Dynamic Classifiers for Neonatal Brain Monitoring

Rehan Ahmed
M.Sc.

NATIONAL UNIVERSITY OF IRELAND, CORK

School of Engineering
Department of Electrical and Electronic Engineering
Irish Centre for Fetal and Neonatal Translational Research
(INFANT)

Thesis submitted for the degree of Doctor of Philosophy

2016

Supervisors: Dr. Gordon Lightbody
Prof. Liam Marnane
Prof. Geraldine Boylan
Dr. Andrey Temko

Head of Department: Prof. Nabeel A. Riza

Research supported by Science Foundation Ireland

Contents

List of Figures

List of Tables

List of Acronyms

MRC Maximal Ratio Combining

AC Alternating Current

AD Atomic Decomposition

AED Anti-epileptic Drugs

aEEG Amplitude-Integrated Electroencephalogram

ANSeR Algorithm for Neonatal Seizure Recognition

AUC Area Under Curve

CHI/b Cerebral Brain Index

CT Computerized Tomography

DTW Dynamic Time Warping

ECG Electrocardiogram

EEG Electroencephalogram

EM Expectation Maximization

FD/h False Detections per hour

FN False Negative

FP False Positive

FFT Fast Fourier Transform

GDR Good Detection Rate

GDTW Gaussian Dynamic Time Warping

GMM Gaussian Mixture Model

HF High Frequency

HIE Hypoxic-Ischemic Encephalopathy

HMM Hidden Markov Model

HR Heart Rate

HRV Heart Rate Variability

IBI Inter Burst Interval

LDA Linear Discriminant Analysis

LF Low Frequency

LOO Leave One Out

MAF Moving Average Filter

MAP Maximum a-Posteriori

MFDD Mean False Detection Duration

ML Maximum Likelihood

MRI Magnetic Resonance Imaging

NICU Neonatal Intensive Care Unit

NN Normalized RR interval

PCA Principal Component Analysis

PDF Probability Density Function

PR Precision Recall

PSD Power Spectral Density

RBF Radial Basis Function

RMS Root Mean Square

ROC Receiver Operating Curve

RR R-R peak interval

SD Standard Deviation

SSA Singular Spectrum Analysis

SVM Support Vector Machine

TH Therapeutic Hypothermia

TINN Triangular Interpolation of NN interval

TN True Negative

TP True Positive

UBM Universal Background Model

VLF Very Low Frequency

Nomenclature

The list is sorted in the order of each symbol's appearance in the thesis.

$\mathbf{X}$	A set of feature vectors
$\mathbf{x}$	A single feature vector
N	Number of feature vectors in the set
n	Dimension of a single feature vector
i	Index of an observation in the training set
y	Class of the feature vector
f	Classifier decision function
$\mathbf{w}$	Slope of the decision hyperplane in SVM
b	Bias of the hyperplane in SVM
d, D	Usually used for the distances
L	Lagrangian function
α	Lagrangian multipliers
Ψ	Mapping function
$\mathbb{F}$	Mapped feature space
C	Regularization parameter of SVM
$\Bbbk$	Kernel function
σ	Parameter of the Gaussian kernel function
P	Posterior Probability
A, B	Parameters of sigmoid function
ξ	Slack variable of SVM
g	Gaussian mixture model

w	Weight of a Gaussian component
$\mathbf{m}$	Mean of the Gaussian component
$\mathbf{\Sigma}$	Covariance matrix of the Gaussian component
j	Index of a Gaussian component in GMM
θ	Represents the parameters (mean, covariance, weight) of a GMM model
M	Total number of Gaussian components in a GMM
p	Likelihood
$\hat{\{\}}$	Estimated values
$\mathfrak{f}$	First moment (MAP estimation)
$\mathfrak{s}$	Second moment (MAP estimation)
c	Count (MAP estimation)
β	Adaptation coefficients for MAP adaptation
δ	Relevance factor of MAP adaptation
S, R	Sequences of feature vectors
F	Fisher score vector
$\mathbf{v}$	Supervector
ζ	The set of all finite strings (string kernel)
W	A warp path in gram matrix (DTW)
K	Length of the warp path (DTW)
k	A vertex of the warp path (DTW)
s, r	A single feature vectors in sequence
ϖ	Distance from the center of the sphere to the input feature space (Polynomial kernel)
Υ	Set of all possible paths in gram matrix (DTW)
$\check{I}$	Number of channels in EEG recording

$\breve{i}$	Index of a EEG channel
$\breve{k}$	Number of classes
$\breve{a}, \breve{b}$	Index of classes in a classification problem
$\breve{W}$	Winning grade/class from a output vector of multiclass classifier
Q	Total number of supervectors in a recording
q	Index of a supervector
Z	Votes vector of multiclass classifier output

I, Rehan Ahmed, certify that this thesis is my own work and I have not obtained a degree in this university or elsewhere on the basis of the work submitted in this thesis.

Rehan Ahmed

Abstract

Brain injury due to lack of oxygen or impaired blood flow around the time of birth, may cause long term neurological dysfunction or death in severe cases. The treatments need to be initiated as soon as possible and tailored according to the nature of the injury to achieve best outcomes. The Electroencephalogram (EEG) currently provides the best insight into neurological activities. However, its interpretation presents formidable challenge for the neurophsiologists. Moreover, such expertise is not widely available particularly around the clock in a typical busy Neonatal Intensive Care Unit (NICU). Therefore, an automated computerised system for detecting and grading the severity of brain injuries could be of great help for medical staff to diagnose and then initiate on-time treatments.

In this study, automated systems for detection of neonatal seizures and grading the severity of Hypoxic-Ischemic Encephalopathy (HIE) using EEG and Heart Rate (HR) signals are presented. It is well known that there is a lot of contextual and temporal information present in the EEG and HR signals if examined at longer time scale. The systems developed in the past, exploited this information either at very early stage of the system without any intelligent block or at very later stage where presence of such information is much reduced. This work has particularly focused on the development of a system that can incorporate the contextual information at the middle (classifier) level. This is achieved by using dynamic classifiers that are able to process the sequences of feature vectors rather than only one feature vector at a time.

An automated system based on Dynamic Time Warping (DTW) and Support Vector Machine (SVM) classifier is proposed to incorporate the characteristics of temporal evolution of neonatal seizures into the developed ANSeR neonatal seizure detector. The system is validated on a large dataset of 261 hours of EEG recordings from 17 neonates. A comparison of a previously developed static-SVM system and the proposed DTW based dynamic-SVM system is presented. The fusion of both techniques is shown to increase the seizure detection rate from 78.6% to 82.6% at a significantly low false alarm rate of 1 false detection per 4 hours. Most importantly, a 12% improvement in the detection of short seizures is obtained.

A novel automated system to classify the severity of HIE in neonates using EEG is also presented. A cross disciplinary method is applied that uses the sequences of short-term features of EEG to grade an hour long recording. Novel post-processing techniques are proposed based on majority voting and prob-

abilistic methods. The proposed system is validated with one-hour-long EEG recordings from 54 full term neonates. The system improved both the accuracy and the confidence/quality of the produced decision. An overall accuracy of 87% is achieved. With a new label 'unknown' assigned to the recordings with lower confidence levels, an accuracy of 96% is attained.

EEG is generally used to assess brain injury but it is neither widely recorded after birth nor is the expertise to interpret it commonly available. A novel system to classify HIE injury using heart rate variability is thus proposed in this study. The system makes decisions based on long-term statistical features extracted from the short-term HR features. The preliminary results show the promising performance and robustness of the proposed method. This tool can serve as a decision support system in remote maternity units to help clinical staff to grade HIE.

Lastly, a pilot study on the fusion of EEG and ECG signals to grade HIE injury is presented. Methods of information fusion at two different levels are investigated; 1) classifier level and 2) decision level. This is a fundamental study in this area. The advantages and disadvantages of both systems are presented through experimental results. Particularly, this work has highlighted different areas that need to be investigated in future to create a multi-modal system that can produce an overall score of the babies brain health based on different sources of information.

To Father, Mother and Amna

Acknowledgements

Firstly I would like to thank Allah and then my parents as without their struggle, sacrifices (both financially and emotionally), I would not have been able to come to this point in my life.

I would like to thank Gordon Lightbody, who was my main mentor in this work. Gordon not only supervised and guided me for my PhD related work but also groomed my character as a researcher and academic. In the desperate times of the Ph.D., I always looked upon him to get motivation and positivity. I learned a lot from his always humble, cooperative and positive nature. Probably if any part of my personality in future will reflect these characteristics then it will be because of Gordon.

Andrey Temko, the second in command but my first contact for any technical help. His contribution to this work is enormous. I would not be wrong to say, that if this work has achieved any mark in quality then its because of Andrey's perfectionism and far-sightedness. I must admit that I had difficult and distressed times in/after his meetings but I always learned the most after those meetings.

Liam Marnane, was the first person who introduced me to this group and I will be always grateful for that first meeting in Alcala. Liam's suggestions towards all the publications and this thesis had far-reaching benefits in making the concepts and overall theme of the work more concise and understandable. I also benefited from his amazing presentation skills of conveying strong and meaningful messages.

Geraldine Boylan, will always be praised for her continuous backbreaking work of building a pioneering neonatal EEG seizure database. One must admire her great efforts for highlighting the ongoing research in neonatal brain monitoring on a global scale. I would like to thank her for taking time for all the reviews, suggestions and prompt feedbacks even under tight schedules.

Thanks to all the postgraduate colleagues (from the first and last half of my PhD). The usual flow-process in a typical Ph.D. cycle is that the seniors usually give you advices, scares you with their difficult times and then successfully finishes leaving you with the mixed feelings of sadness and motivation to finish your own PhD. Second batch(juniors), becomes your good friends as they are the ones who provide you encouragement in your own difficult times. I would like to specially thank the ones (Allison, Brian, Haiyang, James, Oksana, Shima, Wentao) who joined the room in the last two years as that was the most stressful time of my Ph.D. The fun talks, technical discussions, gatherings over tea, all have contributed to relax me and this work indirectly.

Thanks to all Pakistani friends (Asfand, Iftikhar, Junaid, Nasir, Razak, Saim, Usman, Zubair) with whom I have spent most of my leisure time. Most of them are now like a family. I can easily say that the time spent with these guys was and will be most cherished moments of my life.

Last but not the least, I would like to acknowledge the person who bore and pampered me in my most difficult and stressful time, my better-half Amna. Her

contribution to this work might not be direct but indirectly her cooperation and care made this work possible.

Chapter 1

Introduction

The birth of a baby is a moment, enjoyed by many parents. However, in some rare cases these moments can be over shadowed due to the poor health of the new-born. On average 10% (varies with gestational age) of all neonates require admission to the Neonatal Intensive Care Units (NICU) due to problems in delivery (Harisson & Goodman 2015).

Brain injury around the time of birth, due to lack of oxygen (hypoxia) or impaired blood flow in the brain (ischemia), may cause neural dysfunction or death in severe cases. In fact, according to recent surveys and a world health organization report, out of 2.6 million annual neonatal deaths in 2013, Hypoxia Ischemia (HI) is the second (10.5%) leading cause of neonatal deaths globally (UNICEF 2013, Liu et al. 2015).

The severity of brain injury in babies that have suffered HI insult is often very difficult to assess. Electroencephalography (EEG) which measures the electrical activity of the brain, currently offers the best insight into brain function. The clinical assessment methods such as APGAR score or Sarnat grading do not highly correlate with abnormal brain function. Magnetic Resonance Imaging (MRI) although an accurate method for diagnosis, cane however only provide a snapshot of the brain function and does not allow continuous monitoring of the brain. Moreover the availability of MRI equipment is scarce in hospitals. EEG is considered the gold standard for detecting the timing and severity of neonatal brain injuries. Moreover, it also helps in monitoring any progression or improvement in brain injury after treatment is initiated.

Electrodes are attached on the neonate's scalp to record the EEG (Figure 1.1b). The normal neonatal EEG is a very random signal without any obvious pattern (Figure 1.2). EEG becomes abnormal when a certain pattern starts to appear and it loses its natural chaotic behaviour. Clinically, EEG is visually

(a)

(b)

Figure 1.1: (a) A neonate being monitored in the NICU. (b) Electrodes placed on a neonate's head to record the EEG.

analysed to find any abnormal patterns. Typically, an EEG recording can last for more than 24 hours. The neurologist analyses a 10-30 seconds window at a time and then proceeds to the next window. Given this scale, a 3 hour EEG recording, as recorded on the traditional paper trace, would equate to the length of a football pitch. Furthermore, the expertise to interpret the neonatal EEG is very scarce in busy NICU. Amplitude-integrated EEG (aEEG) can be seen as a good alternative as it compresses and highly filters the EEG, it typically misses many short abnormal brain events (Rennie et al. 2004). Thus, an automated system that can detect abnormal brain functions in neonates could significantly help clinical staff in its diagnosis, prognosis and to decide prompt treatment.

1.1 Abnormal brain functions investigated in this work

This work presents automated systems for; 1- classification of Hypoxic Ischemic Encephalopathy (HIE) using EEG & Heart Rate (HR) and 2- detection of neonatal seizures in continuous EEG recordings.

HIE is caused due to lack of oxygen or impaired blood flow to the brain. The long-term outcome depends on the severity of the initial HIE insult. HIE is generally graded into four grades (Figure 1.2b.1). It can be seen, that the HIE affected EEG exhibits various patterns, some of which may be similar across HIE grades; however it is the inter-pattern variability or the way such patterns occur over the whole EEG recording which helps to characterize its grade. The treatment involves, cooling the infant to a body temperature of between 33-34°C for 72 hours without interruption. However to be effective, it must be com-

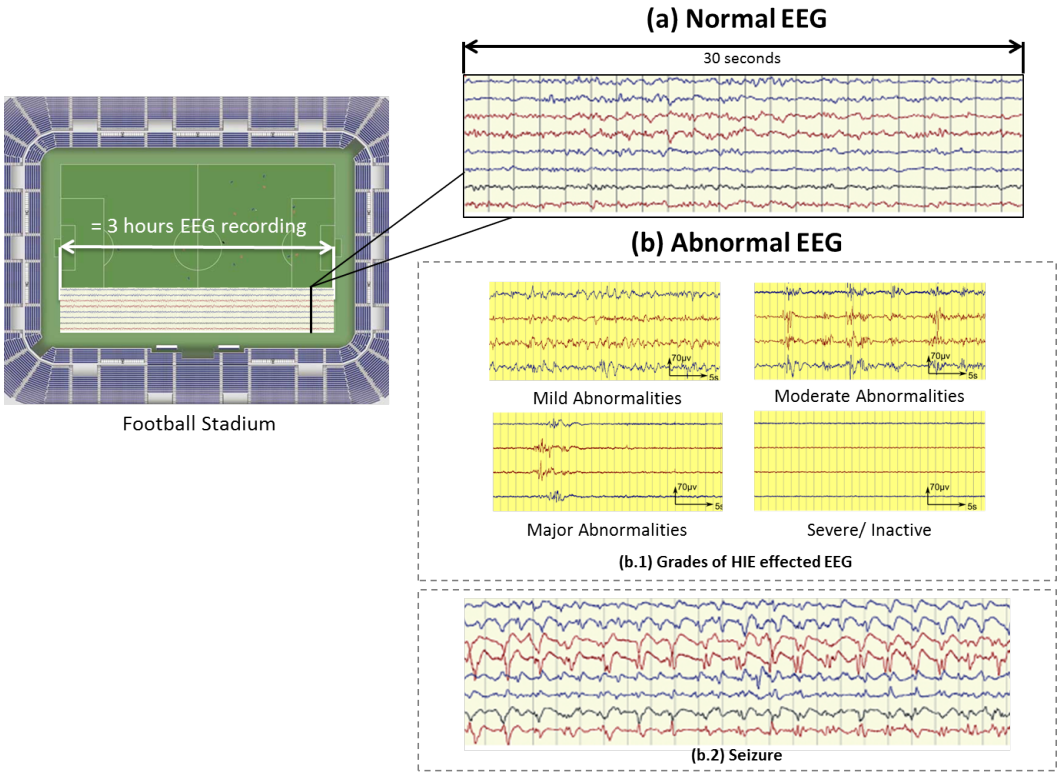


Figure 1.2: A 30 second recording (usual setting for clinicians to look over the EEG recording) of EEG compared to a football stadium. (a) Examples of Normal and (b) Abnormal EEG.

menced within 6 hours of birth. In this narrow window of time the population of neonates who would benefit from treatment must be accurately identified.

Seizures are brief events due to abnormally excessive or synchronous neuronal activity in the brain. On the EEG recording seizures are defined as sudden, repetitive, evolving stereotyped waveforms that have a certain start and ending. Figure 1.2b.2 shows an example of a seizure pattern. Anti-epileptic drugs are used for treatment. If seizures are not detected as early as possible then the resulting lack of treatment could cause severe brain damage or death.

1.2 Automated systems for detecting abnormal brain functions: Literature Review

Automated brain monitoring systems represent a very broad and well researched area. In this section, the past work carried out only in the field of automated systems for neonates is presented. Moreover, it will focus on the systems developed for automated neonatal seizure detection and grading HIE

severity using EEG and ECG signals.

1.2.1 Neonatal seizure detection systems

A number of studies has been done in the past to develop automated systems for the detection/classification of seizures in neonatal EEG. Most of the automated neonatal seizure detection systems can be broadly categorized into rule-based and classifier-based approaches (Thomas et al. 2011). In the rule-based approaches, various features are extracted from the EEG segments and a set of rules and thresholds are used to declare the segment as a seizure or non-seizure. Classifier-based systems on the other hand, employ methods from the machine learning domain that can classify a set of features, extracted from EEG segment, using a data-driven decision rule.

The initial approaches were feature based analysis combined with the heuristic rules. Some of the earliest work on the development of automated systems could be associated with the efforts of Gotman & Gloor (1976) on adult epilepsy detection. This system was used to classify spikes and spike waves (sharp transients with a duration of 10-500ms) by extracting quantitative features from sequences of short segments of EEG. Each feature was then compared to a threshold to make a decision. This research formed the basis for seminal work on an automated neonatal EEG-based seizure detection system by Gotman et al. (1997). This system consisted of three different routines to detect different kinds of seizure activity in the neonatal EEG trace. The first routine, was designed to capture rhythmic discharges using frequency domain features. The second routine detected multiple spikes using a similar method to that proposed in (Gotman & Gloor 1976, Gotman et al. 1979). Lastly, the third routine was used to detect very slow rhythmic discharges by comparing the features of the current epoch to the features of two epochs more than one minute before the current epoch. Another, important early work on neonatal seizure detection was conducted by (Liu et al. 1992) which investigated the time domain method of autocorrelation to capture the seizure activity in the EEG. A method based on analysing the complexity of the background EEG using Singular Spectrum Analysis (SSA) was proposed by (Celka & Colditz 2002). It was argued that SSA can better detect quasi-periodic signals which is the case of seizure activity.

These initial approaches were tested by (Faul et al. 2005) on a dataset of 77 one minute seizure and non-seizure recordings from 13 newborns. The authors showed that the performance of any of these algorithms was not good enough for the use in NICU. Poor performance of these systems was attributed

to artifacts, the high intra patient variability and the resulting large overlap between seizure and non-seizure frequency domain features.

Among other rules-based approaches, Navakatikyan et al. (2006) proposed a system based on wave sequence analysis. The algorithm looked for regularity, amplitudes and shapes of the peaks in a short segment of EEG and used several rules and thresholds to declare a EEG segment as seizure. Deburchgraeve et al. (2008) developed a system with rules that were said to mimic a human observer. The system was divided into two parallel components: One was based on the non-linear energy operator which was used to isolate spike like segments and the other was based on wavelet decomposition, and detected oscillatory type seizures. Correlation based analysis was performed on the output of both components and rules were applied to detect the seizure segments whilst minimizing the false detections. A template matching algorithm based on Dynamic Time Warping (DTW) and spatial dipole clustering was utilized in (Aarabi et al. 2009). This system was designed to capture the temporal and spatial information from EEG recording to detect different abnormal patterns of EEG particularly seizures events. Nagaraj et al. (2014) used Atomic Decomposition (AD) for the seizure detection task. The relative structural complexity of (a measure of the rate of convergence of AD) was obtained for each epoch of EEG and used as the sole feature. A threshold was applied to decide the class of a short EEG segment.

With the availability of more computational power and theoretically strong machine learning methods, the use of classifier based methods has increased recently. A seizure detection system based on Neural Networks (NN) was proposed in (Karayiannis et al. 2006). First, a rule based approach using power spectral features, similar to (Gotman et al. 1997), was used to find a number of short seizure segments from the EEG record. The thresholds were relaxed to detect all possible seizure segments which obviously included many false detections. These segments were then fed to a NN based detector to extract only the most eligible seizure segments. Another NN based system was presented in (Aarabi et al. 2007); in which a feature selection routine based on relevance and redundancy analysis was also employed. A number of rules were set at the decision stage to classify an EEG segment as seizure or non-seizure.

Three classifier models based on linear discriminants, quadratic discriminants and regularized discriminants were tested in a study by Greene et al. (2008). They also examined the early (feature level) and late (decision level) integration of EEG channel information and the effect of the electrode montage

on the seizure detection rate. Thomas et al. (2011) examined Gaussian Mixture Model (GMM) based generative and Support Vector Machines (SVM) based discriminative classification approaches and showed that SVM outperformed the GMM based approach.

Recently, a state of the art patient independent neonatal seizure detection system based on SVM is presented (Temko et al. 2011). Here the EEG was segmented into 8 seconds epochs and 55 features were extracted which were fed to SVM. The output of SVM was converted to probabilities and a moving average filter was applied to smooth the probabilistic output by incorporating the contextual information. The system was validated on the largest dataset of 18 newborns with 267 hours of EEG recordings. Another SVM based patient dependent system was presented recently in (Bogaarts et al. 2014). They used a Kalman filter, to filter both the features and the classifier output. Moreover, a method of baseline feature correction to decrease the inter-patient variability was also introduced. A three minute non-seizure and artifact free EEG segment from the start of each test recording was used to calculate the average non-seizure feature values which was used to estimate an optimal threshold to correct the future EEG feature values.

Finally this section can be concluded with two excerpts that highlights the research of past and present in the area of automated seizure detection. These excerpts are taken from a study in 1979 and more recently from a study in 2015.

"*It is concluded that the final computer displays could only be trusted after visual inspection of the EEG sections provided on paper.*" (Gotman et al. 1979)

"*The seizure detection algorithm (SDA) achieved promising performance and warrants further testing in a live clinical evaluation.*" (Mathieson et al. 2015)

1.2.2 Automated systems for grading HIE using background EEG

Automated classification of background EEG is a relatively new field as compared to seizure detection. The initial approach was made to automate the process of classifying the background EEG of paediatrics (Pasupathy 1994, Si et al. 1998). Features, representing the symmetry, variability and amplitude of the EEG were extracted from a 6 hours recording. A NN based classifier was then used to classify the EEG into 4 main and 7 sub-levels of abnormality (Si

et al. 1998). The system did not show good results and poor accuracy can be attributed to the diversity of the patients in the trial and higher number of EEG grading levels. Moreover, neonatal EEG is significantly different from paediatric EEG. There are patterns representing rapid maturation of brain in early days of life that do not appear at later ages. Therefore such a system developed for paediatrics is not suitable for neonates.

Hathi et al. (2010) proposed a Cerebral Brain Index for neonates (CHI/b). The CHI/b is a score function based on a multi-parameteric algorithm that combines several spectral, temporal and probabilistic features from the EEG to estimate one of three grades of HIE (Hathi et al. 2010).

An algorithm to classify the background EEG based on a Fisher linear discriminant classifier was proposed in Löfhede et al. (2010). The system classified the background EEG into four behavioural ({active, quiet} sleep, {active, quiet} awake) and an abnormal state (burst suppression pattern). However, only quantitative features of inter burst interval were generated as output in the case of burst suppression pattern detection and no score/grade of abnormality was provided. Although automated detection of such patterns is necessary, however this does not translate to a classification of the degree of abnormality of the EEG.

Several quantitative EEG features (e.g. relative delta power, skewness, kurtosis, amplitude, and discontinuity) were tested to grade one-hour long EEG segments in (Korotchikova et al. 2011). Kruskal–Wallis testing with post hoc analysis and multiple linear regression were used for features analysis. It was reported that a linear combination of these features had high correlation with the EEG grade as assigned by a neurophysiologist.

In a study by Stevenson et al. (2013), a non-linear amplitude modulated signal model was assumed to describe a short segment of EEG. Using time-frequency analysis, the EEG signal was decomposed into its amplitude modulated and instantaneous frequency components. Basic statistics (mean, standard deviation, skewness and kurtosis) of these components over a segment of EEG were used as the key features to characterize the EEG. A multi-class linear discriminant classifier was then used to assign one of four HIE grade to an one hour EEG recording. The system showed good accuracy when information about the sleep states was also added.

A Hidden Markov Models (HMM) based EEG diarization approach was proposed to segment and cluster the neonatal EEG into homogeneous states (Temko et al. 2014). Several features were proposed to characterize the re-

sultant state sequence to provide a single measure for a one hour-long EEG recording. These features were aimed at capturing both the statistics and sequentiality of the obtained states. Only statistical analysis to show the capability of the features to discriminate the HIE grades was performed and no accuracy was reported.

In a more recent study, Matic et al. (2014) proposed a tensor-based approach in which continuous EEG was first adaptively segmented and short-term quantized features were extracted. These features were then subsequently used to create a 3D model of a specific grade referred to as the tensor. Features extracted from this model were then fed to a multi-class classifier for classification. Promising results were reported for grading 1 hour EEG files into 3 grades of HIE severity.

1.2.3 Using heart rate variability to classify HIE grade

There has been several attempts on the use of physiological signals other than EEG to classify abnormal brain functions. Some of the basic reasons for using other signals are the difficulty of acquiring the EEG immediately after birth and in cases where EEG is present, other physiological signals may provide extra complementary information regarding brain functions.

The electrocardiogram (ECG) measures the electrical activity generated from the heart. ECG is regularly recorded and is easily accessible to the clinicians after birth. Instantaneous heart rate (HR) is a common biomarker extracted from the ECG signal. Changes in HR during seizures and HIE are reported in a number of recent studies (Goulding et al. 2015, Aliefendioglu et al. 2012, Volpe 2008) and therefore interest has developed to build an automated system that can detect seizures or grade the severity of HIE using the ECG signal.

A system based on HR for detecting seizures was proposed in (Greene et al. 2007b). Several time and frequency domain features were extracted and fed to a linear discriminant classifier. The results of the proposed approach were however, not encouraging. Subsequently, authors proposed a system based on the combination of EEG and ECG signals (Greene et al. 2007a). The integration of information from two signals at features and decision levels was also investigated. Marginal improvements over the only EEG based system were reported by using a late integration method of fusion. A similar combination approach was also tested in (Malarvili & Mesbah 2008). A thorough investigation of HR features for the seizure detection task on a large dataset of 208 hours of ECG

recordings was carried out in (Doyle et al. 2010). The features were classified using SVM. However poor performance was reported due to high variability of feature values among patients and hence the authors advised the use of patient-dependant HR based system.

Temko et al. (2015) presented a study for predicting the neurodevelopmental outcome of neonates with HIE. A large number of HR, EEG and clinical features were tested. SVM was used to classify the features into healthy or abnormal outcome. A comparison of using all features and a subset of features, produced from a feature selection routine, was also performed. The best performance was obtained using 9 EEG, 2 HR and 1 clinical feature. Moreover, encouraging results were reported for predicting the outcome of neonates to normal/abnormal at 2 years of age.

To the best of our knowledge, there is only one previous study on developing a tool for the classification of HIE severity using Hear Rate Variability (HRV) (Matic et al. 2013). They derived 12 features from the HRV signal and then used a simple linear discriminant classifier to classify two hour long ECG recordings into mild and moderate-severe HIE. The dataset included 36 hours of ECG recordings from 18 neonates.

1.3 Aims and scope of the thesis

It is well known that there is a lot of information in the EEG if examined over a long time scale. For example, in the case of seizures, although there is a definite stereotypical pattern of this abnormal brain function; these patterns evolve in frequency and shape through time. An example of such a seizure is shown in Figure 2.9. It can be seen that the frequency and amplitude of this seizure attenuated as it progresses in time. Similarly, for classifying the severity of HIE injury, it is necessary to not only detect the short events like a burst of high frequency activity but also to capture the slow variation of these events in time. Moreover, a clinical expert also analyses the EEG recordings by looking at the bigger picture. S/He goes back and forth through the recording to see what happened before and after a particular event, in order to define its class.

The classifiers used inside most of the above mentioned systems, were only able to classify features extracted from the short independent segments of the EEG that does not capture the temporal context of the EEG signal. Hence, these systems were either not able to capture the slower time varying changes in abnormal brain function or they used methods, before or after the classifier,

that were not sophisticated enough to use this information to its full extent. An example of such a system is neonatal seizure detection system developed by our research group (Temko et al. 2011). The contextual information was explored at the decision level by taking an average of 30 seconds of past and future decisions of the classifier to decide the seizure probability of a given EEG epoch. Nonetheless, this system produced the best performance on a large neonatal dataset for seizure detection but its performance was poor in detecting short seizure events because they were being suppressed by this averaging operation. Similarly, in the case of systems for automated grading of EEG recordings for classifying the HIE (Stevenson et al. 2013), the contextual information was explored by extracting features from a longer EEG epoch of 1 minute.

Classifier level techniques to explore the temporal/contextual information, have shown promising results in other well researched areas of signal processing such as speech recognition (Shimodaira et al. 2002, Smith & Gales 2002), handwriting character recognition (Bahlmann et al. 2002), acoustic events (Temko et al. 2006). However, in the area of detecting/grading neonatal abnormal brain functions, the use of such methods has been relatively scarce.

Thus, the prime objective of this work is to investigate the methods to harness the contextual information inside the classifier (Dynamic/Sequential classifier). The general layout followed for both the automated seizure detection system and the automated HIE grading system is shown in Figure 1.3. Features are extracted from the short segments of EEG. These features are fed into a classifier. Finally the output of the classifier is post-processed to make it meaningful for the clinicians.

Specifically, the DTW based kernel method is used for the seizure detection task. A system based on the fusion of static and sequential classifiers is examined. Moreover, a system for grading the severity of HIE using EEG is investigated that uses a cross disciplinary method of using a supervector kernel inside the SVM. It uses the sequences of short-term features of EEG to classify an hour long recording into one of the four grades of HIE. Furthermore, the usage of HR to classify the grade of HIE is also investigated. A system based on the supervector approach that uses the HR signal to classify HIE into mild and moderate-severe grade is presented.

Additionally, two systems for combining the EEG and ECG signals are proposed for grading the HIE severity. The findings of this study may guide the development of a system that can utilize multiple information sources to get a score of brain health of the neonate.

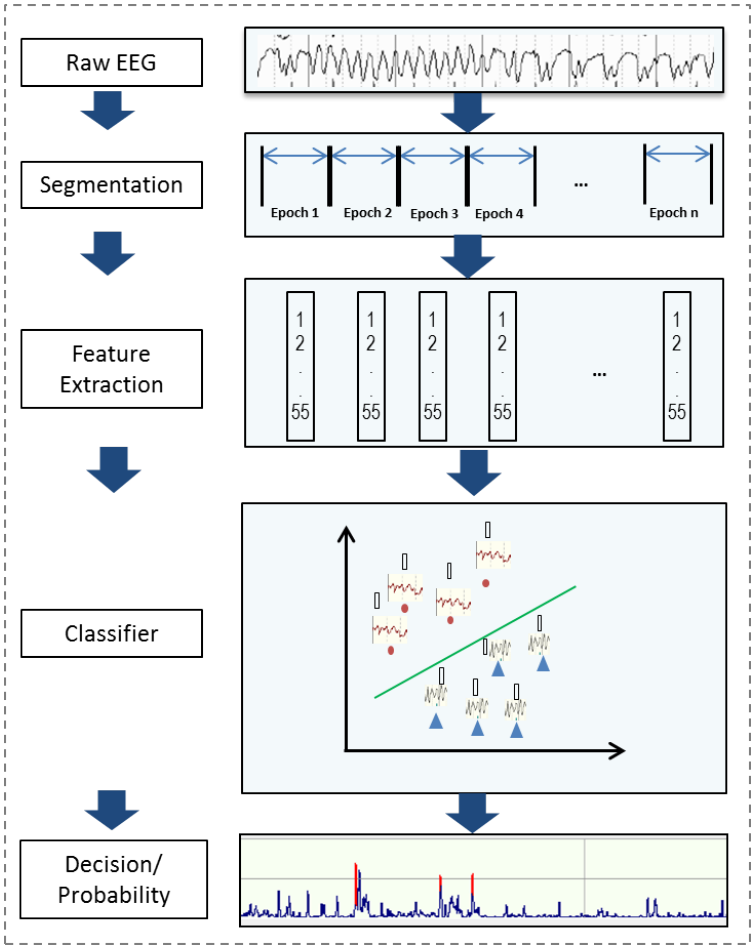


Figure 1.3: Outline of the general automated EEG classification system

1.4 Thesis layout

The layout of rest of thesis is as follows:

Chapter 2 presents the necessary medical background for the work carried out in this thesis. Specifically, it provides a brief introduction to the EEG and its artifacts. It is important to make it clear that the neonatal EEG is different from the adult EEG. Therefore, different patterns that characterize the normal and abnormal neonatal EEG are presented. Moreover, the etiology, diagnosis and treatment methods of HIE and resulting seizures are also discussed. The chapter ends with a brief discussion on heart rate variability and its behaviour during HIE.

Chapter 3 provides a brief introduction to the technical methods used in this work. First, a basic taxonomy of the machine learning classifiers with a special emphasis on static and dynamic classifiers is presented. SVM is used as the main classifier in this work, and hence a theoretical background of the SVM is provided. The SVM is primarily a static classifier and inherently can

not deal with the sequential nature of the data and therefore special kernels are used inside the SVM to classify sequences. This chapter will provide a brief overview of such sequential kernels. Moreover, the supervector kernel used in this work for HIE grading is based on GMM. Therefore a brief description of the GMM approach and techniques to enable them to be used with SVM is also presented.

Chapter 4 presents a system for detecting neonatal seizures. The chapter starts with a literature review of the techniques used in the past to explore the contextual information in neonatal EEG. This is followed by a description of the baseline neonatal seizure detection system which forms the foundation of the current work. The details of DTW technique are presented as a method of comparing sequences. This leads to the explanation of the complete neonatal seizure detection system devised in this study. Lastly, a comparison of the performance achieved by DTW, standard SVM and fusion based seizure detection systems is presented.

Chapter 5 presents the complete description of the developed automated system for grading HIE using the EEG signal. It provides the underlying details of the proposed supervector based approach and motivations for the selection of this technique for the task of HIE grading. Several post processing techniques employed for this task are explained. The chapter closes with a detailed discussion on the obtained results and the analysis of misclassifications.

Chapter 6 presents a novel automated system for the grading of HIE using the HR signal. Details about the HR features and feature extraction process are discussed. A novel system based on the supervector approach is proposed. Results obtained using this approach are compared with other classification techniques. Lastly, a preliminary study on the fusion of EEG and HR based classifiers is presented which outlines the two classification approaches and the initial results obtained by each system.

Chapter 7 concludes the thesis by summarizing the contributions and conclusions drawn from this work. A discussion on the possible future directions is also presented.

1.5 Publications arising from this thesis

Journals:

1. Ahmed, R., Temko, A., Marnane, W., Lighbody, G., Boylan, G., "Grading hypoxic-ischemic encephalopathy severity in neonatal EEG using GMM supervectors and the support vector machine", *Clinical Neurophysiology, Elsevier,* 2015.

2. Ahmed, R., Temko, A., Marnane W., Boylan, G., Lighbody, G., "Exploring temporal information in neonatal seizures using a dynamic time warping based SVM kernel", to be submitted to *Frontiers in Computational Neuroscience*

3. Ahmed, R., Automated system for detection brain injuries, *The Boolean*, 2015

Conferences:

1. Ahmed, R., Temko, A., Marnane, W., Boylan, G., Lighbody, G., "Classification of Hypoxic-Ischemic Encephalopathy Using Long-Term Heart Rate Variability Based Features", *Proceedings of the IEEE Engineering in Medicine and Biology Society Conference, (EMBC)*, 2015

2. Ahmed, R., Temko, A., Marnane, W., Lighbody, G., Boylan, G., Lighbody, G., "Exploring the Temporal Information to Detect the Short Neonatal Seizures", *INFANT research seminar*, 2015

3. Ahmed, R., Multidisciplinary Research in Neonatal Brain Injury, *Proceedings of Baku World Science Forum*, 2014

4. Ahmed, R., Temko, A., Marnane, W., Boylan, G., Lighbody, G., "Grading Brain Injury In Neonatal EEG Using SVM and Supervector Kernel", P*roceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing, (ICASSP)*, 2014, PP: 5894-5898

5. Ahmed, R., Temko, A., Marnane, W., Boylan, G., Lighbody, G., "Dynamic time warping based neonatal seizure detection system," *Proceedings of the IEEE Engineering in Medicine and Biology Society Conference, (EMBC)*, 2012, PP: 4919-4922

References

Aarabi, A., Grebe, R., & Wallois, F. (2007). A multistage knowledge-based system for EEG seizure detection in newborn infants. *Clinical Neurophysiology*, 118(12):2781–2797.

