficient number of trainees or experts have used the system, further analyses are possible on their scorings. For instance, the following questions can be addressed: what is the correlation between the level of experiment and agreement of the user? What is the correlation of wrong/true scorings with the duration of events, etiologies, agreement of experts, types of seizures, etc.?

- The NeoGuard real-time monitor should be modified and tested for other operating systems including Windows and Mac OS. When other algorithms have been validated by the NeoGuard cloud-based system, they should be translated and implemented in this monitor. The GUI will be improved, as well, to be more user-friendly. It should also be able to connect to other EEG recording devices.

- The current version of the NeoGuard cloud-based monitor is able to run the developed algorithm in Matlab and C#. However, it should be equipped with Python and C++ engines for algorithms that may be developed in future. It should also be able to stream real-time data from the EEG devices that support it. In that case, the users are able to monitor EEG signals and alarms in real-time around the clock on the web. Furthermore, the graph of the sleep stages, namely hypnogram, should be made in the GUI. In addition, a native mobile application (for Android/iOS/Windows) can be developed to have more user-friendly interface, safer and faster connection.

Appendix A

Previously developed NeoGuard algorithms

A.1 Heuristic seizure detector

In this section, a previously developed seizure detector mimicking a human EEG observer developed by Deburchgraeve et al. [4] (Version 1), as well as its improved version [7] (Version 2) is described. In order to develop these algorithms, DB1 (see chapter 3) was used [1].

A.1.1 Version 1

This algorithm was developed in our group by Deburchgraeve et al. and its schematic overview is displayed in figure A.1. The comprehensive description of this algorithm is available in the reference paper [4]. Briefly, this algorithm uses two separate procedures for detecting seizures: I) a spike-train seizure detector and II) an oscillatory seizure detector. In the spike-train detection, first, a nonlinear energy operator (NLEO), using the Teager–Kaiser operator, is applied on one channel of EEG data and the output is normalized and smoothed with a moving average (MA) filter with a window size of 120 ms. Then, an adaptive threshold is defined based on (A.1) and applied.

$$Threshold = 0.6[std(epoch) + q_3(epoch)] \tag{A.1}$$

[1] In addition to the recordings used in DB1, the recordings from 4 more neonates with seizure were also used by Deburchgraeve et al. which are not available anymore.

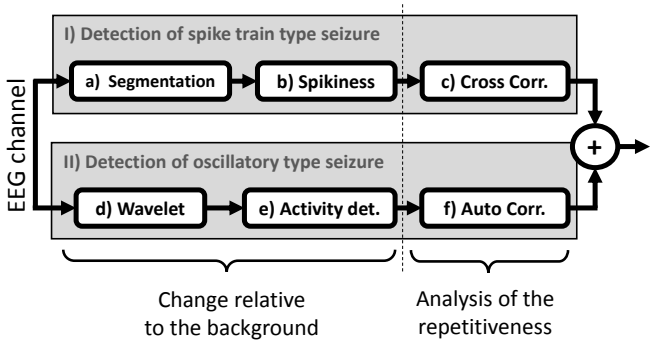


Figure A.1: Schematic overview of the heuristic method. The upper and lower lines show the spike-train type and oscillatory type seizure detectors respectively.

where $std(.)$ and $q_3(.)$ denote the standard deviation and the 3rd quartile.

The potential spikes are defined where the smoothed nonlinear energy is greater than the threshold (figure A.1.a). Next, the selected segments with a duration of more than 60 ms and isolated from the background activity are detected as epileptic spikes (figure A.1.b). Finally, when at least 6 sequential spikes have the overall cross-correlation higher than 0.8, they are considered as a spike-train type seizure (figure A.1.c). Figure A.2 illustrates a definite seizure, which was detected by the spike-train detector. The gray highlighted spikes correspond to the detected spike-trains. In the second detector, the δ (0.5 - 4 Hz) and θ (4 - 8 Hz) frequency bands are extracted from one channel of EEG data using a discrete wavelet transform (figure A.1.d). Then, the potential epileptic activities are defined when 3 seconds of filtered signal has significantly higher energy compared to its previous 30 seconds (figure A.1.e). Next, autocorrelation of the potential activities is calculated and two features including the skewness and the mean of interval differences between the zero-crossing are extracted (figure A.1.f). The activities with the features greater than two pre-defined thresholds are marked as oscillatory type seizures. Figure A.3 illustrates a seizure that was detected by the oscillatory detector. The described two procedures are applied on all channels of the EEG individually. If there is a seizure in at least one channel, that segment is marked as seizure ('OR' operator). If a segment is marked as seizure by both detectors, it is called mixed type seizure. Figure A.4 displays an example of a mixed type seizure.

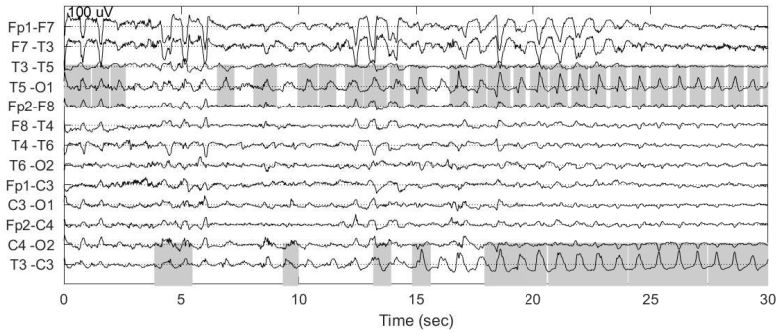


Figure A.2: Example of a definite seizure detected by the spike-train detector. The gray highlighted spikes represent those detected by the algorithm.

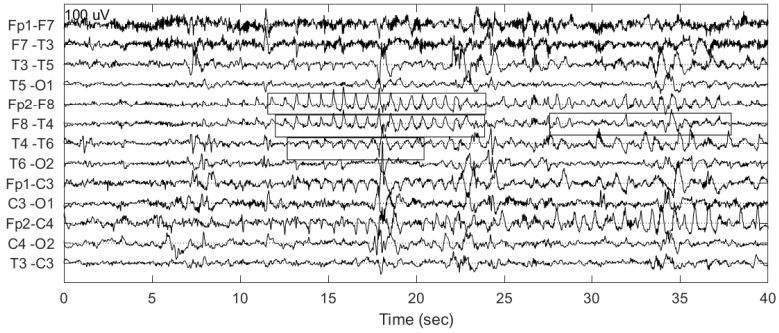


Figure A.3: Example of a definite seizure detected by the oscillatory detector. The rectangles represent the oscillatory seizure activities that were detected by the algorithm.

A.1.2 Version 2

The second version of this algorithm was proposed by Deburchgraeve et al. in [5], as well as Appendix A of [7]. This algorithm is very similar to the 1st version except in two blocks: 1) segmenting the EEG in the spike-train type detector (figure A.1.a), and 2) the autocorrelation analysis in the oscillatory type detector (figure A.1.f).

1. In order to smooth the nonlinear energy, instead of using a moving average filter with a windoes size of 120 ms, a filter bank with 6 moving average filters is applied. In this case, the filter can adequately pass both short

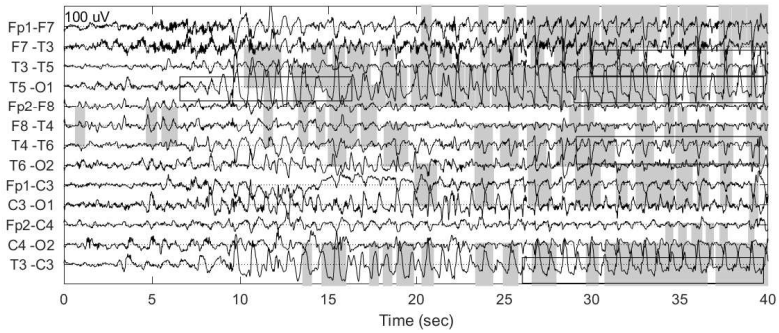


Figure A.4: Example of a definite seizure detected by both spike-train and oscillatory detectors (mixed type). The highlighted spikes and solid rectangles represent respectively the spike-train and oscillatory patterns that were detected by the algorithm.

(50 ms) and long (500 ms) spikes. Furthermore, in order to segment the smoothed nonlinear energy, instead of using the adaptive threshold presented in (A.1), a set of thresholds is applied on the smoothed signal and the threshold with the maximum number of segments is chosen. In order to make this segmentation more robust, the segments that are shorter than 100 ms are excluded from counting.