Aarabi, A., Kazemi, K., Grebe, R., Moghaddam, H. A., & Wallois, F. (2009). Detection of EEG transients in neonates and older children using a system based on dynamic time-warping template matching and spatial dipole clustering. *NeuroImage*, 48(1):50–62.

Aliefendioglu, D., Dogru, T., Albayrak, M., Dibekmısırlıoglu, E., & Sanli, C. (2012). Heart rate variability in neonates with hypoxic ischemic encephalopathy. *Indian Journal of Pediatrics*, 79(11):1468–1472.

Bahlmann, C., Haasdonk, B., & Burkhardt, H. (2002). Online handwriting recognition with support vector machines - a kernel approach. *In Proceedings of 8th International Workshop on Frontiers in Handwriting Recognition, 2002.* pp. 49–54.

Bogaarts, J., Gommer, E., Hilkman, D., van Kranen-Mastenbroek, V. H. J. M., & Reulen, J. P. H. (2014). EEG feature pre-processing for neonatal epileptic seizure detection. *Annals of Biomedical Engineering*, 42(11):2360–2368.

Celka, P. & Colditz, P. (2002). A computer-aided detection of EEG seizures in infants: a singular-spectrum approach and performance comparison. *IEEE Transactions on Biomedical Engineering*, 49(5):455–462.

Deburchgraeve, W., Cherian, P., De Vos, M., Swarte, R., Blok, J., Visser, G. H., Govaert, P., & Van Huffel, S. (2008). Automated neonatal seizure detection mimicking a human observer reading EEG. *Clinical Neurophysiology*, 119(11):2447–2454.

Doyle, O., Temko, A., Marnane, W., Lightbody, G., & Boylan, G. (2010). Heart

rate based automatic seizure detection in the newborn. *Medical Engineering & Physics*, 32(8):829–839.

Faul, S., Boylan, G., Connolly, S., Marnane, L., & Lightbody, G. (2005). An evaluation of automated neonatal seizure detection methods. *Clinical Neurophysiology*, 116(7):1533–1541.

Gotman, J., Flanagan, D., Zhang, J., & Rosenblatt, B. (1997). Automatic seizure detection in the newborn: methods and initial evaluation. *Electroencephalography and Clinical Neurophysiology*, 103(3):356–362.

Gotman, J. & Gloor, P. (1976). Automatic recognition and quantification of interictal epileptic activity in the human scalp EEG. *Electroencephalography and clinical neurophysiology*, 41(5):513–529.

Gotman, J., Ives, J., & Gloor, P. (1979). Automatic recognition of inter-ictal epileptic activity in prolonged EEG recordings. *Electroencephalography and clinical neurophysiology*, 46(5):510–520.

Goulding, R. M., Stevenson, N. J., Murray, D. M., Livingstone, V., Filan, P. M., & Boylan, G. B. (2015). Heart rate variability in hypoxic ischaemic encephalopathy: correlation with EEG grade and two-year neurodevelopmental outcome. *Pediatric Research*, 77(5):681–687.

Greene, B. R., Boylan, G. B., Reilly, R. B., de Chazal, P., & Connolly, S. (2007a). Combination of EEG and ECG for improved automatic neonatal seizure detection. *Clinical neurophysiology*, 118(6):1348–1359.

Greene, B. R., De Chazal, P., Boylan, G. B., Connolly, S., & Reilly, R. B. (2007b). Electrocardiogram based neonatal seizure detection. *IEEE Transactions on Biomedical Engineering*, 54(4):673–682.

Greene, B. R., Marnane, W. P., Lightbody, G., Reilly, R. B., & Boylan, G. B. (2008). Classifier models and architectures for EEG-based neonatal seizure detection. *Physiological Measurement*, 29(10):1157–1178.

Harisson, W. & Goodman, D. (2015). Epidemiologic trends in neonatal intensive care, 2007-2012. *JAMA Pediatrics*, 169(9):855–862.

Hathi, M., Sherman, D., Inder, T., Rothman, N., Natarajan, M., Niesen, C., Korst, L., Pantano, T. et al. (2010). Quantitative EEG in babies at risk for hypoxic ischemic encephalopathy after perinatal asphyxia. *Journal of Perinatology*, 30(2):122–126.

Karayiannis, N. B., Mukherjee, A., Glover, J. R., Ktonas, P. Y., Frost Jr, J. D., Hrachovy, R., & Mizrahi, E. M. (2006). Detection of pseudosinusoidal epileptic seizure segments in the neonatal EEG by cascading a rule-based algorithm with a neural network. *IEEE Transactions on Biomedical Engineering*, 53(4):633–641.

Korotchikova, I., Stevenson, N. J., Walsh, B. H., Murray, D. M., & Boylan, G. B. (2011). Quantitative EEG analysis in neonatal hypoxic ischaemic encephalopathy. *Clinical Neurophysiology*, 122(8):1671–1678.

Löfhede, J., Thordstein, M., Löfgren, N., Flisberg, A., Rosa-Zurera, M., Kjellmer, I., & Lindecrantz, K. (2010). Automatic classification of background EEG activity in healthy and sick neonates. *Journal of neural engineering*, 7(1):016007.

Liu, A., Hahn, J., Heldt, G., & Coen, R. (1992). Detection of neonatal seizures through computerized EEG analysis. *Electroencephalography and Clinical Neurophysiology*, 82(1):30–37.

Liu, L., Oza, S., Hogan, D., Perin, J., Rudan, I., Lawn, J. E., Cousens, S., Mathers, C. et al. (2015). Global, regional, and national causes of child mortality in 2000–13, with projections to inform post-2015 priorities: an updated systematic analysis. *The Lancet*, 385(9966):430 – 440.

Malarvili, M. & Mesbah, M. (2008). Combining newborn EEG and HRV information for automatic seizure detection. *In Proceedings of IEEE Annual International Conference of Engineering in Medicine and Biology Society, (EMBC) 2008*. IEEE, pp. 4756–4759.

Mathieson, S. R., Stevenson, N. J., Low, E., Marnane, W. P., Rennie, J. M., Temko, A., Lightbody, G., & Boylan, G. B. (2015). Validation of an automated seizure detection algorithm for term neonates. *Clinical Neurophysiology*, in press.

Matic, V., Cherian, P. J., Koolen, N., Naulaers, G., Swarte, R. M., Govaert, P., Huffel, S. V., & Vos, M. D. (2014). Holistic approach for automated background EEG assessment in asphyxiated full-term infants. *Journal of Neural Engineering*, 11(6):066007.

Matic, V., Cherian, P. J., Widjaja, D., Jansen, K., Naulaers, G., Van Huffel, S., & De Vos, M. (2013). Heart rate variability in newborns with hypoxic brain injury. *In Oxygen Transport to Tissue XXXV*. Springer, pp. 43–48.

Nagaraj, S. B., Stevenson, N. J., Marnane, W. P., Boylan, G. B., & Lightbody, G. (2014). Neonatal seizure detection using atomic decomposition with a novel dictionary. *IEEE Transactions on Biomedical Engineering*, 61(11):2724–2732.

Navakatikyan, M. A., Colditz, P. B., Burke, C. J., Inder, T. E., Richmond, J., & Williams, C. E. (2006). Seizure detection algorithm for neonates based on wave-sequence analysis. *Clinical Neurophysiology*, 117(6):1190–1203.

Pasupathy, A. K. (1994). *An Expert System for EEG Monitoring in the Pediatric ICU*. Ph.D. thesis, McGill University, Montreal.

Rennie, J. M., Chorley, G., Boylan, G. B., Pressler, R., Nguyen, Y., & Hooper, R. (2004). Non-expert use of the cerebral function monitor for neonatal seizure detection. *Archives of Disease in Childhood - Fetal and Neonatal Edition*, 89(1):F37–F40.
URL: *http://fn.bmj.com/content/89/1/F37.abstract*

Shimodaira, H., Noma, K.-i., Nakai, M., & Sagayama, S. (2002). Dynamic time-alignment kernel in support vector machine. *In Advances in Neural Information Processing Systems 14*. MIT Press, pp. 921–928.

Si, Y., Gotman, J., Pasupathy, A., Flanagan, D., Rosenblatt, B., & Gottesman, R. (1998). An expert system for EEG monitoring in the pediatric intensive care unit. *Electroencephalography and clinical Neurophysiology*, 106(6):488–500.

Smith, N. & Gales, M. J. (2002). Using SVMs and discriminative models for speech recognition. *In Proceedinds of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2002*, vol. 1. IEEE, pp. 77–80.

Stevenson, N. J., Korotchikova, I., Temko, A., Lightbody, G., Marnane, W. P., & Boylan, G. B. (2013). An automated system for grading EEG abnormality in term neonates with hypoxic-ischaemic encephalopathy. *Annals of Biomedical Engineering*, 41(4):775–785.

Temko, A., Doyle, O., Murray, D., Lightbody, G., Boylan, G., & Marnane, W. (2015). Multimodal predictor of neurodevelopmental outcome in newborns with hypoxic-ischaemic encephalopathy. *Computers in Biology and Medicine*, 63:169–177.

Temko, A., Marnane, W., Boylan, G., & Lightbody, G. (2014). EEG 'diarization'for the description of neonatal brain injuries. *In Proceedinds of IEEE In-*

ternational Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2014. IEEE, pp. 5844–5848.

Temko, A., Monte, E., & Nadeu, C. (2006). Comparison of sequence discriminant support vector machines for acoustic event classification. *In Proceedinds of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2006*, vol. 5. IEEE, pp. 721–724.

Temko, A., Thomas, E., Marnane, W., Lightbody, G., & Boylan, G. (2011). EEG-based neonatal seizure detection with support vector machines. *Clinical Neurophysiology*, 122(3):464–473.

Thomas, E. M., Temko, A., Lightbody, G., Marnane, W. P., & Boylan, G. B. (2011). Advances in automated neonatal seizure detection. *In New Advances in Intelligent Signal Processing*, vol. 372 of *Studies in Computational Intelligence*. Springer Berlin Heidelberg, pp. 93–113.

UNICEF (2013). The UN Inter-Agency Group for Child Mortality Estimation: Levels and trends in child mortality. Tech. rep., UNICEF, New York.

Volpe, J. J. (2008). *Neurology of the Newborn*. Elsevier Health Sciences.

Chapter 2

Medical Background

The presented thesis work is related to the two most common neurological problems in full term neonates i.e. hypoxic-ishcemic encephalopathy and neonatal seizures. This chapter provides a brief background knowledge on the etiology, diagnosis, and the current clinical treatments for these brain injuries. Hence, it will equip the reader with the essential medical information of the problem area dealt within this thesis. Lastly, a brief overview on the relationship between heart rate and brain injury is also presented in the last section.

2.1 Brain and Electroencephalography

The brain is the central command system of the human body which instructs other organs according to the input it gets from the five sensory systems. It is composed of two types of cells called the glial cells and the neurons. Glial cells provide the support to the neurons in different forms. Neurons are the most important cells of the brain and the body because of their unique ability to generate signals that can travel to other neurons and also to the different parts of the body. The human brain consists of billions of such neurons and every neuron is interconnected with billions of other neurons through axons and dendrites. It is estimated that an adult brain has more than 60 trillion neuronal connections (Stiles & Jernigan 2010). The transmission of the signals occurs at the end of an axon on a very complex chemical junction called the synapse.

The electrochemical process used by the neurons to communicate with each other give rise to an electric field in the brain tissue. This electric field is very small to be detected on the scalp. However, synchronous generation of such electric potentials from many neurons can be captured from the scalp by a

process known as electroencephalography (EEG). In this procedure, electrodes (usually disposable flat surface electrodes) are attached to the scalp of the subject at specific places. The scalp is first prepared using a conductive gel to reduce the impedance. The voltage detected by the EEG electrodes at the scalp is of the order of a few micro volts. Therefore, the signal from these electrodes is first amplified using a differential amplifier to make it useable for viusualization and signal processing purposes.

Different parts of the brain generate different types of activity so the electrodes need to be applied in a specific manner. The electrodes are placed usually over the central region of the brain and over the frontal, temporal, parietal and occipital lobes, using the standardized system of 10-20 electrode placement as shown in Figure 2.1. The name of each channel starts with the part of the brain where it is attached; e.g. F7 refers to frontal part of the brain. The number identifies the hemisphere of the brain (even number for right hemisphere and odd for left hemisphere). Less electrodes are used for neonatal EEG because of the smaller head size of infants. These electrodes are (F4, F3, C4, C3, Cz, T4, T3, P4, P3, O1, O2).

The voltage represented by the EEG trace (channel) is actually the potential difference between two electrodes, in which one electrode is used as reference. There are different systems, called the montages, for selecting a pair of electrodes to create an EEG channel; the most common ones are referential and the bipolar montage.

In the referential montage, one common reference electrode is fixed. A channel is then derived by subtracting one electrode's potential from this common reference electrode. Other channels from all the other electrodes are derived in a similar way. One of the problems with the referential montage is finding the quiet reference electrode. If a reference electrode is contaminated with the artifacts then all the resulting channels will take on the associated noise.

In the bipolar montage the channels are created in the form of a chain. An electrode that is used as input for one channel, serves as the reference for the next. The advantage of this technique is that the activity of the neighboring electrodes can be distinguished.

The bipolar montage is most commonly used in neonatal EEG monitoring. Therefore, in this study the bipolar montage with the following 8 EEG bipolar pairs was used: F4-C4, C4-O2, F3-C3, C3-O1, T4-C4, C4-Cz, Cz-C3 and C3-T3. Figure 2.1 shows the placement of the electrodes on scalp. The red arrows

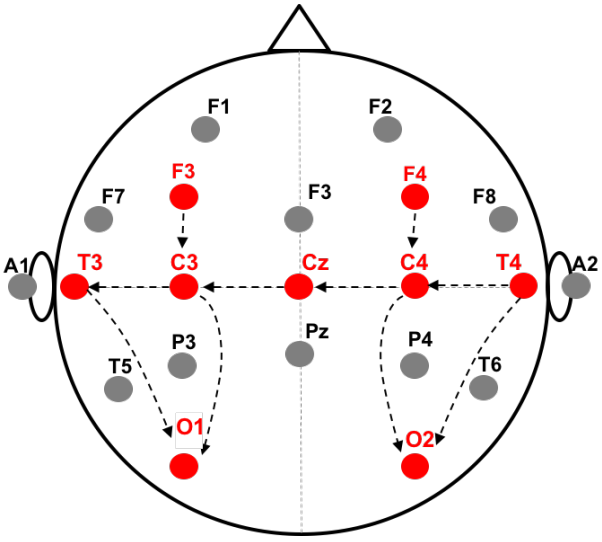


Figure 2.1: The 10-20 system of EEG electrodes placement. The electrodes in red color show the modified 10-20 system for neonates.

indicate the links of electrodes that form a chain on the scalp according to the bipolar montage.

2.1.1 Neonatal EEG

The EEG of neonates is unique and different from that of adults. The inter-neuronal connections are constantly changing to adapt to the newly changed surrounding environment of the body (Stiles & Jernigan 2010). The neonatal EEG show predictable patterns of this rapid maturation of the brain in early days of life. Some of these patterns do not appear at later ages. The electrical activity in EEG is usually divided into four frequency bands: delta (0-3.5Hz), theta (4-7.5Hz), alpha (8-13Hz) and beta (13-30Hz) (Niedermeyer & Silva 2005). Figure 2.2 shows examples of such EEG waveforms. An example of background neonatal EEG is shown in Figure 2.3. It can be seen that normal EEG consists of random patterns which do not correlate with each other.

2.1.2 Abnormal neonatal EEG

The neonatal EEG is very different to that of older children and adults. Neonatal EEG could include a range of frequencies, amplitudes and features such as focal attenuation, focal slowing, burst suppression (Walsh et al. 2011). Persistent and excessive presences of such EEG patterns could have the grim prognosis (Patrizi et al. 2003). A brief description of some patterns is given below.

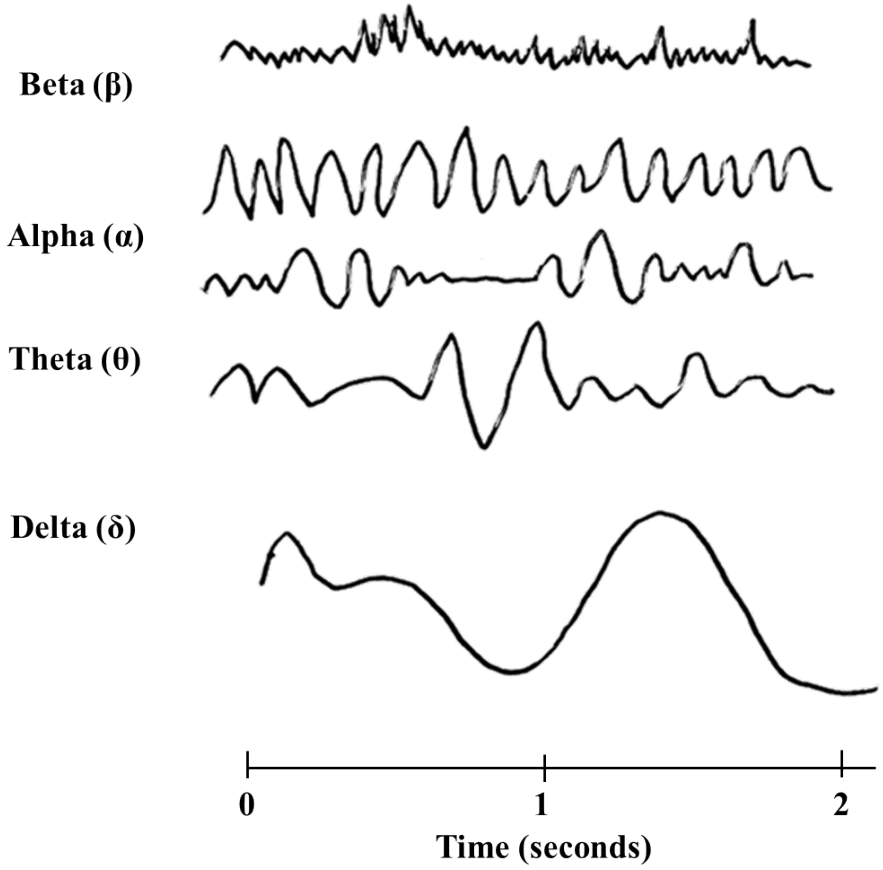


Figure 2.2: Examples of different types of EEG waveforms according to their frequency bands.

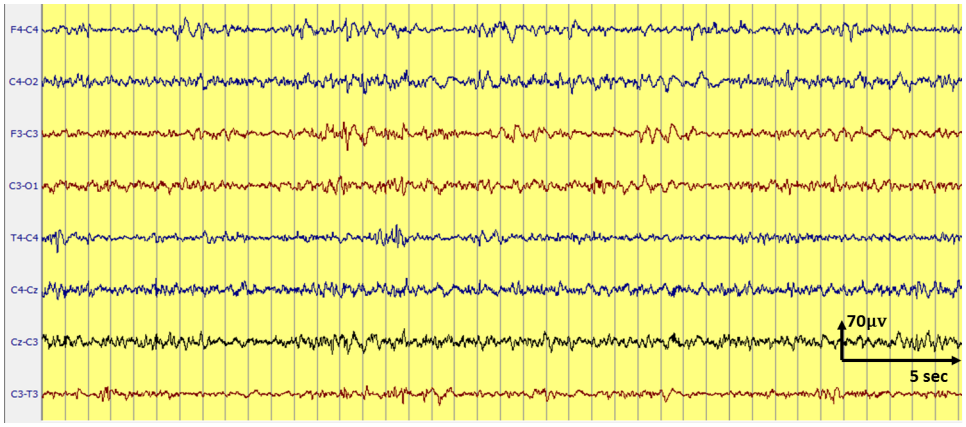


Figure 2.3: Background EEG of a healthy neonate.

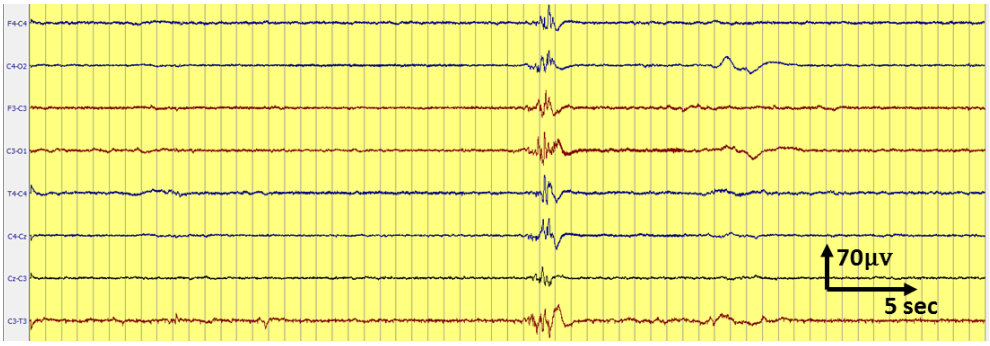
- **Burst suppression**: is characterized by a sudden burst of high voltage and frequency of electrical activity followed by an interval of low voltage ($<$10-15 μ V) called the Inter Burst Interval (IBI). Figure 2.4a shows an example of a burst suppression pattern. The duration of the IBI is broadly accepted as a function of age in preterm neonates. The normal IBI is long for very preterm infants and decreases with the age. In full term neonates, burst supression is always abnormal. The presence of long periods of discontinuty ($>$10s) in full term neonates is considered abnormal and is usually recognized as an effect of a brain injury (Walsh et al. 2011). Furthermore, excessive and persistent presence of burst suppression patterns is also considered abnormal regardless of the age.

- **Asymmetry**: The consitituent elements of normal neonatal EEG e.g. voltage, frequencies and the resulting patterns are usually the same for left and right hemispheres of the brain. The normal EEG will show mirror electrographic images of both hemispheres and patterns will be symmetric. Although, occasional asymmetry is allowed in the neonatal EEG, their presence with a 2:1 difference in electrographic elements (amplitude, frequency etc.) is considered abnormal (Tsuchida et al. 2013). Figure 2.4b shows an example of asymmetrical EEG pattern. It can be seen in the highlighted box that the EEG patterns in the channels from left hemisphere does not appear in the channels obtained from right hemisphere.

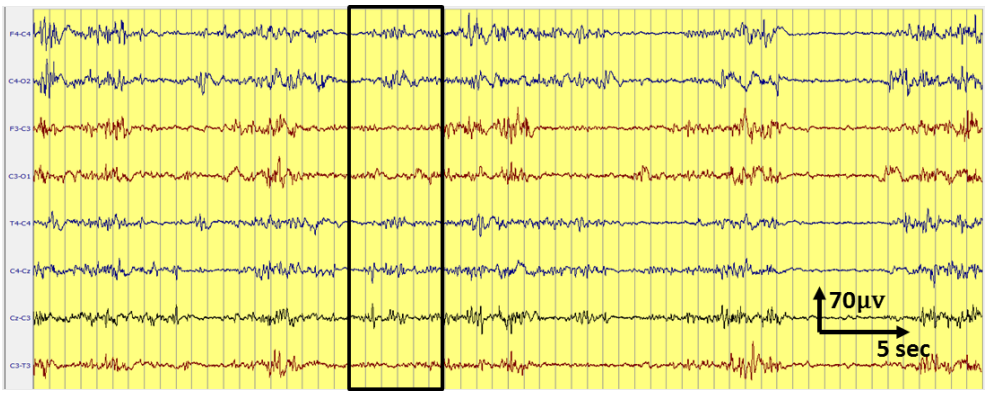
- **Asynchrony**: Similar to the normal neonatal EEG being symmetrical, it should also be synchronous. If a certain pattern appears in two EEG channels that represent two different brain hemispheres then they would be considered synchronous if the time between their onsets is less than 1.5 seconds. However, some degree of asynchrony is allowed for preterm babies. Figure 2.4c presents an example of an asynchronous EEG pattern where the onset of a EEG pattern in channel F4-C4 does not match the pattern in channel C4-O2. The EEG is considered abnormal if there is excessive amount of asynchrony present between EEG patterns (Tsuchida et al. 2013).

- **Sharp waves and spikes**: A sharp wave is defined as a sudden predominant transient deflection whose polarity could either be negative (negative sharp wave) or positive (positive sharp wave). They should be distinctly different from the background EEG. Sharp waves lasting $<$100ms are called spikes whereas if the duration is greater than 100ms then they

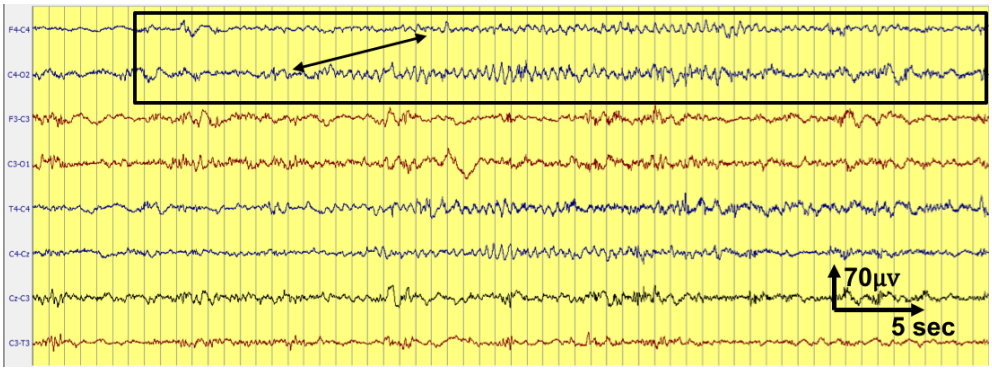
are simply called sharp waves. An example of sharp waves is presented in Figure 2.4d. The sharp waves are indicated with arrows. Persistent excessive presence of sharp waves in neonatal EEG could indicate localized hemorrhage or other brain injury and thus are considerd as abnormal (Tsuchida et al. 2013).



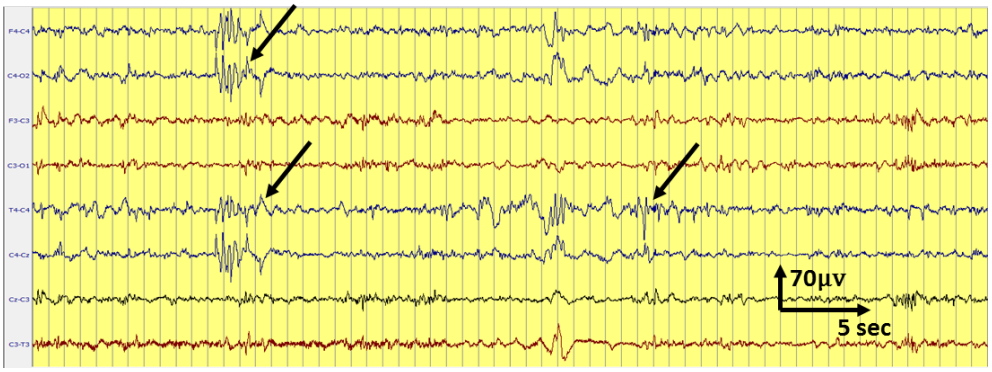
(a) Burst Suppression



(b) Asymmetry



(c) Asynchrony



(d) Sharp Waves

Figure 2.4: Some examples of abnormal EEG patterns.

2.1.3 EEG artifacts

Electrical activities represented in an EEG trace whose source is not brain are considered artifacts. Broadly, these non-cerebral signals can be divided into two categories; I- physiological artifacts which are the signals generated by the human body. II- extra-physiological artifacts which are generated by the nearby instruments that effects the electric field around the EEG monitoring units. Most of the artifacts specially physiological artifacts mimic brain activities and thus present the most formidable challenge in developing classification systems for detecting the brain injuries.

A brief description of some artifacts that are related to this work along with figures is given below.

I- Physiological artifacts

- **Muscle artifacts:** are caused by the electrical activity generated by the muscles. Although muscle artifacts are the most common EEG artifacts, they are the easiest to identify. Generally these artifacts can be characterised by the very short duration recurring potentials. Figure 2.5a shows an example of muscle artifact appear in multiple EEG channels. The EEG corrupted by artifacts shows high frequency components that does not match the background EEG.

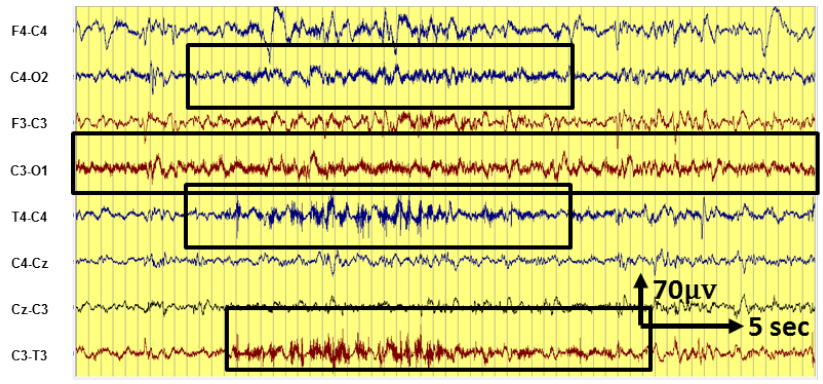
- **Respiration artifacts:** are caused by the human respiratory system. The respiration artifacts are categorized in two kinds. 1- caused by the body movement while breathing which appears as a slow rythmic wave on the EEG trace. 2- appears as a sharp or slow wave that is synchronous to the exhalation and inhalataion process. Figure 2.5b shows an exmple of the later type of respiration artifact. The EEG corrupted by artifacts is highlighted with a box. It can be clearly seen that EEG in channel F3-C3 and Cz-C3 correlates with the respiration trace pattern at the bottom. Respiration artifacts could mimic a seizure pattern.

- **ECG and pulse artifacts:** are caused by the electric potentials generated by the heart. These artifacts occur more frequently in neonates that have short or wide necks. The EEG corrupted by the ECG can be recognised by the rythmic activity that correlates with the ECG trace. The pulse artifact is similar to ECG artifact and is caused by placing an electrode on a pulsating vessel. Figure 2.5c shows an example of pulse artifact where a loose electrode on channel T4-C4, resulted in high frequency noise and

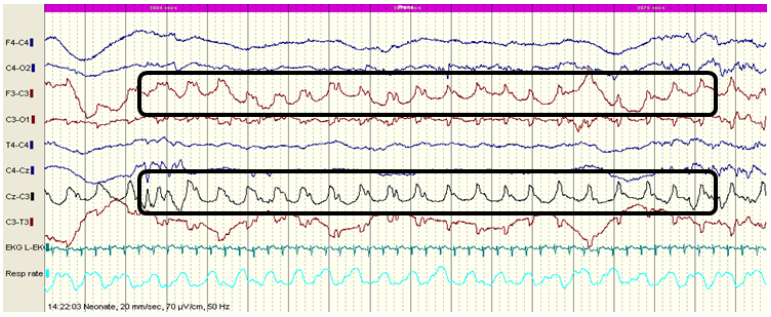
pulse artefact. The ECG trace at the bottom in red shows the QRS complex preceding the ECG by 200-300 ms.

II- Extra-Physiological artifacts

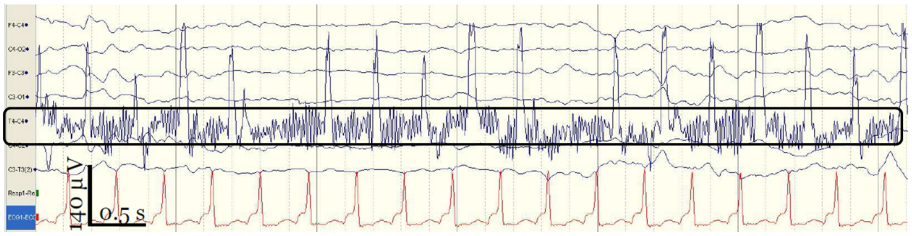
These are the artifacts that appear on the EEG trace due to near-by changes in the enviornment. The most common of these are electrode-pop/detachment and 50 Hz AC contamination of the EEG signal. The electrode detachement, for example occured due to respiration or head movement, produces a sudden change in the impedance which causes a slow wave to appear on the EEG trace without changing the background EEG (Figure 2.6a). AC artifact (Figure 2.6b) are easy to identify and can be removed easily with a notch filter. Other artifacts that can corrupt the EEG signals includes, rocking cot, near-by ventilator movements etc.



(a) Muscle artifact indicated with boxes



(b) Respiration artifact on channel Cz-C3 and F3-C3

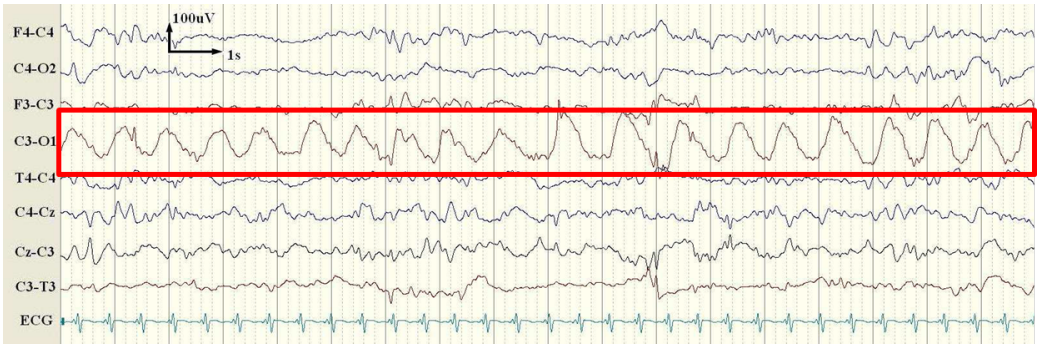


(c) Pulse artifact visible on channel T4-C4

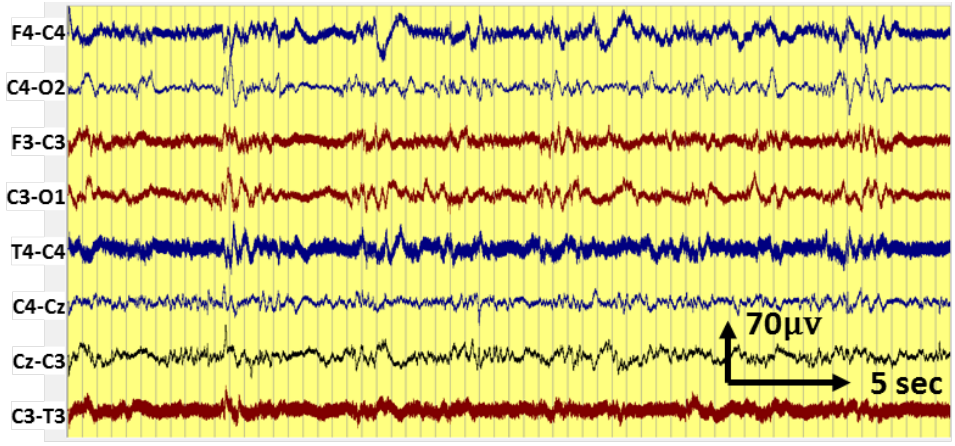
Figure 2.5: Examples of physiological artifacts

2.2 Hypoxic-Ischemic encephalopathy (HIE)

Perinatal Hypoxia ischemia (HI) is a condition caused by the lack of oxygen in the blood or simply the lack of blood flow in the brain prior, during or after the birth. This situation can arise due to a number of reasons with labour or delivery complications being the most common (O'Brien et al. 1966). It is reported that HI is responsible for a third of all neonatal deaths globally and among survivors can lead to progressive encephalopathy (brain damage) called Hypoxic-Ischemic Encephalopathy (HIE) (Volpe 2008). The long-term outcome of HIE depends on the severity of the initial HI insult. Mild encephalopathy may have a normal outcome, 20-40% of neonates with moderate encephalopathy can have an abnormal outcome and severe encephalopathy generally leads to



(a) Electrode detachment visible on C3-O1



(b) 50 Hz AC noise clearly visible on the F4-C4, F3-C3, C3-O1, T4-C4, and C3-T3

Figure 2.6: Examples of Extra-physiological artifacts

neurological disability or death (Gray et al. 1993) in the majority of neonates.

This condition occurs because the amount of oxygen and glucose needed for neurons to function properly is decreased. However the actual damage occurs after the normal blood flow and oxygen is restored in the brain, and is due to toxins released from the damaged cells. This phenomenon is called excitotoxicity (Fatemi et al. 2009, Johnston et al. 2001, Choi & Rothman 1990).

2.2.1 Diagnosis of HIE

The physical symptoms shown by the neonate suffering from HIE depends on the severity of injury (Wu 2015). Current neonatal practice relies on the initial assessment of the infant's clinical state and other clinical markers to grade the severity of encephalopathy following delivery. It was previously shown that these markers are often not helpful in differentiating between grades of encephalopathy, and do not vary between mild, moderate or severe grades (Murray et al. 2006). In addition sedative drugs can confound the clinical assess-

Table 2.1: HIE Grades

Grade	Description
0	**Normal**: Continuous background pattern with normal physiologic features e.g. anterior slow waves
1	**Mild abnormalities**: Continuous background pattern with mild asymmetric patterns, mild voltage depression or poorly defined sleep wave cycles
2	**Moderate abnormalities**: Discontinuous activity with IBI $\leq$ 10s, no clear sleep wake cycling, clear asymmetry or asynchrony
3	**Major abnormalities**: Discontinuous activity with IBI 10-60s, severe fading background patterns no sleep wake cycling
4	**Inactive**: Background activity of $\leq 10\mu V$ or severe discontinuity of IBI $\geq$ 60s

ment. Brain imaging (MRI) can be used to identify the severity of brain injury but it is very difficult to obtain in the first 24-48 hours after birth. EEG is considered to be one of the best methods for HIE diagnosis as it allows continous monitoring of the brain. Moreover, it can be used immediately after birth, can objectively grade the severity of the HIE injury and most importantly enables the monitoring of the evolution of the encephalopathy.

2.2.2 HIE effected EEG (HIE-EEG)

HIE-EEG is classified into four main grades (Murray et al. 2009). Figure 2.7 shows examples of ideal EEG epochs of the 4 grades of HIE. Some of the main features that differentiate the grade of HIE-EEG are the inter-burst-interval (IBI), amplitude and the discontinuity of the background EEG (Boylan et al. 2008). Moreover, absence or poorly defined sleep-wake cycles can also be seen in HIE. Table 5.1 outlines the classification criteria of the HIE EEG introduced by Murray et al. (2009). This classification scheme can be used to predict the outcome of HIE at 2 years (Murray et al. 2009).

2.2.3 Treatment for HIE

The results of several international trials has shown that early induced Therapeutic Hypothermia (TH) is beneficial in HIE, improving the survival rate and reducing the neurological disability (Azzopardi et al. 2009). The treatment involves cooling the infant to a body temperature of between 33-34 °C for 72 hours without interruption. Cooling slows down the metabolic rate in the brain

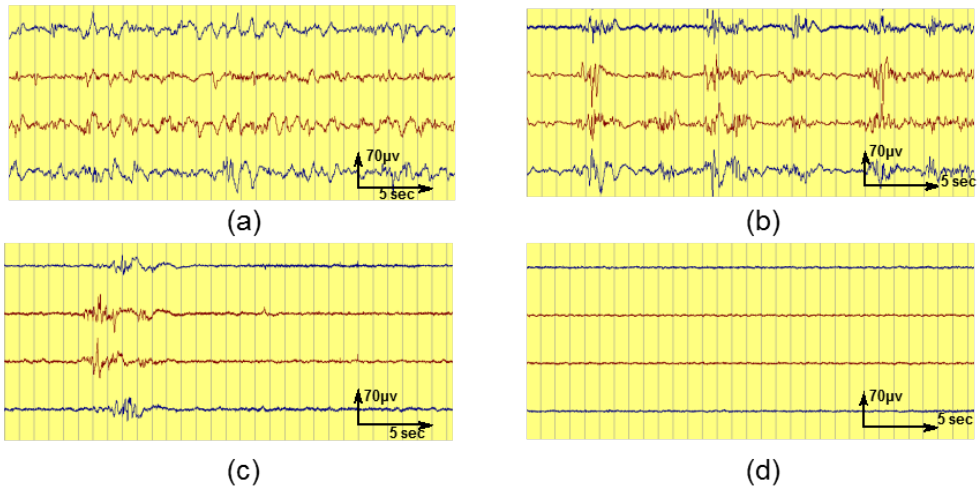


Figure 2.7: Ideal examples of EEG waveforms in the 4 grades of HIE. (a) mild (b) moderate abnormalities, (c) major abnormalities, (d) severe HIE /inactive.

and prevents the secondary injury (Erecinska et al. 2003). TH has now become a standard of care for moderate and severe HIE. However to be effective, it must be commenced within 6 hours of delivery. In this narrow window of time the population of neonates who would benefit from treatment (those with moderate or severe encephalopathy) must be accurately identified.

2.3 Neonatal seizures

A seizure is excessive synchronous electrical discharges of neurons in the brain and these discharges can be measured using the EEG. The neonatal period (up until the 28^{th} day from birth) is considered the most vulnerable period of life to have seizures (Volpe 2008). Moreover seizures occur more frequently during the first 72 hours after birth (Lynch et al. 2012, AL-Naddawi et al. 2011). It is reported that 3 out of 1000 full term neonates experience seizures (Lanska et al. 1995) whereas this figure grows to 58 to 132 per thousand in infants with immature birth weight (Kohelet et al. 2004, Watkins et al. 1988). Failure to detect seizures and the resulting lack of treatment may result in brain damage and in severe cases, death.

Clinically observable neonatal seizures are generally classified into the following types, according to the scheme introduced by Volpe (2008).

1. Subtle seizures
2. Tonic seizures

3. Clonic seizures

4. Myoclonic seizures

Subtle seizures are the most frequent occurring seizures in neonates and include physiological symptoms such as horizontal deviation of eyes with or without jerking, eye lid blinking and some lower body rhythmic movements. They are overlooked the most because of the behavioral similarities with the normal neonate. The effects of tonic seizures could be seen as the contraction of limb, facial, axial and other muscles. Clonic seizures are revealed as the rhythmic shakes in some muscles of the face, limbs, axial and diaphragm or could be multifocal with jerks appearing in more than one muscle group. Myoclonic seizures are manifested by the rapid, single or repetitive jerks. They could affect one part of the body or the whole body.

2.3.1 Etiology of neonatal seizures

Neonatal seizures are caused by many different problems such as hypoxia ischemia, brain hemorrhage, stroke, infections during pregnancy or after birth, poor health of the mother during pregnancy and metabolic disturbances. However, the most common cause of neonatal seizures is HI during or after the birth. HI is estimated to be the root cause of 80-85% of neonatal seizures (Volpe 2008, Sabzehei et al. 2014).

2.3.2 Diagnosis of neonatal seizures

The diagnosis is clinically made by observation of the neonate for the abnormal, repetitive and stereotypical behaviors. Brain imaging using tools like Magnetic Resonance Imaging (MRI) or Computed Tomography (CT) scan could be employed to detect brain hemorrhage and other structural abnormalities.

The possibility of continous and non-invasive monitoring of the neonate's brain makes the EEG the tool-of-choice. EEG provides the best understanding of the nature of paroxysmal movements and also in differentiating epileptic seizures from non-epileptic seizures (Clancy 1996). It is reported that only one third of all seizures are clinically visible (Murray et al. 2008) and the rest need to be detected using EEG. Moreover, the initiation of anti-epileptic drugs (AED) could make it difficult to observe a seizure clinically. However, AEDs do not suppress the EEG ictal discharges (Weiner et al. 1991, Rennie & Boylan

2007) which makes EEG-monitoring essential after their introduction. Therefore, clinical diagnosis of seizures, without EEG confirmation, could have severe limitations (Murray et al. 2008).

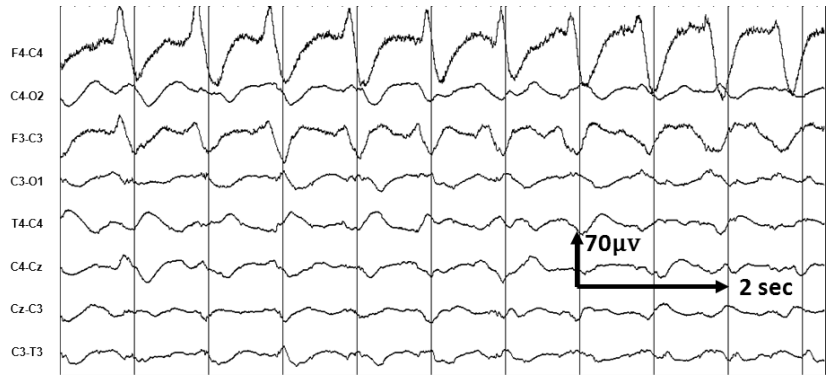
The mean duration of neonatal seizures as reported by a number of studies, is between 1 minute to 2.5 minutes (Clancy & Legido 1987, Scher et al. 1993, Low et al. 2012, Murray et al. 2008, Boylan et al. 2013). The seizure could be repetitive in-between. The minimum length of a valid neonatal seizure is 10 seconds (Clancy & Legido 1987).

2.3.3 Neonatal seizures and EEG

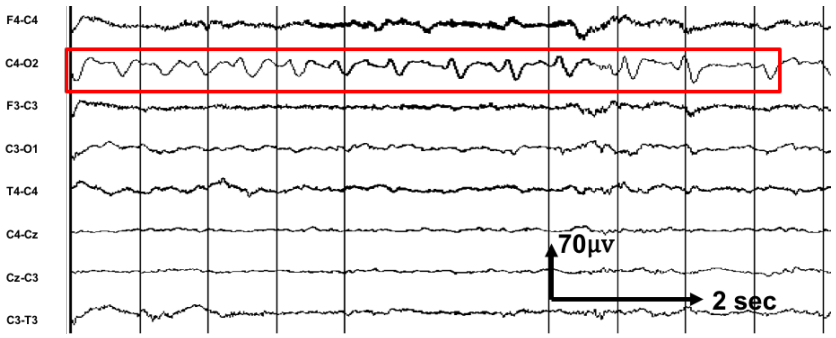
The seizures that are measured using EEG are called electrographic seizures. These seizures may or may not be accompanied with clinical activity. The EEG of a neonatal seizure contains repetitive rhythmic waveforms of alpha, beta, theta and delta frequencies, sharp and slow waves and spikes wave discharge (Patrizi et al. 2003, Lombroso 1996). These waveforms may accelerate or decelerate in speed. The seizure patterns vary from one neonate to another and could be different in the EEG recording of a single neonate at different time points.

Neonatal seizures can be focal, multifocal or generalized. This means that a seizure can occur only in one part of brain or many parts of brain or may travel from one part to another. This leads to the seizure appearing in one or many channels of the EEG. Figure 2.8 shows examples of such global and focal seizures. It can be seen in Figure 2.8a that, the seizure waveform is present in many channels, whereas the seizure in Figure 2.8b is localized only in channel C4-O2. Figure 2.8c shows an example of a seizure pattern that moves from the left to right hemisphere. This change in location could be abrupt during the seizure progression. Consistent focal seizures are usually associated with focal brain damage. The EEG after the seizure can return to the pre-ictal state immediately or with a slow transient.

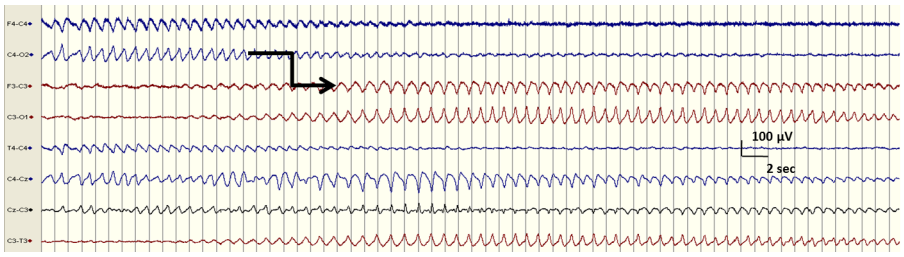
A single neonatal seizure may change in frequency, morphology and propagation (Patrizi et al. 2003, Abend & Wusthoff 2012). It was shown in (Patrizi et al. 2003) that changes in morphology and frequency were present in more than a half of all neonatal ictal discharges. It can be seen in the example shown in Figure 2.9, where a seizure event starts with high amplitude spikes and ends up with very low amplitude spikes. The lack of strong inhibitory factors in the immature brain likely contributes to the propensity for the spread of the discharges. The changing morphology of the discharges may be the result of



(a) Global seizure



(b) Local seizure



(c) Seizure changing location from left hemisphere to right hemisphere

Figure 2.8: Examples of seizure patterns

a slow recruitment of additional neuronal networks during ictus (Patrizi et al. 2003).

2.4 Heart rate and brain damage

An injury to the brain could result in malfunctioning of other organs of the body. It has been shown that changes in respiratory control, blood pressure and heart rate variability (HRV) can occur after brain injury (Volpe 2008). Heart rate is a common parameter used to assess the health of neonates around the world. In healthy neonates, the variability in heart rate is high in the early days

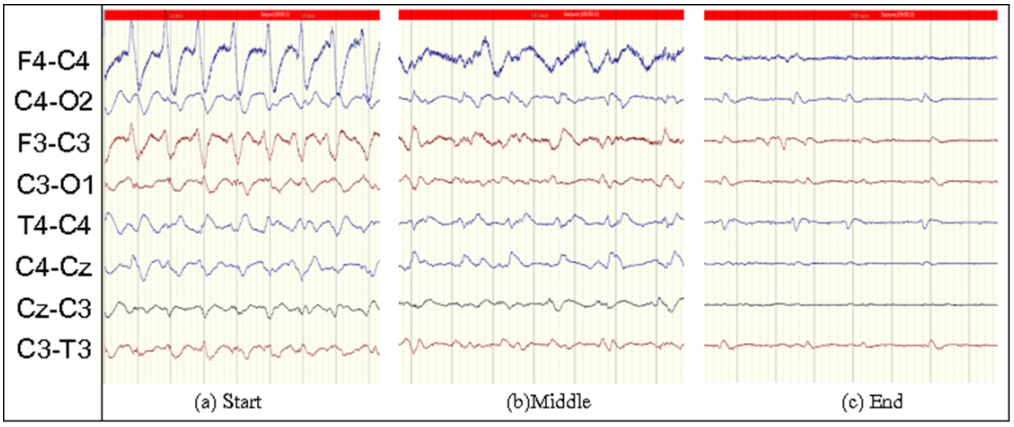


Figure 2.9: Evolution of a single seizure event. (a) Slow wave activity at start with sharp/spike components involved high in amplitude, phase reversal. (b) As the seizure progresses the EEG becomes lower in amplitude. (c) only very low amplitude discharges are seen with a flattening of the background.

of life because the body is trying to adapt to the new enviornment (Selig et al. 2011). Neonates who have suffered a brain injury or were preterm have less variability in heart rate during the neonatal period (Selig et al. 2011, Goulding et al. 2015).

The neonatal heart rate is clinically monitored using Electrocardiography (ECG) which is a noninvasive method of measuring the electrical activity of heart. In a typical clinical setting for adults, 12 electrodes are placed on the body to measure the tiny voltage generated from the heart. However for neonates, only 2 electrodes, one on each shoulder, are used. A typical waveform of ECG is shown in Figure 2.10. The ECG waveform is commonly divided into three sub-parts, namely, P wave, QRS complex and a T wave. Details on the etiology of these sub-waveforms can be found in (Schwartz et al. 2002). Heart rate is commonly estimated by calculating the number of R peaks in a minute. The variability in the interval between RR peaks is used to measure the HRV and can help in observing a number of body functions including the sympathetic and parasympathetic nervous system.

2.4.1 Heart rate variability and HIE

EEG is considered the gold standard for diagnosis and continous monitoring of HIE affected neonates. However, there are situations when EEG or expertise to interpret it are not readily available after birth e.g. in remote maternity hospitals. Therefore, interest has grown recently to develop new biomarkers for detecting and grading HIE injury.

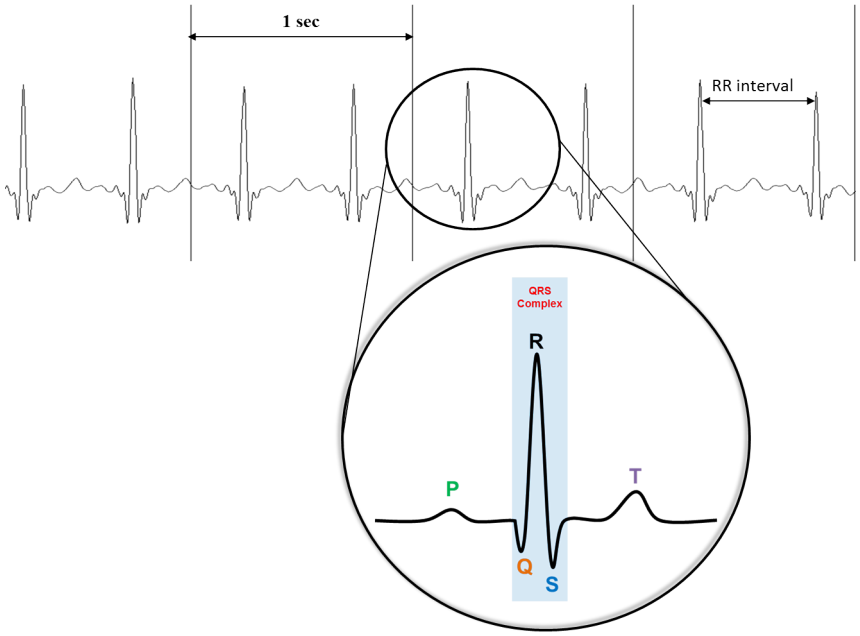


Figure 2.10: An example of ECG waveform of a healthy neonate. Encircled zoomed part shows the typical construction of ECG of one complete heart beat.

Recently many studies have been conducted to find the relationship between HRV and HIE in neonates and have shown that HRV could be a good indicator of the grade of HIE injury in neonates (Massaro et al. 2014, Goulding et al. 2015, Aliefendioglu et al. 2012, Matic et al. 2013). Figure 2.11 shows the instantaneous RR interval generated from 3 min of neonatal ECG of a neonate with mild HIE along with that of a neonate with severe HIE. It can be seen that the variability of the RR interval (HRV) decreases as the grade of the HIE worsens. It is believed that this lower variability in heart rate is either due to brain stem injury which controls the heart rate rhythm (Volpe 2008, Novak et al. 1995) or because of an immature nervous system which is the result of premature birth (Selig et al. 2011). A number of features derived from the HRV signal e.g. mean RR interval and standard deviation of RR interval, have been shown to discriminate between the HIE grades and later neurodevelopmental outcomes (Goulding et al. 2015).

This association of HRV with HIE makes it a good candidate for using it in conjunction with the EEG. This could help the clinical staff to better understand or validate the cause of brain damage and grade the severity of the HIE injury.

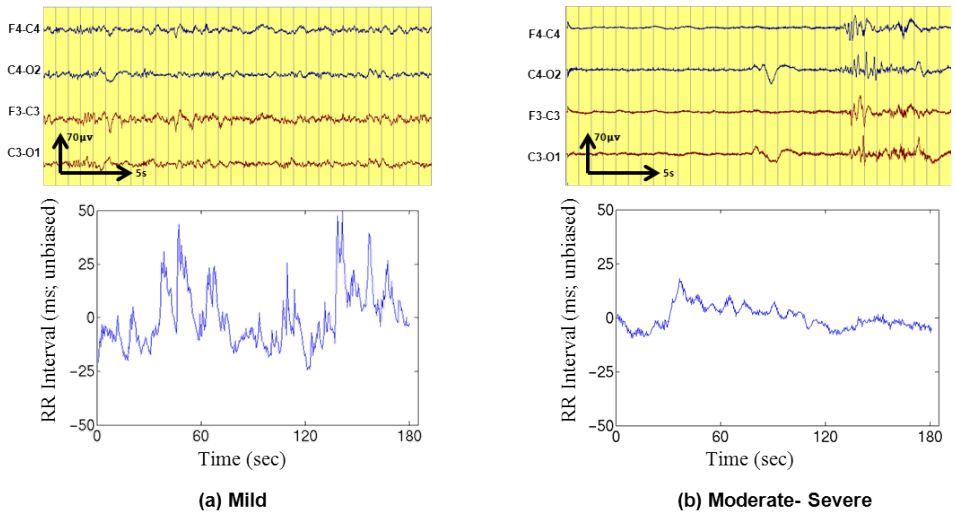


Figure 2.11: Examples of the Mild and moderate - severe HIE grade on EEG and its related HRV in RR interval. The HRV signals are 3 minute long. The signals are normalized by subtracting the mean HRV in this time duration. The corresponding EEG samples shown are 1 minute long.

2.5 Datasets

As mentioned earlier in chapter 1, this work dealt with the development of two automated systems; one for seizure detection in EEG recordings and the second for grading HIE using EEG and ECG recordings. Therefore two separate datasets are used to train and test the developed systems. Both datasets were collected in the NICU of Cork University Maternity Hospital. Table 2.2 provides an overview of these datasets.

Dataset used for seizure detection system: The dataset used in the automated seizure detection work consisted of long EEG recordings from 17 neonates. Each recording's length ranged from 5 to 26 hours. Seizures in these recordings were anotated by a neurophysiologist. Further details on this data can be found in section 4.3.

Dataset used for HIE grading system: For automated HIE grading using EEG, the dataset comprised of one hour EEG recordings of 54 neonates. Each recording was given a grade (mild, moderate, major and severe) with the consensus of two neurophysiologists. Furter details on this dataset are presented in section 5.3. Apart from EEG signals, the ECG signals from these neonates were also recorded. These ECG signals were used to develop an automated HIE grading system using the heart rate variability information.

Table 2.2: An overview of the datasets used in this study.

Seizure detection dataset	**HIE Grading Dataset**
· Continuous EEG recordings (5-26 hours) · Recordings from 17 neonates · Seizures annotated by two neurophysiologists · Recorded in NICU of Cork Maternity University Hospital	· 1 hour segments of EEG and ECG recordings · Recordings from 54 neonates · Segments graded by two neurophysiologists into (mild, moderate, severe, inactive) HIE · Recorded in NICU of Cork Maternity University Hospital

2.6 Conclusion

This chapter presented the basics of neonatal EEG. Particularly, the normal, abnormal patterns and artifacts that exist in the neonatal EEG are discussed. The etiology, diagnosis methods and treatments of the HIE injury and seizures are presented to inform the reader about the current clinical practices. Lastly, the effect of HIE injury on heart rate is briefly explained. Hence this chapter forms the basis of the functions of an automated system (what needs to be classfied) for neonatal brain monitoring.

References

Abend, N. S. & Wusthoff, C. J. (2012). Neonatal seizures and status epilepticus. *Journal of Clinical Neurophysiology : official publication of the American Electroencephalographic Society*, 29(5):441–448.

AL-Naddawi, M. N., Hameed, N. N., Kadum, M. J., & Al-Dabbas, N. W. (2011). Clinical types and possible etiologies of neonatal seizures: A hospital based study. *Journal of the Faculty of Medicine*, 53(1):1–5.

Aliefendioglu, D., Dogru, T., Albayrak, M., Dibekmısırlıoglu, E., & Sanli, C. (2012). Heart rate variability in neonates with hypoxic ischemic encephalopathy. *Indian Journal of Pediatrics*, 79(11):1468–1472.

Azzopardi, D. V., Strohm, B., Edwards, A. D., Dyet, L., Halliday, H. L., Juszczak, E., Kapellou, O., Levene, M. et al. (2009). Moderate hypothermia to treat perinatal asphyxial encephalopathy. *New England Journal of Medicine*, 361(14):1349–1358.

Boylan, G. B., Murray, D. M., & Rennie, J. M. (2008). The normal EEG and aEEG. *In Neonatal Cerebral Investigation*. Cambridge University Press, pp. 83–91.

Boylan, G. B., Stevenson, N. J., & Vanhatalo, S. (2013). Monitoring neonatal seizures. *Seminars in Fetal and Neonatal Medicine*, 18(4):202 – 208.

Choi, D. W. & Rothman, S. M. (1990). The role of glutamate neurotoxicity in hypoxic-ischemic neuronal death. *Annual Review of Neuroscience*, 13(1):171–182.

Clancy, R. R. (1996). The contribution of EEG to the understanding of neonatal seizures. *Epilepsia*, 37 Suppl 1:52–59.

Clancy, R. R. & Legido, A. (1987). The exact ictal and interictal duration of electroencephalographic neonatal seizures. *Epilepsia*, 28(5):537–541.

Erecinska, M., Thoresen, M., & Silver, I. A. (2003). Effects of Hypothermia on Energy Metabolism in Mammalian Central Nervous System. *Journal of Cerebral Blood Flow & Metabolism*, 23(5):513–530.
URL: *http://jcb.sagepub.com/content/23/5/513.abstract*

Fatemi, A., Wilson, M. A., & Johnston, M. V. (2009). Hypoxic-ischemic encephalopathy in the term infant. *Clinics in perinatology*, 36(4):835–858.

Goulding, R. M., Stevenson, N. J., Murray, D. M., Livingstone, V., Filan, P. M., & Boylan, G. B. (2015). Heart rate variability in hypoxic ischaemic encephalopathy: correlation with EEG grade and two-year neurodevelopmental outcome. *Pediatric Research*, 77(5):681–687.

Gray, P. H., Tudehope, D. I., Masel, J. P., Burns, Y. R., Mohay, H. A., O'Callaghan, M. J., & Williams, G. M. (1993). Perinatal hypoxic-ischaemic brain injury: prediction of outcome. *Developmental Medicine and Child Neurology*, 35(11):965–973.

Johnston, M. V., Trescher, W. H., Ishida, A., & Nakajima, W. (2001). Neurobiology of hypoxic-ischemic injury in the developing brain. *Pediatric research*, 49(6):735–741.

Kohelet, D., Shochat, R., Lusky, A., & Reichman, B. (2004). Risk factors for neonatal seizures in very low birthweight infants: population-based survey. *Journal of Child Neurology*, 19(2):123–128.

Lanska, M. J., Lanska, D. J., Baumann, R. J., & Kryscio, R. J. (1995). A population-based study of neonatal seizures in Fayette County, Kentucky. *Neurology*, 45(4):724–732.

Lombroso, C. T. (1996). Neonatal seizures: a clinician's overview. *Brain and Development*, 18(1):1–28.

Low, E., Boylan, G. B., Mathieson, S. R., Murray, D. M., Korotchikova, I., Stevenson, N. J., Livingstone, V., & Rennie, J. M. (2012). Cooling and seizure burden in term neonates: an observational study. *Archives of Disease in Childhood - Fetal and Neonatal Edition*, 97(4):267–272.

Lynch, N. E., Stevenson, N. J., Livingstone, V., Murphy, B. P., Rennie, J. M., & Boylan, G. B. (2012). The temporal evolution of electrographic seizure burden in neonatal hypoxic ischemic encephalopathy. *Epilepsia*, 53(3):549–557.

Massaro, A. N., Govindan, R. B., Al-Shargabi, T., Andescavage, N. N., Metzler, M., Chang, T., Glass, P., & du Plessis, A. J. (2014). Heart rate variability in encephalopathic newborns during and after therapeutic hypothermia. *Journal of Perinatology*, 34(11):836–841.

Matic, V., Cherian, P. J., Widjaja, D., Jansen, K., Naulaers, G., Van Huffel, S., & De Vos, M. (2013). Heart rate variability in newborns with hypoxic brain injury. *In Oxygen Transport to Tissue XXXV*. Springer, pp. 43–48.

Murray, D. M., Boylan, G. B., Ali, I., Ryan, C. A., Murphy, B. P., & Connolly, S. (2008). Defining the gap between electrographic seizure burden, clinical expression and staff recognition of neonatal seizures. *Archives of Disease in Childhood-Fetal and Neonatal Edition*, 93(3):F187–F191.

Murray, D. M., Boylan, G. B., Ryan, C. A., & Connolly, S. (2009). Early EEG findings in hypoxic-ischemic encephalopathy predict outcomes at 2 years. *Pediatrics*, 124(3):e459–e467.

Murray, D. M., Ryan, C. A., Boylan, G. B., Fitzgerald, A. P., & Connolly, S. (2006). Prediction of seizures in asphyxiated neonates: correlation with continuous video-electroencephalographic monitoring. *Pediatrics*, 118(1):41–46.

Niedermeyer, E. & Silva, F. H. L. d. (2005). *Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*. Lippincott Williams & Wilkins.

Novak, V., Novak, P., deMarchie, M., & Schondorf, R. (1995). The effect of severe brainstem injury on heart rate and blood pressure oscillations. *Clinical Autonomic Research*, 5(1):24–30.

O'Brien, J. R., Usher, R. H., & Maughan, G. B. (1966). Causes of birth asphyxia and trauma. *Canadian Medical Association Journal*, 94(21):1077–1085.

Patrizi, S., Holmes, G. L., Orzalesi, M., & Allemand, F. (2003). Neonatal seizures: characteristics of EEG ictal activity in preterm and fullterm infants. *Brain and Development*, 25(6):427–437.

Rennie, J. & Boylan, G. (2007). Treatment of neonatal seizures. *Archives of Disease in Childhood. Fetal and Neonatal Edition*, 92(2):F148–F150.

Sabzehei, M. K., Basiri, B., & Bazamamoun, H. (2014). The Etiology, Clinical Type, and Short Outcome of Seizures in NewbornsHospitalized in Besat Hospital/Hamadan/ Iran. *Iranian Journal of Child Neurology*, 8(2):24–28.
URL: *http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4058061/*

Scher, M. S., Hamid, M. Y., Steppe, D. A., Beggarly, M. E., & Painter, M. J. (1993). Ictal and interictal electrographic seizure durations in preterm and term neonates. *Epilepsia*, 34(2):284–288.

Schwartz, P., Garson, A., Paul, T., Stramba-Badiale, M., Vetter, V., Villain, E., & Wren, C. (2002). Guidelines for the interpretation of the neonatal electrocardiogram. *European heart journal*, 17(23):1329–1344.

Selig, F. A., Tonolli, E. R., Silva, E. V. C. M. d., & Godoy, M. F. d. (2011). Heart rate variability in preterm and term neonates. *Arquivos brasileiros de cardiologia*, 96(6):443–449.

Stiles, J. & Jernigan, T. L. (2010). The basics of brain development. *Neuropsychology review*, 20(4):327–348.

Tsuchida, T. N., Wusthoff, C. J., Shellhaas, R. A., Abend, N. S., Hahn, C. D., Sullivan, J. E., Nguyen, S., Weinstein, S. et al. (2013). 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. *Journal of Clinical Neurophysiology*, 30(2):161–173.

Volpe, J. J. (2008). *Neurology of the Newborn*. Elsevier Health Sciences.

Walsh, B. H., Murray, D. M., & Boylan, G. B. (2011). The use of conventional EEG for the assessment of hypoxic ischaemic encephalopathy in the newborn: A review. *Clinical Neurophysiology*, 122(7):1284 – 1294.

Watkins, A., Szymonowicz, W., Jin, X., & Yu, V. V. Y. (1988). Significance of seizures in very low-birthweight infants. *Developmental Medicine & Child Neurology*, 30(2):162–169.

Weiner, S. P., Painter, M. J., Geva, D., Guthrie, R. D., & Scher, M. S. (1991). Neonatal seizures: electroclinical dissociation. *Pediatric Neurology*, 7(5):363–368.

Wu, Y. (2015). Clinical features, diagnosis, and treatment of neonatal encephalopathy.
URL: *http://www.uptodate.com/contents/clinical-features-diagnosis-and-treatment-of-neonatal-encephalopathy, Accessed on:2015-01-14*

Chapter 3

From Static to Dynamic Classification

This chapter provides an overview of the different classification methods, and hence will act as a reference for all the classification tasks and algorithms detailed in the chapters to follow. Specifically, it will introduce Support Vector Machine (SVM) which is used as the primary classifier in this work. Moreover, sequential kernel methods are presented that can enable SVM to classify variable length sequences of feature vectors to make decisions based on both discrete and the contextual information.

The rest of the chapter is arranged as follows. The first section provides an overview and motivation of sequential/dynamic classification. The operation and construction of the standard Support Vector Machine (SVM) classifier is described in section 2. Lastly section 3 presents an overview of different sequential kernels along with a brief introduction to the Gaussian Mixture Models (GMM) as used by some sequential kernels.

3.1 Static and dynamic classifiers

A typical supervised machine learning task is to create a classifier that can correctly predict the classes of new observations given training examples of the old observations (Mitchell 1997). This task can typically be formalized as follows.

Let $\mathbf{x}$ represent the set of features of a single observation, and y denote the class associated with this observation. A training example is normally formed as a pair $(\mathbf{x}, y)$ containing an observation and its corresponding class. A classifier is a function f that maps from observations to classes. The aim of the learning process is to find a f that can correctly predict the class $y = f(\mathbf{x_{new}})$ of a

new observation $\mathbf{x_{new}}$. This is achieved by optimizing the parameters of this function f, so that it gives good results on the training data without over-fitting (Dietterich 2002).

Usually in many classification problems, it is assumed that each observation is independently and identically drawn from a joint distribution $P(\mathbf{x}, y)$ which means that there is no relationship between observations. Although a classifier built on this assumption may produce good results, there are classification cases when there is some correlation present between the observations. For example, in hand writing recognition for the English language, if the classifier determined that the first letter of a certain word is 'Q' then it is almost certain that the next letter would be U which means that there exists some relationship between both letters. Another example is hand gesture recognition using image data where an individual image may represent a static hand position in space. However, in order to classify a complete gesture, the whole motion of hand through a number of images needs to be tracked (Mitra & Acharya 2007). Clearly, a classifier that can exploit this kind of contextual information is expected to achieve better performance than the classifier that utilizes only a single observation.

The case of EEG signals is similar to these examples. Usually, features are extracted from a short segment of EEG called an epoch (Figure 3.1a). However, as mentioned in Chapter 2, EEG also contains information over a longer time scale. For example, in the case of seizures, although there is a definite stereotypical pattern of this injury; these patterns evolve in frequency and shape through time. An example of such a seizure is shown in Figure 3.1a. It can be seen, that the seizure changes its morphology (as highlighted by the colored regions) over a longer time scale. This information may also be used in detecting seizures by classifying a *sequence* of feature vectors. Similarly, for classifying the severity of HIE injury, it is necessary to not only detect the short events like a burst of high frequency activity but also to capture the slow variation of these events in time.

Not all classifiers can inherently classify sequences of observation/feature vectors. Such classifiers are called static classifiers, whereas the ones that can classify variable length sequences of feature vectors are called sequential/dynamic classifiers. An example of a static classifier is SVM, which inherently can not deal with arbitrary sequence length. However the use of special dynamic kernel functions inside the SVM could make it a dynamic classifier.

The purpose of this study is to investigate the sequential classification approaches for seizure detection and classification of background EEG, and hence

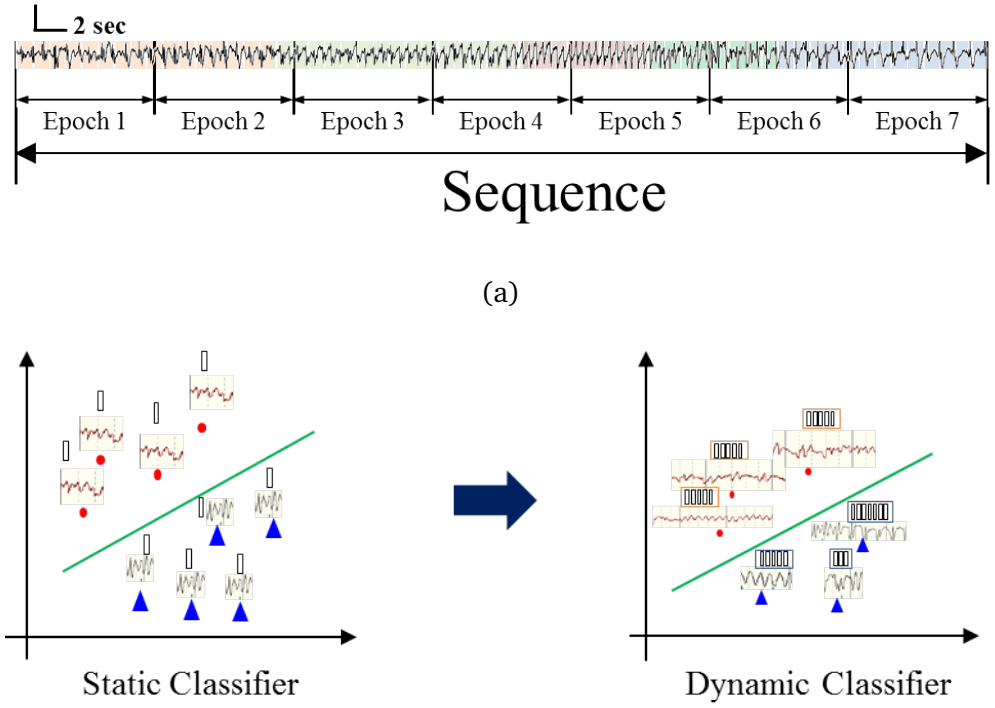


Figure 3.1: (a) An Example of an evolving seizure sequence. (b) Static Vs Dynamic Classifier.

as shown in Figure 3.1b, it is intended to transform the static classifier used in the earlier work by (Temko et al. 2011) which only processed a single feature vector to a dynamic classifier that can classify sequences of feature vectors. As SVM is used as the primary classifier in this work, the rest of chapter will first describe the working of SVM and then will present different methods that can enable it to explore contextual and sequential information.

3.2 Support Vector Machine (SVM)

SVM is a binary discriminative classifier and uses a hyperplane to classify the data (Vapnik & Kotz 1982). The SVM has shown state of the art performance in many pattern recognition areas and has become a popular classifier in the machine learning community.