2. In order to select seizures among the potential activities, instead of the mentioned two features, three following features are extracted:
   - Regularity of the distances between the zero crossings (figure A.5.A)
   - Regularity of the distances between the peaks (figure A.5.B)
   - Regularity of the normalized RMS values of the peaks which are delimited by the zero crossings (figure A.5.C)

A.2 Amplitude-integrated EEG

Amplitude-integrated EEG (aEEG), which is a compressed, rectified version of the EEG, is the most common signal representing brain activity used in most NICUs. Cerebral function monitor (CFM^{TM}) is the best known and widely used monitor for plotting the aEEG using 2 to 4 EEG channels. The CFM generally displays about 5 hours of a recording on one page and is utilized in many NICUs because of its ease of use and support. Usually, with a simple

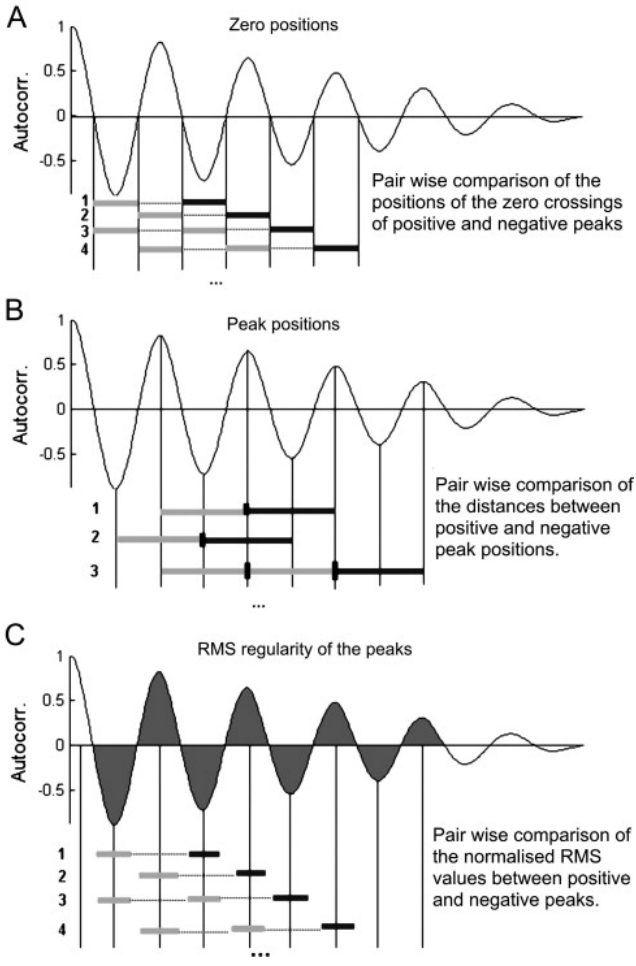


Figure A.5: Illustration of the new features of the autocorrelation function used in version 2 [7].

Table A.1: Asymmetrical filter characteristics

Frequency	Slop
0-2 Hz	+60 db / decade
2-12 Hz	+12 db / decade
12-16 Hz	Flat at +1.2 db
16-50 Hz	-120 db / decade
>50 Hz	-60 db / decade

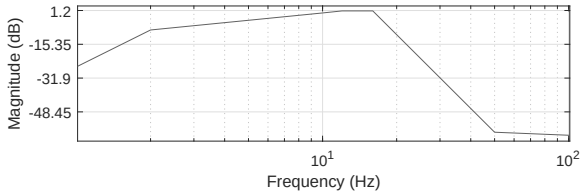


Figure A.6: The bod plot of the asymmetrical filter
s

training, even non-expert EEG readers can interpret the aEEG and detect some frequent EEG patterns such as seizures. It makes the CFM as an important monitor in practice and, therefore, prospectively developed brain monitors are expected to be able to display aEEG besides other signals. However, the exact aEEG algorithm plotted in the CFM is not shared, and it should be estimated. Zhang and Ding have proposed an algorithm to estimate the aEEG [232] based on the principles explained in [231]. This algorithm was re-implemented and boosted in the NeoGuard group as described in the following steps:

1. Bipolar signal: the bipolar signal is made by the subtraction of the two target channels (usually C3-C4, or O1-O2).

2. Asymmetrical filtering: the bipolar signal is filtered with an asymmetrical band-pass filter. The characteristics of this filter are shown in table A.1 and its Bod plot is displayed in figure A.6.

3. Full-wave rectification: the filter signal is rectified using an absolute function.

4. Envelope detection: the envelope of the rectified signal are detected using leaky peak detection filter as follows:

$$y(n) = \begin{cases} y(n-1)e^{-\frac{1}{T_r F s}} & \text{if } x(n) < y(n-1)e^{-\frac{1}{T_r F s}} \\ x(n) - (x(n) - y(n-1))e^{-\frac{1}{T_d F s}} & \text{otherwise} \end{cases} \tag{A.2}$$

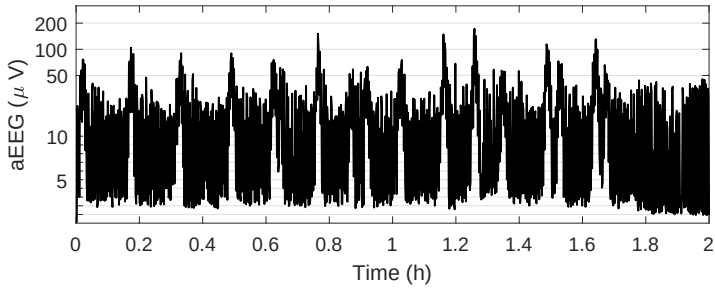


Figure A.7: Example of an estimated aEEG. The patterns with high amplitude correspond to electroencephalographic seizures.

where Fs denotes the sampling frequency, and T_r and T_d are respectively the rising and decaying time constant, which are equal to 0.5 and 1.5 respectively.

5. Down-sample: since the aEEG is plotted normally at 6 cm per hour, the high temporal resolution of the EEG signal is not needed. The signal can be compressed to be plotted faster with less memory need. To this end, only the global maximum and minimum values of the envelopes within every 15s epochs (with no overlap) are used. As a result, every hour of EEG will be compressed to 480 points, regardless of the EEG sampling frequency.

To illustrate, figure A.7 displays an estimated aEEG. As is clear in this figure, the vertical axis of aEEG is usually semi-logarithmic so that the amplitudes below and above $10\mu V$ are plotted linearly and logarithmically respectively.

A.3 Inter-burst interval

In order to calculate the inter-burst intervals (IBI), as an important biomarker in different problems, such as abnormal background assessment and maturation analysis, two previously developed algorithms proposed in [8] and [9] are used respectively for preterm and term infants. These algorithms are briefly described in the following subsections.

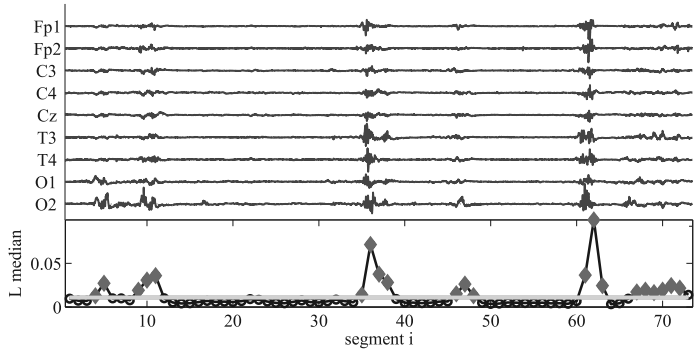


Figure A.8: Example of 65 seconds multi-channel EEG recording from a preterm infant and automatically detected IBI. The upper traces are the raw EEG and the lower one is the line length. The diamonds and the horizontal gray line respectively correspond to the burst segments and the threshold [8].

A.3.1 IBI for preterm neonates based on line length

1. The multi-channel EEG is filtered by a 50 Hz and 100 Hz notch filters followed by a 1-20 Hz band-pass filter.

2. The filtered signal is split into 1s epochs with 0.12 s overlap for each EEG channel.

3. The line length is calculated and normalized by (A.3) and (A.4) respectively.

$$L(i,ch) = \sum_{j=1}^{Fs-1} |x_{j+1} - x_j| \tag{A.3}$$

$$\bar{L}(i,ch) = \frac{L(i,ch)}{\sum_i L(i,ch)} \tag{A.4}$$

   where Fs denotes the sampling frequency, i and ch are the segment index and channel respectively.

4. The median of the line length through the channels is calculated by

$$\widehat{L}(i) = median(\bar{L}(i,ch)) \tag{A.5}$$

5. Finally, the moments in which the median line-length is smaller than the threshold defined by equation (A.6) are marked as IBI.

$$Thr = F.mean(\widehat{L}(i)) \tag{A.6}$$

where F is an arbitrary parameter to make a trade-off between high sensitivity and specificity. In this implementation, 0.8 was chosen.

Figure A.8 displays an example of a multi-channel EEG and corresponding detected IBIs.

A.3.2 dynamic IBI for term neonates

1. The multi-channel EEG signal is filtered with a 0.7-20 Hz band-pass filter.
2. Each channel of the filtered signal is adaptively split into quasi-stationary segments base on the algorithm described in [233]. Briefly, two fixed concatenated moving average windows slide on the EEG and are compared together. Where the amplitude and frequency of two windows are significantly different, the segment is split.
3. The segments are clustered into 3 categories as low, medium, and high amplitude with respect to their averaged signal power and frequency.
4. Finally, the multi-channel EEG segments in which the absolute majority of channels (50%+1) are clustered as low amplitude are marked as dynamic IBI. Figure A.9 illustrates a multi-channel EEG in which 3 IBIs are detected.

This method has also been extended with a post-processing technique using LS-SVM, which was not implemented in the monitors described in this thesis.

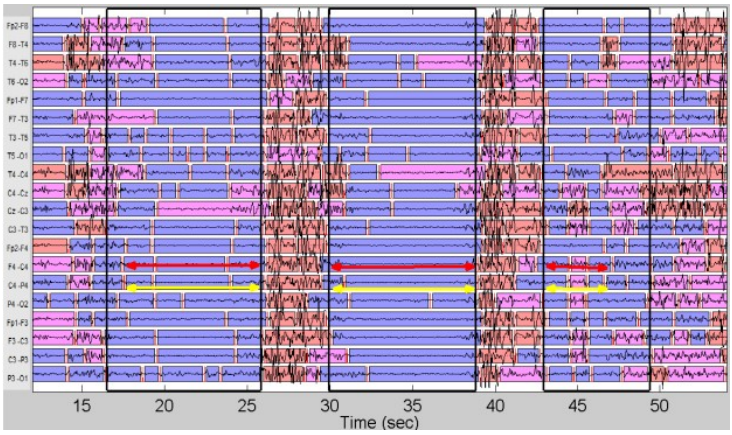


Figure A.9: Example of three detected dynamic IBIs in a multi-channel EEG. The blue, pink, and red rectangles respectively correspond to the low, medium, and high amplitude clusters. The thick solid rectangles show the detected IBIs after majority voting. The red and yellow arrows show the IBIs marked by an expert EEG reader [9].