In order to explain the construction/working of SVM, we will consider the simplest case of two linearly separable classes. Suppose a training set of N feature vectors $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ with labels $\{y_1, y_2, \ldots, y_N\}$ where $\mathbf{x}_i \in \Re^n$ and

$y_i \in \{-1, 1\}$. Lets suppose there exists a linear hyperplane

$$\mathbf{w} \cdot \mathbf{x} + b = 0 \tag{3.1}$$

which separates the dataset into two regions: $\mathcal{R}_+$ where positive samples reside and the region $\mathcal{R}_-$ where negative samples reside. The data points that lie on the hyperplane will satisfy $\mathbf{w} \cdot \mathbf{x} + b = 0$, where $\mathbf{w}$ is the normal vector to the hyperplane, $\frac{|b|}{||\mathbf{w}||}$ is the perpendicular distance from the hyperplane to the origin and $||\mathbf{w}||$ is the Euclidean norm of $\mathbf{w}$. Figure 3.2 shows an illustration of the explained case. Let $d_+(d_-)$ be the shortest distance to the nearest +ive(-ive) data point from the hyperplane. The margin of a separating hyperplane can be defined as $(d_+ + d_-)$. It can be appreciated from the Figure 3.2 that there are many possible hyperplanes which could separate the two classes of data. However not all of them provide the maximum gap between the two classes. The SVM training algorithm strives for the hyperplane which will separate the data with largest margin. For a linearly separable case, where all data points are correctly classified, this can be formulated as follows:

$$\mathbf{w} \cdot \mathbf{x_i} + b \geq +1 \quad for \quad y_i = +1 \tag{3.2}$$

$$\mathbf{w} \cdot \mathbf{x_i} + b \leq -1 \quad for \quad y_i = -1 \tag{3.3}$$

or

$$y_i((\mathbf{w} \cdot \mathbf{x_i}) + b) \geq 1 \quad where \quad i = 1, \cdots, N \tag{3.4}$$

The data points for which these two equalities hold, lie on either of the two parallel hyperplanes called the supporting hyperplanes, $H_+ : (\mathbf{w} \cdot \mathbf{x}_i + b = +1)$ where +ive training data points lie and $H_- : (\mathbf{w} \cdot \mathbf{x}_i + b = -1)$ where -ive training data points lie. The distance D, called the *margin*, between these two supporting hyperplanes is

$$D = \frac{\mathbf{w} \cdot \mathbf{x}_+ + b}{||\mathbf{w}||} + \frac{\mathbf{w} \cdot \mathbf{x}_- + b}{||\mathbf{w}||} = \frac{2}{||\mathbf{w}||} \tag{3.5}$$

In order to maximize the generalization performance of the classifier, the margin/gap $D = \frac{2}{||\mathbf{w}||}$ needs to be increased which implies that $||\mathbf{w}||$ needs to minimized. This optimization problem can be formed as follows

$$\min(\frac{1}{2}||\mathbf{w}||^2) \quad \text{subject to} \quad y_i((\mathbf{w} \cdot \mathbf{x_i}) + b) \geq 1 \quad \forall \quad i \in \{1, \cdots, N\} \tag{3.6}$$

Equation 3.6 is called the primal formulation of the SVM training algorithm.

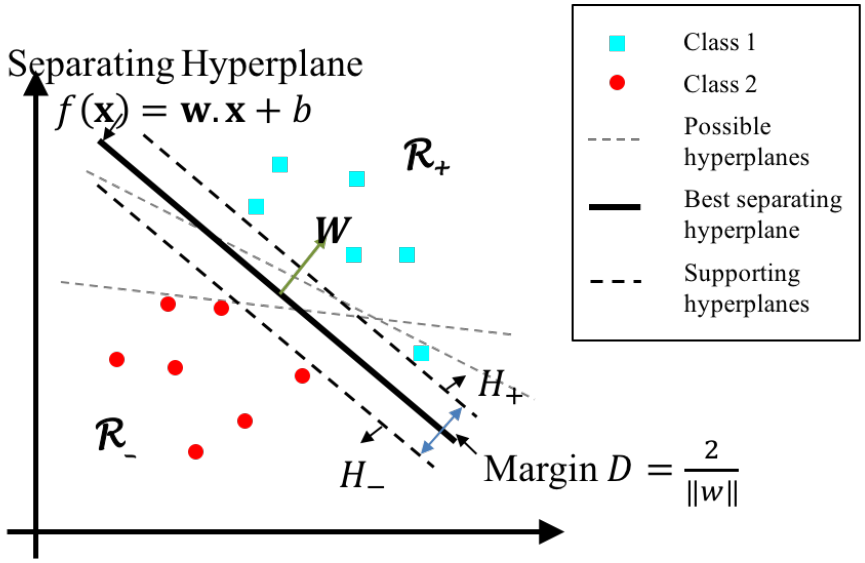


Figure 3.2: An illustration of a linear SVM given two perfectly separable classes.

To solve the primal, it is re-formulated into a dual optimization problem as a Lagrangian function $L(\mathbf{w}, b, \alpha)$.

$$\begin{aligned} L(\mathbf{w}, b, \alpha) &= \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{N} \alpha_i \left[y_i(\mathbf{w} \cdot \mathbf{x_i} + b) - 1\right] \\ &= \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{N} \alpha_i y_i(\mathbf{w} \cdot \mathbf{x}_i) - b\sum_{i=1}^{N} \alpha_i y_i + \sum_{i=1}^{N} \alpha_i \end{aligned} \tag{3.7}$$

where $\alpha_i \geq 0, i = 1, \cdots, N$ are the Lagrange multipliers and the optimization problem becomes

$$\max_{\alpha} \left(\min_{\mathbf{w},b} L(\mathbf{w}, b, \alpha) \right) \tag{3.8}$$

This is called the dual formulation of the actual primal formulation of the SVM optimization problem. There are two benefits of representing the primal in this way; the constraints in Equation 3.6 are replaced by the Lagrange multipliers, which are easier to handle and the input data appears in the form of a dot product which allows for the use of kernels to map the data in higher dimensional spaces (Burges 1998). It will be shown later that the kernel trick is a very important attribute of SVM particularly, in the non-linear case where data may not be linearly separable in the input feature space.

In order to minimize the Lagrangian function, its derivative with respect to

$\mathbf{w}$ and b are set to zero as follows

$$\begin{aligned} \frac{dL(\mathbf{w}, b, \alpha)}{db} &= -\sum_{i=1}^{N} \alpha_i y_i = 0 \\ \frac{dL(\mathbf{w}, b, \alpha)}{d\mathbf{w}} &= \mathbf{w} - \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i = 0 \end{aligned} \tag{3.9}$$

which yields

$$\sum_{i=1}^{N} \alpha_i y_i = 0 \quad \text{and} \quad \mathbf{w} = \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i \tag{3.10}$$

substituting these derivations in equation 3.7

$$L(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) \tag{3.11}$$

and hence the optimization problem becomes (substitute 3.11 in 3.8)

$$\begin{aligned} &\max_{\alpha} \left(\sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) \right) \\ &\text{subject to} \\ &\sum_{i=1}^{N} \alpha_i y_i = 0 \quad \text{and} \quad \alpha_i \geq 0, i = 1, \cdots, n \end{aligned} \tag{3.12}$$

By solving the equation 3.12 for α, the weight $\mathbf{w}$ can be determined using equation 3.10. Substituting $\mathbf{w}$ in the SVM function of equation 3.1 for the hyperplane yields the SVM decision function

$$f_{svm}(\mathbf{x}) = sign\left(\sum_{i=1}^{N} \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{x}_j) + b \right). \tag{3.13}$$

This means that the classification problem consists of assigning to any input vector $\mathbf{x}$ one of the two classes according to the sign of Equation 3.13. The above equation also indicates another interesting property of SVM that it does not need to save all the training data and requires only those training samples for which their α_i is non-zero (samples that lie on the supporting hyperplanes called the support vectors).

3.2.1 Soft margin SVM

The above mentioned SVM formulation is for the perfectly separable classes. However, in most real-world classification problems the classes are not perfectly separable (linearly) due to presence of outliers. Therefore, it is necessary to extend the linearly separable SVM to linearly non-separable case by softening the margin and allowing some errors. In order to do so, a non-negative slack variable ξ is added to the equalities in Equation 3.4 for each training vector as

$$\mathbf{w} \cdot \mathbf{x_i} + b \geq +1 - \xi_i \quad for \quad y_i = +1 \tag{3.14}$$

$$\mathbf{w} \cdot \mathbf{x_i} + b \leq -1 + \xi_i \quad for \quad y_i = -1 \tag{3.15}$$

$$\tag{3.16}$$

Thus, for an error to occur, the corresponding ξ_i must exceed 1, so $\sum_{i=1}^{N} \xi_i$ is an upper bound on the number of training errors (Burges 1998). An extra cost for errors is introduced in the objective function which results in following optimization problem

$$\begin{aligned} &\min(\frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{N}\xi_i) \\ &\text{subject to} \quad y_i((\mathbf{w} \cdot \mathbf{x_i}) + b) \geq 1 - \xi_i \quad \forall \quad i \in \{1, \cdots, N\} \end{aligned} \tag{3.17}$$

where C is the regularization parameter which needs to be chosen by the user and controls the degree of penalization to the slack variables. The dual optimization problem for the Equation 3.17 remains the same except for the change in the constraints as follows.

$$\begin{aligned} &\max_{\alpha}\Big(\sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j(\mathbf{x}_i \cdot \mathbf{x}_j)\Big) \\ &\text{subject to} \\ &\sum_{i=1}^{N}\alpha_i y_i = 0 \quad \text{and} \quad 0 \leq \alpha_i \leq C, \quad i = 1, \cdots, N \end{aligned} \tag{3.18}$$

Figure 3.3 shows an illustration of soft margin SVM where classes are not completely separable. SVM creates a decision boundary that best separates the two classes by allowing some data points to fall outside their respective supporting hyperplanes. The errors ξ_i are equal to the distance from the corresponding supporting hyperplane of the class to the data point (Hastie et al. 2005). These

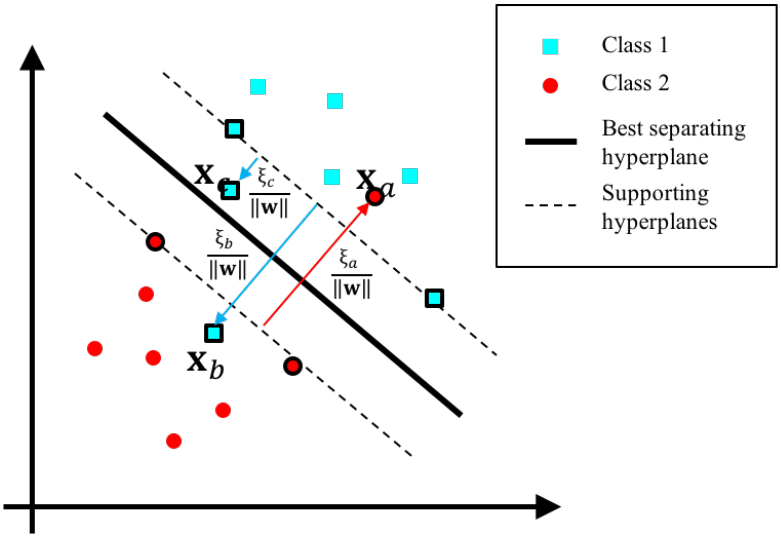


Figure 3.3: An illustration of the soft margin SVM. (Adopted from (Hastie et al. 2005))

data points then become the support vectors along with the ones that lie on the supporting hyperplanes (circle with black outline).

3.2.2 Kernels in SVM

Very often the classes are not linearly separable in the input feature space. Figure 3.4 shows an example of such a case. The data lies in a two dimensional space and it is obvious that no linear hyperplane can separate this data. In such cases, the SVM formulation can be generalized for the case where the decision function is not a linear function of the data (Burges 1998). It can be noted that, the only way the data appears in the equations of optimization problems (Equation 3.18) and the SVM decision function (Equation 3.13) is in the form of a dot product $(\mathbf{x}_i \cdot \mathbf{x}_j)$. Now suppose, the n dimensional data is mapped to some other (possibly infinite) Euclidean space $\mathbb{F}$ using a mapping function Ψ:

$$\Psi : \mathbf{R}^d \rightarrow \mathbb{F} \tag{3.19}$$

The reason for this mapping is that, in the high dimensional space the mapped data may become more linearly separable compared to the input feature space. It can be seen in Figure 3.4b that when a non-linear transformation is applied to the data, such that in the new transformed space the data is defined by three

dimensions rather than two dimensions as follows,

$$\Psi_{3d}(\mathbf{x}) = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \\ \mathbf{x}_a^2 + \mathbf{x}_b^2 \end{pmatrix}, \tag{3.20}$$

then the data can be separated using a linear hyperplane.

Now, in order for the SVM to create a hyperplane, all the data will need to be first transformed using the $\Psi_{3d}(\mathbf{x})$. The dot product in the Equation 3.13 will thus be replaced by $\Psi_{3d}(\mathbf{x_i}) \cdot \Psi_{3d}(\mathbf{x_j})$. However, this could be a very challenging task when the input feature space is also high dimensional. Therefore, SVM employs a special *kernel function* $\mathbb{k}$ to perform the non-linear mapping such that $\mathbb{k}(\mathbf{x}_i, \mathbf{x}_j) = \Psi(\mathbf{x}_i) \cdot \Psi(\mathbf{x}_j)$. Thus, rather than first transforming the data into a new feature space, the SVM will never need to explicitly know the transformation Ψ and only $\mathbb{k}$ would be needed in the training algorithm whose result will be the same as taking the dot product in the mapped feature space (Burges 1998, Camastra et al. 2009). With this new kernel function, the optimization problem can be formulated as

$$\begin{aligned} &\max_{\alpha} \Big(\sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \Psi(\mathbf{x}_i) \cdot \Psi(\mathbf{x}_j) \Big) \\ &= \max_{\alpha} \Big(\sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \mathbb{k}(\mathbf{x}_i, \mathbf{x}_j) \Big) \end{aligned} \tag{3.21}$$

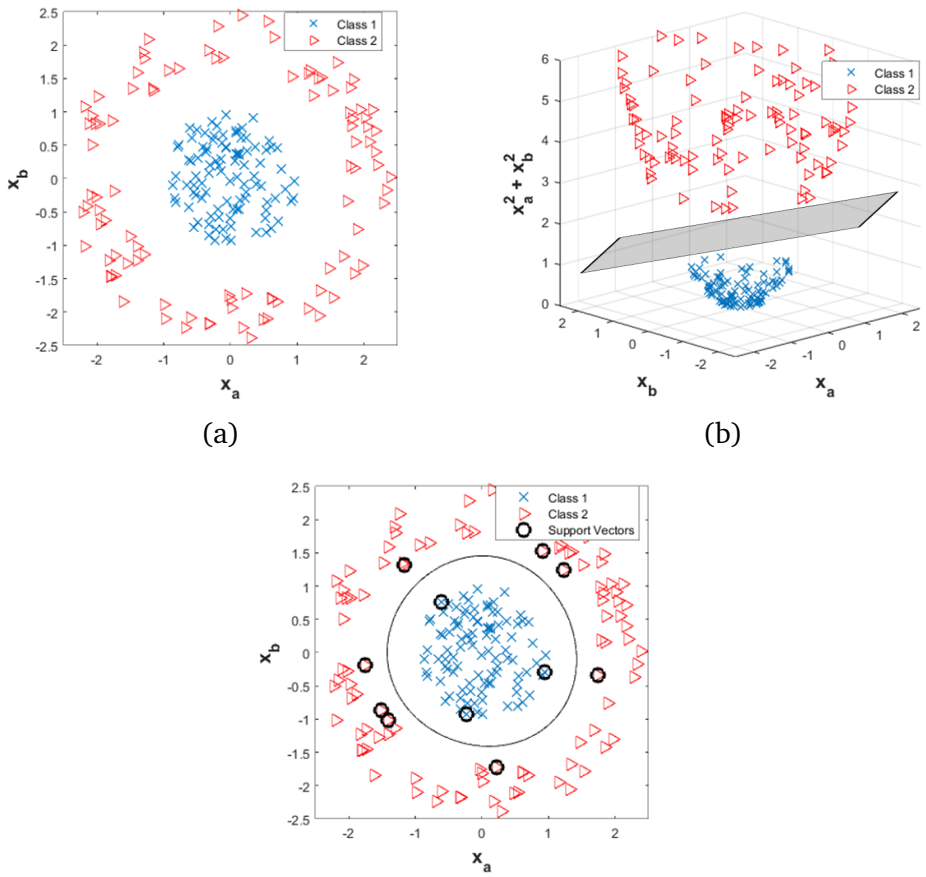
subject to

$$\sum_{i=1}^{N} \alpha_i y_i = 0 \quad \text{and} \quad 0 \leq \alpha_i \leq C, \quad i = 1, \cdots, N \tag{3.22}$$

The resulting decision function of the SVM is

$$f_{svm}(\mathbf{x}) = sign\left(\sum_{i=1}^{N} \alpha_i y_i \mathbb{k}(\mathbf{x}_i, \mathbf{x}_j) + b \right). \tag{3.23}$$

There are a number of kernel functions that can be used. It should be noted however, that in order for a kernel to be valid for SVM, it should be symmetric and positive semi definite. This condition guarantees that SVM objective function remains convex and will converge to the unique optimal solution. Some of the popular kernel functions are as follows.



(a) (b)

(c)

Figure 3.4: (a) A classification problem where data is not linearly separable in the input space. (b) Mapping the data to a higher dimensional space could make it linearly separable. (c) An SVM trained using a polynomial kernel. It can be seen that the decision boundary is no longer a linear hyperplane in the input two dimensional space.

- *Gaussian radial basis function:*

$$\mathbb{k}(\mathbf{x}_i, \mathbf{x}_j) = \exp(\frac{-||\mathbf{x}_i - \mathbf{x}_j||^2}{2\sigma^2}) \tag{3.24}$$

- *Polynomial:*

$$\mathbb{k}(\mathbf{x}_i, \mathbf{x}_j) = \Big((\mathbf{x}_i \cdot \mathbf{x}_j) + \theta\Big)^{deg} \tag{3.25}$$

3.2.3 Converting the SVM output to probabilities

Output of the SVM function is a distance measure along with a sign that determines its class. This distance can be converted into the posterior probabilistic

measure using a sigmoid function defined as,

$$P_{svm}(y = 1|f_{svm}(\mathbf{x})) = \frac{1}{1 + \exp(Af_{svm}(\mathbf{x}) + B)}. \tag{3.26}$$

The parameters A and B of this function are calculated over the training data using an iterative method of maximum likelihood estimation. The training data could be the same as used for training the SVM or a separate hold-out set.

In this work, the data used for training the SVM is also utilized for calculating A and B parameters.

3.2.4 Making a sequential SVM

SVM has shown state of the art performance in many areas, but inherently it is not built for classifying sequences of the feature vectors.

The most obvious and straight forward approach is to concatenate all the feature vectors of a fixed number into one single feature vector. However, as the basic kernels developed for SVM are designed to classify single data points and even if the sequences of fixed length are available those kernels could not exploit the hidden contextual and sequential information in the sequences.

A second approach is to use the sequential kernels. These specialized kernel functions map the sequences to a higher dimensional space and then return a measure of similarity between them by exploiting the contextual information in that space. Some such sequential kernels for SVM are

- Fisher kernel
- Supervector kernel
- Gaussian dynamic time warping kernel
- Polynomial time warping kernel
- Dynamic time alignment kernel
- String kernel

A brief description of these kernel functions is provided in section 3.2.7.

The choice of the kernel function usually depends on the type of information present in the sequence or the type of information needed to be explored from a given sequence. For example in some cases the *sequentiality* of feature vectors may provide significant information for classification. An example could be

speech recognition where the task is to identify words and sentences in spoken language and convert them to text or machine-readable format. Human speech is composed of many components ranging from smaller phonemes, words, sentences to longer conversations. Usually, features are extracted from the a short window of few seconds/milliseconds to get the information about the most basic component i.e. phoneme. However, in order to classify and make sense of a word or sentence, a sequence of short feature vectors needs to be classified together. Fisher kernel with Hidden Markov Models (HMM), string kernel and Dynamic Time Warping (DTW) based kernels are popular choices in such scenarios because they classify sequences based on the sequentiality information.

On the other hand, in some classification problems it is not important to conserve or explore the sequentiality of the feature vectors. For example in speaker verification/identification problem it is not important to know the order of the words uttered by the speaker. In such cases, the underlying *statistics* of the constituting feature vectors in a sequence is explored i.e. what feature values appeared or how many of such feature values are present in the sequence. The kernel functions used in such cases are usually based on generative models e.g. supervector kernel and Fisher kernel with Gaussian Mixture Models (GMM).

3.2.5 Sequential kernels used in this work

In the context of this work, the Gaussian Dynamic Time Warping (GDTW) kernel is used for the seizure detection problem and a supervector kernel is used for the classification of background EEG. The reason for using a GDTW kernel to detect seizures is that the neonatal seizure events undergo certain change in pattern from the start to the end (Figure 2.9,3.1a). Therefore it is important to use a kernel function that can explore this contextual and temporal information to detect seizure events. The classification of background EEG however, does not require to detect a single event or pattern but to find the amount of short events (bursts) or the distance between these events (Inter burst interval) in a given sequence to classify it into one of four grades of HIE. Details on both of these kernels will be presented in subsequent chapters.

As the supervector kernel is based on GMM, therefore it is worthwhile giving a brief introduction of GMM as used in this work. This will be followed by a concise description of some sequential kernels for SVM.

3.2.6 Gaussian Mixture Models

The GMM is a probabilistic model which assumes that the data representing the feature space can be modelled using a finite number of Gaussian distributions (Reynolds 2009). It is represented as a Probability Density Function (PDF), which is a weighted sum of M Gaussian components:

$$p(\mathbf{x}) = \sum_{j=1}^{M} w_j \, g(\mathbf{x}|\mathbf{m}_j, \mathbf{\Sigma}_j). \tag{3.27}$$

where $\mathbf{x}$ is a feature vector of *n* dimensions, $w_j, j = 1, \dots M$ are the mixture weights and $g(\mathbf{x}|\mathbf{m_j}, \mathbf{\Sigma_j})$ are the component densities. Each Gaussian component is a n dimensional Gaussian function described as

$$g(\mathbf{x}|\mathbf{m}_j, \mathbf{\Sigma}_j) = \frac{1}{(2\pi)^{\frac{n}{2}}|\mathbf{\Sigma}_j|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{m_j})^T \mathbf{\Sigma}_j^{-1}(\mathbf{x} - \mathbf{m_j}) \right) \tag{3.28}$$

where $\mathbf{m}_j$ is the mean and $\mathbf{\Sigma}_j$ is the covariance matrix of the j^{th} Gaussian component (Reynolds 2009). The weights of all the Gaussian components should sum to unity. Figure 3.5 shows a GMM model generated from the data which obviously looks for three clusters. The peaks of each individual component represents the density of data in the input space. The height of each peak in the GMM represent the weight of each component, which in turn represents the density/amount of data in each cluster. It can be seen that the amount of data in cluster 3 was the smallest therefore it has the lowest peak in the GMM.

There are a number of GMM parameters estimation methods; however, Maximum Likelihood remains the most common and well established method (Reynolds 2009). Maximum A-Posteriori (MAP) estimation is another method to estimate the GMM parameters which is commonly used to adapt the parameters of a well trained GMM (by ML algorithm). The details of both these algorithms will be described below.

3.2.6.1 Maximum likelihood estimation

As the name implies, this algorithm tries to find a model which maximizes the likelihood of the training data. Given a set of training data $\mathbf{X} = \{\mathbf{x_1}, \cdots, \mathbf{x_N}\}$ and a GMM model with parameters $\vartheta = \{w_j, \mathbf{m}_j, \mathbf{\Sigma}_j\} \quad j = 1, \cdots, M$, the likelihood of the data towards this GMM model can be calculated as

$$p(\mathbf{X}|\vartheta) = \prod_{n=1}^{N} p(\mathbf{x}_n, \vartheta) \tag{3.29}$$

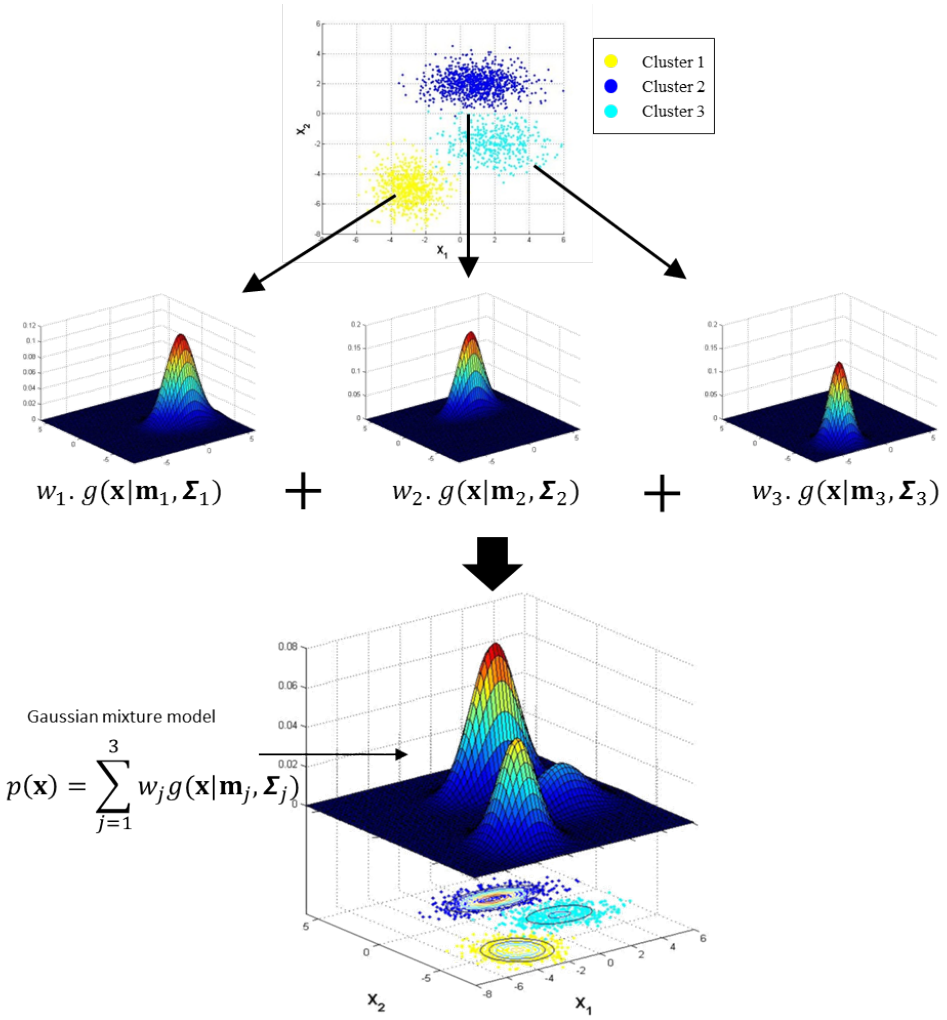


Figure 3.5: Construction of a GMM model. Here the data is used to make a GMM with 3 Gaussian components. The lowest plot shows the PDF of the GMM.

Now in order to find the parameters ϑ that maximize the likelihood $p(\mathbf{X}|\vartheta)$, an algorithm called the Expectation Maximization (EM) (Dempster et al. 1977) is used. The main concept of EM algorithm is to start with some initial parameters ϑ, that could have been guessed or derived from another algorithm such as K-means. Then iterate and keep looking for a new set parameters $\hat{\vartheta}$ with the constraint $p(\mathbf{X}|\vartheta) \geq p(\mathbf{X}|\hat{\vartheta})$ until a certain stop condition. The stop condition could be number of iterations or a convergence threshold where no more change in likelihood happens.

Therefore, the EM algorithm can be divided into two steps; 1-**Expectation**: where a-posteriori probability is calculated and 2- **Maximization** where the new parameters $\hat{\vartheta}$ are calculated. In the expectation step, the a-posterior prob-

ability is calculated using the Bayes theorem as,

$$p_{jn}^{\sim} = \tilde{p}(j|\mathbf{x}_n, \vartheta) = \frac{w_j g(\mathbf{x_n}|\mathbf{m}_j, \mathbf{\Sigma}_j)}{\sum_{m=1}^{M} w_m g(\mathbf{x_n}|\mathbf{m}_m, \mathbf{\Sigma}_m)} \tag{3.30}$$

for each of the M Gaussian component and each feature vector $\mathbf{x}_n \in \mathbf{X}$.

In the maximization step, the parameters of each of the M Gaussian components are updated using the following equations

$$\text{Weights} \qquad \hat{w}_j = \frac{1}{N} \sum_{n=1}^{N} \tilde{p}_{jn} \tag{3.31}$$

$$\text{Means} \qquad \hat{\mathbf{m}}_j = \frac{\sum_{n=1}^{N} \tilde{p}_{jn} \mathbf{x}_n}{\sum_{n=1}^{N} \tilde{p}_{jn}} \tag{3.32}$$

$$\text{Covariance} \qquad \hat{\sigma}_j^2 = \frac{\sum_{n=1}^{N} \tilde{p}_{jn} \mathbf{x}_n^2}{\sum_{n=1}^{N} \tilde{p}_{jn}} - \mathbf{m}_j^2 \tag{3.33}$$

Figure 3.6 shows the evolution of Gaussian components in different iterations of EM algorithm. The data is the same as that used to produce Figure 3.5. The means of the Gaussian components are initialized in the first iteration and then move towards their respective clusters. Moreover, it can be seen that the variance of each component also changes as the algorithm proceeds.

3.2.6.2 Maximum a-posteriori adaptation (MAP)

Another method to estimate the GMM parameters is by use of MAP adaptation. MAP adaptation could also be considered as a ML regularization scheme because the estimation of new parameters is usually from a well trained GMM model via EM (Reynolds 2009). The specifics of this method are as follows.

Given a set of training observations $\mathbf{X} = \{\mathbf{x_1}, \cdots, \mathbf{x_N}\}$ and a prior model ϑ_p, the posterior probabilities $\tilde{p}_{jn}$ of this data are calculated using Equation 3.30. Then sufficient statistics measures are calculated. These statistical measures are the count, the first and the second moments which are required to compute the weight, mean and variance respectively for the new model. These sufficient statistics can be calculated as

$$\text{Count} \qquad c_j = \sum_{n=1}^{N} \tilde{p}_{jn} \tag{3.34}$$

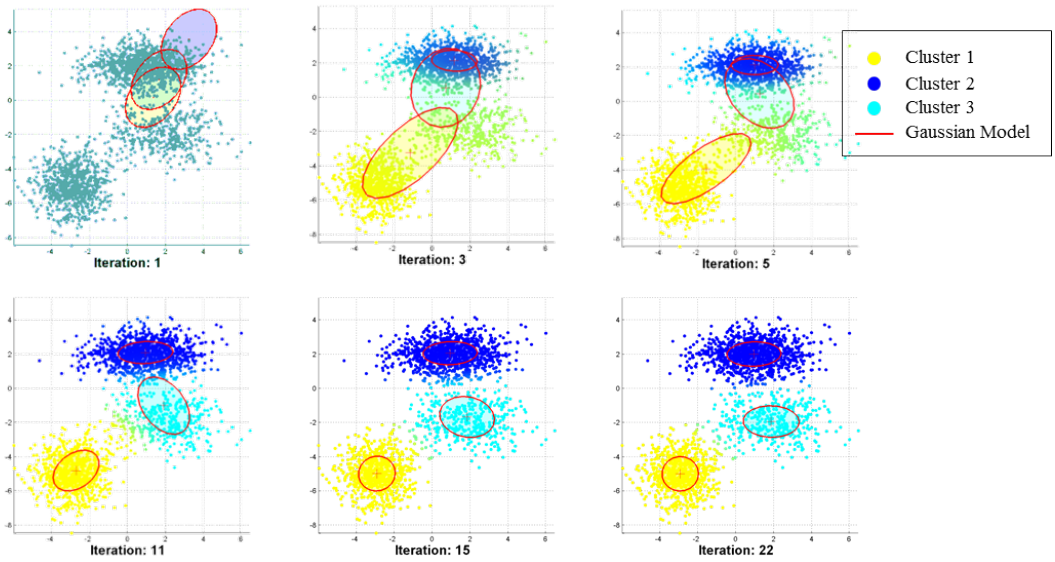


Figure 3.6: Visualization of the Expectation Maximization algorithm for training a GMM model using Maximum Likelihood.

$$\text{1st moment} \qquad \mathfrak{f}_j(\mathbf{x}) = \frac{1}{c_j}\sum_{n=1}^{N}\tilde{p}_{jn}\mathbf{x}_n \tag{3.35}$$

$$\text{2nd moment} \qquad \mathfrak{s}_j(\mathbf{x}) = \frac{1}{c_j}\sum_{n=1}^{N}\tilde{p}_{jn}\mathbf{x}_n^2 \tag{3.36}$$

Using the measures in the above equations the new model parameters are calculated as follows

$$\hat{w}_j = [\beta_j^w(\frac{c_j}{N}) + (1-\beta_j^w)w_j]\lambda \tag{3.37}$$

$$\hat{\mathbf{m}}_j = \beta_j^m\mathfrak{f}_j(\mathbf{x}) + (1-\beta_j^m)\mathbf{m}_j \tag{3.38}$$

$$\hat{\sigma}_j^2 = \beta_j^\sigma\mathfrak{s}_j(\mathbf{x}) + (1-\beta_j^\sigma)(\sigma_j^2+\mathbf{m}_j^2) - \hat{\mathbf{m}}_j^2 \tag{3.39}$$

where $\beta_j^w, \beta_j^m, \beta_j^\sigma$ are the adaptation coefficients that control the amount of adaptation from the old parameters. λ is the scale factor to ensure the weights of all mixtures sum to one. The adaptation coefficients can be calculated as

$$\beta_j^{\{w,m,\sigma\}} = \frac{c_j}{c_j+\delta^{\{w,m,\sigma\}}} \tag{3.40}$$

where $\delta^{\{w,m,\sigma\}}$ is called the relevance factor for each of the parameter sets which can be fixed after trying certain values according to the input data. Moreover, in most GMM applications there is only one value used for all adaptation parameters i.e. $\beta_f = \beta_j^w = \beta_j^m = \beta_j^\sigma$. Therefore it comes down to only one parameter

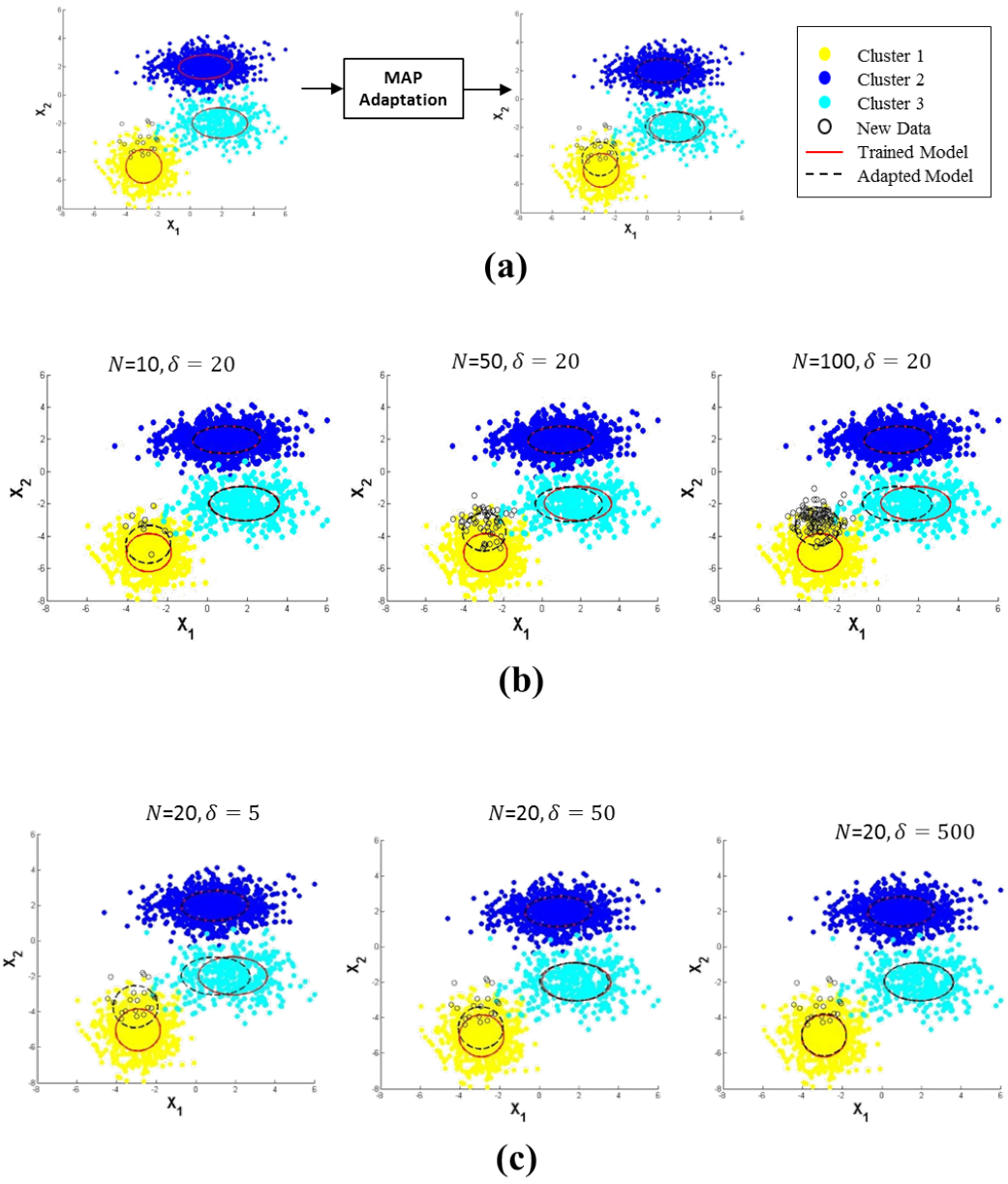


Figure 3.7: An illustration of MAP adaptation of a general model towards some new training data.