Bibliography

[1] *http://www.mhhe.com/cls/psy/ch02/brain.mhtml*. january 2018.

[2] M. Okamoto, H. Dan, K. Sakamoto, K. Takeo, K. Shimizu, S. Kohno, I. Oda, S. Isobe, T. Suzuki, K. Kohyama, *et al.*, "Three-dimensional probabilistic anatomical cranio-cerebral correlation via the international 10–20 system oriented for transcranial functional brain mapping," *Neuroimage*, vol. 21, no. 1, pp. 99–111, 2004.

[3] A. Temko, E. Thomas, W. Marnane, G. Lightbody, and G. Boylan, "EEG-based neonatal seizure detection with Support Vector Machines," *Clinical Neurophysiology*, vol. 122, pp. 464–473, 2011.

[4] W. Deburchgraeve, P. J. Cherian, M. De Vos, R. M. Swarte, J. H. Blok, G. H. Visser, P. Govaert, and S. Van Huffel, "Automated neonatal seizure detection mimicking a human observer reading EEG," *Clinical Neurophysiology*, vol. 119, pp. 2447–2454, 2008.

[5] W. Deburchgraeve, *Development of an automated neonatal EEG seizure monitor*. PhD Thesis, Katholieke Universiteit of Leuven, Leuven, 2010.

[6] A. H. Ansari, P. J. Cherian, A. Dereymaeker, V. Matic, K. Jansen, L. De Wispelaere, C. Dielman, J. Vervisch, R. M. Swarte, P. Govaert, G. Naulaers, M. De Vos, and S. Van Huffel, "Improved multi-stage neonatal seizure detection using a heuristic classifier and a data-driven post-processor," *Clinical Neurophysiology*, vol. 127, pp. 3014–3024, 2016.

[7] M. De Vos, W. Deburchgraeve, P. J. Cherian, V. Matic, R. M. Swarte, P. Govaert, G. H. Visser, and S. Van Huffel, "Automated artifact removal as preprocessing refines neonatal seizure detection," *Clinical Neurophysiology*, vol. 122, pp. 2345–2354, 2011.

[8] N. Koolen, K. Jansen, J. Vervisch, V. Matic, M. De Vos, G. Naulaers, and S. Van Huffel, "Line length as a robust method to detect high-

activity events: Automated burst detection in premature EEG recordings," *Clinical Neurophysiology*, vol. 125, no. 10, pp. 1985–1994, 2014.

[9] V. Matic, P. J. Cherian, K. Jansen, N. Koolen, G. Naulaers, R. M. Swarte, P. Govaert, G. H. Visser, S. Van Huffel, and M. De Vos, "Automated eeg inter-burst interval detection in neonates with mild to moderate postasphyxial encephalopathy," in *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, pp. 17–20, IEEE, 2012.

[10] J. M. Stern, *Atlas of EEG patterns*. Lippincott Williams & Wilkins, 2005.

[11] E. Niedermeyer and F. L. da Silva, *Electroencephalography: basic principles, clinical applications, and related fields*. Lippincott Williams & Wilkins, 2005.

[12] A. Varsavsky, I. Mareels, and M. Cook, *Epileptic seizures and the EEG: measurement, models, detection and prediction*. CRC Press, 2016.

[13] W. O. Tatum IV, *Handbook of EEG interpretation*. Demos Medical Publishing, 2014.

[14] K. Majumdar, *A Brief Survey of Quantitative EEG*. CRC Press, 2017.

[15] L. Marshall, H. Helgadóttir, M. Mölle, and J. Born, "Boosting slow oscillations during sleep potentiates memory," *Nature*, vol. 444, no. 7119, pp. 610–613, 2006.

[16] J. L. Cantero, M. Atienza, R. Stickgold, M. J. Kahana, J. R. Madsen, and B. Kocsis, "Sleep-dependent θ oscillations in the human hippocampus and neocortex," *Journal of Neuroscience*, vol. 23, no. 34, pp. 10897–10903, 2003.

[17] S. R. Benbadis, "Introduction to sleep electroencephalography," *Sleep: A comprehensive handbook*, pp. 989–1024, 2006.

[18] J.-P. Banquet, "Spectral analysis of the eeg in meditation," *Electroencephalography and clinical neurophysiology*, vol. 35, no. 2, pp. 143–151, 1973.

[19] W. Klimesch, "Eeg alpha and theta oscillations reflect cognitive and memory performance: a review and analysis," *Brain research reviews*, vol. 29, no. 2, pp. 169–195, 1999.

[20] S. Sanei and J. A. Chambers, *EEG signal processing*. John Wiley & Sons, 2013.

[21] A. R. Clarke, R. J. Barry, R. McCarthy, and M. Selikowitz, "Age and sex effects in the eeg: development of the normal child," *Clinical neurophysiology*, vol. 112, no. 5, pp. 806–814, 2001.

[22] E. Orekhova, T. Stroganova, I. Posikera, and M. Elam, "Eeg theta rhythm in infants and preschool children," *Clinical neurophysiology*, vol. 117, no. 5, pp. 1047–1062, 2006.

[23] D. K. Stevenson, W. E. Benitz, P. Sunshine, M. L. Druzin, and S. R. Hintz, *Fetal and neonatal brain injury.* Cambridge University Press, 2017.

[24] L. De Vries and L. Hellström-Westas, "Role of cerebral function monitoring in the newborn," *Archives of Disease in Childhood-Fetal and Neonatal Edition*, vol. 90, no. 3, pp. F201–FF207, 2005.

[25] R. Swarte, *The clinical value of intensive monitoring in term asphyxiated newborns.* PhD thesis, Erasmus MC: University Medical Center Rotterdam, 2010.

[26] E. Hämäläinen, "New trends in brain hypoxia ischemia research," 2007.

[27] P. J. Cherian, *Improvements in Neonatal Brain Monitoring after Perinatal Asphyxia.* 2010.

[28] S. Y. Proskuryakov, A. G. Konoplyannikov, and V. L. Gabai, "Necrosis: a specific form of programmed cell death?," *Experimental cell research*, vol. 283, no. 1, pp. 1–16, 2003.

[29] G. I. Evan, L. Brown, M. Whyte, and E. Harrington, "Apoptosis and the cell cycle," *Current opinion in cell biology*, vol. 7, no. 6, pp. 825–834, 1995.

[30] S. Kobayashi, "Choose delicately and reuse adequately: the newly revealed process of autophagy," *Biological and Pharmaceutical Bulletin*, vol. 38, no. 8, pp. 1098–1103, 2015.

[31] A. Fatemi, M. A. Wilson, and M. V. Johnston, "Hypoxic-ischemic encephalopathy in the term infant," *Clinics in perinatology*, vol. 36, no. 4, pp. 835–858, 2009.

[32] N. Mizushima and M. Komatsu, "Autophagy: renovation of cells and tissues," *Cell*, vol. 147, no. 4, pp. 728–741, 2011.

[33] C. M. Cotten and S. Shankaran, "Hypothermia for hypoxic–ischemic encephalopathy," *Expert review of obstetrics & gynecology*, vol. 5, no. 2, pp. 227–239, 2010.

[34] M.-C. Lai and S.-N. Yang, "Perinatal hypoxic-ischemic encephalopathy," *BioMed Research International*, vol. 2011, 2010.

[35] D. Ellison, S. Love, L. M. C. Chimelli, B. Harding, J. S. Lowe, H. V. Vinters, S. Brandner, and W. H. Yong, *Neuropathology E-Book: A Reference Text of CNS Pathology*. Elsevier Health Sciences, 2012.

[36] P. Plouin and A. Kaminska, *Pediatric Neurology Part I: Chapter 51. Neonatal seizures*. Handbook of Clinical Neurology, Elsevier Science, 2013.

[37] R. Pressler, "Neonatal seizures," in *Introduction to Epilepsy*, pp. 142–149, 2012.

[38] J. J. Volpe, *Neurology of the Newborn*. Philadelphia: Saunder WB, 5th ed., 2008.

[39] B. R. Greene, S. Faul, W. P. Marnane, G. Lightbody, I. Korotchikova, and G. B. Boylan, "A comparison of quantitative EEG features for neonatal seizure detection," *Clinical Neurophysiology*, vol. 119, pp. 1248–1261, 2008.

[40] F. E. Jensen, "Neonatal Seizures: An Update on Mechanisms and Management," *Clinics in Perinatology*, vol. 36, pp. 881–900, 2009.

[41] B. Chang and D. Lowenstein, "Mechanisms of disease: Epilepsy," *New England Journal of Medicine*, vol. 349, no. 13, pp. 1257–1266, 2003. cited By 420.

[42] O. Devinsky, D. Gazzola, and W. C. LaFrance, "Differentiating between nonepileptic and epileptic seizures," *Nature Reviews Neurology*, vol. 7, no. 4, pp. 210–220, 2011.

[43] F. W. Drislane, *Status epilepticus: a clinical perspective*. Springer, 2005.

[44] F. Al-Mufti and J. Claassen, "Neurocritical care: status epilepticus review," *Critical care clinics*, vol. 30, no. 4, pp. 751–764, 2014.