δ, that controls the amount of adaptation required for the new model. Furthermore, the adaptation parameters are data dependent which means if the number of data points N is greater than the probabilistic count c_j of the data will then be high which in turn will cause the new parameters $\hat{\vartheta}$ to shift more towards the training data. If N is low then new parameters will be more similar to the older model parameters ϑ (Reynolds 2009).

Figure 3.7 shows an example of MAP adaptation. The red lined ellipses represents the already trained model as shown in Figure 3.6a with the EM al-

gorithm. The black dashed line ellipses represent the adapted model given the new data. It can be seen that means of cluster 1 and 3 moved towards the new data. For illustration purposes only the means of the trained model are adapted. Figure 3.7b and 3.7c shows the effect of the number of data points N and relevance factor δ on the rate of adaptation. It can be seen that adaptation rate is higher when the number of data points are increased given a fixed value of δ. However, if the number of data points will remain the same in an experiment then δ can be used to control the rate of adaptation. As δ is in the denominator of Equation 3.40, therefore the lower the value of δ, the higher the adaptation rate.

MAP estimation is widely used in the speaker recognition area, where a general model of many speakers is trained using ML and then a speaker specific model is created by adapting the general model using MAP adaptation.

A note on GMMs as used in this work: There are a few things that need to be highlighted on the use of GMMs as used in the work carried out in this thesis. First, k-means clustering was used to initialize the model parameters in ML estimation. Secondly, the log of likelihood is used instead of simple likelihood (Equation 3.29) in the ML estimation. Finally, MAP adaptation in this work is used to create a specific class model by adapting the parameters from a general EEG model. This general EEG model is called the Universal Background Model (UBM) and is created by the ML estimation process. More details on UBM will be presented in chapter 5.

3.2.7 Sequential Kernels for SVM

This section provides a brief overview of some sequential kernels. First, some common notations used throughout the following section will be defined.

Consider two sequences $S = \{s_1, ..., s_{N_S}\}$ and $R = \{r_1, ..., r_{N_R}\}$ with variable lengths N_S and N_R. s_i and r_j could be a time series data point or a n dimensional feature vector in the respective sequences. In order to measure similarity between these sequences, the following kernels can be used:

- **Fisher kernel:** is a popular approach in the area of speech, audio and speaker classification (Wan & Renals 2005, Temko et al. 2006). The main idea is to make a kernel function from a generative model. It uses a Fisher score function that maps the sequence of feature vectors of any length on to a score space and returns a fixed length feature vector that is usable in

the discriminative classifier (Jaakkola & Haussler 1999).

Given a sequence S, and a generative model e.g. GMM $g(\phi)$, where $\phi = \{w_i, \mathbf{m_i}, \mathbf{\Sigma}_i\}$, the Fisher score-vector can be computed as

$$F(S) = \Delta_\phi \log P(S|g, \phi) \tag{3.41}$$

where Δ is the Fisher-score operator and $P(X|g, \phi)$ is called the Fisher score argument which is the likelihood of the sequence S to the model $g(\phi)$. Each element of the Fisher score-vector $F(S)$ is a derivative of log-likelihood of P $(S|g, \phi)$ with respect to one of the GMM model parameter $\{w_i, \mathbf{m_i}, \mathbf{\Sigma}_i\}$. This is a similar process to the ML algorithm where the new parameters are similarly updated by taking the derivatives. When the derivatives are small, it means the ML algorithm is close to the desired model or maximum likelihood is about to be achieved (Wan & Renals 2005). Similarly, the Fisher score-vector captures the information on how close the sequence is to the given model using derivatives.

The resulting kernel function for two sequences S and R is

$$\Bbbk_f(S, R) = (F(S) \cdot F(R)) \tag{3.42}$$

- **Supervector kernel:** also relies on a generative model. However, rather than using the log-likelihood of a sequence coming from a given GMM model, it uses the parameters of an *adapted GMM model* to represent a variable length sequence.

  Given a sequence S and a GMM model g_o, a new GMM model g_a is created by adapting the means $\mathbf{m}_i \quad i = 1, \ldots, M$, of G_o using the MAP adaptation technique. A GMM supervector $\mathbf{v}$ is created by concatenating all the means of the new model. It is shown in (Campbell et al. 2006) that an inner product between two supervectors, $\mathbf{v_1}$ and $\mathbf{v_2}$, is an upper bound of the Kullback-Leibler divergence between the two GMMs (Campbell et al. 2006). Thus, given two sequences S and R, the supervector kernel is

defined as

$$\begin{aligned}\mathbb{k}_{sv}(S,R) &= \sum_{i=1}^{M} w_i (\mathbf{m}_i^S)^T \mathbf{\Sigma}_i^{-1} \mathbf{m}_i^R \\ &= \sum_{i=1}^{M} (\sqrt{w_i}\mathbf{\Sigma}_i^{-\frac{1}{2}} \mathbf{m}_i^S)^T (\sqrt{w_i}\mathbf{\Sigma}_i^{-\frac{1}{2}} \mathbf{m}_i^R). \end{aligned} \tag{3.43}$$

where w_i are the weights, $\mathbf{\Sigma}_i$ covariance matrix and M is the number of Gaussian components of the GMM model g_a. This approach has shown state of the art results in speaker verification (Campbell et al. 2006).

- **String kernel:** The string kernel is an interesting approach which can be implemented for different types of signals by first quantizing the sequences using different symbols and then applying the method below to get information about the sequentiality and repetitiveness of these symbols in a sequence.

  The string kernel was originally developed for document classification (Lodhi et al. 2002). The main idea is to divide any strings/signals/sequences into sub-strings/sub-signals/sub-sequences and then to produce a score that is weighted according to the number of matched sub-sequences and their frequency of occurrence in the other sequence. Moreover, the gaps in the parts of sub sequences are also weighted.

  Given a sequence S which contains u sub-sequences, the mapping is defined as

$$\psi_m(S) = \sum_{i:m=s[i]} \lambda^{l(i)} \tag{3.44}$$

  where λ is the weight according to the length $l(i)$ of the sub-sequence m and $s[i]$ is the index of all the possible sub-sequences of the sequence S. Now the string kernel function between two sequences S_1 and S_2 is defined as

$$\begin{aligned}\mathbb{k}_s(S,R) &= \sum_{u\in\zeta} \langle \psi_m(S) \cdot \psi_m(R) \rangle \\ &= \sum_{u\in\zeta} \sum_{i:m=s[i]} \sum_{j:m=t[j]} \lambda^{l(i)+l(j)} \end{aligned} \tag{3.45}$$

$$\text{where} \quad \zeta = \text{the set of all finite strings} \tag{3.46}$$

Intuitively, it means that mapped feature $\psi_m(S)$ measures the number of occurrences of sub-sequences in the sequence S weighted according to their length. In the case of kernel function the inner product of two sequences S and R, will provide a sum over all common sub-sequences weighted according to their frequency of occurrence.

In order to elaborate a bit more on its working, assume two words *fat* and *far*. There are 3 sub-strings in each word i.e. fat={fa,at,fat} and far={fa,ar,far}. Now the only common sub-string is *fa* and its length is 2, therefore the $\Bbbk(far, fat)$ will be $\lambda^{2+2} = \lambda^4$.

- **Gaussian Dynamic Time Warping (GDTW) kernel:** Dynamic Time Warping (DTW) is a popular technique to compare the shapes of two variable length time series (Muller 2007). Specifically, it calculates distance (e.g. Euclidean distance) d_l between each time point of the two series S & R and creates a gram matrix of these local distances.

  Consider a warp path $W = \{w_1, \cdots, w_k, \cdots, w_K\} \in \mathbf{W}$ of length K through this gram matrix, where $w_k = (n_k, m_k)$ represents the k^{th} vertex on this path. This path is constructed under the monotonic constraints $m_{k+1} \geq m_k$ and $n_{k+1} \geq n_k$ to preserve the shape, continuity and restrained from the backward propagation. Then an overall global DTW distance D_{dtw} between the series is computed by finding the shortest path in this gram matrix using Dynamic programming (Sakoe & Chiba 1978), The distance D_{dtw} could be defined as

$$D_{dtw}(S, R) = \min\Big\{\frac{1}{K}\sum_{k=1}^{K} d_l(n_k, m_k)\Big\}, \tag{3.47}$$

  where

$$d_l(n_k, m_k) = ||s_{n_k} - r_{m_k}||^2 \tag{3.48}$$

  In order to use this approach in the SVM for comparing variable length sequences, the popular RBF kernel function as defined in Equation 3.24 is used. However, the Euclidean distance in the kernel function is replaced by the DTW distance.

  Hence, given two sequence S and R the kernel is defined as

$$\Bbbk_{dtw}(S, R) = \exp(\frac{-D_{dtw}(S, R)}{2\sigma^2}) \tag{3.49}$$

  The resulting kernel is called Gaussian dynamic time warping kernel. This

approach was first developed for the purpose of online handwriting recognition (Bahlmann et al. 2002) and has shown good results in many other pattern recognition areas (Temko et al. 2006).

- **Polynomial time warping kernel:** This kernel shares the same idea of using DTW to compare sequences except for two changes (Wan & Carmichael 2005). Instead of using the original feature vectors, it performs a spherical normalization which maps them onto the surface of a unit sphere. The function to perform this normalization is

$$\hat{s}_i = \frac{1}{\sqrt{s_i^2 + \varpi^2}} \begin{bmatrix} s_i \\ \varpi \end{bmatrix} \tag{3.50}$$

where s_i is the i^{th} feature vector of a sequence S and ϖ is a constant which is the distance from the center of the sphere to the input space. For more details on these parameters please see (Wan 2003). Next, the local distance in the DTW gram matrix is replaced by the cosine distance d_{cos} and DTW global distance D_{Cdtw} is calculated by finding the shortest path in the gram matrix. Thus the global distance D_{Cdtw} between two sequences S and R could be defined as

$$D_{Cdtw}(S, R) = \min\Big\{\frac{1}{K}\sum_{k=1}^{K} d_{cos}(\hat{n}_k, \hat{m}_k)\Big\} \tag{3.51}$$

where

$$d_{cos}(\hat{n}_k, \hat{m}_k) = arccos(\hat{s}_{\hat{n}_k} - \hat{r}_{\hat{m}_k}) \tag{3.52}$$

Lastly, cosine of the D_{Cdtw} is taken and raised to the pw^{th} power. The kernel function is then defined as

$$\Bbbk_{poly}(S, R) = \cos^{pw}(D_{Cdtw}) \tag{3.53}$$

- **Dynamic time alignment kernel:** This kernel (Shimodaira et al. 2002) also uses a slightly modified version of the original DTW kernel. Here, the Euclidean distance between two feature vectors is replaced by a kernel function. This local kernel could be a dot product or any other SVM kernel. As the kernel gives a measure of similarity (its value is higher if two vectors are similar thus instead of finding the path that minimizes the global distance), this approach looks for the path that maximizes the

global distance D_{dtak} using the criterion formulated in (Shimodaira et al. 2002). Hence given two sequences S and R, the global distance can be calculated as

$$D_{dtak}(S,R) = \max\Big\{\frac{1}{K}\sum_{k=1}^{K}\Bbbk_l(n_k,m_k)\Big\}, \tag{3.54}$$

where $\Bbbk_l$ is the local kernel function and can be defined as $\Bbbk_l = \exp(\frac{(s_{n_k}-r_{m_k})}{2\sigma^2})$. The definition of the global kernel function for the two sequences becomes

$$\Bbbk_{dtak}(S,R) = D_{dtak}(S,R) \tag{3.55}$$

- **Global alignment DTW kernel:** All the DTW based kernel methods defined above looked for an optimum path in the gram matrix. This approach however does not follow the same method and instead uses the sum of all the possible alignment paths. It was argued that the paths closer to the diagonal or optimum path also contribute towards a better alignment of the two sequences (Cuturi et al. 2007). Any of the above mentioned local distances can be used to create the gram matrix. If the local distance is assumed to be d_{dtak} as defined in Equation 3.54 between the feature vectors of two sequences S and R, then the global distance D_{GA} can be calculated as

$$D_{GA}(S,R) = \sum_{k^*\in\Upsilon}\Big\{\frac{1}{K}\sum_{k=1}^{K}\Bbbk_l(n_k,m_k)\Big\}, \tag{3.56}$$

where Υ is the set of all possible alignment paths in the gram matrix. Then the kernel function can be written as

$$\Bbbk_{GA}(S,R) = D_{GA}(S,R) \tag{3.57}$$

This kernel when used with D_{dtak} distance, has shown better performance in classifying spoken letters, when compared to the HMM and simple dynamic time alignment kernel defined in Equation 3.55 (Cuturi et al. 2007).

A note on DTW based kernels:

In order to make a valid SVM kernel, it is necessary that it should be Positive Semi Definite (PSD). This condition guarantees, that kernel will be able

to reproduce the Hilbert space and the resulting optimization problem for SVM training will be convex. However all the DTW based kernels described above have been argued and criticized to be non-PSD (Gudmundsson et al. 2008). Despite this fact, they have been used in many studies and shown good performance for different sequential classification problems.

3.3 Conclusion

This chapter presented the basic operation of SVM and its extension to the classification of variable length sequences via different kernel methods. The main motivation to use sequential kernel is to explore the contextual information that is present in the EEG and other physiological signals on a longer time scale. Different kernel methods based on generative modelling and dynamic time warping techniques are presented. Although all kernels have strong theoretical basis and have shown to improve results, not one kernel however can be singled out as superior to the other kernels.

References

Bahlmann, C., Haasdonk, B., & Burkhardt, H. (2002). Online handwriting recognition with support vector machines - a kernel approach. *In Proceedings of 8th International Workshop on Frontiers in Handwriting Recognition, 2002.* pp. 49–54.

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167.

Camastra, F., Vinciarelli, A., & Yu, J. (2009). Machine learning for audio, image and video analysis. *Journal of Electronic Imaging*, 18(2):029901–029901.

Campbell, W., Sturim, D., Reynolds, D., & Solomonoff, A. (2006). SVM based speaker verification using a GMM supervector kernel and NAP variability compensation. *In Proceedinds of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2006*, vol. 1. pp. 97–100.

Cuturi, M., Vert, J.-P., Birkenes, O., & Matsui, T. (2007). A kernel for time series based on global alignments. *In Proceedinds of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2007*, vol. 2. IEEE, pp. 413–416.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Dietterich, T. G. (2002). Machine learning for sequential data: A review. *In Structural, Syntactic, and Statistical Pattern Recognition*. Springer, pp. 15–30.

Gudmundsson, S., Runarsson, T. P., & Sigurdsson, S. (2008). Support vector machines and dynamic time warping for time series. *In Proceedings of IEEE International Joint Conference on Neural Networks, (IJCNN) 2008.* IEEE, pp. 2772–2776.

Hastie, T., Tibshirani, R., Friedman, J., & Franklin, J. (2005). *The elements of statistical learning: data mining, inference and prediction*, vol. 1 of *Springer Series in Statistics*. Springer-Verlag New York, 2nd ed.

Jaakkola, T. & Haussler, D. (1999). Exploiting generative models in discriminative classifiers. *In Advances in Neural Information Processing Systems II*. MIT Press, pp. 487–493.

Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., & Watkins, C. (2002). Text classification using string kernels. *The Journal of Machine Learning Research*, 2:419–444.

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill Boston, MA:.

Mitra, S. & Acharya, T. (2007). Gesture recognition: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 37(3):311–324.

Muller, M. (2007). Dynamic Time Warping. *In Information Retrieval for Music and Motion*. Springer Berlin Heidelberg, pp. 69–84.

Reynolds, D. (2009). Gaussian mixture models. *In Encyclopedia of Biometrics*. Springer US, pp. 659–663.

Sakoe, H. & Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 26(1):43–49.

Shimodaira, H., Noma, K.-i., Nakai, M., & Sagayama, S. (2002). Dynamic time-alignment kernel in support vector machine. *In Advances in Neural Information Processing Systems 14*. MIT Press, pp. 921–928.

Temko, A., Monte, E., & Nadeu, C. (2006). Comparison of sequence discriminant support vector machines for acoustic event classification. *In Proceedinds of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2006*, vol. 5. IEEE, pp. 721–724.

Temko, A., Thomas, E., Marnane, W., Lightbody, G., & Boylan, G. (2011). EEG-based neonatal seizure detection with support vector machines. *Clinical Neurophysiology*, 122(3):464–473.

Vapnik, V. N. & Kotz, S. (1982). *Estimation of dependences based on empirical data*, vol. 40. Springer-verlag New York.

Wan, V. (2003). *Speaker verification using support vector machines*. Ph.D. thesis, University of Sheffield, Sheffield.

Wan, V. & Carmichael, J. (2005). Polynomial dynamic time warping kernel support vector machines for dysarthric speech recognition with sparse training data. *In Proceedings of the 9th European Conference on Speech Communication and Technology (INTERSPEECH) 2005*. pp. 3321–3324.

Wan, V. & Renals, S. (2005). Speaker verification using sequence discriminant support vector machines. *IEEE Transactions on Speech and Audio Processing*, 13(2):203–210.

Chapter 4

Automated system for neonatal seizure detection using dynamic classifier

This chapter will give a detailed presentation of the methods employed for the neonatal seizure detection task, along with the obtained results and observations. Specifically, the Dynamic Time Warping (DTW) based kernel method that allows sequential classification to explore the contextual and temporal information in neonatal seizures is discussed here.

The chapter is organized in three sections. First a literature review of the previously used methods to make use of contextual information is presented. Then the DTW technique is explained in section 2 followed by the details of the baseline and proposed neonatal seizure detection system. Results are discussed in the last section.

4.1 Exploring the contextual information for seizure detection

A typical neonatal seizure detection system comprises of the following main stages as shown in Figure 4.1. The signal representation stage (feature-level) is the process where relevant features are extracted from the EEG signal. In the classification stage (classifier level) the extracted feature or feature vectors are assigned to the seizure or non-seizure class using a set rules and thresholds which are either automatically derived from the data (classifier) or manually adjusted following or mimicking expert neurologists reasoning. The post-

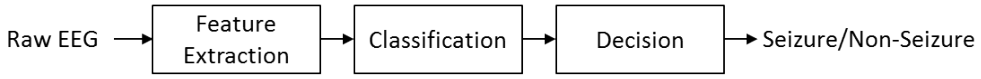


Figure 4.1: An overview of genereal architecture of neonatal seizure detection systems

processing stage (decision level) involves smoothing or other transformation to offer a clinician a support in decision making.

In contrast to background EEG and artifacts, seizure events in newborns change their frequency, morphology, and propagation. Figure 4.2 shows an example of a neonatal seizure where the attenuating amplitudes and evolving frequency of the spikes can be clearly seen. Contextual information also termed as the temporal evolution, signal dynamics or sequentiality of neonatal seizures has been explored in various ways in most automated detection systems reported to date (Aarabi et al. 2009, Gotman et al. 1997, Navakatikyan et al. 2006, Liu et al. 1992, Deburchgraeve et al. 2008, Aarabi et al. 2007).

At the feature-level, the temporal evolution of the EEG signal within the window has been captured using template extraction and matching (Aarabi et al. 2009), wavelet transformation (Ocak 2009) or EEG wave sequence analysis (Navakatikyan et al. 2006). Concatenation of the consecutive feature vectors in time was proposed in (Shoeb & Guttag 2010). The Kalman filter was exploited to enhance the contrast of the extracted features to the past background activity (Bogaarts et al. 2014).

At the decision-level, the temporal structure of neonatal seizure is incorporated through the use of smoothing filters applied over the classifier output, such as moving average (Temko et al. 2011a), median filter (Nagaraj et al. 2014) or Kalman filter (Bogaarts et al. 2014).

The classifier-level techniques typically find the temporal and contextual matching between dynamic length sequences of feature vectors. Examples are support vector machines with sequential kernels (Chaovalitwongse & Pardalos 2008) or Hidden Markov model (HMM) (Wong et al. 2007).

Incorporation of contextual information at each level has its advantages and drawbacks. The feature-level methods typically consider temporal evolution of EEG on a short-term scale. The filtering methods introduced at decision level perform smoothing over a large window and the short seizures get suppressed and are frequently missed as a result (Thomas et al. 2011). The exploration of temporal evolution on the classifier-level has been relatively scarce in the area of neonatal seizure detection whereas the classifier level techniques have shown

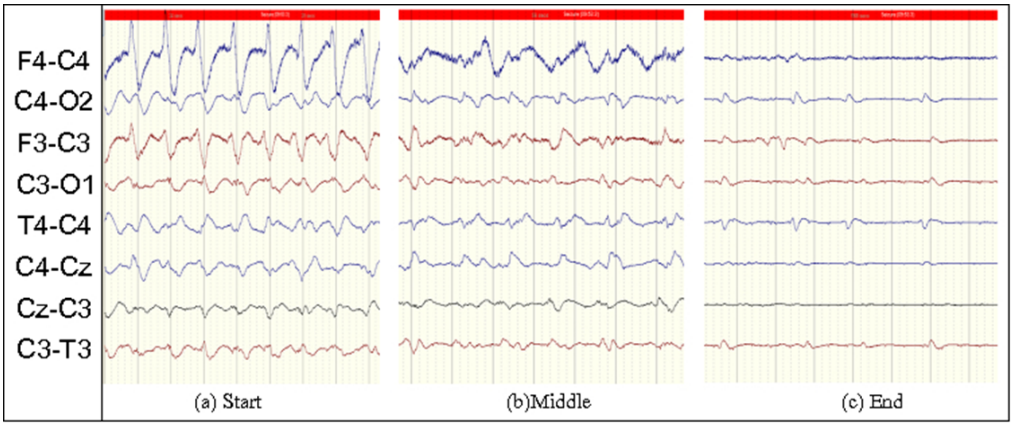


Figure 4.2: Evolution of a single seizure event. (a) Slow wave activity at start with sharp/spike components involved high in amplitude, phase reversal. (b) As the seizure progresses the EEG becomes lower in amplitude. (c) only very low amplitude discharges are seen with a flattening of the background.

promising results in other areas of signal processing such as speech recognition (Shimodaira et al. 2002, Smith & Gales 2002), handwriting character recognition (Bahlmann et al. 2002), acoustic events (Temko et al. 2006).

A state of the art neonatal seizure detection system based on the SVM has been reported in (Temko et al. 2011a). The system was developed by the Neonatal Brain Research Group at University College Cork. It was able to classify fixed short-term EEG epochs of 8 seconds (Temko et al. 2011a) (detailed in section 4.2) and used an averaging filter at the decision level to incorporate the contextual information and subsequently decrease the false alarm rate. However, it was shown in (Thomas et al. 2011) that although the system achieved a very high detection rate for longer seizures, shorter seizures of length less than one minute were frequently missed because of a large averaging filter window. Therefore, it was hypothesized that incorporating a classifier that can implicitly use the contextual information and classify long sequences of short-term feature vectors, could improve the detection rate of short seizures while keeping down the number of false detections.

4.2 Baseline neonatal seizure detection systems

Figure 4.3 shows an overview of the complete system. This system has shown state of the art results for neonatal seizure detection (Temko et al. 2011a). Each channel of the EEG recording is classified separately. The raw EEG from each channel is first down-sampled and then segmented into 8s epochs with a 50% overlap. A set of 55 features (outlined in Table 4.2) is extracted from each EEG

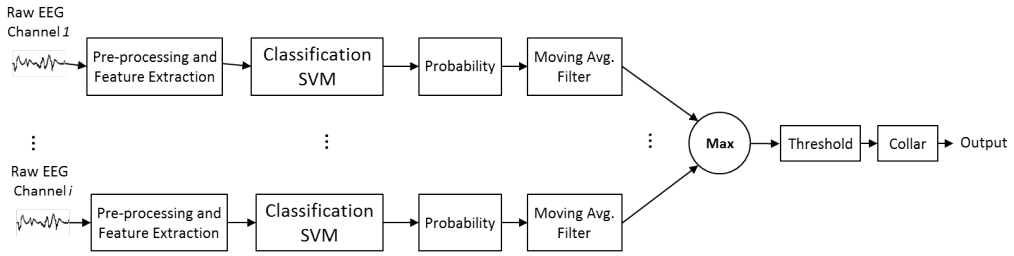


Figure 4.3: An overview of the state of the art neonatal seizure detection system based on static SVM.

epoch and fed to the classifier. A brief description of this set of features will follow in a subsequent section.

A SVM classifier with Radial Basis Function (RBF) kernel is used to classify each epoch. The output of the SVM is converted to probabilities using the method described in (Platt 1999). Hence each epoch for each channel is finally represented by a probability of seizure.

Three post-processing techniques are then applied to these probabilities. First a central moving averaging filter with a length of 15 epochs is used to smooth the probabilistic output. This filter also acts to include contextual information where information from the past and future epochs is used to decide the probability of a central epoch. Up to this block all the channels are handled separately. A MAX function is then applied on the probabilities across all the channels to get a single output which is then compared to a threshold. The epochs whose probabilities are greater than the threshold are considered as seizure epochs. Lastly, a collar is applied to the epochs classified as 'seizure epochs', which extends the decisions on either side by a further 40 seconds.

The performance of the system is assessed using the epoch-based, Receiver Operating Curve (ROC) area which measures the number of correctly classified epochs. Likewise the event-based Good Detection Rate (GDR) at the cost of False Detection per hour (FD/h) is used which determines the percentage of detected seizure events at the expense of making false alarms per hour. A dataset of 17 patients was used in this work. A mean ROC of 96.3% and a mean GDR of 89% at 1 FD/h was reported using the Leave One Out (LOO) patient cross validation method. As mentioned earlier the system frequently missed seizure events of length less than 60s. The false alarms were usually generated by seizure like activities in the background EEG or due to artifacts.

Recently, some new post-processing techniques for the above mentioned system have been investigated and have shown to improve the performance of the system. All of these methods are decision level techniques and are applied on

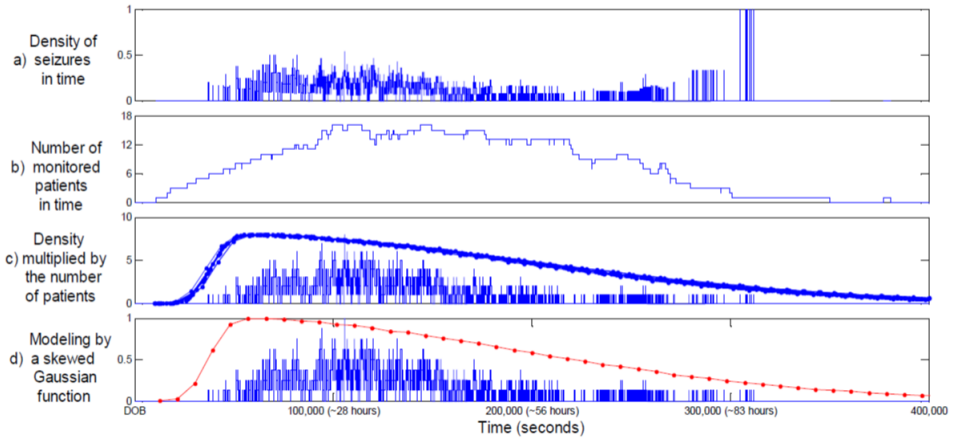


Figure 4.4: Different steps of modelling the probability density function.

the output of the SVM classifier. In (Temko et al. 2012b), the maximum output of the SVM classified across all channels was multiplied by data driven weights. These weights were derived from a Probability Density Function (PDF) that represented seizure burden in time (Figure 4.4). It was reported that the probability of a seizure in time is highest at 30 hours after birth so the value of weights will be highest around this time point and the probability will be lower before/after this time resulting in lower value of weights. A relative improvement of 23% was reported. Another method to incorporate the patient specific prior information was introduced in the form of channel weighting (Temko et al. 2012a). The weights were selected using a Bayesian probabilistic framework which finds the most likely channel to have a seizure in the future using the past information. These adaptive weights were calculated for each new epoch and then multiplied before combining the channels. This method resulted in a 22% relative increase in the ROC area averaged over 17 patients. Lastly, a technique of incorporating the knowledge of the background EEG activity was proposed in (Temko et al. 2013). The probability of non-seizure epochs for each channel was averaged for a certain time period to provide the estimated level of background. The probability of next epoch was then compared to the (threshold + estimated background level). Hence an adaptive threshold is created for every new epoch which helped to decrease the number of false alarms due to "seizure-like" activities (Figure 4.5). An improvement of 0.8% in overall ROC and a 5% increase in GDR at 1 FD/h was observed when compared to the above mentioned baseline system of (Temko et al. 2011a). This system named ANSeR (Algorithm for Neonatal Seizure Recognition) is currently undergoing clinical trials. In this chapter the results of the proposed system are com-

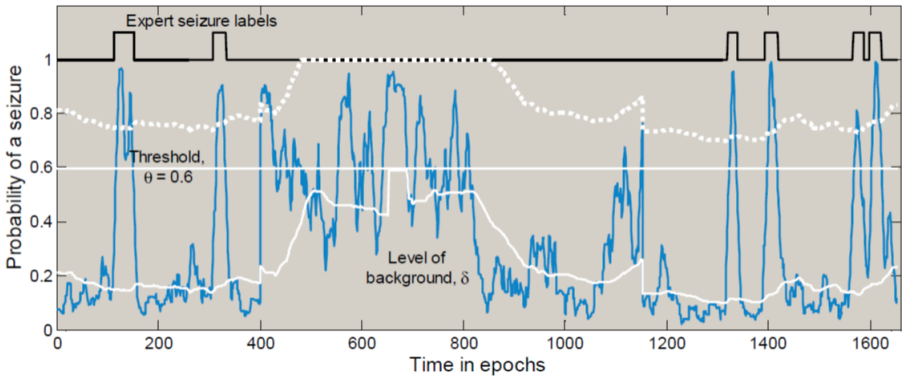


Figure 4.5: Example of the probabilistic output (in blue) for 1 hour 50 minutes of EEG from a single patient. The expert seizure labels are on the top. Here, the threshold θ, is set to 0.6 as indicated by the white horizontal line. The adaptively modelled level of background,δ is indicated below the horizontal line, in white. The dynamic threshold for the current level of background, $(\theta + \delta)$, is indicated above the horizontal line, in white.

pared to the baseline system defined in (Temko et al. 2011a) with and without a post-processing stage.

4.3 Neonatal seizure datasets

EEG data from 17 neonates was used in this study. These neonates were full term with the gestational age ranging from 39-42 weeks. The data was collected in the neonatal intensive care unit of Cork University Maternity Hospital. A written consent from the parents was taken to use the data for research purposes. The multichannel EEG was recorded using a Carefusion NicOne video EEG monitor with a sampling rate of 256Hz. The 10-20 system of electrodes placement was used. Eight bipolar EEG channels were derived from these electrodes (F4-C4, C4-O2, F3-C3, C3-O1, T4-C4, C4-Cz, Cz-C3, C3-T3). Seizures in these recordings were annotated with the consensus of two expert neurophysiologists using EEG and simultaneous video recordings. These seizures were a secondary injury due to HIE, however neonates were not cooled for treating HIE. Two datasets were derived from the complete EEG recordings of the neonates as follows.

1. A smaller dataset of only short seizure and non-seizure events of less than one minute.

2. A larger dataset comprising the continuous recordings of several hours of each patient.

The proposed system is first tested on a smaller dataset of shorter seizures and non-seizures. Validation of the achieved results is then carried out on the larger dataset of continuous recordings. The details of these datasets are as follows.

4.3.1 Per channel annotations

Seizures in the EEG recordings were annotated without specifying the channel in which they were present (Global annotations). In order to train the classifier however, it is essential to annotate the seizures with the channel information. The reason being, a seizure may not be present in all the channels and if only global annotations are used to get the seizure epochs, we may end up with a number of non-seizure epochs as well. Therefore, the neurophysiologist also annotated 2 minutes of seizure events in the continuous EEG recordings with channel information for every neonate (Per channel annotations).

4.3.2 Dataset 1 (short events only)

This dataset only contained seizures of lengths less than 64s (15 epochs) from 16 patients. As mentioned earlier, not all seizures in the dataset were annotated with the channel information. Moreover, even the seizures annotated with the channel information were not all short seizures or complete events. So, it was not directly possible to train the system on short seizure events only. Therefore, to get around this problem, the already developed neonatal seizure detection system was used to generate the per channel annotations of all the short seizures in the dataset with the help of the global annotations.

In order to do so, the probabilistic output of our best performing state of the art neonatal seizure detection system (Temko et al. 2013) was used as an annotator. First the short seizures epochs from all the channels were extracted using the global annotations. Then the channel where the system's probability was the highest for most of the event's duration was selected. This way a new set of per channel annotations was generated in order to train the classifier. Figure 4.6 shows an example of a short seizure event obtained using the above mentioned process. It can be seen that the second seizure has the highest probability in the channel F3-C3 (Top plot). Now using this information and the global annotation, the seizure event from this channel was extracted. It is obvious from the EEG trace that indeed the seizure pattern was most prominent in the channel F3-C3.

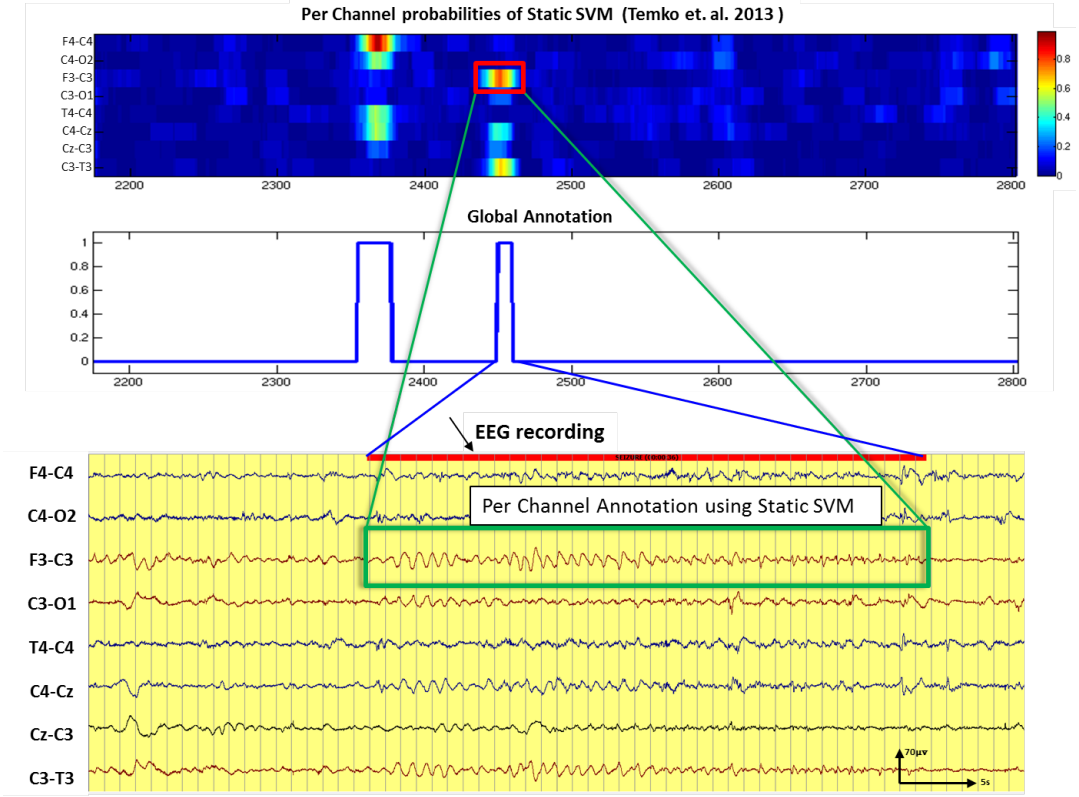


Figure 4.6: An example of creating the per channel annotation process using a Static SVM output. The top plot shows the per channel probabilities of the static SVM. The probability of the second seizure is highest in channel F3-C3 (red box). The middle plot shows the global annotations. The last plot shows the EEG recording and the global annotation of the second seizure event by the neurophysiologist and the green box shows the obtained seizure event with the channel information using the SVM probability.

Lastly, sequences of 15 epochs were created for both seizure and non-seizures. The seizures of length less than 15 epochs were padded with the non-seizure epochs, located immediately before and after the event, on both sides. For example if a seizure was of only 9 epochs then 3 epochs in front and 3 epochs at the end were added to make it a complete sequence. For each patient 9-times more non-seizure sequences were randomly extracted compared to the number of short seizures coming from the same patient. For example, if a patient had 5 short seizures, then 45 non-seizure sequences were randomly extracted from the same patient for training and testing the classifiers. This proportion was chosen to match the real life EEG data where the presence of seizures is very minimal compared to the non-seizure/background EEG.