[45] R. Lawrence and T. Inder, "Neonatal status epilepticus," in *Seminars in pediatric neurology*, vol. 17, pp. 163–168, Elsevier, 2010.

[46] J. Engel, "Report of the ilae classification core group," *Epilepsia*, vol. 47, no. 9, pp. 1558–1568, 2006.

[47] J. M. Rennie, G. Chorley, G. B. Boylan, R. Pressler, Y. Nguyen, and R. Hooper, "Non-expert use of the cerebral function monitor for neonatal seizure detection," *Archives of Disease in Childhood-Fetal and Neonatal Edition*, vol. 89, no. 1, pp. F37–F40, 2004.

[48] J. Campistol, "Advances in neonatal neurology," *Revista de neurologia*, vol. 31, no. 7, pp. 601–604, 2000.

[49] S. Shorvon, R. Guerrini, M. Cook, and S. Lhatoo, *Oxford textbook of epilepsy and epileptic seizures*. OUP Oxford, 2012.

[50] D. M. Kaufman, H. L. Geyer, and M. J. Milstein, *Kaufman's Clinical Neurology for Psychiatrists E-Book*. Elsevier Health Sciences, 2016.

[51] G. P. Prigatano and N. H. Pliskin, *Clinical neuropsychology and cost outcome research: A beginning*. Psychology Press, 2003.

[52] M. Milstein and D. Kaufman, *Clinical Neurology for Psychiatrists*. Major Problems in Neurology, Elsevier Health Sciences, 2006.

[53] H. Morris, *Neurology Update: Reviews for Continuing Professional Development*. Radcliffe Publishing, 2006.

[54] W. O. Tatum IV, P. W. Kaplan, and P. Jallon, *Epilepsy A to Z: a concise encyclopedia*. Demos Medical Publishing, 2009.

[55] P. J. Cherian, R. M. Swarte, and G. H. Visser, "Technical standards for recording and interpretation of neonatal electroencephalogram in clinical practice," *Annals of Indian Academy of Neurology*, vol. 12, no. 1, pp. 58–70, 2009.

[56] A. Dereymaeker, K. Pillay, J. Vervisch, S. Van Huffel, G. Naulaers, K. Jansen, and M. De Vos, "An automated quiet sleep detection approach in preterm infants as a gateway to assess brain maturation," *International Journal of Neural Systems*, p. 1750023, 2017.

[57] O. De Wel, M. Lavanga, A. C. Dorado, K. Jansen, A. Dereymaeker, G. Naulaers, and S. Van Huffel, "Complexity analysis of neonatal eeg using multiscale entropy: Applications in brain maturation and sleep stage classification," *Entropy*, vol. 19, no. 10, p. 516, 2017.

[58] M. Lavanga, O. De Wel, A. Caicedo, E. Heremans, K. Jansen, A. Dereymaeker, G. Naulaers, and S. Van Huffel, "Automatic quiet sleep detection based on multifractality in preterm neonates: Effects of maturation," in *Engineering in Medicine and Biology Society (EMBC), 2017 39th Annual International Conference of the IEEE*, pp. 2010–2013, IEEE, 2017.

[59] A. Piryatinska, G. Terdik, W. A. Woyczynski, K. A. Loparo, M. S. Scher, and A. Zlotnik, "Automated detection of neonate EEG sleep stages," *Computer Methods and Programs in Biomedicine*, vol. 95, pp. 31–46, 2009.

[60] D. Mantini, M. G. Perrucci, S. Cugini, A. Ferretti, G. L. Romani, and C. Del Gratta, "Complete artifact removal for eeg recorded during continuous fmri using independent component analysis," *Neuroimage*, vol. 34, no. 2, pp. 598–607, 2007.

[61] S. Tong and N. V. Thakor, *Quantitative EEG analysis methods and clinical applications*. Artech House, 2009.

[62] E. Harmon-Jones and J. Beer, *Methods in Social Neuroscience*. Guilford Press, 2012.

[63] A. Husain, *Practical Epilepsy*. Springer Publishing Company, 2015.

[64] R. Cooper, J. W. Osselton, and J. C. Shaw, *EEG technology, 2nd edition*. Butterworth, 1974.

[65] T. Yamada and E. Meng, *Practical guide for clinical neurophysiologic testing: EEG*. Lippincott Williams & Wilkins, 2012.

[66] B. Robertson, B. Marshall, and M.-A. Carno, *Polysomnography for the sleep technologist: instrumentation, monitoring, and related procedures*. Elsevier Health Sciences, 2013.

[67] P. J. Cherian, W. Deburchgraeve, R. M. Swarte, M. De Vos, P. Govaert, S. Van Huffel, and G. H. Visser, "Validation of a new automated neonatal seizure detection system: A clinician's perspective," *Clinical Neurophysiology*, vol. 122, pp. 1490–1499, 2011.

[68] M. I. Levene, "Measurement of the growth of the lateral ventricles in preterm infants with real-time ultrasound.," *Archives of Disease in Childhood*, vol. 56, no. 12, pp. 900–904, 1981.

[69] A. Dereymaeker, A. H. Ansari, K. Jansen, P. J. Cherian, J. Vervisch, P. Govaert, L. De Wispelaere, C. Dielman, V. Matic, A. C. Dorado, M. De Vos, S. Van Huffel, and G. Naulaers, "Interrater agreement in visual scoring of neonatal seizures based on majority voting on a web-based system: The Neoguard EEG database," *Clinical Neurophysiology*, vol. 128, pp. 1737–1745, 2017.

[70] Y. Ben-Ari and G. L. Holmes, "Effects of seizures on developmental processes in the immature brain," *The Lancet Neurology*, vol. 5, pp. 1055–1063, 2006.

[71] G. L. Holmes, "The Long-Term Effects of Neonatal Seizures," *Clinics in Perinatology*, vol. 36, pp. 901–914, 2009.

[72] F. E. Jensen, "Developmental factors in the pathogenesis of neonatal seizures," *Journal of Pediatric Neurology*, vol. 7, pp. 5–12, 2009.

[73] G. Ramantani, "Neonatal epilepsy and underlying aetiology: to what extent do seizures and EEG abnormalities influence outcome?," *Epileptic Disorders*, vol. 15, pp. 365–375, 2013.

[74] S. P. Miller, J. Weiss, A. Barnwell, D. M. Ferriero, B. Latal-Hajnal, A. Ferrer-Rogers, N. Newton, J. C. Partridge, D. V. Glidden, D. B. Vigneron, and A. J. Barkovich, "Seizure-associated brain injury in term newborns with perinatal asphyxia," *Neurology*, vol. 58, pp. 542–548, 2002.

[75] D. K. Shah, C. J. Wusthoff, P. Clarke, J. S. Wyatt, S. M. Ramaiah, R. J. Dias, J.-C. Becher, O. Kapellou, and J. P. Boardman, "Electrographic seizures are associated with brain injury in newborns undergoing therapeutic hypothermia," *Archives of Disease in Childhood - Fetal and Neonatal Edition*, vol. 99, pp. F219–F224, 2014.

[76] F. Pisani, C. Cerminara, C. Fusco, and L. Sisti, "Neonatal status epilepticus vs recurrent neonatal seizures Clinical findings and outcome," *Neurology*, vol. 69, pp. 2177–2185, 2007.

[77] L. G. M. v. Rooij, L. S. d. Vries, S. Handryastuti, D. Hawani, F. Groenendaal, A. C. v. Huffelen, and M. C. Toet, "Neurodevelopmental Outcome in Term Infants With Status Epilepticus Detected With Amplitude-Integrated Electroencephalography," *Pediatrics*, vol. 120, pp. e354–e363, 2007.

[78] F. Pisani, C. Facini, E. Pavlidis, C. Spagnoli, and G. Boylan, "Epilepsy after neonatal seizures: Literature review," *European Journal of Paediatric Neurology*, vol. 19, pp. 6–14, 2015.

[79] F. Pisani and C. Spagnoli, "Neonatal Seizures: A Review of Outcomes and Outcome Predictors," *Neuropediatrics*, vol. 47, pp. 012–019, 2016.

[80] D. M. Murray, G. B. Boylan, I. Ali, C. A. Ryan, B. P. Murphy, and S. Connolly, "Defining the gap between electrographic seizure burden, clinical expression and staff recognition of neonatal seizures," *Archives of Disease in Childhood - Fetal and Neonatal Edition*, vol. 93, pp. F187–F191, 2008.

[81] A. Malone, C. Anthony Ryan, A. Fitzgerald, L. Burgoyne, S. Connolly, and G. B. Boylan, "Interobserver agreement in neonatal seizure identification," *Epilepsia*, vol. 50, no. 9, pp. 2097–2101, 2009.

[82] S. O. Wietstock, S. L. Bonifacio, J. E. Sullivan, K. B. Nash, and H. C. Glass, "Continuous Video Electroencephalographic (EEG) Monitoring for Electrographic Seizure Diagnosis in Neonates: A Single-Center Study," *Journal of Child Neurology*, vol. 31, pp. 328–332, 2016.

[83] G. B. Boylan, N. J. Stevenson, and S. Vanhatalo, "Monitoring neonatal seizures," *Seminars in Fetal and Neonatal Medicine*, vol. 18, pp. 202–208, 2013.

[84] C. J. Wusthoff, "Diagnosing Neonatal Seizures and Status Epilepticus," *Journal of Clinical Neurophysiology*, vol. 30, pp. 115–121, 2013.

[85] R. A. Shellhaas, "Continuous long-term electroencephalography: The gold standard for neonatal seizure diagnosis," in *Seminars in Fetal and Neonatal Medicine*, vol. 20, pp. 149–153, Elsevier, 2015.