4.3.3 Dataset 2 (continuous recordings)

Table 4.1 shows the particulars of each neonate's recording in the dataset. There were a total of 261h of EEG data with mean EEG recording time of 15h per patient. These recordings contained 821 different types of seizure events with a mean seizure length of 4.4 minutes. The recordings were not edited and no artefacts have been removed. So, this dataset is true representation of the real time situation in hospitals.

Table 4.1: EEG Dataset

			Seizures length		
Patient ID	**Record length (h)**	**Seizure events**	**Mean**	**Min**	**Max**
1	24.1	17	1' 40"	28"	4' 4"
2	24.7	3	6' 19"	1' 4"	11' 19"
3	22.7	170	2' 18"	24"	10' 55"
4	26.1	65	1' 30"	40"	3' 28"
5	24.0	51	6' 32"	28"	31' 11"
6	5.7	44	1' 15"	31"	2' 4"
7	13.2	71	1' 58"	28"	10' 31"
8	24.5	17	6' 6"	40"	19' 23"
9	24.0	157	5' 27"	28"	37' 16"
10	10.4	28	6' 0"	24"	34' 55"
11	6.2	14	5' 40"	55"	7' 47"
12	12.0	37	2' 31"	31"	10' 16"
13	12.1	25	4' 29"	1' 19"	12' 48"
14	5.5	13	8' 58"	1' 55"	39' 12"
15	12.2	58	2' 17"	19"	7' 19"
16	7.6	31	10' 34"	2' 23"	34' 47"
17	6.6	20	5' 36"	36"	23' 23"
Total	**261.7**	**821**			

4.4 Extending the SVM to become a dynamic classifier using Dynamic Time Warping

SVM is a binary classifier and the recent state of the art results suggest that it is well suited to the two class seizure detection task. However, as discussed in Chapter 3 that SVM can not inherently classify variable length sequences of feature vectors and thus specialized kernel functions are required to enable SVM to classify sequences. In the context of our seizure detection task, it was

important to use a kernel that cared about the sequentiality of feature vectors to preserve the temporal information. DTW is a popular technique to measure similarity between two variable length sequences. This section will provide details of the DTW method and how it can be used inside SVM to make it a dynamic classifier.

Consider two sequences $S = \{s_1, ..., s_{N_S}\}$ and $R = \{r_1, ..., r_{N_R}\}$ with variable lengths N_S and N_R. Where s_i and r_j could be a time series data point or a t dimensional feature vector in the respective sequences. A local distance measure like euclidean distance $d_l = ||s_i - r_j||^2$ between each element of the two sequences can be calculated to get a distance matrix of size $N_S \times N_R$. Consider a warp path $W = \{w_1, \cdots, w_k, \cdots, w_K\} \in \mathbf{W}$ of length K through this gram matrix, where $w_k = (n_k, m_k)$ represents the c^{th} vertex on this path. This path is constructed under the monotonic constraints $m_{k+1} \geq m_k$ and $n_{k+1} \geq n_c$ to preserve the shape, continuity and to be restrained from the backward propagation. The cost D_W of a particular path $W \in \mathbf{W}$ is

$$D_W(S, R) = \frac{1}{K} \sum_{k=1}^{\mathbb{k}} d_l(n_k, m_k) \tag{4.1}$$

where

$$d_l(n_k, m_k) = ||s_{n_k} - r_{m_k}||^2 \tag{4.2}$$

In order to find the optimal alignment path that gives the shortest distance $D_\phi = min\{D_W(S, R)\}$, an accumulated cost matrix D_{cost} is calculated. To compute the shortest distance to point (i, j) in the matrix (D_{cost}), the algorithm seeks the minimum value using $\min\{(d_{acc(i-1,j-1)} + d_{l(i,j)}), (d_{acc(i-1,j)} + d_{l(i,j)}), (d_{acc(i,j-1)} + d_{l(i,j)})\}$, where d_{acc} is the accumulated distance at the node. The dynamic programming algorithm, as defined in (Sakoe & Chiba 1978) is used to find the path that gives the shortest DTW distance in this matrix. Figure 4.7 presents an illustration of the formulated DTW algorithm using a toy example of two variable length sequences. This simple example provides an insight into the DTW algorithm and how it computes the shortest path and consequently finds a measure of similarity between the sequences.

Similarly, Figure 4.8 shows an example of comparing an actual seizure sequence to both another seizure and a non-seizure sequence. When a seizure sequence is compared to a non-seizure sequence, the warp path stays away from the diagonal and becomes longer which results in higher DTW distance. Whereas, the warping path is near the diagonal and hence shorter when a seizure is compared to another seizure sequence which results in lower DTW

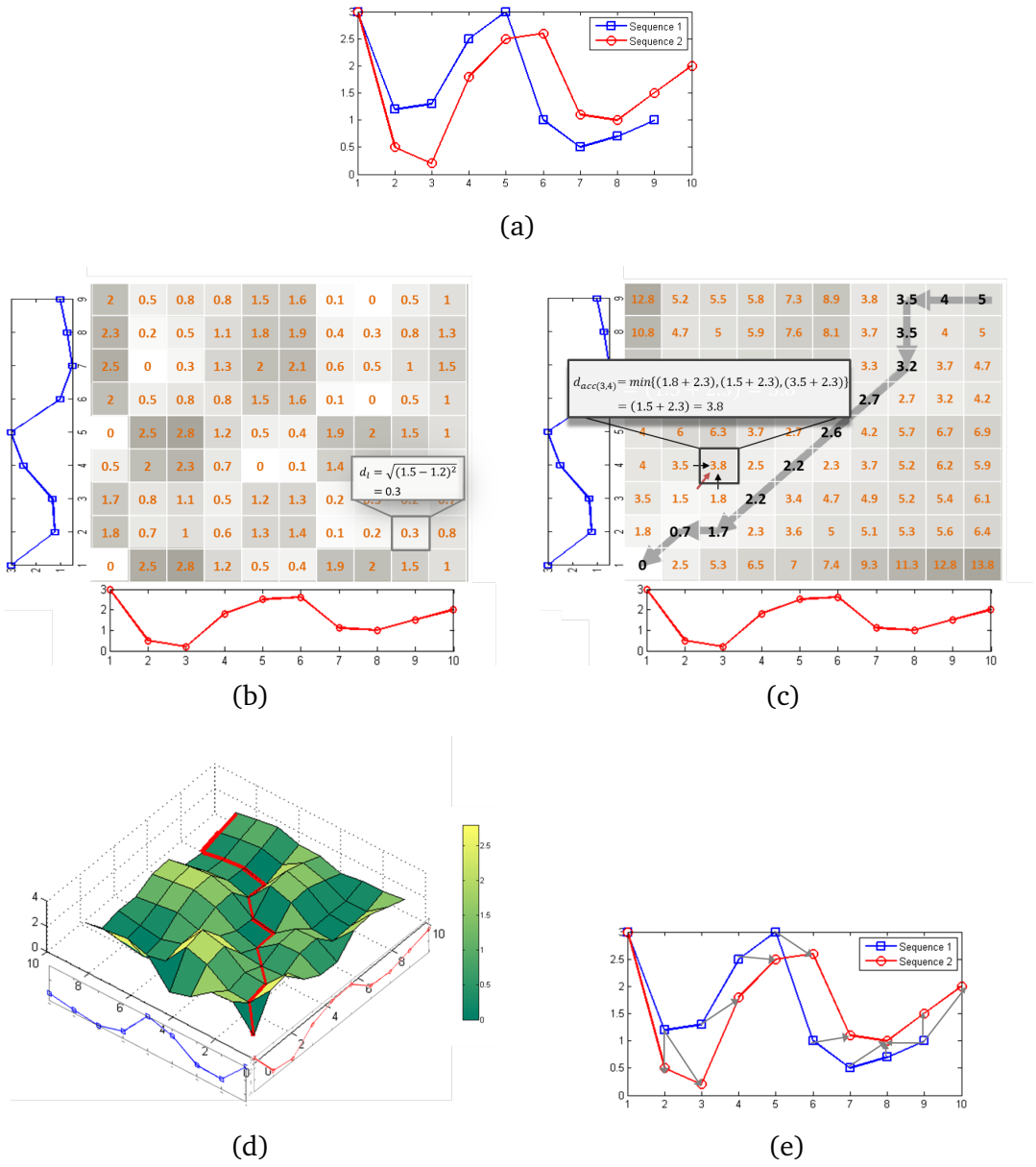


Figure 4.7: An illustration of DTW algorithm using a toy example of two variable length sequences. (a) Consider two variable length sequences with values $Sequence_1 = \{3, 1.2, 1.3, 2.5, 3, 1, 0.5, 0.7, 1\}$ and $Sequence_2 = \{3, 0.5, 0.2, 1.8, 2.5, 2.6, 1.1, 1, 1.5, 2\}$. (b) Gram matrix: is constructed by calculating euclidean distance between each datapoint of the two sequences. An example of local distance calculation is highlighted here for points $(9, 2)$. (c) Accumulated distance matrix: is computed by calculating the shortest accumulated distance to reach a point in the matrix. An example is highlighted at the point $(3, 4)$ where the shortest distance to reach this point is from the diagonal element (highlighted by red arrow). The DTW distance D_ϕ between two sequences is the last point in accumulated distance matrix, which is $(10, 9)$ in this case. The optimal path in this example is highlighted using grey coloured arrows. (d) A 3D view of the gram matrix. The red line represents the path found by the DTW algorithm. (e) DTW alignment of 2 variable length sequences.

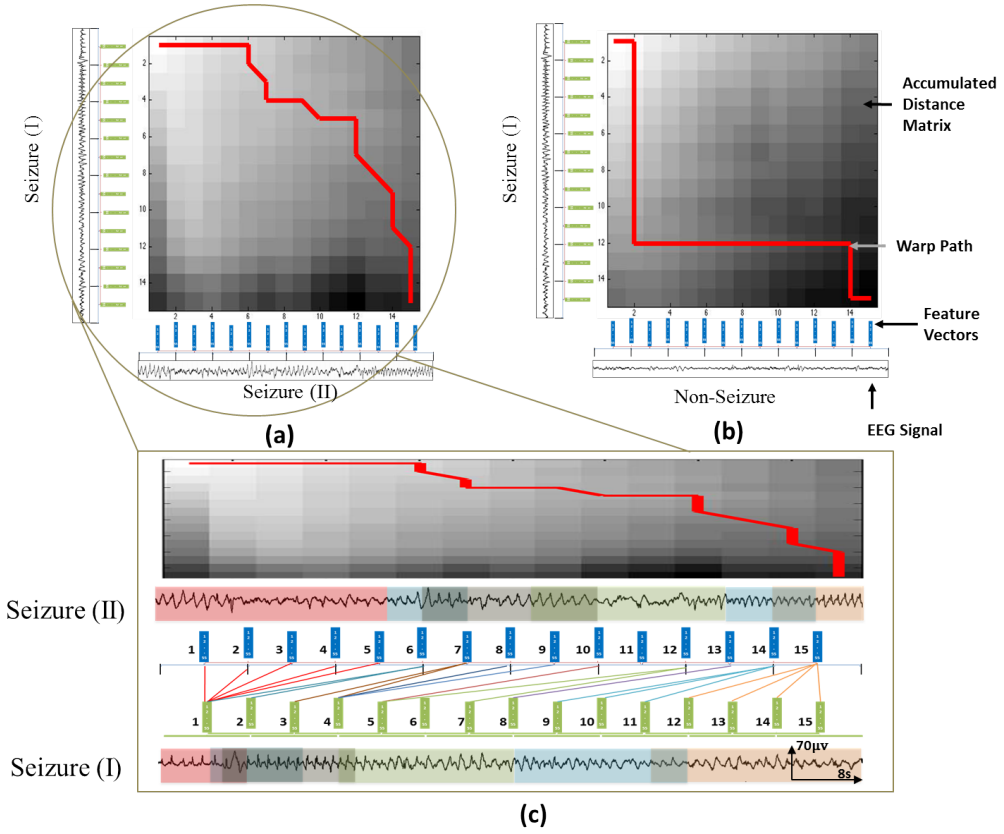


Figure 4.8: An illustration of the dynamic time warping process. Here sequences of 15 feature vectors (1 minute) of two different seizures and a non-seizure are compared. The red line indicates the warp path in the accumulated cost matrix. (a) seizure to seizure (b) seizure to non-seizure (c) A detailed view of the warping process between the two seizure sequences. The coloured parts of the each EEG signal shows the matched parts of the other sequence.

distance. It should be noted that warp path W continues to move in a straight direction until it finds a good match, in which case it starts to move diagonally. Figure 4.8c shows the detailed view of the sequence matching process using DTW. It can be seen that seizure sequence (II) is attenuated after first 8 seconds whereas seizure sequence (I) kept its repetitiveness and because of this mismatch the warp path moved in the straight direction for the first 6 blocks.

4.4.1 Gaussian DTW kernel in SVM

The classical SVM classifier uses a hyperplane to separate the input data. Given a two class problem, with a pre-labelled training set $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)$ where $y_i \in \{-1, +1\}$ and $x_i \in \Re$. In SVM classification, a test vector $\mathbf{u}$ is assigned a class y_i by evaluating

$$f_{svm}(u) = sign\left(\sum_{i \in I_{sv}} \alpha_i y_i \mathbb{k}(\mathbf{u}, \tilde{\mathbf{x}}_i) + b\right) \tag{4.3}$$

where α_i are Lagrange multipliers, b is the bias and $\tilde{x}_i (i \in I_{sv})$ are the n_{sv} retained support vectors. $\mathbb{k}$ is the kernel of the SVM that maps the input data into a higher dimensional feature space. A commonly used kernel function is the Gaussian radial basis kernel defined as

$$\mathbb{k}(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \tag{4.4}$$

If P and Q are two sequences, then a kernel function can be found by replacing the euclidean distance in Equation 4.4 by the DTW distance D_ϕ, (Bahlmann et al. 2002);

$$\mathbb{k}(S, R) = \exp\left(-\frac{D_\phi(S, R)}{2\sigma^2}\right) \tag{4.5}$$

In this manner, SVM with a GDTW kernel will be able to classify the variable length sequences according to their DTW distances. It should be noted that the kernel defined in 4.5 is not a Positive Semi Definite (PSD) kernel. However, the better performance provided by this kernel for contextual classification problems has enabled its widespread use in many pattern recognition areas.

4.5 Neonatal Seizure Detection System

The seizure detection system of Figure 6.4 is used in this work. This is an extension of the state of the art system (Temko et al. 2011a) presented in section 4.2. In this work the classifier for each channel can be a static-SVM, a GDTW-SVM or indeed a fusion of both static and dynamic classifier.

4.5.1 Pre-processing and Feature Extraction

The raw EEG is first down-sampled from 256Hz to 32Hz with an anti-aliasing filter set at 12.8Hz. Filtered EEG in each channel is then segmented into 8s epochs with a 50% overlap using a sliding window. A set of 55 different features is then extracted from each epoch. These features (Table 4.2) are derived from the frequency, time and information theory domains as described below.

- **Frequency domain features** For the frequency domain features the Power Spectrum Density (PSD) of each epoch is obtained using a 256 point Fast Fourier Transform (FFT). Thus an EEG epoch e_i is represented by a vector $\Omega_i = [\Omega(0), \Omega(1), \cdots, \Omega(\Lambda) \cdots \Omega(128)]$. Where Ω is the amplitude of a particular frequency Λ. A number of features are extracted from the PSD of the epoch. Total power represents the total energy in the full spectrum (0 - 12Hz) in the EEG epoch. Normalized and non-normalized power in the sub-bands of 2Hz range were also used as a feature. It was reported in (Thomas 2011) that power in the lower frequency bands was high during seizure activity. The power in the sub-bands normalized by the total power in the spectrum represents not only the influence of a particular frequency band but also the change in the overall spectrum. The peak frequency is the frequency with highest amplitude in the vector Ω_i. This feature has however been shown not to provide significant discrimination between seizure and non-seizure classes (Thomas et al. 2011). The spectral edge frequency is the frequency below which certain percentage of total power of the spectrum lies. Additionally, the EEG is decomposed into 8 coefficients using the Daubechies 4 wavelet, the energy in the 5th coefficient corresponding to 1-2Hz is used as a feature.

- **Time domain features:** Many time domain features were extracted by simple statistical analysis of each EEG epoch. Root Mean Squared (RMS) amplitude represents the average absolute amplitude of the epoch. The number of zero crossings is the number of times the EEG signal crosses zero (amplitude). The Hjorth parameters, proposed in (Hjorth 1970), quantifies the complexity of EEG signal in terms of the variance of the raw EEG epoch. These parameters are activity, mobility and complexity. Two features, skewness and kurtosis represent the shape of the distribution of data in EEG epoch. Non-linear energy quantifies the amplitude of the EEG signal along with the change in the amplitude in the given epoch. Line length for each EEG epoch is the sum of all absolute distances between the consecutive samples. The assumption with this feature is that line length of a non-seizure epoch will be less due to more low amplitude activity as compared to the seizure epoch. Some features were extracted from the first Δ and second derivative $\Delta\Delta$ of the EEG signal (variance, zero-crossings). The autoregressive (AR) modelling based features were extracted by splitting an EEG epoch into two halves. The parameters of AR model are estimated on the first half and then the fit of the estimated

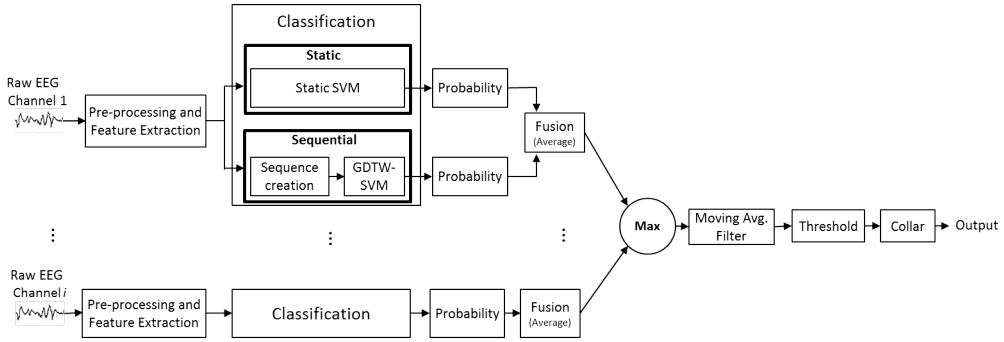


Figure 4.9: An overview of the complete proposed neonatal seizure detection system.

model is examined on the second half.

- **Information theory based features:** Features based on information theory were chosen based on an analysis of such features by (Faul et al. 2005). Entropy or specifically Shannon entropy is a measure of the amount of uncertainty of an outcome. It is obtained by computing the distribution/histogram of the EEG signal. This concept of entropy can be extended to other forms of information extracted from the EEG epoch, for example frequency spectrum, singular values of the embedding matrix using singular value decomposition algorithm and the Fisher information.

The usability of these features has been previously validated in number of previous studies on neonatal seizure detection (Gotman et al. 1997, Greene et al. 2008, Temko et al. 2011a, Thomas et al. 2010), adult seizure detection (Faul et al. 2009) and neurological outcome prediction (Doyle et al. 2010). This set of simple features provide a detailed picture of the signal statistics required for the neonatal seizure detection problem. The extracted feature vectors are then fed to the classification stage.

Further details about each feature is beyond the scope of this thesis and can be studied in (Thomas 2011).

4.5.2 Classification

In this study, three classifier configurations; GDTW-SVM only, Static SVM with RBF kernel only and lastly fusion of both classifiers have been investigated to get insight and compare performance obtained by each system. Here, only the working of the fusion based system will be described. The construction and working of the individual classifiers is the same, except their outputs are not fused after classification.

Table 4.2: Short-term features used in this study.

Domain	Features
Frequency	• Total power • Peak frequency • Spectral edge frequency • Power in 2 Hz width sub-bands((0-2Hz), (1-3Hz),. . . ,(10-12Hz)) • Normalized power in subbands • Wavelet energy
Time	• Non-linear line length • Number of maxima and minima • Root mean squared amplitude • Hjorth parameters • Zero crossings (raw epoch, Δ, $\Delta\Delta$) • Autoregressive modelling error (model order (1-9)) • Skewness • Kurtosis • Non-linear energy • Variance (Δ and $\Delta\Delta$)
Information Theory	• Shannon entropy • Singular value decomposition entropy • Fisher information • Spectral entropy

The fusion based classification uses two separate classifiers, a static-SVM based classifier and a sequential GDTW-SVM based classifier. The input stream of the feature vectors is duplicated to feed to each individual classifier. Each channel is classified separately. The static classifier processes a single feature vector at a time. A Gaussian RBF kernel is used inside the static-SVM classifier.

In the case of the sequential GDTW-SVM based classifier, sequences of 15 feature vectors are created. This corresponds to an EEG signal of 64s. These sequences are then classified by the sequential classifier. A shift of one epoch is used to make the next sequence. Hence the output of both classifiers remain synchronized to fuse them at a later stage.

4.5.3 Post-processing

The output of each classifier is in the form of SVM distance. These distances are then converted into posterior probabilities using a sigmoid function

$$P(y = 1|f_{svm}(\mathbf{x})) = \frac{1}{1 + \exp(Af_{svm}(\mathbf{x}) + B)} \tag{4.6}$$

Here d is the SVM distance to the separating hyperplane, A and B are the parameters of the sigmoid function estimated on the training dataset using the method described in (Platt 1999). The probabilistic outputs from both classifiers are fused together using a simple averaging function. The same process is repeated for all the channels. The effects of this step can be seen in Figure 4.10. The fused probabilistic outputs from each channel is then combined by applying a MAX operator. This process is similar to the way seizures are clinically annotated; if a seizure event is found in one channel then the whole epoch is marked as seizure. A moving average filter (MAF) is applied to this MAXed single probabilistic stream. The output from this stage is then compared to a threshold and then a binary decision is made i.e. 0: non-seizure and 1: seizure. To present the performance of the system the threshold is gradually varied from 0 to 1. If a seizure is detected, the decision is extended to either side using the collar technique similar to the one used in (Temko et al. 2011a).

4.5.4 Performance Assessment

In order to assess the performance of the proposed system on continuous recordings (Dataset II), a LOO patient cross validation was used. The classification model is created using the training data of 16 patients and the resulting

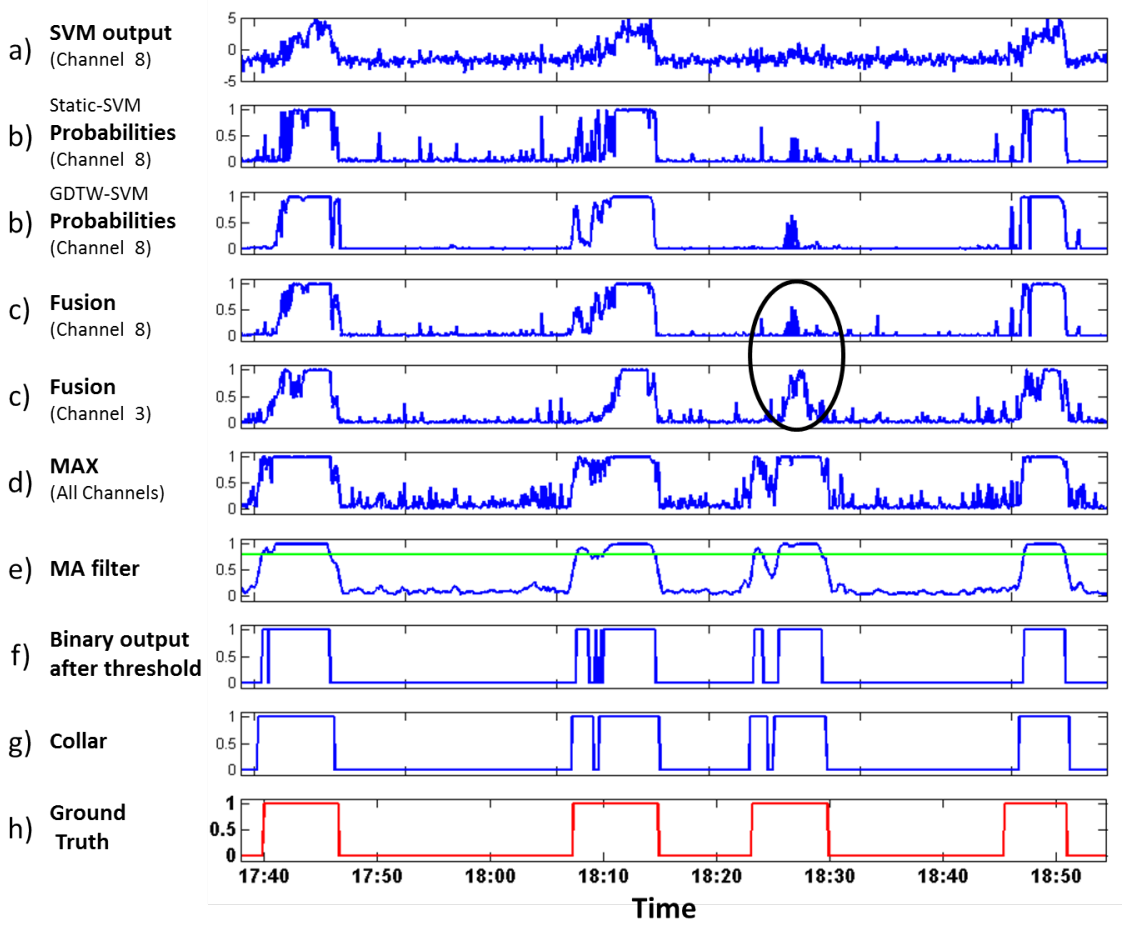


Figure 4.10: Effects of different post processing steps. (a) The raw output of the SVM classifier. (b) The raw outputs of Static-SVM and GDTW-SVM converted to probability using the sigmoid function. (c) Fusion of the probabilities of both classifiers in channel 8 and channel 3. It can be seen that the encircled seizure event is not completely present in channel 8. (d) Therefore a MAX operation is performed on all the channels. (e) The smoothed probabilities after a 9-tap moving average filter is applied. (f) The binary output resulting from applying a threshold of 0.8 to the filtered probabilities of seizure. (g) The collar operation is performed on the detected seizure events which increases the duration of all seizures on either sides by 7 epochs (h) The neurphysiologist annotations, where 1 indicates seizure.

model is then tested on the remaining one unseen patient's recording. It is well known that LOO provides the most unbiased assessment of the system performance to match real-life scenarios. For the purpose of reporting and comparing the performance of the system, two types of metrics are used: 1) epoch based sensitivity, specificity and precision, 2) Event based Good detection Rate (GDR) at the expense of False detections per hour (FD/h).

The binary decisions produced by the system are compared to the overall

global annotations (ground truth) and then can be assigned one of the four labels using the confusion matrix shown in Figure 4.11; True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN) as shown in Figure 4.12a. The epoch based metrics are then calculated as follows.

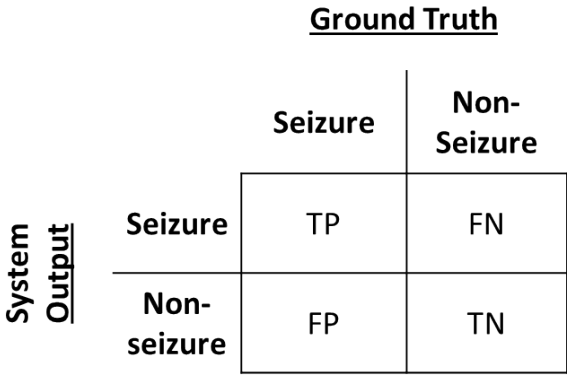


Figure 4.11: Confusion matrix.

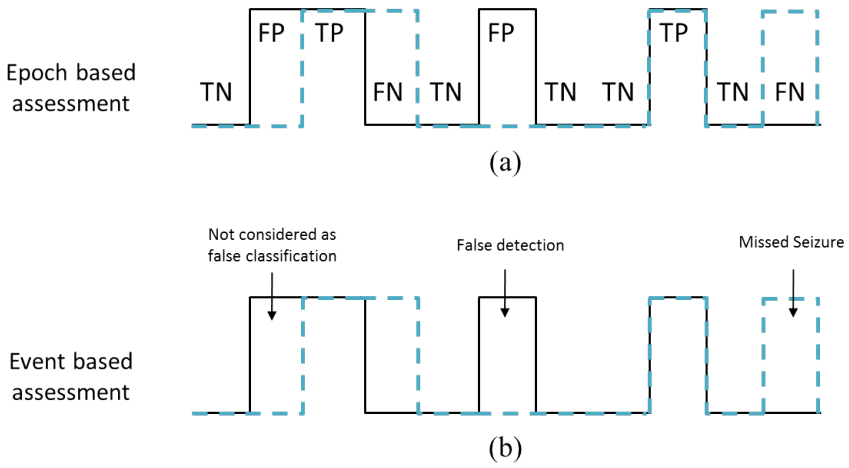


Figure 4.12: Method of calculating epoch and event based metrics.

Sensitivity is the percentage of correctly identified seizure epochs and can be represented as

$$Sesitivity = \frac{\sum TP}{\sum TP + \sum FN} \times 100\% \tag{4.7}$$

Specificity is the percentage of correctly detected non-seizure epochs.

$$Specificity = \frac{\sum TN}{\sum TN + \sum FP} \times 100\% \tag{4.8}$$

Precision is defined as the percentage of identified seizure epochs that are cor-

rect.

$$Precision = \frac{\sum TP}{\sum TP + \sum FP} \times 100\% \tag{4.9}$$

The area under the Receiver Operating Characteristic (ROC) curve is reported by computing the sensitivity and specificity at different thresholds applied on the final probability. The threshold is gradually increased from 0 to 1. Moreover, for the automated system to be of any clinical significance, the specificity needs to be very high. Therefore ROC90 is also reported, as adopted by our earlier studies (Temko et al. 2013). ROC90 is the area under the ROC curve where specificity is greater than 90%. A disadvantage of the ROC curve is its insensitivity to the amount of data in each class (Davis & Goadrich 2006) which is usually the case in seizure detection problem where non-seizure epochs are significantly greater in number as compared to seizure epochs. Therefore, the Precision-Recall/sensitivity (PR) curve is also reported which better represents the capability of a classifier to detect seizure events taking into account the amount of non-seizure data present in a particular recording.

The events based metrics allow for the assessment of the system's performance from the perspective of the number of detected individual seizure events. It could happen that system achieves a very high ROC areas because the testing data contains many longer seizures. However, it will not reflect on the system's ability to detect individual events, especially shorter seizure events. In this work the good detection rate is reported which is the percentage of detected seizure events at the cost of false detections per hour.

The same criteria of LOO patient cross validation was used in the case of Dataset I of short seizure and non-seizure sequences. This means that seizure and non-seizure sequences of the 15 patients were first used to train the system and a model was created. This trained model is then tested on the sequences of the remaining unseen patient. The output of the GDTW-SVM is a single probabilistic value for each sequence, whereas static-SVM produced 15 probabilistic values, one for each epoch which are then averaged to get a single probability value for each sequence. The ROC and PR curves are then obtained using the similar method of thresholding as mentioned above. As each probabilistic value represents an event itself, therefore no separate event based metric is reported in this case. All of the above mentioned configurations (non-seizure number, LOO and seizure sequence completion) were adopted to keep the testing procedure close to the real scenarios.

4.5.5 Model Selection

There are two SVM parameters that need to be tuned according to the nature of the data it is presented with; the kernel parameter σ and the generalization parameter C. The static classifier uses a radial basis function kernel inside the SVM and a five fold internal cross validation was performed to tune the parameters (Temko et al. 2011a). Data of 16 patients was used to train the classifier. To get the training examples of the seizure class, 2 minute seizure events from each patient were annotated along with the channel information.

In order to find the best parameters for the GDTW-based SVM a 3-fold cross validation is applied on the training data. It was empirically observed that the kernel parameter for GDTW-based SVM did not significantly change so a smaller number of folds were used for it. Sequences of seizure and non-seizure classes were created by concatenating 15 consecutive feature vectors. It should be noted that the sequentiality of feature vectors in a single sequence was not altered or randomized. Hence, each sequence of feature vector was representative of a 64s EEG signal in time. The sequence length was fixed for both seizure and non-seizure classes. In cases where a seizure was less than 15 feature vectors, then the neighbouring non-seizure epochs were padded on both sides. Moreover, the choice of a 15 feature vector length sequences was motivated by the fact, that the duration of most seizures is around 1 minute. Figure 4.13 shows the histogram of number of seizures and their respective lengths in our dataset. It can be seen clearly that most of the seizure's length is around 10 to 15 epochs.

4.5.6 Software Implementation

The above mentioned neonatal seizure detection system's main algorithm from feature extraction to post-processing was implemented in MATLAB. For classification, the SVM^{light} version 6.02 implementation was used which is available for free from *http://svmlight.joachims.org/*. This SVM library is written in the 'C' language by Thorsten Joachims, Cornell University (Joachims 1999). A small part of this library was rewritten in 'C' in this work to allow sequence classification using GDTW kernel.

A typical training cycle for Dataset 2 with data of 16 patients using static-SVM and RBF kernel took 2-3 hours whereas it took around 12 hours to complete the same training cycle with GDTW-based SVM. The bottleneck was the kernel computation time. The testing phase was dependant on the recording

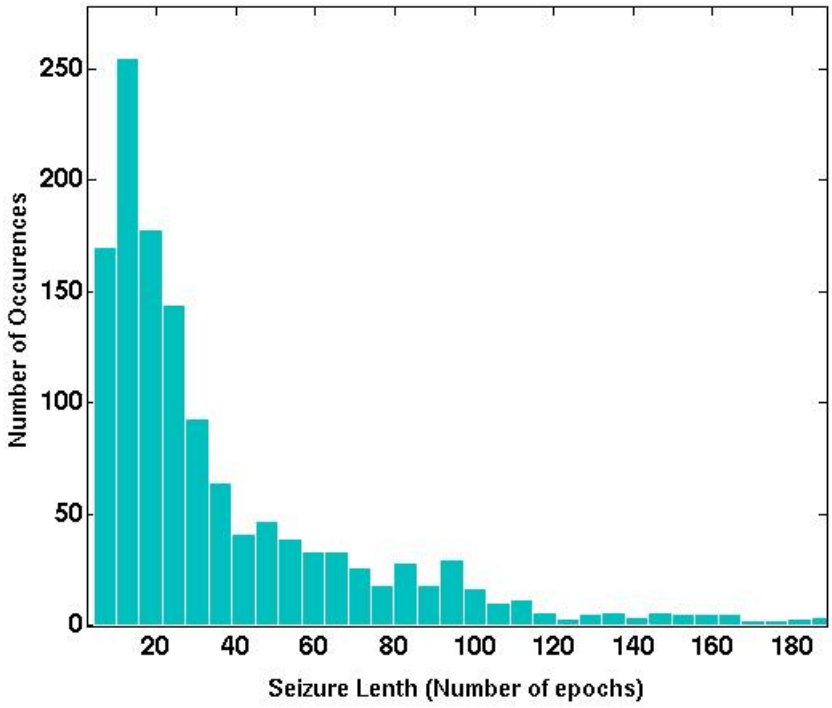


Figure 4.13: Histogram showing the length of seizures in the dataset.

length. The GDTW based classifier took an average 5 times more time during testing the unseen recordings. The machine used for these experiments was Intel Xeon E5345 with a processing power of 2.33GHz and 8GB of RAM memory.

4.6 Performance on Dataset 1 (Short seizures classification)

The main motivation of the developed system was to increase the detection rate of short seizures. Therefore the proposed system is first tested on the dataset of short seizures only.

The mean ROC and PR curves of static, sequential and fusion based classifiers are shown in Figure 4.14. It can be seen that both static and sequential classifiers performed differently on the ROC curve in Figure 4.14a. The GDTW-SVM classified more seizure sequences in the lower specificity region whereas the performance of the static classifier was better in the higher specificity region. This fact is quantified by ROC90 metric which is lower for GDTW-SVM (36.8%) as compared to static-SVM (41.1%). It is also evident from the PR curves that although the GDTW-SVM classified more seizure sequences, its pre-

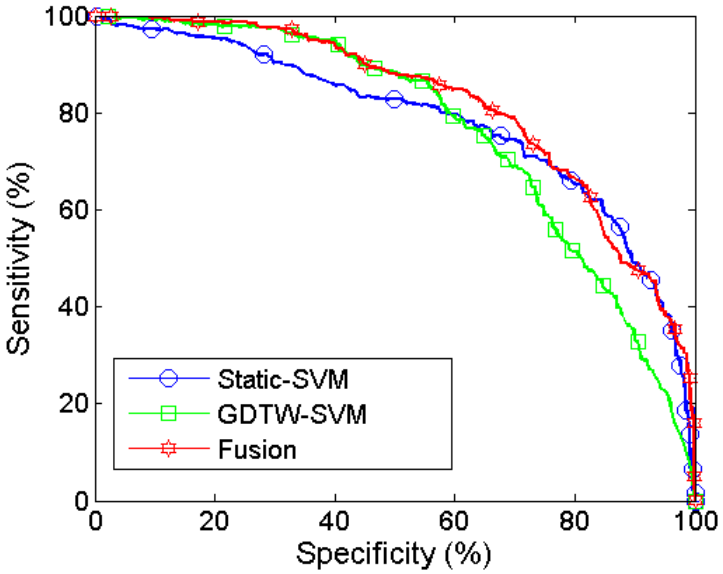
cision was however lower than the static-SVM and therefore is more prone to false alarms. Overall, GDTW-SVM achieved slightly higher ROC area of 81.9% as compared to 80.8% by the static-SVM. A significant increase of 14.2% in ROC90 and 5.32% in the ROC area is obtained over the individual static-SVM method, when both classifier's outputs are fused using the averaging operation. It can be seen that the output of the fusion based classifier takes advantage of the complementary behaviour and pushes the ROC curve towards the ideal point (100,100) on the ROC graph. Moreover, improvement in the precision at higher sensitivity values using the fusion approach can also be seen in the PR curves in Figure 4.14b. The results on the individual patients performance are shown in Table 4.3. The GDTW-SVM achieved higher mean ROC and PR areas for 9/16 and 8/16 patients respectively. It can be seen that no single classifier is better for all the patients. However, the fusion based system beats the individual classifiers performance for 11/16 patients. Furthermore, it increases the precision for most of the high density short seizure patients i.e. {3,7,12,16}.