[86] N. S. Abend, C. J. Wusthoff, E. M. Goldberg, and D. J. Dlugos, "Electrographic seizures and status epilepticus in critically ill children and neonates with encephalopathy," *The Lancet Neurology*, vol. 12, pp. 1170–1179, 2013.

[87] M. A. Navakatikyan, P. B. Colditz, C. J. Burke, T. E. Inder, J. Richmond, and C. E. Williams, "Seizure detection algorithm for neonates based on wave-sequence analysis," *Clinical Neurophysiology*, vol. 117, pp. 1190–1203, 2006.

[88] J. Mitra, J. R. Glover, P. Y. Ktonas, A. Thitai Kumar, A. Mukherjee, N. B. Karayiannis, J. D. Frost, R. A. Hrachovy, and E. M. Mizrahi, "A Multi-stage System for the Automated Detection of Epileptic Seizures in Neonatal EEG," *Journal of clinical neurophysiology: official publication of the American Electroencephalographic Society*, vol. 26, pp. 218–226, 2009.

[89] A. Temko, E. Thomas, W. Marnane, G. Lightbody, and G. B. Boylan, "Performance assessment for EEG-based neonatal seizure detectors," *Clinical Neurophysiology*, vol. 122, pp. 474–482, 2011.

[90] S. Vanhatalo, "Development of neonatal seizure detectors: An elusive target and stretching measuring tapes," *Clinical Neurophysiology*, vol. 122, pp. 435–437, 2011.

[91] A. Ansari, V. Matic, M. De Vos, G. Naulaers, P. Cherian, and S. Van Huffel, "Improvement of an automated neonatal seizure detector using a post-processing technique," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 5859–5862, 2015.

[92] S. R. Mathieson, N. J. Stevenson, E. Low, W. P. Marnane, J. M. Rennie, A. Temko, G. Lightbody, and G. B. Boylan, "Validation of an automated seizure detection algorithm for term neonates," *Clinical Neurophysiology*, vol. 127, pp. 156–168, 2016.

[93] R. R. Clancy and A. Legido, "The Exact Ictal and Interictal Duration of Electroencephalographic Neonatal Seizures," *Epilepsia*, vol. 28, pp. 537–541, 1987.

[94] T. N. Tsuchida, C. J. Wusthoff, and R. A. Shellhaas, "American Clinical Neurophysiology Society critical care monitoring committee," *American clinical neurophysiology society standardized EEG terminology and categorization for the description of continuous EEG monitoring in neonates: report of the American Clinical Neurophysiology Society critical care monitoring committee. J Clin Neurophysiol*, vol. 30, pp. 161–73, 2013.

[95] L. S. Smit, R. J. Vermeulen, W. P. F. Fetter, R. L. M. Strijers, and C. J. Stam, "Neonatal Seizure Monitoring Using Non-Linear EEG Analysis," *Neuropediatrics*, vol. 35, pp. 329–335, 2004.

[96] D. K. Shah, M. T. Mackay, S. Lavery, S. Watson, A. S. Harvey, J. Zempel, A. Mathur, and T. E. Inder, "Accuracy of Bedside Electroencephalographic Monitoring in Comparison With Simultaneous Continuous Conventional Electroencephalography for Seizure Detection in Term Infants," *Pediatrics*, vol. 121, pp. 1146–1154, 2008.

[97] N. S. Abend, A. Gutierrez-Colina, H. Zhao, R. Guo, E. Marsh, R. R. Clancy, and D. J. Dlugos, "Interobserver reproducibility of electroencephalogram interpretation in critically ill children," *Journal of clinical neurophysiology: official publication of the American Electroencephalographic Society*, vol. 28, pp. 15–19, 2011.

[98] N. J. Stevenson, R. R. Clancy, S. Vanhatalo, I. Rosén, J. M. Rennie, and G. B. Boylan, "Interobserver agreement for neonatal seizure detection using multichannel EEG," *Annals of Clinical and Translational Neurology*, vol. 2, pp. 1002–1011, 2015.

[99] C. J. Wusthoff, J. Sullivan, H. C. Glass, R. A. Shellhaas, N. S. Abend, T. Chang, and T. N. Tsuchida, "Interrater agreement in the interpretation of neonatal electroencephalography in hypoxic-ischemic encephalopathy," *Epilepsia*, vol. 58, pp. 429–435, 2017.

[100] C. A. Lantz and E. Nebenzahl, "Behavior and interpretation of the κ statistic: Resolution of the two paradoxes," *Journal of Clinical Epidemiology*, vol. 49, pp. 431–434, 1996.

[101] A. J. Viera and J. M. Garrett, "Understanding interobserver agreement: the kappa statistic," *Fam Med*, vol. 37, no. 5, pp. 360–363, 2005.

[102] M. S. Scher, M. Y. Hamid, D. A. Steppe, M. E. Beggarly, and M. J. Painter, "Ictal and interictal electrographic seizure durations in preterm and term neonates," *Epilepsia*, vol. 34, no. 2, pp. 284–288, 1993.

[103] J. L. Fleiss, B. Levin, and M. C. Paik, "Statistical methods for rates and proportions," in *Statistical Methods for Rates and Proportions*, pp. 598–626, John Wiley & Sons, Inc., 2003.

[104] J. Sim and C. C. Wright, "The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements," *Physical Therapy*, vol. 85, pp. 257–268, 2005.

[105] S. B. Wilson, M. L. Scheuer, C. Plummer, B. Young, and S. Pacia, "Seizure detection: correlation of human experts," *Clinical Neurophysiology*, vol. 114, pp. 2156–2164, 2003.

[106] H. C. Glass, R. A. Shellhaas, C. J. Wusthoff, T. Chang, N. S. Abend, C. J. Chu, M. R. Cilio, D. V. Glidden, S. L. Bonifacio, S. Massey, T. N. Tsuchida, F. S. Silverstein, and J. S. Soul, "Contemporary Profile of Seizures in Neonates: A Prospective Cohort Study," *The Journal of Pediatrics*, vol. 174, pp. 98–103.e1, 2016.

[107] M. Olischar, A. J. Davidson, K. J. Lee, and R. W. Hunt, "Effects of morphine and midazolam on sleep-wake cycling in amplitude-integrated electroencephalography in post-surgical neonates $\geq$ 32 weeks of gestational age," *Neonatology*, vol. 101, no. 4, pp. 293–300, 2012.

[108] V. Vapnik, "Pattern recognition using generalized portrait method," *Automation and remote control*, vol. 24, pp. 774–780, 1963.

[109] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[110] V. Vapnik, *Statistical learning theory*, vol. 1. Wiley New York, 1998.

[111] H. W. Kuhn and A. W. Tucker, "Nonlinear Programming," pp. 481–492, The Regents of the University of California, 1951.

[112] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines: And Other Kernel-based Learning Methods*. New York, NY, USA: Cambridge University Press, 2000.

[113] M. Renardy and R. C. Rogers, *An introduction to partial differential equations*, vol. 13. Springer Science & Business Media, 2006.

[114] J. A. Suykens, T. Van Gestel, and J. De Brabanter, *Least squares support vector machines*. World Scientific, 2002.

[115] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[116] S. Abe, *Support vector machines for pattern classification*, vol. 53. Springer, 2005.

[117] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[118] J. Bouvrie, "Notes on convolutional neural networks," 2006.

[119] T. Serre, A. Oliva, and T. Poggio, "A feedforward architecture accounts for rapid categorization," *Proceedings of the national academy of sciences*, vol. 104, no. 15, pp. 6424–6429, 2007.

[120] T. Serre, M. Kouh, C. Cadieu, U. Knoblich, G. Kreiman, and T. Poggio, "A theory of object recognition: computations and circuits in the feedforward path of the ventral stream in primate visual cortex," tech. rep., AI MEMO MIT, 2005.

[121] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984.

[122] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

[123] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[124] A. E. Hoerl and R. W. Kennard, "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, vol. 12, pp. 55–67, 1970.

[125] A. E. Hoerl and R. W. Kennard, "Ridge Regression: Applications to Nonorthogonal Problems," *Technometrics*, vol. 12, pp. 69–82, 1970.

[126] D. W. Marquardt and R. D. Snee, "Ridge Regression in Practice," *The American Statistician*, vol. 29, no. 1, pp. 3–20, 1975.

[127] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[128] R. Tibshirani, "The lasso method for variable selection in the Cox model," *Statistics in medicine*, vol. 16, no. 4, pp. 385–395, 1997.

[129] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*, vol. 1. Springer series in statistics New York, 2001.

[130] L. El Ghaoui, V. Viallon, and T. Rabbani, "Safe feature elimination for the LASSO," *Submitted, April*, 2011.

[131] D. Ghosh and A. M. Chinnaiyan, "Classification and selection of biomarkers in genomic data using LASSO," *BioMed Research International*, vol. 2005, no. 2, pp. 147–154, 2005.

[132] N. Meinshausen and P. Bühlmann, "High-dimensional graphs and variable selection with the lasso," *The annals of statistics*, pp. 1436–1462, 2006.

[133] N.-J. Hsu, H.-L. Hung, and Y.-M. Chang, "Subset selection for vector autoregressive processes using Lasso," *Computational Statistics & Data Analysis*, vol. 52, pp. 3645–3657, 2008.

[134] M. Lavanga, O. De Wel, A. Caicedo Dorado, K. Jansen, A. Dereymaeker, G. Naulaers, and S. Van Huffel, "Linear and nonlinear functional connectivity methods to predict brain maturation in preterm babies," in *8th international workshop on biosignal interpretation*, pp. 37–40, 2016.

[135] B. Efron and R. J. Tibshirani, *An introduction to the bootstrap*. CRC press, boca raton, fl: chapman & hall ed., 1993.

[136] A. C. Davison and D. V. Hinkley, *Bootstrap Methods and Their Application*. Cambridge University Press, 1997.

[137] B. Efron, "Second Thoughts on the Bootstrap," *Statistical Science*, vol. 18, pp. 135–140, 2003.

[138] J. S. Hahn and D. M. Olson, "Etiology of Neonatal Seizures," *NeoReviews*, vol. 5, pp. e327–e335, 2004.

[139] J. Connell, R. Oozeer, L. de Vries, L. M. Dubowitz, and V. Dubowitz, "Clinical and EEG response to anticonvulsants in neonatal seizures," *Archives of Disease in Childhood*, vol. 64, pp. 459–464, 1989.

[140] M. S. Scher, J. Alvin, L. Gaus, B. Minnigh, and M. J. Painter, "Uncoupling of EEG-clinical neonatal seizures after antiepileptic drug use," *Pediatric Neurology*, vol. 28, pp. 277–280, 2003.

[141] A. M. E. Bye and D. Flanagan, "Spatial and Temporal Characteristics of Neonatal Seizures," *Epilepsia*, vol. 36, pp. 1009–1016, 1995.

[142] J. Gotman, "Automatic seizure detection: improvements and evaluation," *Electroencephalography and Clinical Neurophysiology*, vol. 76, pp. 317–324, 1990.

[143] D. M. Eaton, M. Toet, J. Livingston, I. Smith, and M. Levene, "Evaluation of the Cerebro Trac 2500 for monitoring of cerebral function in the neonatal intensive care," *Neuropediatrics*, vol. 25, pp. 122–128, 1994.

[144] A. Liu, J. S. Hahn, G. P. Heldt, and R. W. Coen, "Detection of neonatal seizures through computerized EEG analysis," *Electroencephalography and Clinical Neurophysiology*, vol. 82, pp. 30–37, 1992.

[145] J. Gotman, D. Flanagan, J. Zhang, and B. Rosenblatt, "Automatic seizure detection in the newborn: methods and initial evaluation," *Electroencephalography and Clinical Neurophysiology*, vol. 103, pp. 356–362, 1997.

[146] M. Roessgen, A. Zoubir, and B. Boashash, "Seizure detection of newborn EEG using a model-based approach," *IEEE Transactions on Biomedical Engineering*, vol. 45, pp. 673–685, 1998.

[147] P. Celka and P. Colditz, "A computer-aided detection of EEG seizures in infants: a singular-spectrum approach and performance comparison," *IEEE Transactions on Biomedical Engineering*, vol. 49, pp. 455–462, 2002.

[148] N. J. Stevenson, J. M. O'Toole, L. J. Rankine, G. B. Boylan, and B. Boashash, "A nonparametric feature for neonatal EEG seizure detection based on a representation of pseudo-periodicity," *Medical Engineering & Physics*, vol. 34, pp. 437–446, 2012.

[149] S. Nagaraj, N. Stevenson, W. Marnane, G. Boylan, and G. Lightbody, "Neonatal Seizure Detection Using Atomic Decomposition With a Novel Dictionary," *IEEE Transactions on Biomedical Engineering*, vol. 61, pp. 2724–2732, 2014.

[150] H. Hassanpour, M. Mesbah, and B. Boashash, "Time–frequency based newborn EEG seizure detection using low and high frequency signatures," *Physiological Measurement*, vol. 25, p. 935, 2004.

[151] A. Temko, E. Thomas, G. Boylan, W. Marnane, and G. Lightbody, "An SVM-based system and its performance for detection of seizures in neonates," in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2009. EMBC 2009*, pp. 2643–2646, 2009.

[152] E. M. Thomas, A. Temko, G. Lightbody, W. P. Marnane, and G. B. Boylan, "Gaussian mixture models for classification of neonatal seizures using EEG," *Physiological Measurement*, vol. 31, pp. 1047–1064, 2010.

[153] H. Li and A. Jeremic, "Neonatal seizure detection using blind distributed detection with correlated decisions," in *2011 Annual*

International Conference of the IEEE Engineering in Medicine and Biology Society,EMBC, pp. 6580–6584, 2011.

[154] J. G. Bogaarts, E. D. Gommer, D. M. W. Hilkman, V. H. J. M. v. Kranen-Mastenbroek, and J. P. H. Reulen, "EEG Feature Pre-processing for Neonatal Epileptic Seizure Detection," *Annals of Biomedical Engineering*, vol. 42, pp. 2360–2368, 2014.

[155] A. Zwanenburg, P. Andriessen, R. K. Jellema, H. J. Niemarkt, T. G. Wolfs, B. W. Kramer, and T. Delhaas, "Using trend templates in a neonatal seizure algorithm improves detection of short seizures in a foetal ovine model," *Physiological Measurement*, vol. 36, no. 3, p. 369, 2015.

[156] A. Aarabi, R. Grebe, and F. Wallois, "A multistage knowledge-based system for EEG seizure detection in newborn infants," *Clinical Neurophysiology*, vol. 118, pp. 2781–2797, 2007.

[157] P. W. Holland and R. E. Welsch, "Robust regression using iteratively reweighted least-squares," *Communications in Statistics - Theory and Methods*, vol. 6, pp. 813–827, 1977.

[158] F. Mormann, K. Lehnertz, P. David, and C. E. Elger, "Mean phase coherence as a measure for phase synchronization and its application to the EEG of epilepsy patients," *Physica D: Nonlinear Phenomena*, vol. 144, pp. 358–369, 2000.

[159] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003.

[160] J. C. Platt, "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods," *ADVANCES IN LARGE MARGIN CLASSIFIERS*, vol. 10, no. 3, pp. 61–74, 1999.

[161] H.-T. Lin, C.-J. Lin, and R. C. Weng, "A note on Platt's probabilistic outputs for support vector machines," *Machine Learning*, vol. 68, pp. 267–276, 2007.

[162] A. Ansari, P. Cherian, A. Caicedo, M. De Vos, G. Naulaers, and S. Van Huffel, "Improved neonatal seizure detection using adaptive learning," in *Engineering in Medicine and Biology Society (EMBC), 2017 39th Annual International Conference of the IEEE*, pp. 2810–2813, IEEE, 2017.

[163] H. C. Glass, D. Glidden, R. J. Jeremy, A. J. Barkovich, D. M. Ferriero, and S. P. Miller, "Clinical Neonatal Seizures are Independently Associated with Outcome in Infants at Risk for Hypoxic-Ischemic Brain Injury," *The Journal of Pediatrics*, vol. 155, pp. 318–323, 2009.

[164] G. L. Holmes, "Seizure-induced neuronal injury Animal data," *Neurology*, vol. 59, pp. S3–S6, 2002.

[165] S. Mitra, G. Bale, S. Mathieson, C. Uria-Avellanal, J. Meek, I. Tachtsidis, and N. J. Robertson, "Changes in Cerebral Oxidative Metabolism during Neonatal Seizures Following Hypoxic–Ischemic Brain Injury," *Frontiers in Pediatrics*, vol. 4, 2016.

[166] E. C. Wirrell, E. A. Armstrong, L. D. Osman, and J. Y. Yager, "Prolonged seizures exacerbate perinatal hypoxic-ischemic brain damage," *Pediatric Research*, vol. 50, pp. 445–454, 2001.

[167] J. Y. Yager, E. A. Armstrong, H. Miyashita, and E. C. Wirrell, "Prolonged neonatal seizures exacerbate hypoxic-ischemic brain damage: correlation with cerebral energy metabolism and excitatory amino acid release," *Developmental Neuroscience*, vol. 24, no. 5, pp. 367–381, 2002.

[168] A. H. Ansari, P. Cherian, A. Caicedo, G. Naulaers, M. De Vos, and S. Van Huffel, "Neonatal seizure detection using deep convolutional neural networks," *International Journal of Neural Systems*, 2018 In Press.

[169] J. M. Rennie, "Neonatal seizures," *European Journal of Pediatrics*, vol. 156, pp. 83–87, 1997.

[170] R. Ahmed, A. Temko, W. P. Marnane, G. Boylan, and G. Lightbody, "Exploring temporal information in neonatal seizures using a dynamic time warping based SVM kernel," *Computers in Biology and Medicine*, vol. 82, pp. 100–110, 2017.

[171] A. Aarabi, F. Wallois, and R. Grebe, "Automated neonatal seizure detection: A multistage classification system through feature selection based on relevance and redundancy analysis," *Clinical Neurophysiology*, vol. 117, pp. 328–340, 2006.

[172] Y. Bengio, "Learning Deep Architectures for AI," *Foundations and Trends® in Machine Learning*, vol. 2, pp. 1–127, 2009.

[173] S. Ghosh-Dastidar and H. Adeli, "Improved spiking neural networks for eeg classification and epilepsy and seizure detection," *Integrated Computer-Aided Engineering*, vol. 14, no. 3, pp. 187–212, 2007.

[174] S. Ghosh-Dastidar, H. Adeli, and N. Dadmehr, "Principal component analysis-enhanced cosine radial basis function neural network for robust epilepsy and seizure detection," *IEEE Transactions on Biomedical Engineering*, vol. 55, no. 2, pp. 512–518, 2008.

[175] S. Ghosh-Dastidar and H. Adeli, "A new supervised learning algorithm for multiple spiking neural networks with application in epilepsy and seizure detection," *Neural networks*, vol. 22, no. 10, pp. 1419–1431, 2009.