From the above presented results it can be deduced that both static and sequential classifiers provide complementary information and it is clear from the results that fusion of both techniques increases the detection rate of the shorter seizures while keeping down the number of false alarms.

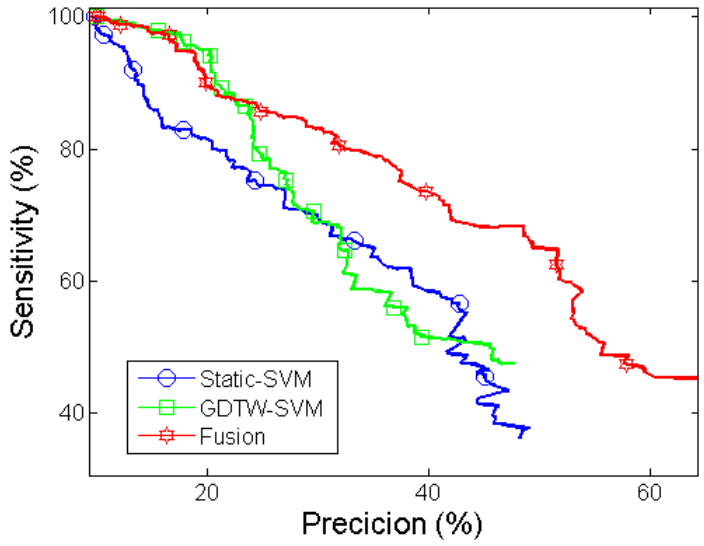
4.7 Performance on Dataset 2 (Continuous recordings)

In this section the performance of the classifiers obtained on the continuous recordings of 17 patients (as described in Table 4.1) is presented. The use of post processing methods on the probabilistic output of the classifiers does not allow us to observe the real behaviour of a classifier. Therefore, the raw performance of each classifier will be described first and then the performance of the proposed fusion based system with the post processing stage (MA filter and collar) will be presented.

The ROC curves obtained from the individual and the fusion of both classifier's raw outputs are shown in Figure 4.15. It can be seen that the GDTW-SVM outperforms the static-SVM in terms of epoch based metrics. The GDTW-SVM based system achieves an ROC90 area of 71.9% as compared to 69.8% attained by the static classifier. Moreover, the fusion of the raw probabilities yields the highest ROC90 area of 75.2%.



(a) ROC curves: The AUCs are Static-SVM=80.8%, GDTW-SVM=81.9% and Fusion=86.1%. ROC90 are Static-SVM=41.1%, GDTW-SVM=36.8% and Fusion=55.3%.



(b) PR curves: The AUCs are Static-SVM=47.6%, GDTW-SVM=46.3% and Fusion=59.7%.

Figure 4.14: ROC and PR curves of different classifier combinations for classifying the short seizures.

Table 4.3: Per patient results of different classifiers combinations for classifying the short seizures.

Pat. ID	# of Seizures	ROC Area (%)			PR Area (%)		
		SVM-static	GDTW-SVM	Fusion	SVM-static	GDTW-SVM	Fusion
1	5	87.1	88.4	**95.6**	51.7	27.0	**67.1**
3	83	86.9	80.9	**89.0**	50.7	52.0	**63.5**
4	10	95.9	91.1	**100**	57.8	82.4	**100**
5	7	63.5	**82.8**	79.8	**48.8**	13.7	40.6
6	16	91.1	95.5	**97.3**	90.9	72.1	**93.2**
7	42	94.4	97.5	**99.3**	82.6	61.8	**95.2**
8	2	100	97.2	**100**	87.5	100	**100**
9	32	**89.6**	71.8	84.8	16.8	**56.6**	40.3
10	14	57.9	**61.5**	61.3	13.8	20.8	**25.4**
11	16	80.5	92.0	**93.4**	43.5	28.4	**54.4**
12	46	86.8	87.7	**93.7**	51.9	50.8	**74.9**
13	5	84.1	91.2	**94.0**	62.3	23.6	**63.9**
14	3	40.7	**70.4**	67.9	**15.6**	7.7	13.3
15	45	**77.4**	59.5	70.3	12.2	**38.6**	21.5
16	57	78.6	73.5	**79.1**	32.5	35.6	**40.5**
17	9	**78.3**	69.4	72.7	42.2	**69.1**	61.3
Total	392						
Mean		80.8	81.9	**86.1**	47.6	46.3	**59.7**

* The underlined values shows the higher value among Static-SVM and GDTW-SVM. The bolded values show the highest value among Static-SVM, GDTW-SVM and fusion based classifier.

Figure 4.16 shows the PR curves of the classifiers. The GDTW-based system achieved higher PR area of 79.6% as compared to 77.9% of the static-SVM based system. Moreover, the fusion of the both classifier's probabilities provided a further increase of 3% in the PR area.

Performance of the classifiers using event based metric is presented in Figure 4.17. Although individually, GDTW-SVM based classifier performed better in epoch based metrics, it however did not achieve higher detection rate in event based metric. For example, given a threshold of 0.5 FD/h, the GDTW based classifier detected 69% of the total seizure events as compared to 73% detected by the static-SVM. However, the fusion of both classifiers significantly increased

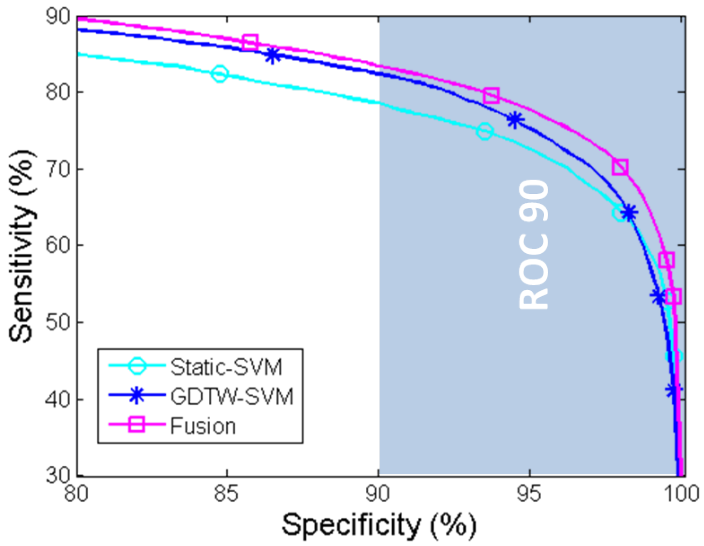


Figure 4.15: ROC curves with highlighted ROC90 area without post-processing. AUC: RBF-SVM=69.8%, GDTW-SVM=71.9%, Fusion=75.2%

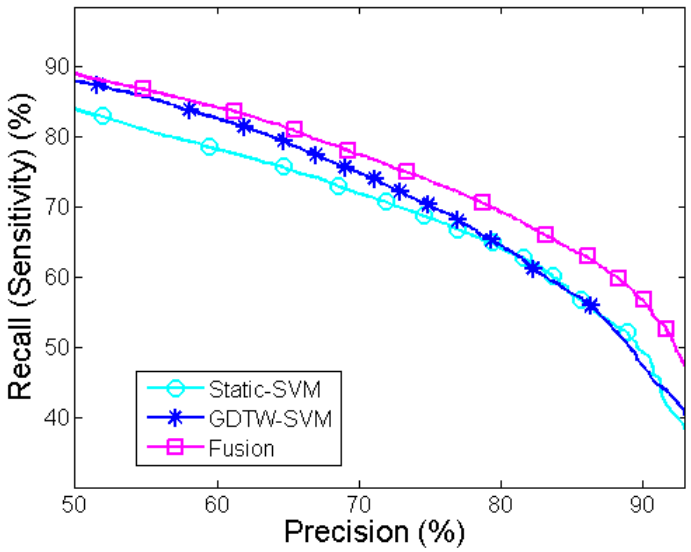


Figure 4.16: Precision-Recall curves without post-processing. AUC: RBF-SVM=77.9%, GDTW-SVM=79.6%, Fusion=82.7%

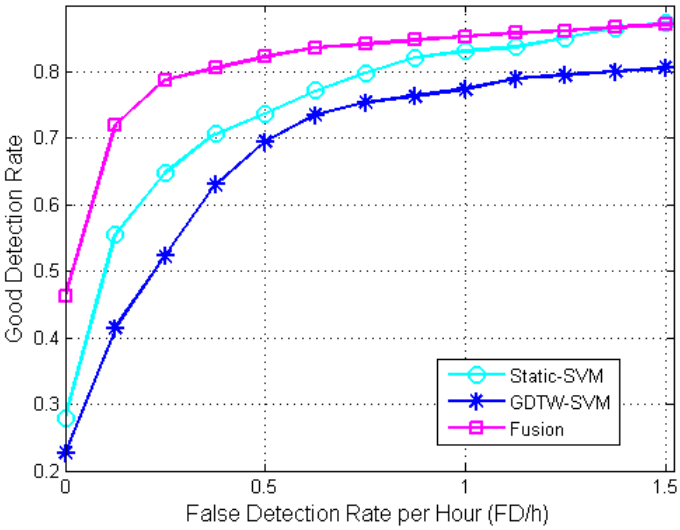


Figure 4.17: Good detection rate at the expense of false detections per hour of the system without post-processing.

the detection rate to 82% at 0.5 FD/h. This improvement using fusion of two classifiers indicates that although the detection rate of the GDTW-SVM classifier is low, it however detected events which were unique to the static SVM.

In order to get a better insight on this complementary behaviour, the detections of both classifiers were further examined. Table 4.4 presents the mean percentage of unique seizure events detected by the either classifier at 0.5 FD/h. The GDTW-SVM detected 58.5% of total seizure events, of which 15.5% were unique to the events detected by the static-SVM system. Whereas the static-SVM detected 68.2% of total seizure events, of which 25% were unique new event detections as compare to the GDTW-SVM. With the fusion of both systems, the total detection rate substantially increases to 83.2%. Moreover, the fusion system detected all the events that were produced uniquely by the individual classifiers.

Table 4.4: Percentage of detected events unique to other classifier.

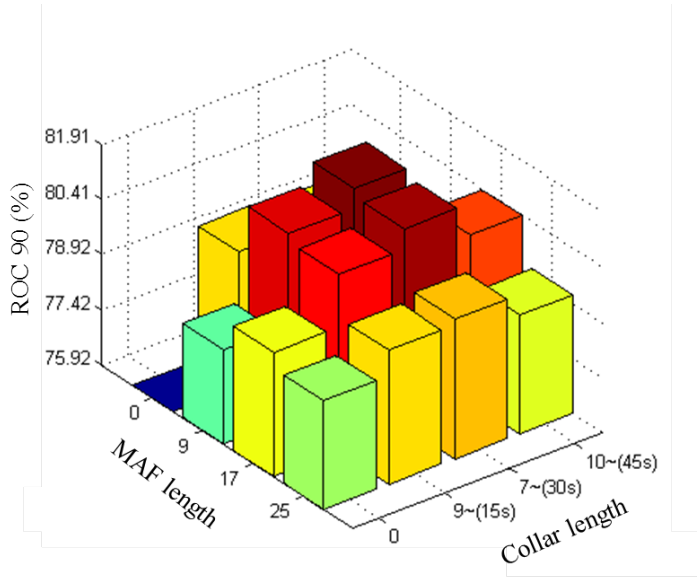
		Unique decisions to (%)	
	Detected events (%)	**Static-SVM**	**GDTW-SVM**
GDTW-SVM	58.5	15.4	0
Static-SVM	68.2	0	25
Fusion	83.2	17.4	24.9

* The threshold was set at 0.5FD/h

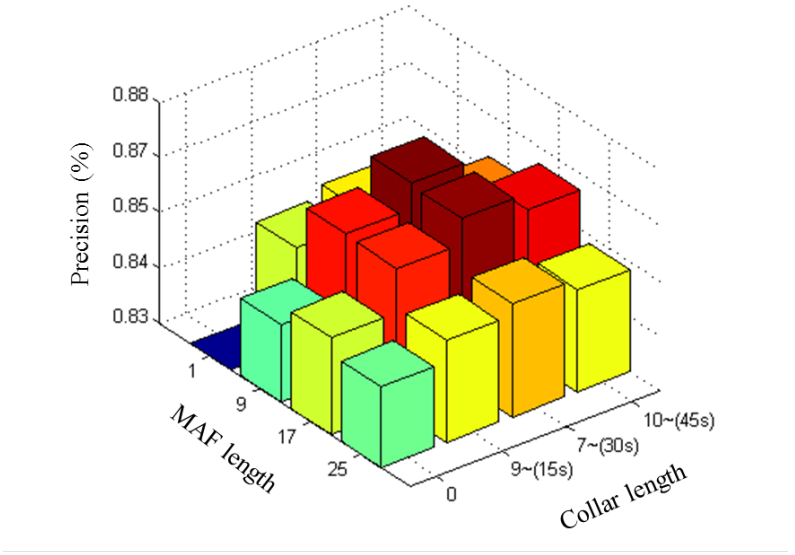
The results of the current system are also compared to the system using post processing stage as described in (Temko et al. 2011a). A MAF and collar was applied on the fusion based system's output. The length of MAF and collar for this system were chosen to maximize the ROC90 and PR areas. Figure 4.18 shows the effect of different values of MAF and collar on the system's performance. It can be seen that MAF of 9 epochs (40 seconds) and a collar of 7 epochs (32 seconds) resulted in best ROC90 and PR areas. For the static-SVM, a MAF of 15 epochs and a collar of 7 epochs was used. These parameters for static-SVM were chosen in (Temko et al. 2011b) and were reported to give best performance.

With the addition of a post-processing stage, both systems achieved an ROC90 area of 82.6%. The PR area for the fusion based system and the static-SVM with post-processing was 88.8% and 88.7% respectively. Although the presented fusion approach achieved similar performance to static-SVM as proposed in (Temko et al. 2011a) on epoch based metrics, it did however improve the event based metric and managed to achieve slightly higher detection rate at the cost of significantly lower false detections per hour. It can be seen in Figure 4.19 that the system detects 65% seizures at no false detections which is an 6% improvement from the previous system. The proposed system achieves 82.6% GDR at 0.25 FD/h which means it correctly detects 39 more seizure events as compared to the static-SVM based system of (Temko et al. 2011a). The proposed system correctly detects 90% of the seizure events at 1 FD/h. There is no significant improvement in GDR after 1 FD/h. It was argued in (Temko et al. 2011b) that the performance of event based metrics for any system can be easily improved by increasing the collar length. Therefore it is necessary to also present the Mean False Detection Duration (MFDD). As the GDTW based system processes a sequence of 15 epochs so it may be expected that MFDD of the proposed approach will be increased. However, it can be seen that the MFDD of the proposed approach remains similar to the system described in (Temko et al. 2011b).

Table 4.5 shows the comparison of the static-SVM and the fusion based system performances (with/without post-processing) for each iteration of the LOO's unseen patient. The GDTW-SVM based system outperforms the static-SVM in ROC90 and PR areas in 11/17 and 10/17 patients respectively when no post-processing was used. The fusion-based system without post processing gets higher ROC90 and PR areas for 12/17 and 13/17 patients. In terms of GDR, although the overall performance using the proposed system, is signifi-



(a) ROC90 area



(b) PR area

Figure 4.18: Effect of MAF and collar on ROC90 and PR area.

cantly improved however it is not consistent over all the patients. Specifically, 8/17 patient's GDR improves while the GDR of 6 patients was decreased. In particular, the GDR was poor in two patients (3,9) where the density of seizures was very high in the recordings. This could be due to the fact that GDTW-SVM based system is more conservative towards detecting seizures whereas its good performance in other patients may be attributed to its lower false alarm rate. Therefore if the presence of seizures is very high in a recording then the GDR

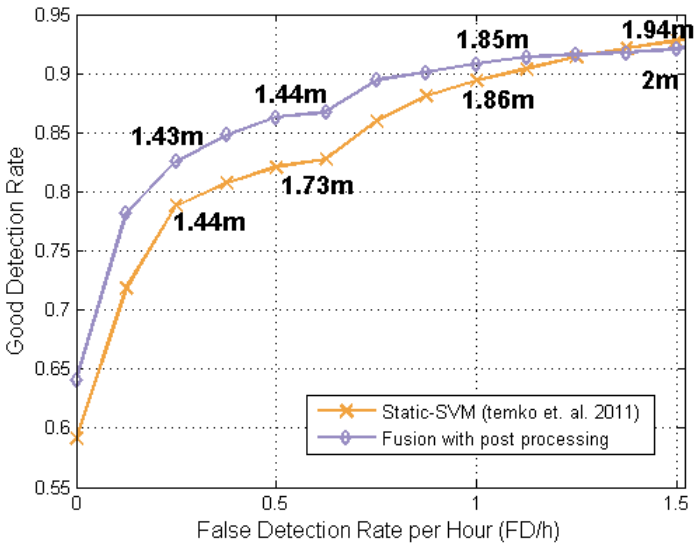


Figure 4.19: Good detection rate of the systems at the expense of false detections per hour with the post-processing stage. The text indicate the MFDD in minutes.

could be decreased using the proposed approach.

Overall the fusion based system had positive impact and the mean GDR was increased to 78.7% at 0.25 FD/h. Similarly, the statistics of per patient performance remains the same for the systems where post processing techniques were used.

As the main motivation of the proposed technique was to improve the detection of short seizure events therefore it is important to see how the system performed on the different lengths of seizures. Figure 4.20 shows a comparison of the percentage of detected seizures with/without fusion of two classifiers (using MAF and collar) at a stringent 0.25 FD/h. Seizures are categorized according to their lengths. It can be seen that the proposed fusion based system improved the detection rate in all categories. Most importantly, there was a 12% improvement in the detection of short seizures of length less than 1 minute is achieved which were previously being missed by the static-SVM based system. The proposed system also detected 9% more seizures of lengths between 1-2 minutes.

4.8 Conclusions

This chapter presented an approach to exploit the temporal and contextual information in the neonatal seizures and background EEG. A DTW based kernel

Table 4.5: Per patient comparison of static, sequential and fusion based systems with/without post-processing.

Pat. ID	# of Seiz	Raw Probabilities									With post-processing					
		ROC90 Area			PR Area			GDR @ 0.25 Fd/h			ROC90 Area		PR Area		GDR @ 0.25 Fd/h	
		S-SVM	G-SVM	S+G	S-SVM	G-SVM	S+G	S-SVM	G-SVM	S+G	S-SVM	S+G	S-SVM	S+G	S-SVM	S+G
1	17	40.9	52.0	**55.1**	25.7	35.7	**38.9**	52.9	58.8	**70.5**	**78.3**	73.3	**61.2**	59.1	58.8	**70.6**
2	3	75.3	78.0	**82.1**	51.7	55.9	**65.9**	8.6	56.2	**100**	**91.4**	89.6	79.6	**79.9**	100	100
3	170	72.9	65.5	**73.0**	88.9	86.7	**90.7**	**81.7**	5.5	68.8	80.5	**83.0**	**94.0**	93.9	61.2	**70.4**
4	65	66.2	68.2	**70.8**	64.8	67.6	**70.2**	48.9	73.8	**82.7**	**83.9**	82.7	**81.3**	80.8	**82.3**	80
5	51	51.8	55.0	**56.9**	74.7	77.3	**78.6**	70.5	**74.5**	70.6	64.3	**65.0**	84.1	**84.4**	78.4	78.4
6	44	59.3	**68.0**	66.0	71.7	77.4	**78.0**	41.5	68.2	**84.1**	63.8	**70.1**	77.2	**81.6**	88.6	**95.5**
7	71	69.7	68.4	**74.4**	82.8	82.3	**86.7**	67.3	24.1	**71.8**	72.3	**83.7**	86.7	**92.0**	40.8	**67.6**
8	17	68.1	68.9	**76.5**	68.1	68.2	**77.6**	67.0	22.5	**70.6**	87.3	**87.6**	**87.3**	86.9	70.6	**77.2**
9	157	78.6	77.3	**81.7**	95.9	95.9	**96.9**	**87.9**	78.4	86.6	**88.5**	88.0	**98.7**	98.6	94.9	94.9
10	28	**58.3**	48.9	56.1	**79.5**	72.0	78.0	**42.9**	32.1	39.3	**76.1**	65.7	**90.9**	85.5	35.7	35.7
11	14	83.8	86.3	**87.1**	91.4	93.8	93.8	92.9	51.6	92.9	92.0	**92.6**	96.6	**97.1**	92.9	92.9
12	37	69.8	70.1	**73.5**	78.4	79.7	**82.0**	**74.3**	34.8	71.2	80.5	**80.9**	88.0	**88.3**	75.7	75.7
13	25	76.8	83.1	83.1	85.7	90.6	**90.9**	29.1	58.2	**96.0**	87.4	**89.2**	93.8	**94.4**	96	96
14	13	78.9	**86.2**	86.1	92.2	**95.5**	95.4	92.3	76.9	92.3	88.6	**89.0**	96.6	**96.8**	100	100
15	58	77.4	69.6	**79.5**	86.6	81.8	**89.0**	**91.4**	21.5	79.3	**91.8**	89.2	**96.4**	95.0	**87.9**	86.2
16	31	86.3	89.6	**90.1**	98.4	98.4	**98.8**	96.8	96.8	96.8	**87.9**	84.5	**99.1**	98.9	96.8	**100**
17	20	72.5	**86.5**	85.9	87.7	94.2	94.2	55.0	55.0	**65.0**	**90.4**	90.2	**96.7**	96.4	65	**75**
Mean		69.8	71.9	**75.2**	77.9	79.6	**82.7**	64.8	52.3	**78.7**	82.6	82.6	88.7	**88.8**	78	**82.1**

' All the values are in percentages.

* S-SVM= Static-SVM, G-SVM=GDTW SVM and S+G= Fusion of static-SVM and GDTW-SVM

º The underlined values shows the higher value in comparison between S-SVM and G-SVM. The bolded values show the highest value among S-SVM, G-SVM and fusion of both.

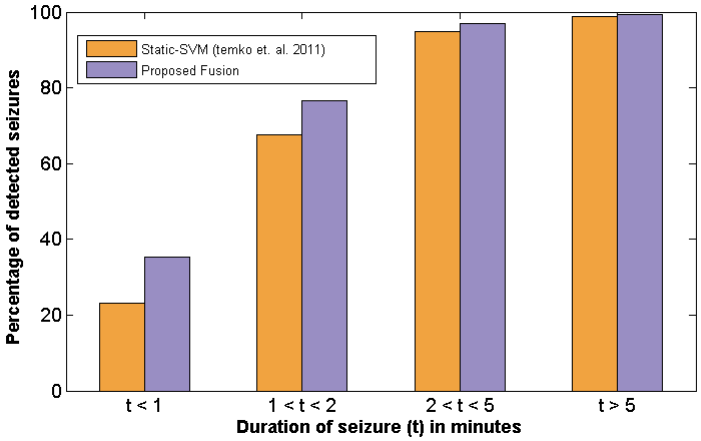


Figure 4.20: Percentage of seizures detected according to their lengths with FD/h < 0.25. Number of seizures in each group are 156, 223, 253, 189 respectively.

function is used which enabled the SVM to process the sequences of short term feature vectors extracted from 8s EEG epoch. A patient-independent neonatal seizure detection system based on the combination of a static (static-SVM) and a sequential (GDTW-SVM) classifier was thus proposed. Results indicate that the fusion of the output of both classifier leads to significant improvement in both epoch and event based metrics. The proposed fusion based system was also tested with a post-processing stage proposed in (Temko et al. 2011a). The performance was compared to the static-SVM with the same post-processing stage. The results show promising improvement in the detection rate of the seizures at significantly low false detections per hour. Particularly, the system increased the detection rate of short seizure events of less than 1 minute by 12%. However, it is expected that this improvement can be accentuated with better post-processing methods tailored to the proposed fusion based method.

References

Aarabi, A., Grebe, R., & Wallois, F. (2007). A multistage knowledge-based system for EEG seizure detection in newborn infants. *Clinical Neurophysiology*, 118(12):2781–2797.

Aarabi, A., Kazemi, K., Grebe, R., Moghaddam, H. A., & Wallois, F. (2009). Detection of EEG transients in neonates and older children using a system based on dynamic time-warping template matching and spatial dipole clustering. *NeuroImage*, 48(1):50–62.

Bahlmann, C., Haasdonk, B., & Burkhardt, H. (2002). Online handwriting recognition with support vector machines - a kernel approach. *In Proceedings of 8th International Workshop on Frontiers in Handwriting Recognition, 2002.* pp. 49–54.

Bogaarts, J., Gommer, E., Hilkman, D., van Kranen-Mastenbroek, V. H. J. M., & Reulen, J. P. H. (2014). EEG feature pre-processing for neonatal epileptic seizure detection. *Annals of Biomedical Engineering*, 42(11):2360–2368.

Chaovalitwongse, W. & Pardalos, P. (2008). On the time series support vector machine using dynamic time warping kernel for brain activity classification. *Cybernetics and Systems Analysis*, 44(1):125–138.

Davis, J. & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. *In Proceedings of the 23rd international conference on Machine learning*. ACM, pp. 233–240.

Deburchgraeve, W., Cherian, P., De Vos, M., Swarte, R., Blok, J., Visser, G. H., Govaert, P., & Van Huffel, S. (2008). Automated neonatal seizure detection mimicking a human observer reading EEG. *Clinical Neurophysiology*, 119(11):2447–2454.

Doyle, O., Temko, A., Murray, D., Lightbody, G., Marnane, W., & Boylan, G. (2010). Predicting the neurodevelopmental outcome in newborns with

hypoxic-ischaemic injury. *In Proceedings of IEEE Annual International Conference of Engineering in Medicine and Biology Society, (EMBC) 2010*. IEEE, pp. 1370–1373.

Faul, S., Boylan, G., Connolly, S., Marnane, W., & Lightbody, G. (2005). Chaos theory analysis of the newborn EEG - is it worth the wait? *In 2005 IEEE International Workshop on Intelligent Signal Processing*. pp. 381–386.

Faul, S., Temko, A., & Marnane, W. (2009). Age-independent seizure detection. *In Proceedings of IEEE Annual International Conference of Engineering in Medicine and Biology Society, (EMBC) 2009*, vol. 2009. pp. 6612–6615.

Gotman, J., Flanagan, D., Zhang, J., & Rosenblatt, B. (1997). Automatic seizure detection in the newborn: methods and initial evaluation. *Electroencephalography and Clinical Neurophysiology*, 103(3):356–362.

Greene, B. R., Faul, S., Marnane, W. P., Lightbody, G., Korotchikova, I., & Boylan, G. B. (2008). A comparison of quantitative EEG features for neonatal seizure detection. *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology*, 119(6):1248–1261.

Hjorth, B. (1970). EEG analysis based on time domain properties. *Electroencephalography and Clinical Neurophysiology*, 29(3):306 – 310.

Joachims, T. (1999). Making large-Scale SVM Learning Practical. *In* B. Schölkopf, C. Burges, & A. Smola, eds., *Advances in Kernel Methods - Support Vector Learning*. MIT Press, Cambridge, MA, pp. 169–184.

Liu, A., Hahn, J., Heldt, G., & Coen, R. (1992). Detection of neonatal seizures through computerized EEG analysis. *Electroencephalography and Clinical Neurophysiology*, 82(1):30–37.

Nagaraj, S. B., Stevenson, N. J., Marnane, W. P., Boylan, G. B., & Lightbody, G. (2014). Neonatal seizure detection using atomic decomposition with a novel dictionary. *IEEE Transactions on Biomedical Engineering*, 61(11):2724–2732.

Navakatikyan, M. A., Colditz, P. B., Burke, C. J., Inder, T. E., Richmond, J., & Williams, C. E. (2006). Seizure detection algorithm for neonates based on wave-sequence analysis. *Clinical Neurophysiology*, 117(6):1190–1203.

Ocak, H. (2009). Automatic detection of epileptic seizures in EEG using discrete wavelet transform and approximate entropy. *Expert Systems with Applications*, 36(2, Part 1):2027 – 2036.

Platt, J. C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *In Advances in Large Margin Classifiers*, Neural Information Processing Series. MIT Press, pp. 61–74.

Sakoe, H. & Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 26(1):43–49.

Shimodaira, H., Noma, K.-i., Nakai, M., & Sagayama, S. (2002). Dynamic time-alignment kernel in support vector machine. *In Advances in Neural Information Processing Systems 14*. MIT Press, pp. 921–928.

Shoeb, A. H. & Guttag, J. V. (2010). Application of machine learning to epileptic seizure detection. *In Proceedings of the 27th International Conference on Machine Learning (ICML-10)*. pp. 975–982.

Smith, N. & Gales, M. J. (2002). Using SVMs and discriminative models for speech recognition. *In Proceedinds of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2002*, vol. 1. IEEE, pp. 77–80.

Temko, A., Boylan, G., Marnane, W., & Lightbody, G. (2013). Robust neonatal EEG seizure detection through adaptive background modeling. *International Journal of Neural Systems*, 23(04):1350018.

Temko, A., Lightbody, G., Thomas, E. M., Boylan, G. B., & Marnane, W. (2012a). Instantaneous measure of EEG channel importance for improved patient-adaptive neonatal seizure detection. *IEEE Transactions on Biomedical Engineering*, 59(3):717–727.

Temko, A., Monte, E., & Nadeu, C. (2006). Comparison of sequence discriminant support vector machines for acoustic event classification. *In Proceedinds of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2006*, vol. 5. IEEE, pp. 721–724.

Temko, A., Stevenson, N., Marnane, W., Boylan, G., & Lightbody, G. (2012b). Inclusion of temporal priors for automated neonatal EEG classification. *Journal of neural engineering*, 9(4):046002.

Temko, A., Thomas, E., Marnane, W., Lightbody, G., & Boylan, G. (2011a). EEG-based neonatal seizure detection with support vector machines. *Clinical Neurophysiology*, 122(3):464–473.

Temko, A., Thomas, E., Marnane, W., Lightbody, G., & Boylan, G. (2011b). Performance assessment for EEG-based neonatal seizure detectors. *Clinical Neurophysiology*, 122(3):474–482.

Thomas, E. M. (2011). *A machine learning framework for neonatal seizure detection*. Ph.D. thesis, University College Cork, Cork, Ireland.

Thomas, E. M., Temko, A., Lightbody, G., Marnane, W. P., & Boylan, G. B. (2010). Gaussian mixture models for classification of neonatal seizures using EEG. *Physiological Measurement*, 31(7):1047.

Thomas, E. M., Temko, A., Lightbody, G., Marnane, W. P., & Boylan, G. B. (2011). Advances in automated neonatal seizure detection. *In New Advances in Intelligent Signal Processing*, vol. 372 of *Studies in Computational Intelligence*. Springer Berlin Heidelberg, pp. 93–113.

Wong, S., Gardner, A. B., Krieger, A. M., & Litt, B. (2007). A stochastic framework for evaluating seizure prediction algorithms using hidden Markov models. *Journal of Neurophysiology*, 97(3):2525–2532.

Chapter 5

Grading HIE Severity using Neonatal EEG

This chapter will describe the methods for the automated classification of the Hypoxic-Ischemic Encephalopathy (HIE) injury into four grades of severity using EEG. A cross disciplinary method is applied that uses the sequences of short-term features of EEG to grade an hour long recording. Specifically, it will present the supervector kernel approach employed for this task. Novel post-processing techniques are proposed based on majority voting and probabilistic methods. The proposed system is validated with one-hour-long EEG recordings from 54 full term neonates

5.1 Hypoxic-Ischemic Encephalopathy (HIE)

HIE is a common cause of neonatal death and long-term neurological disability with reported incidences of 3-5 per 1000 births (Volpe 2008). Perinatal asphyxia occurs when there is a lack of oxygen (hypoxia) and decreased blood supply (ischemia) to the neonatal brain around the time of birth (Berger & Garnier 1999). If this is prolonged, hypoxic-ischemic encephalopathy develops. The treatment involves cooling the infant to a body temperature of between 33-34°C for 72 hours without interruption (Azzopardi et al. 2009) - this is called therapeutic hypothermia. However to be effective, it must be commenced within 6 hours of delivery. In this narrow window of time the population of neonates who would benefit from treatment (those with moderate or severe encephalopathy) must be accurately identified.

In clinical practice, the EEG is visually analysed to grade the severity of HIE. Mostly, the HIE is classified into four main grades (Murray et al. 2009) as shown in Figure 5.1. Some of the main features inspected are the inter-burst-interval

(IBI), sleep-wake cycling, amplitude and the discontinuity of the background EEG (Boylan et al. 2008). However, grading HIE using the EEG requires the presence of a highly qualified neurophysiologist. This expertise is not widely available particularly around the clock in a typical busy Neonatal Intensive Care Unit (NICU), and hence an automated system for grading HIE could be of great help for medical staff.

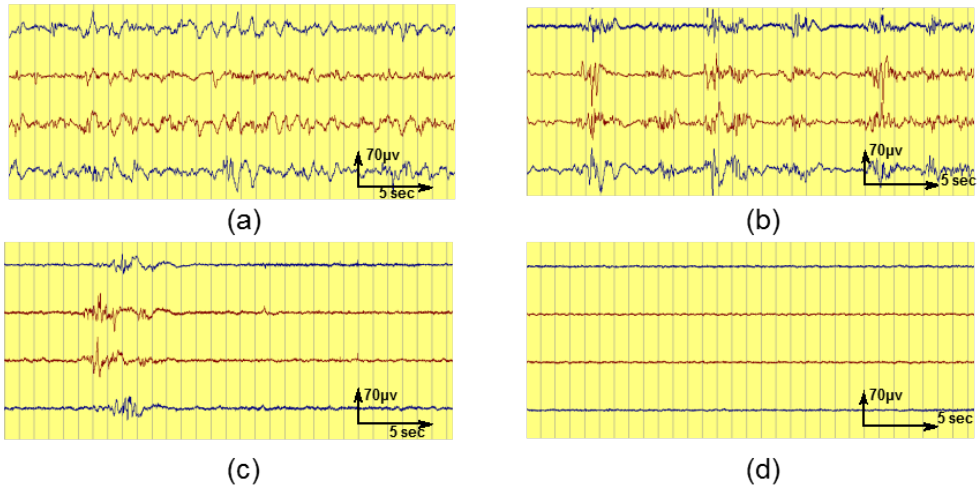


Figure 5.1: Ideal examples of EEG waveforms in 4 grades of HIE. (a) mild abnormalities (b) moderate abnormalities, (c) major abnormalities, (d) severe HIE /inactive

5.2 Related work and the basis of the proposed system

Automatic grading of HIE using EEG is a relatively new area. Some of the earlier studies, to assess the HIE grade using EEG, were focused on developing quantitative features. The goal of such studies was to identify, using statistical tests of significance, the key features that can either help to distinguish different grades of HIE or predict the outcome of HIE injury at a later age. Various methods were used to analyze these features such as linear regression analysis or classification trees (Ambalavanan et al. 2006, Korotchikova et al. 2011).

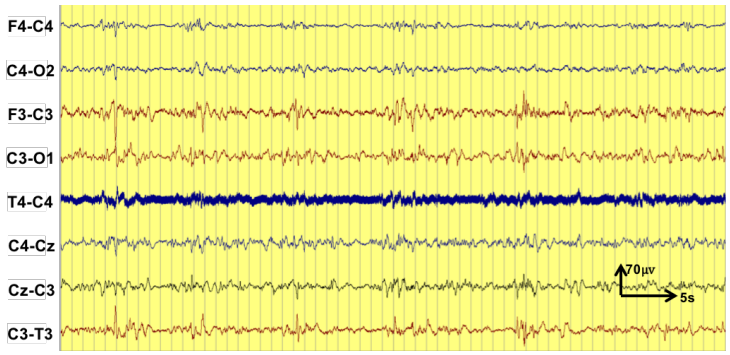
Some ideal examples of HIE grades is shown in Figure 5.1. As can be seen, the HIE-EEG exhibits various patterns, some of which may be similar across HIE grades; it is the inter-pattern variability or the way such patterns occur over the EEG recording which helps to characterize its grade. Moreover, these patterns are not always as obvious as shown in Figure 5.1. For example as shown in

Figure 5.2a, the amplitude of the EEG in channel F4-C4 is very low in a Grade 1 recording whereas channel T4-C4 is corrupted with artifacts. Grade 2 recording in Figure 5.2b does not have a clear burst suppression pattern. In Figure 5.2c, although there is a burst suppression pattern present, bursts are more stretched out. And lastly, a grade 4 recording in Figure 5.2d is modulated by a slow wave. Thus, grading the background EEG requires an amount of data that is large enough to capture the slower time scale components and the classification approach needs to be robust enough to comprehend different kinds of patterns appearing in the background EEG to discriminate between the HIE grades.