[176] A. Gabor, R. Leach, and F. Dowla, "Automated seizure detection using a self-organizing neural network," *Electroencephalography and clinical Neurophysiology*, vol. 99, no. 3, pp. 257–266, 1996.

[177] W. Webber, R. P. Lesser, R. T. Richardson, and K. Wilson, "An approach to seizure detection using an artificial neural network (ann)," *Electroencephalography and clinical Neurophysiology*, vol. 98, no. 4, pp. 250–272, 1996.

[178] Y. LeCun, Y. Bengio, and others, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, no. 10, p. 1995, 1995.

[179] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A Fast Learning Algorithm for Deep Belief Nets," *Neural Comput.*, vol. 18, pp. 1527–1554, 2006.

[180] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion," *Journal of Machine Learning Research*, vol. 11, no. Dec, pp. 3371–3408, 2010.

[181] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[182] B. Hutchinson, L. Deng, and D. Yu, "Tensor deep stacking networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1944–1957, 2013.

[183] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, and G. Penn, "Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pp. 4277–4280, IEEE, 2012.

[184] D. C. Ciresan, U. Meier, J. Masci, L. Maria Gambardella, and J. Schmidhuber, "Flexible, high performance convolutional neural networks for image classification," in *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, vol. 22, p. 1237, Barcelona, Spain, 2011.

[185] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[186] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back, "Face recognition: A convolutional neural-network approach," *IEEE transactions on neural networks*, vol. 8, no. 1, pp. 98–113, 1997.

[187] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[188] H. Cecotti and A. Graeser, "Convolutional neural network with embedded fourier transform for EEG classification," in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pp. 1–4, IEEE, 2008.

[189] H. Cecotti and A. Graser, "Convolutional neural networks for P300 detection with application to brain-computer interfaces," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 3, pp. 433–445, 2011.

[190] P. W. Mirowski, Y. LeCun, D. Madhavan, and R. Kuzniecky, "Comparing SVM and convolutional networks for epileptic seizure prediction from intracranial EEG," in *Machine Learning for Signal Processing, 2008. MLSP 2008. IEEE Workshop on*, pp. 244–249, IEEE, 2008.

[191] U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, and H. Adeli, "Deep convolutional neural network for the automated detection and diagnosis of seizure using eeg signals," *Computers in biology and medicine*, 2017.

[192] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pp. 4580–4584, IEEE, 2015.

[193] M. Tan, C. d. Santos, B. Xiang, and B. Zhou, "Lstm-based deep learning models for non-factoid answer selection," *arXiv preprint arXiv:1511.04108*, 2015.

[194] S. Bai, "Growing random forest on deep convolutional neural networks for scene categorization," *Expert Systems with Applications*, vol. 71, pp. 279–287, 2017.

[195] X.-X. Niu and C. Y. Suen, "A novel hybrid CNN–SVM classifier for recognizing handwritten digits," *Pattern Recognition*, vol. 45, no. 4, pp. 1318–1325, 2012.

[196] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: an astounding baseline for recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 806–813, 2014.

[197] A. Temko, G. Boylan, W. Marnane, and G. Lightbody, "Robust neonatal eeg seizure detection through adaptive background modeling," *International journal of neural systems*, vol. 23, no. 04, p. 1350018, 2013.

[198] S. Faul, G. Boylan, S. Connolly, W. Marnane, and G. Lightbody, "Chaos theory analysis of the newborn eeg-is it worth the wait?," in *IEEE Int. Workshop on Intelligent Signal Processing*, pp. 381–386, 2005.

[199] A. Vedaldi and K. Lenc, "MatConvNet - Convolutional Neural Networks for MATLAB," in *Proceeding of the ACM Int. Conf. on Multimedia*, 2015.

[200] C. Nosarti, K. W. Nam, M. Walshe, R. M. Murray, M. Cuddy, L. Rifkin, and M. P. Allin, "Preterm birth and structural brain alterations in early adulthood," *NeuroImage: Clinical*, vol. 6, pp. 180–191, 2014.

[201] R. H. Pickler, J. M. McGrath, B. A. Reyna, N. McCain, M. Lewis, S. Cone, P. Wetzel, and A. Best, "A Model of Neurodevelopmental Risk and Protection for Preterm Infants," *The Journal of Perinatal & Neonatal Nursing*, vol. 24, no. 4, pp. 356–365, 2010.

[202] M. S. Scher, "Ontogeny of EEG-sleep from neonatal through infancy periods," *Sleep Medicine*, vol. 9, no. 6, pp. 615–636, 2008.

[203] A. Dereymaeker, K. Pillay, J. Vervisch, M. De Vos, S. Van Huffel, K. Jansen, and G. Naulaers, "Review of sleep-EEG in preterm and term neonates," *Early Human Development*, 2017.

[204] K. A. Allen, "Promoting and protecting infant sleep.," *Advances in neonatal care*, vol. 12, no. 5, pp. 288–291, 2012.

[205] A. van den Hoogen, C. J. Teunis, R. A. Shellhaas, S. Pillen, M. Benders, and J. Dudink, "How to improve sleep in a neonatal intensive care unit: A systematic review," *Early Human Development*, 2017.

[206] N. Koolen, L. Oberdorfer, Z. Rona, V. Giordano, T. Werther, K. Klebermass-Schrehof, N. Stevenson, and S. Vanhatalo, "Automated classification of neonatal sleep states using EEG," *Clinical Neurophysiology*, vol. 128, pp. 1100–1108, 2017.

[207] K. Palmu, T. Kirjavainen, S. Stjerna, T. Salokivi, and S. Vanhatalo, "Sleep wake cycling in early preterm infants: Comparison of polysomnographic recordings with a novel EEG-based index," *Clinical Neurophysiology*, vol. 124, no. 9, 2013.

[208] N. J. Stevenson, K. Palmu, S. Wikström, L. Hellström-Westas, and S. Vanhatalo, "Measuring brain activity cycling (BAC) in long term EEG

monitoring of preterm babies," *Physiological Measurement*, vol. 35, no. 7, pp. 1493–1508, 2014.

[209] V. Krajca, S. Petranek, K. Paul, M. Matousek, J. Mohylova, and L. Lhotska, "Automatic Detection of Sleep Stages in Neonatal EEG Using the Structural Time Profiles," in *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, pp. 6014–6016, IEEE, 2005.

[210] V. Krajča, S. Petránek, J. Mohylová, K. Paul, V. Gerla, and L. Lhotská, "Neonatal EEG Sleep Stages Modelling by Temporal Profiles," in *In: Moreno Díaz R., Pichler F., Quesada Arencibia A. (eds) Computer Aided Systems Theory – EUROCAST 2007. EUROCAST 2007. Lecture Notes in Computer Science*, vol. 4739, (Berlin), pp. 195–201, 2007.

[211] J. S. Barlow, "Computer characterization of tracé alternant and REM sleep patterns in the neonatal EEG by adaptive segmentation–an exploratory study.," *Electroencephalography and clinical neurophysiology*, vol. 60, pp. 163–73, feb 1985.

[212] L. Fraiwan, K. Lweesy, N. Khasawneh, M. Fraiwan, H. Wenz, and H. Dickhaus, "Time frequency analysis for automated sleep stage identification in fullterm and preterm neonates.," *Journal of Medical Systems*, vol. 35, no. 4, pp. 693–702, 2011.

[213] M. Lavanga, O. De Wel, A. Caicedo Dorado, E. Heremans, K. Jansen, A. Dereymaeker, G. Naulaers, and S. Van Huffel, "Automatic quiet sleep detection based on multifractality in preterm neonates: effects of maturation," in *Proc. 39th Annual International Conference of the IEEE Engineering in Medicine & Biology Society*, (Seogwipo, South Korea), 2017.

[214] D. Shen, G. Wu, and H.-I. Suk, "Deep Learning in Medical Image Analysis.," *Annual review of biomedical engineering*, vol. 19, pp. 221–248, jun 2017.

[215] U. R. Acharya, H. Fujita, S. L. Oh, Y. Hagiwara, J. H. Tan, and M. Adam, "Application of deep convolutional neural network for automated detection of myocardial infarction using ECG signals," *Information Sciences*, vol. 415-416, pp. 190–198, 2017.

[216] O. Tsinalis, P. M. Matthews, Y. Guo, and S. Zafeiriou, "Automatic Sleep Stage Scoring with Single-Channel EEG Using Convolutional Neural Networks," *ArXiv*, oct 2016.

[217] A. Supratak, H. Dong, C. Wu, and Y. Guo, "DeepSleepNet: a Model for Automatic Sleep Stage Scoring based on Raw Single-Channel EEG,"

IEEE transactions on neural systems and rehabilitation engineering, mar 2017.

[218] K. Mikkelsen and M. De Vos, "Personalizing deep learning models for automatic sleep staging." arXiv:1801.02645 [q-bio.NC].

[219] I. Korotchikova, N. J. Stevenson, V. Livingstone, C. A. Ryan, and G. B. Boylan, "Sleep-wake cycle of the healthy term newborn infant in the immediate postnatal period.," *Clinical neurophysiology*, vol. 127, no. 4, pp. 2095–2101, 2016.

[220] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, pp. 861–874, 2006.

[221] A. H. Ansari, P. Cherian, A. Caicedo, K. Jansen, A. Dereymaeker, L. De Wispelaere, C. Dielman, J. Vervisch, P. Govaert, M. De Vos, *et al.*, "Weighted performance metrics for automatic neonatal seizure detection using multi-scored eeg data," *IEEE journal of biomedical and health informatics*, 2017 In Press.

[222] E. Thomas, A. Temko, W. Marnane, G. Boylan, and G. Lightbody, "Discriminative and Generative Classification Techniques Applied to Automated Neonatal Seizure Detection," *IEEE Journal of Biomedical and Health Informatics*, vol. 17, pp. 297–304, 2013.

[223] A. W. Bowman and A. Azzalini, *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*. OUP Oxford, 1997.

[224] J. A. K. Suykens, J. De Brabanter, L. Lukas, and J. Vandewalle, "Weighted least squares support vector machines: robustness and sparse approximation," *Neurocomputing*, vol. 48, pp. 85–105, 2002.

[225] E. M. Mizrahi and P. Kellaway, "Characterization and classification of neonatal seizures," *Neurology*, vol. 37, pp. 1837–1837, 1987.

[226] M. S. Scher, M. J. Painter, I. Bergman, M. A. Barmada, and J. Brunberg, "EEG diagnoses of neonatal seizures: clinical correlations and outcome," *Pediatric Neurology*, vol. 5, pp. 17–24, 1989.

[227] R. A. Shellhaas, A. I. Soaita, and R. R. Clancy, "Sensitivity of Amplitude-Integrated Electroencephalography for Neonatal Seizure Detection," *Pediatrics*, vol. 120, pp. 770–777, 2007.

[228] R. A. Shellhaas and R. R. Clancy, "Characterization of neonatal seizures by conventional EEG and single-channel EEG," *Clinical Neurophysiology*, vol. 118, pp. 2156–2161, 2007.

[229] D. K. Shah, G. B. Boylan, and J. M. Rennie, "Monitoring of seizures in the newborn," *Archives of Disease in Childhood - Fetal and Neonatal Edition*, vol. 97, pp. F65–F69, 2012.

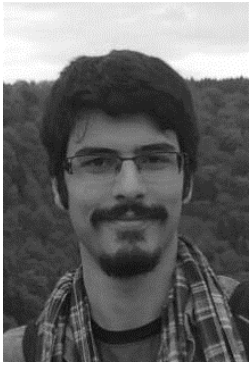
[230] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, pp. 159–174, 1977.

[231] D. Maynard, P. F. Prior, and D. Scott, "Device for continuous monitoring of cerebral activity in resuscitated patients.," *British medical journal*, vol. 4, no. 5682, p. 545, 1969.

[232] D. Zhang and H. Ding, "Calculation of compact amplitude-integrated eeg tracing and upper and lower margins using raw eeg data," *Health*, vol. 5, no. 05, p. 885, 2013.

[233] K. Paul, V. Krajča, Z. Roth, J. Melichar, and S. Petránek, "Comparison of quantitative eeg characteristics of quiet and active sleep in newborns," *Sleep Medicine*, vol. 4, no. 6, pp. 543–552, 2003.

Curriculum vitae



Amir H. Ansari was born on September 12, 1986 in Tehran, Iran. He received his B.Sc. degree in electrical engineering with distinction in control engineering from the University of Shahed in 2009 and M.Sc. in bio-electrical engineering from University of Tehran, Iran, in 2012. Then, he joined the biomedical data processing research group at the department of Electrical Engineering (ESAT-STADIUS) of the KU Leuven in Belgium as a pre-doctoral student in January 2013. Since October 2013, he has been a PhD student under the supervision of Prof. Dr. Ir. Sabine Van Huffel and Prof. Dr. Gunnar Naulaers. As being part of the NeoGuard consortium, he had a collaboration with the Department of Development and Regeneration and the NICU of the University Hospitals Leuven, the Department of Pediatrics at Koningin Paola Kinderziekenhuis, Antwerp, and the department of Neurology at Erasmus MC Rotterdam, the Netherlands. He also was collaborating with the institute of biomedical engineering in the University of Oxford, UK, specifically with the Computational Intelligence In Biomedical Monitoring Laboratory led by Prof. Dr. Maarten De Vos. He received FWO support for a one-month research visit in this lab.

His current research interests are centered on machine learning techniques including neuro-fuzzy networks, convolutional neural networks, deep learning, and support vector machines, as well as their applications on biomedical problems, such as seizure detection and sleep stage classification.

List of publications

Papers in international journals

1. Matic, V., Cherian, P. J., Koolen, N., **Ansari, A. H.**, Naulaers, G., Govaert, P., Van Huffel, S., De Vos, M., Vanhatalo, S. (2015). Objective differentiation of neonatal EEG background grades using detrended fluctuation analysis. Frontiers in human neuroscience, 9(189), 1-14.

2. **Ansari, A. H.**, Cherian, P. J., Dereymaeker, A., Matic, V., Jansen, K., De Wispelaere, L., Dielman, C., Vervisch, J., Swarte, R. M., Govaert, P., Naulaers, G., De Vos, M., Van Huffel, S., (2016). Improved multi-stage neonatal seizure detection using a heuristic classifier and a data-driven post-processor. Clinical Neurophysiology, 127(9), 3014-3024.

3. **(Dereymaeker, A. *, Ansari, A. H. *)**, Jansen, K., Cherian, P. J., Vervisch, J., Govaert, P., De Wispelaere, L., Dielman, C., Matic, V., Caicedo, A., De Vos, M., Van Huffel, S., Naulaers, G., (2017). Interrater agreement in visual scoring of neonatal seizures based on majority voting on a web-based system: The Neoguard EEG database. Clinical Neurophysiology, 128(9), 1737-1745.
   *** Joint first author.**

4. **Ansari, A. H.**, Cherian, P. J., Caicedo, A., Jansen, K., Dereymaeker, A., De Wispelaere, L., Dielman, C., Vervisch, J., Govaert, P., De Vos, M., Naulaers, G., Van Huffel, S., (2018). Weighted performance metrics for automatic neonatal seizure detection using multi-scored EEG data. IEEE Journal of Biomedical and Health Informatics, In Press: 22(4), July 2018.

5. **Ansari, A. H.**, Cherian, P. J., Caicedo, A., De Vos, M., Naulaers, G., Van Huffel, S., (2018). Neonatal seizure detection using deep convolutional neural networks. Accepted for publication in International Journal of Neural Systems. In Press.

6. Dereymaeker, A., Matic, V., **Ansari, A. H.**, Cherian, P. J., Vervisch, J., Govaert, P., De Wel, O., Caicedo, A., De Vos, M., Van Huffel, S., Naulaers, G., Jansen, K., (2018). Automated EEG background analysis to identify neonates with hypoxic-ischemic encephalopathy treated with hypothermia at risk for adverse outcome: a pilot study. Accepted for publication in Pediatrics & Neonatology. In Press.

Papers submitted to international journals

7. **Ansari, A. H.**, Cherian, P. J., Caicedo, A., Dereymaeker, A., Jansen, K., De Wispelaere, L., Dielman, C., Vervisch, J., Govaert, P., De Vos, M., Naulaers, G., Van Huffel, S., (2017). NeoGuard: a public, online learning platform for neonatal seizures. Submitted to the Clinical EEG and Neuroscience.

8. **(Ansari, A. H. *, De Wel, O. *)**, Lavanga, M., Caicedo, A., Dereymaeker, A., Jansen, K., Vervisch, J., De Vos, M., Naulaers, G., Van Huffel, S., (2018). Sleep stage classification in preterm infants using deep convolutional neural networks. Submitted to the Journal of Neural Engineering.
   *** Joint first author.**

Papers in proceedings of international conferences

9. **Ansari, A. H.**, Matic, V., De Vos, M., Naulaers, G., Cherian, P. J., Van Huffel, S. (2015, August). Improvement of an automated neonatal seizure detector using a post-processing technique. In Engineering in Medicine and Biology Society (EMBC), 37th Annual International Conference of the IEEE (pp. 5859-5862). IEEE.

10. **Ansari, A. H.**, Cherian, P. J., Caicedo, A., De Vos, M., Naulaers, G., Van Huffel, S. (2017, July). Improved neonatal seizure detection using adaptive learning. In Engineering in Medicine and Biology Society (EMBC), 39th Annual International Conference of the IEEE (pp. 2810-2813). IEEE.

Abstracts in proceedings of international conferences

11. **A. H. Ansari**, P. J. Cherian, V. Matic, A. Dereymaeker, L. De Wispelaere, C. Dielman, J. Vervisch, P. Govaert, G. Naulaers, M. De Vos, S. Van Huffel, "Performance measurement of an automated neonatal seizure detector in EEG seizure data showing inter-rater disagreement," in the 9th International Conference on Brain Monitoring and Neuroprotection in the Newborn, Cork, Ireland, October 2015.

12. V. Matic, P. J. Cherian, **A. H. Ansari**, N. Koolen, G. Naulaers, P. Govaert, S. Van Huffel, M. De Vos "Metrics for the background EEG assessment in asphyxiated infants," in IEEE Engineering in Medicine and Biology Society, Milan, Italy, August 2015.

13. **A. H. Ansari**, P. J. Cherian, V. Matic, A. Dereymaeker, L. De Wispelaere, C. Dielman, J. Vervisch, P. Govaert, G. Naulaers, M. De Vos, S. Van Huffel, "Online neonatal seizure learning platform," in the 10th International Conference on Brain Monitoring and Neuroprotection in the Newborn, Killarney, Ireland, October 2017.



FACULTY OF ENGINEERING SCIENCE
DEPARTMENT OF ELECTRICAL ENGINEERING (ESAT)
Stadius Center for Dynamical Systems,
Signal Processing and Data Analytics
Kasteelpark Arenberg 10, bus 2446
B-3001 Leuven
Belgium
amirhossein.ansari@kuleuven.be
http://www.esat.kuleuven.be/stadius