In a study by Stevenson et al. (2013), the EEG was first segmented into 64 second epochs. A non-linear amplitude modulated signal model was assumed to describe the EEG over this duration. Using time-frequency analysis, the EEG signal was decomposed into its amplitude modulated and instantaneous frequency components. Basic statistics (mean, standard deviation, skewness and kurtosis) of these components over the 64 second window were used as the key features to characterize the EEG. A multiclass linear discriminant classifier was then used to assign HIE grade to an one hour EEG recording.

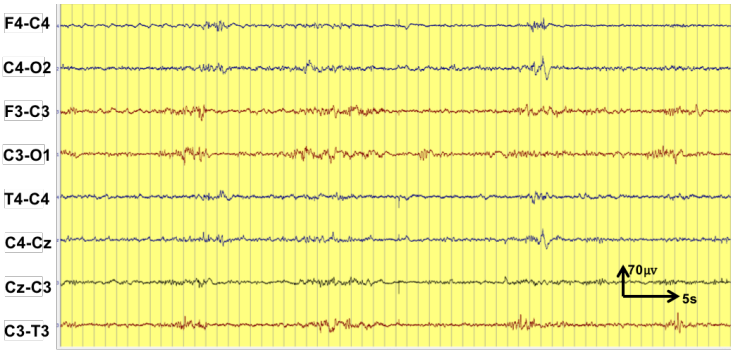
In a more recent study, Matic et al. (2014) proposed a tensor-based approach in which continuous EEG was first adaptively segmented and short-term quantized features were extracted. These features were then subsequently used to create a 3D model of a specific grade referred to as the tensor. Features extracted from this model were then fed to a multiclass classifier for classification.

The work presented in this study takes a different approach. First a set of short-term features are extracted from 8 second EEG epochs on which stationarity can be assumed. The sequence of short-term feature vectors is used to extract statistical model based long-term features. A multi-class classifier based on the support vector machine (SVM) is then used to exploit this information and classify these sequences into one of the four grades. Figure 5.3 shows an overview of this process. This approach combines a generative statistical modeling technique known as the Gaussian mixture model (GMM) with the discriminative SVM and is known as the supervector approach (Campbell et al. 2006). This method has previously shown promising results in many pattern recognition areas where similarly a decision has to be made over a long data segment and where the statistical feature distribution within a segment is more important than the sequentiality of the data (Hu et al. 2007, You et al. 2009, Zhuang et al. 2010).

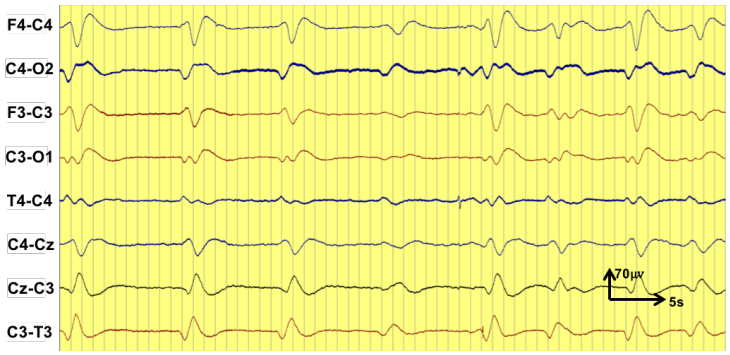


(a) Grade 1

(b) Grade 2



(c) Grade 3



(d) Grade 4

Figure 5.2: Representative examples of EEG of different grades of HIE

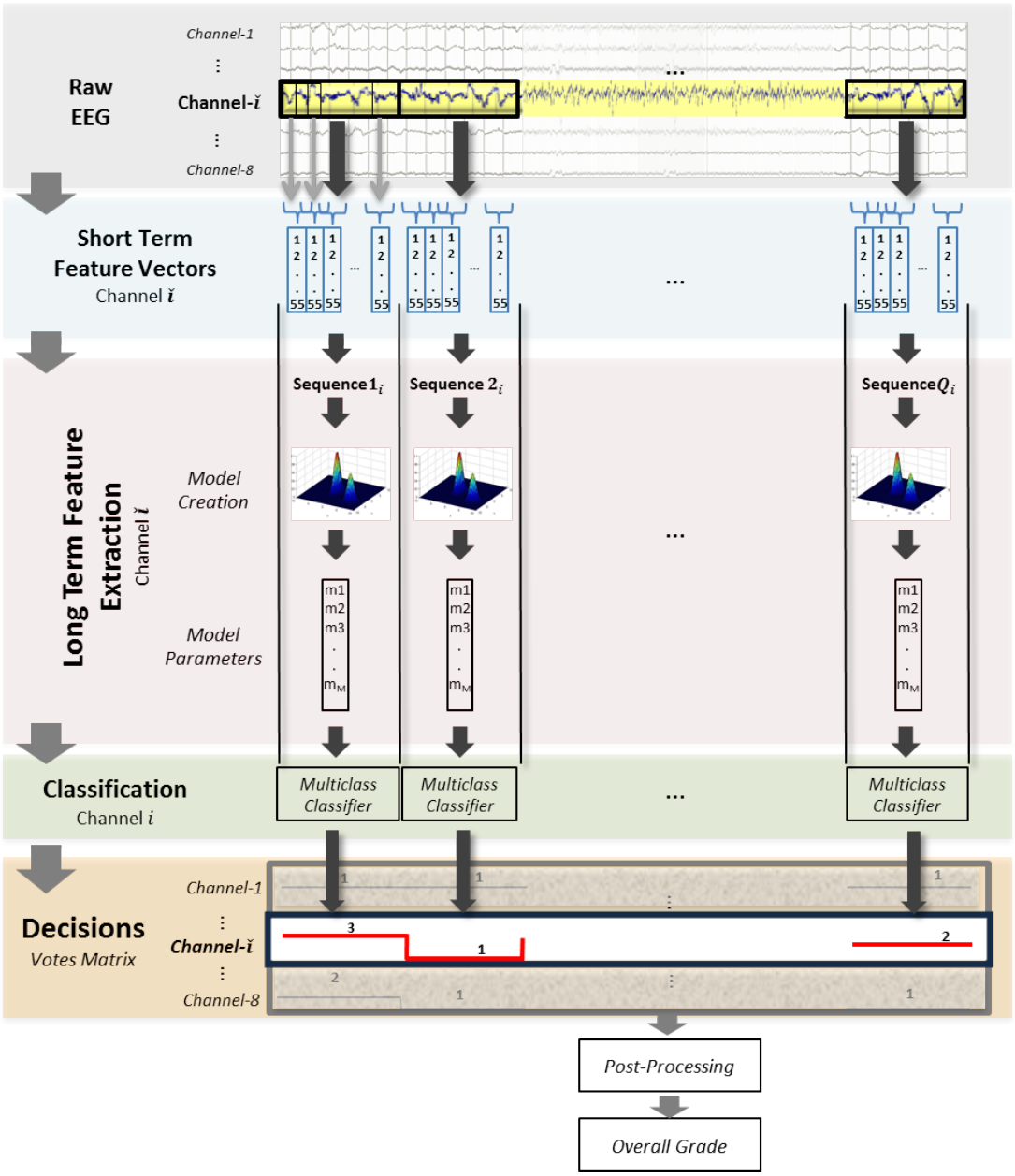


Figure 5.3: Overview of the GMM-supervector based HIE-EEG grading system

5.3 Dataset

Continuous video-EEG was recorded in 54 full term neonates using a NicoletOne EEG system in the NICU of Cork University Maternity Hospital, Cork, Ireland. The EEG recordings were started within 6 hours of birth and continued for 24-72 hours to monitor the evolution of the developing encephalopathy and for seizure surveillance. The neonates in this cohort were not treated with therapeutic hypothermia. The standard protocol for EEG recording in the NICU required the following 9 active electrodes: T4, T3, O1, O2, F4, F3, C4, C3,

Table 5.1: HIE Grades

Grade	Description
0	**Normal**: Continuous background pattern with normal physiologic features e.g. anterior slow waves
1	**Mild abnormalities**: Continuous background pattern with mild asymmetric patterns, mild voltage depression or poorly defined sleep wave cycles
2	**Moderate abnormalities**: Discontinuous activity with IBI $\leq$ 10s, no clear sleep wake cycling, clear asymmetry or asynchrony
3	**Major abnormalities**: Discontinuous activity with IBI 10-60s, severe fading background patterns no sleep wake cycling
4	**Inactive**: Background activity of $\leq 10\mu V$ or severe discontinuity of IBI $\geq$ 60s

and Cz. The following 8 EEG bipolar pairs were used to annotate the data: F4-C4, C4-O2, F3-C3, C3-O1, T4-C4, C4-Cz, Cz-C3 and C3 - T3. Segments of EEG data of approximately one hour in length were selected from each neonate at different time points in the EEG recordings. These segments had the continuous presence of a specific HIE grade and were free of seizures and major movement artefacts (amplitude higher than $250\mu V$ lasting 3 seconds). Two independent EEGers annotated each segment into one of the four HIE grades using the guidelines defined in (Murray et al. 2009) and summarized in Table 5.1. This grading criteria was specifically developed for full-term neonates with HIE. The same dataset has previously been used in (Stevenson et al. 2013) and thus a direct comparison of the results is possible.

5.4 Automated system for grading HIE severity

An overview of the complete system is shown in the Figure 5.3. The raw EEG is first pre-processed and segmented into epochs. Short-term features are extracted from these EEG epochs. The feature vectors are then combined into sequences. The statistical model-based long-term features are extracted from the sequences and then fed to the classifier. The decisions are post processed to form the final HIE grade.

5.4.1 Preprocessing and Short-Term Feature Extraction

The EEG from each channel is down-sampled from 256Hz to 32Hz with an anti-aliasing filter set at 12.8Hz. The down-sampled and filtered EEG is then segmented into 8s epochs with 50% overlap. A total of 55 short-term features outlined in Table 5.2, are extracted from each EEG epoch. This feature set provides a generic EEG description, which is computed from the frequency, time and information theory domains. Moreover, Table 5.2 also shows this features-set categorized into 3 overlapping groups depending on the type of information they capture from the EEG signal (spectral, structural and energy) following the approach previously developed in this research group (Faul et al. 2009).

The discriminative capability of these features for EEG has been discussed in previous work on neonatal seizure detection (Gotman et al. 1997, Greene et al. 2008, Temko et al. 2011, Thomas et al. 2010), adult seizure detection (Faul et al. 2009), and neurological outcome prediction (Doyle et al. 2010). The frequency domain features were limited up to 12 Hz as it was noted that there was very little power in the EEG signal in the frequency bands after 13 Hz. In a separate study (Stevenson et al. 2013), the features up to 64Hz were used and it was observed that this frequency range was prone to the effects of artifacts in the EEG signal which degraded the performance. Figure 5.4 shows the behavior of each feature given a segment of EEG from each grade. It can be seen that the frequency domain feature values gradually decrease from grade 1 to 4, indicating that the brain becomes less active as HIE injury develops to a higher grade. Similarly, time and information domain features also show significant change with changing grade of HIE. This set of simple features is therefore regarded as a generic descriptor of short EEG epochs.

5.4.2 Long-Term Feature Extraction

An overview of long-term feature extraction is shown in Figure 5.5. A generic statistical model of EEG background activity, referred to as the Universal Background Model (UBM), is first created from short-term features as a weighted mixture of Gaussians. The sequence of short-term feature vectors is then used to adapt the UBM parameters and a new model is created. The parameters of this new model, called the supervector, indicate how much the constituent Gaussian components of the UBM were shifted by this sequence. The supervector is then fed to the discriminative classifier. Details of the GMM, UBM and supervector creation are given below.

Table 5.2: Short-term features used in this study. There are three overlapping feature groups indicated by the superscript: Spectral [1], Energy [2], Structural [3]

Domain	Features
Frequency	• Total power[1,2] • Peak frequency[1] • Spectral edge frequency[1] • Power in 2 Hz width sub-bands((0-2Hz), (1-3Hz),... ,(10-12Hz))[1] • Normalized power in subbands[1] • Wavelet energy[1]
Time	• Non-linear line length[1,2] • Number of maxima and minima[1] • Root mean squared amplitude[2] • Hjorth parameters[2] • Zero crossings (raw epoch, Δ, $\Delta\Delta$)[1] • Autoregressive modelling error (model order (1-9))[3] • Skewness[3] • Kurtosis[3] • Non-linear energy[1,2] • Variance (Δ and $\Delta\Delta$)[1,2]
Information Theory	• Shanon entropy[3] • Singular value decomposition[3] • Fisher information[3] • Spectral entropy[1,3]

I. Gaussian Mixture Model:

The GMM is a probabilistic model which assumes that the data representing the feature space can be modelled using a finite number of Gaussian distributions (Reynolds 2009). It is represented as a Probability Density Function (PDF), which is a weighted sum of M Gaussian components

$$p(\mathbf{x}) = \sum_{j=1}^{M} w_j \, g(\mathbf{x}|\mathbf{m}_j, \mathbf{\Sigma}_j). \tag{5.1}$$

Here $\mathbf{x}$ is a feature vector of *n* dimensions, $w_j, j = 1, \ldots M$ are the mix-

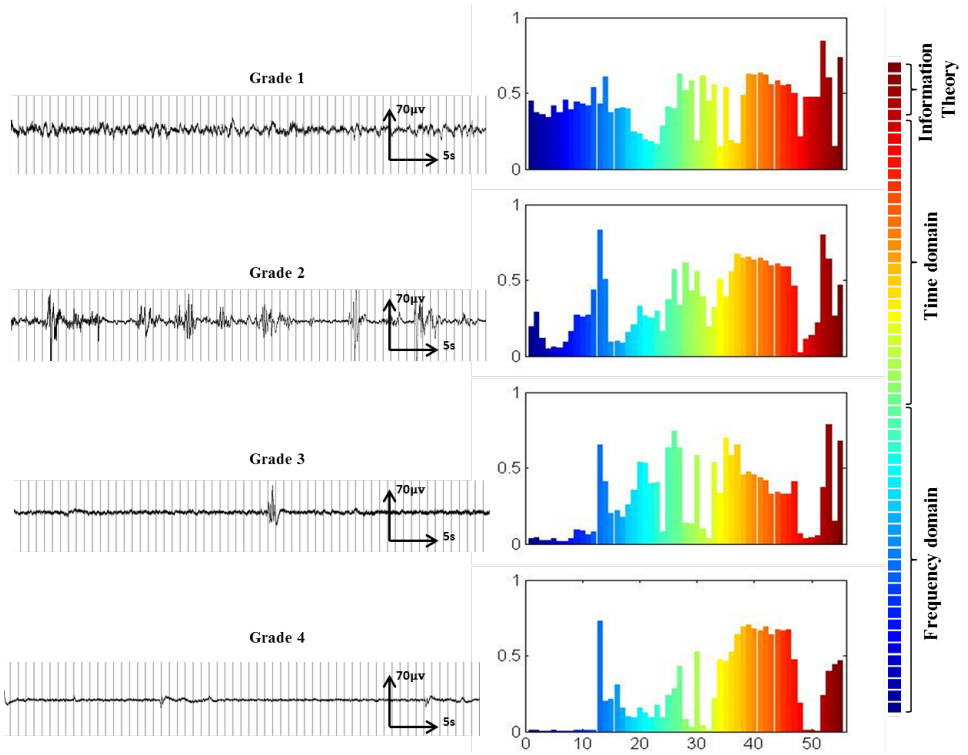


Figure 5.4: Examples of sequences of different grades of HIE EEG and their respective effect on the features from different domains. (For each example, the features from a sequence of 20 epochs (80 sec) averaged and then normazlized to 0-1 range.)

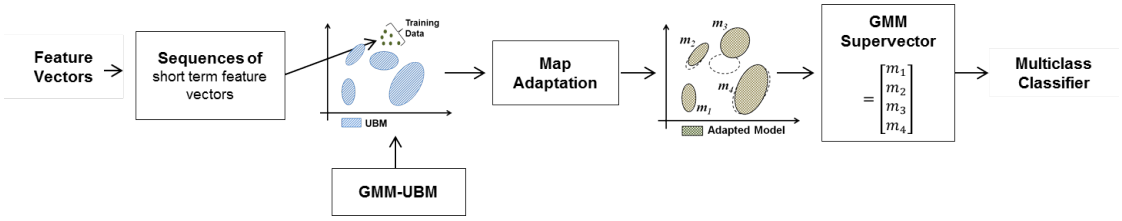


Figure 5.5: The process of creating supervectors from long-term sequences

ture weights and $g(\mathbf{x}|\mathbf{m_j}, \mathbf{\Sigma_j})$ are the component densities. Each Gaussian component is a n dimensional Gaussian function described as

$$g(\mathbf{x}|\mathbf{m}_j, \mathbf{\Sigma}_j) = \frac{1}{(2\pi)^{\frac{n}{2}}|\mathbf{\Sigma}_j|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{m_j})^T\mathbf{\Sigma}_j^{-1}(\mathbf{x}-\mathbf{m_j})\right) \quad (5.2)$$

where $\mathbf{m}_j$ is the mean and $\mathbf{\Sigma}_j$ is the covariance matrix of the j^{th} Gaussian component. During training, the parameters, $w_j, \mathbf{m}_j$ and $\mathbf{\Sigma}_j$ are optimised iteratively via the Expectation Maximization (EM) algorithm (Dempster et al. 1977) in order to maximize the log-likelihood of the model to the

input feature space. Figure 5.6 illustrates a probability density plot of a GMM of 4 components created using two features (AR modeling error and non-linear line length) over all available HIE-EEG data. The projection of the probability densities is shown below as a heatmap. The peaks show the Gaussian components of the GMM and the height of the peaks represent the weight of the component, which in turn represents the density of data of that component.

The GMM can be thought of as an advanced form of the k-means clustering method which also encompasses the information about the distribution of features using the covariance matrix.

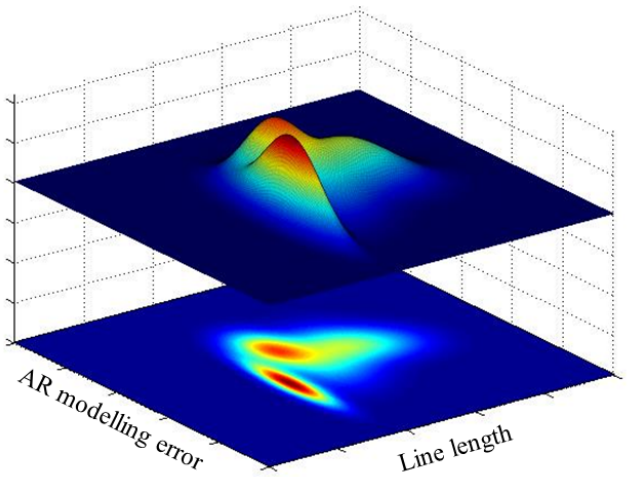


Figure 5.6: Probability density plot of a Gaussian mixture model of 4 components using 2 features (non-linear line length, AR modeling error (order 1)) over all HIE-EEG data.

II. The Universal Background Model (UBM):

Grading the HIE-EEG not only requires detecting a wide range of brain activities but also tracking the evolution of these activities in time. To achieve this, a general model called the UBM is created, to represent all the classes and background activities of the EEG without taking into account the class labels. The concept of the UBM evolved out of the speaker identification area where it was used to create a general model of human speech (Reynolds et al. 2000). Similarly it is used here to create a general model of neonatal EEG.

The UBM is generated by pooling the EEG data of all the grades and then training a GMM using the EM algorithm. The main advantage of using

a UBM is that this model represents the complete diversity of EEG activities across all HIE grades including artefacts that would not have been modelled otherwise. Specifically, the UBM allows for the control of the activity that would have been unseen if the model of a particular grade was trained directly using the data of this grade. Additionally, the UBM compensates for the lack of training data available for direct training of individual GMMs for each HIE grade.

The core of the developed automated HIE grading system is the UBM model. The short-term feature vectors used to create the UBM are standardized to have zero mean and unit variance. The Principal Component Analysis (PCA) transformation is applied to de-correlate the short-term features. PCA transforms the original set of features (possibly correlated) to another feature subspace of linearly uncorrelated variables called the principal components. The method works in a way that the maximum variance of the original space is along the first principal component. The next components will be orthogonal and have less variance to the ones preceding them. The consequent vectors are hence uncorrelated. This allows a diagonal covariance matrix to be used in the GMM. Both the de-correlation and dimensionality reduction are considered conventional techniques of feature preprocessing for the GMM as they simplify the PDF estimation problem (Reynolds et al. 2000).

Two configurations for the UBM creation were tested; number of Gaussian components {4, 8, 16, 32, 64} and percentage of energy retained of the original feature space after PCA transformation {97, 98, 99, 100}%. Best performance was obtained with 4 Gaussian components and 98% of the cumulative energy of the original space. It was observed empirically that using more Gaussian components did not improve the results due to the limited amount of data. Additionally, 98% of the cumulative energy of the original space is retained, reducing the original 55 dimensional feature space to an average of 23 dimensions.

III. UBM Adaptation, Supervectors and SVM kernel:

The long-term features are extracted through the process of UBM adaptation. Once the UBM is trained, any sequence of short-term feature vectors can then be converted to a vector of long-term features. The maximum a-posteriori (MAP) adaptation technique (Gauvain & Lee 1994, Reynolds 2009) is used to adapt the parameters of the UBM.

Given a UBM and a set of N feature vectors (sequence) $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$, a new GMM model is created by adapting the means m_j of the UBM by MAP adaptation. A GMM supervector $\mathbf{v} = [\hat{\mathbf{m}}_1, \cdots, \hat{\mathbf{m}}_M]^T$, is then created by concatenating all the means of the new adapted model.

The process of UBM adaptation is illustrated in Figure 5.7. The top plane in each sub plot shows the 2D heatmap of a UBM created from the data of all the four grades using two features, non-linear line length and the Auto-Regressive (AR) modeling error (order 1). The bottom plane shows the adapted model of each individual grade. It can be seen that the means (center vectors) of the constituent Gaussian component of the UBM move towards the distribution of the data of the particular grade. This information from the long EEG sequences can now be used by a discriminative classifier to identify the EEG grade. In order to make this information useable for a discriminative classifier, it should be transformed into a feature vector. The Supervector approach creates a long-term feature vector by concatenating the means of the constituent Gaussians (Figure 5.5). In this work SVM is used as a classifier to classify the supervectors.

SVM is a binary classifier. Consider a two class problem, with a pre-labelled training set of supervectors $\{(\mathbf{v}_1, y_1), \ldots, (\mathbf{v}_n, y_n)\}$ where $y_t \in \{-1, +1\}$. In SVM classification a test supervector v is classified by evaluating

$$d(\mathbf{v}) = \sum_{t \in I_{sv}} \alpha_t y_t \mathbb{k}(\mathbf{v}, \tilde{\mathbf{v}}_t) + b \tag{5.3}$$

Here there are n_{sv} retained support vectors from the training data; $\tilde{\mathbf{v}}_t$ each with weight α_t and associated target y_t, for $t \in I_{sv}$(the set of indices of the n_{sv} retained support vectors). The kernel $\mathbb{k}$ in the SVM can be used to measure the similarity between two supervectors that represent two EEG sequences. It is shown in (Campbell et al. 2006) that an inner product between two supervectors is an upper bound of the Kullback-Leibler divergence between the PDFs of the two sequences modelled by a mixture of Gaussians. Thus, the SVM kernel function which yields a measure of similarity between two supervectors $\mathbf{v}$ and the t^{th} retained support vector $\tilde{\mathbf{v}}_t$ from training data is,

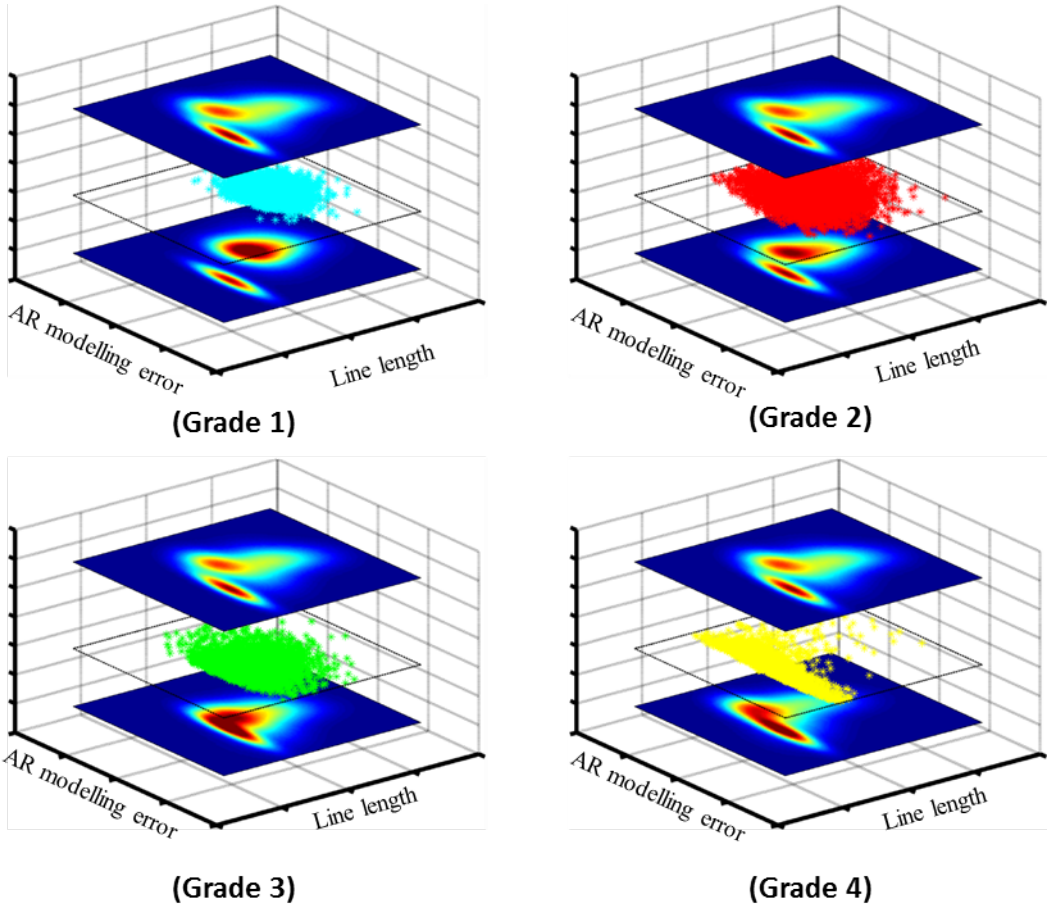


Figure 5.7: Illustration of GMM-UBM adaptation. Top planes show the GMM UBM, center plane shows the data of two features (Non-linear Line length, AR modeling error (order 1)) of each specific grade. The bottom plane shows the adapted UBM model.

$$
\begin{aligned}
K(\mathbf{v}, \tilde{\mathbf{v}}_t) &= \sum_{j=1}^{M} w_j (\hat{\mathbf{m}}_\mathbf{j})^T \boldsymbol{\Sigma}_j^{-1} \tilde{\mathbf{m}}_\mathbf{j}^\mathbf{t} \\
&= \sum_{j=1}^{M} \left(\sqrt{w_j} \boldsymbol{\Sigma}_j^{-\frac{1}{2}} \hat{\mathbf{m}}_j \right)^T \left(\sqrt{w_j} \boldsymbol{\Sigma}_j^{-\frac{1}{2}} \tilde{\mathbf{m}}_j^t \right) \\
&= \psi^T \tilde{\psi}
\end{aligned}
\tag{5.4}
$$

where

$$
\psi = \begin{bmatrix} \sqrt{w_1} \boldsymbol{\Sigma}_1^{-\frac{1}{2}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{w_M} \boldsymbol{\Sigma}_M^{-\frac{1}{2}} \end{bmatrix} \begin{bmatrix} \hat{m}_1 \\ \vdots \\ \hat{m}_M \end{bmatrix} = \Gamma v \tag{5.5}
$$

and

$$\tilde{\psi} = \Gamma \tilde{v}_i \tag{5.6}$$

It can be seen that equation (5.4) represents the dot product between the two supervectors scaled by a factor of Γ. Note that the scaling terms, weight w_j and covariance matrix $\mathbf{\Sigma}_j$, are the same for all sequences and can be computed beforehand. This will allow the use of a simple linear kernel inside the SVM. In this work, sequences of 20 short-term feature vectors without overlap are used to create one supervector. This corresponds to a span of 80 seconds of the EEG signal. A similar duration of EEG was chosen for feature extraction in (Stevenson et al. 2013).

5.4.3 Classification

Classification is a two stage process. In the first stage, a classifier model is created using the training data. The parameters of this model are iteratively tuned with an internal cross-validation routine. In the second stage the developed classifier is tested on the unseen testing data.

A multiclass-SVM is used in this work to classify HIE grades. There are many approaches reported in the literature to make the SVM a multiclass classifier (Hsu & Lin 2002). In this study, the one-against-one approach is used. A total of $\breve{k}(\breve{k}-1)/2$ SVM classifier models are required for $\breve{k}$ classes. Since there are 4 possible grades, therefore 6 separate 2-class SVM classifiers are created to classify between each pair of grades. Two-fold cross validation over the training data is used to find the regularization parameter C for the linear SVM which controls the amount of errors permitted during training. In the testing phase, the test supervectors are passed through the trained SVM model. Every channel is classified separately. The output of the multi-class SVM for a supervector $\mathbf{v}_{q\breve{i}}$ extracted from the q^{th} sequence of the $\breve{i}^{th}$ channel, is a 6-dimensional vector $\mathbf{Z}_{q\breve{i}} = [Z_{1\breve{i}} \dots Z_{6\breve{i}}]^T$, which contains the winning grades produced by each of the 6 classifiers as shown in Figure 5.9. The decisions of all the sequences from all the channels are stored in a decision matrix $\mathbf{L} \in \mathfrak{R}^{\breve{I} \times Q}$, where $\breve{I}$ is the number of channels and Q is the number of sequences in this channel.

5.4.4 Post-processing

I. One-Step Majority Voting:

Given a decision matrix **L**, an overall majority vote is then performed on

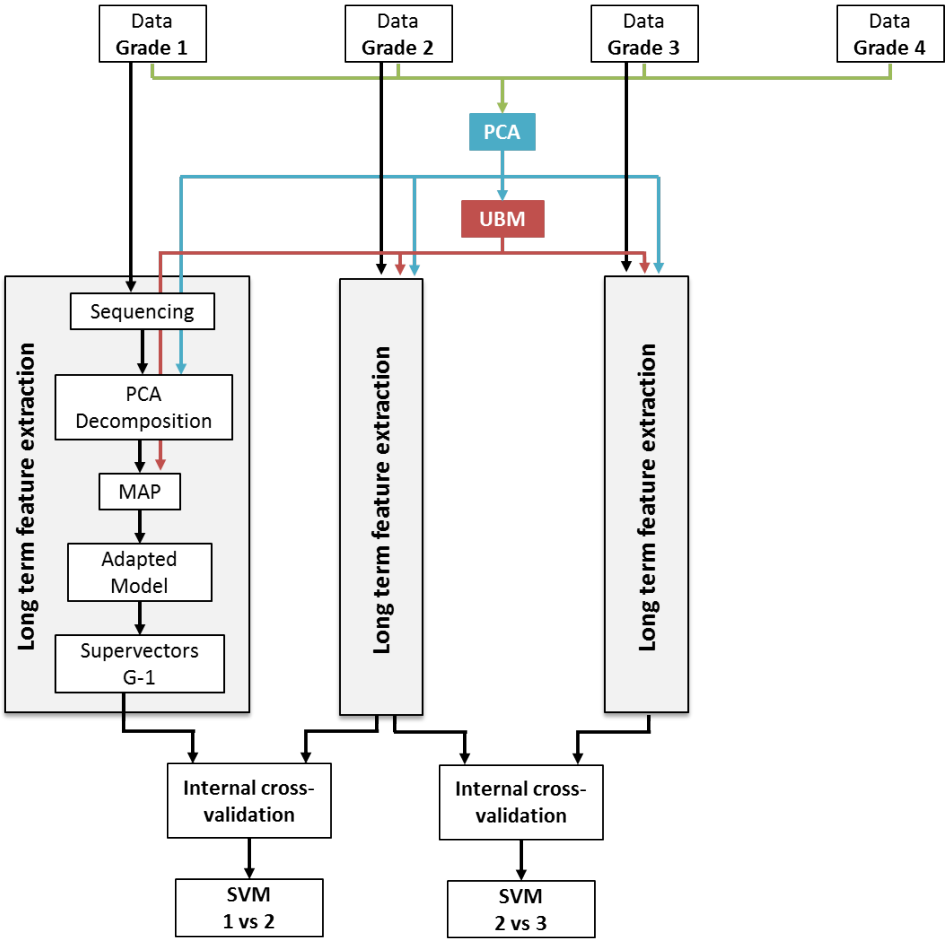


Figure 5.8: The whole process of training and creating SVM models. In this illustration, only two SVM models (1 vs 2) and (2 vs 3) are shown. However, other models can be similarly created.

this matrix across all sequences and all channels to determine the HIE grade for a complete EEG recording. This process is illustrated in Figure 5.9.

II. Two-step voting:

This method has two stages; in the first stage majority voting is used to identify two winning grades, $\breve{W}_1$ and $\breve{W}_2$. In the second stage only the decisions from the corresponding binary SVM classifier 'SVM-($\breve{W}_1$ vs $\breve{W}_2$)' are selected and the final majority voting is performed on them to make a decision. The advantage of the two-step voting is that the first step provides a general overview of the overall decisions on an hour-long EEG and then second stage validates the decision using the specific local expert classifier. This can be beneficial in case there is a tie between two classes in the first step or the difference in the number of votes of two grades is

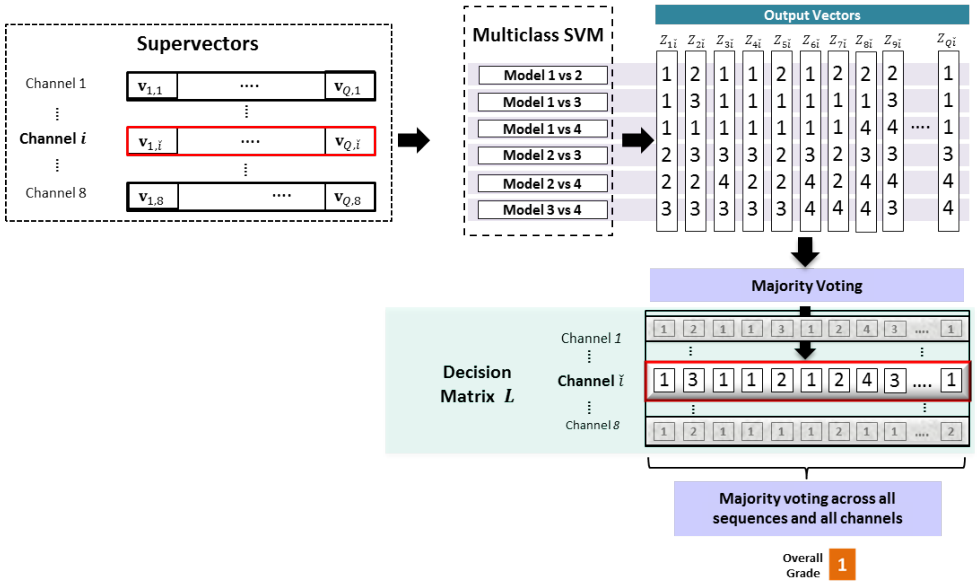


Figure 5.9: One-step majority voting for assigning an overall grade.

very low. An example of such a situation is shown in Figure 5.10. It can be seen that although grade 1 has a greater number of votes in 1^{st} step of voting but grade 2 wins when decisions of the expert classifier '***SVM 1 vs 2***' was used to make a final decision.

III. Probabilistic Methods:

As an alternative to making a binary decision, this study has also investigated the use of posterior probabilities derived from the SVM distances provided by each of the classifiers in the multiclass SVM. Probabilistic outputs are generally more intuitive to understand by the clinical staff and have been used in many previous EEG analysis systems (Temko et al. 2011). The probabilistic information is helpful in both assigning a grade to an individual sequence and creating a continuous probabilistic trace of the severity of the EEG recording to be used in the clinical environment. The posterior probability of every supervector's SVM distance can be obtained using a sigmoid function

$$P\Big(y = \breve{a}|f_{svm}(\mathbf{v}_{q\breve{i}})\Big) = \frac{1}{\Big(1 + \exp\Big(Af_{svm}(\mathbf{v}_{q\breve{i}}) + B\Big)\Big)} \tag{5.7}$$

where f_{svm} is the distance to the separating hyperplane, A and B are the parameters of the sigmoid function estimated on the training dataset using the method described in (Platt 1999). Parameters A and B will be

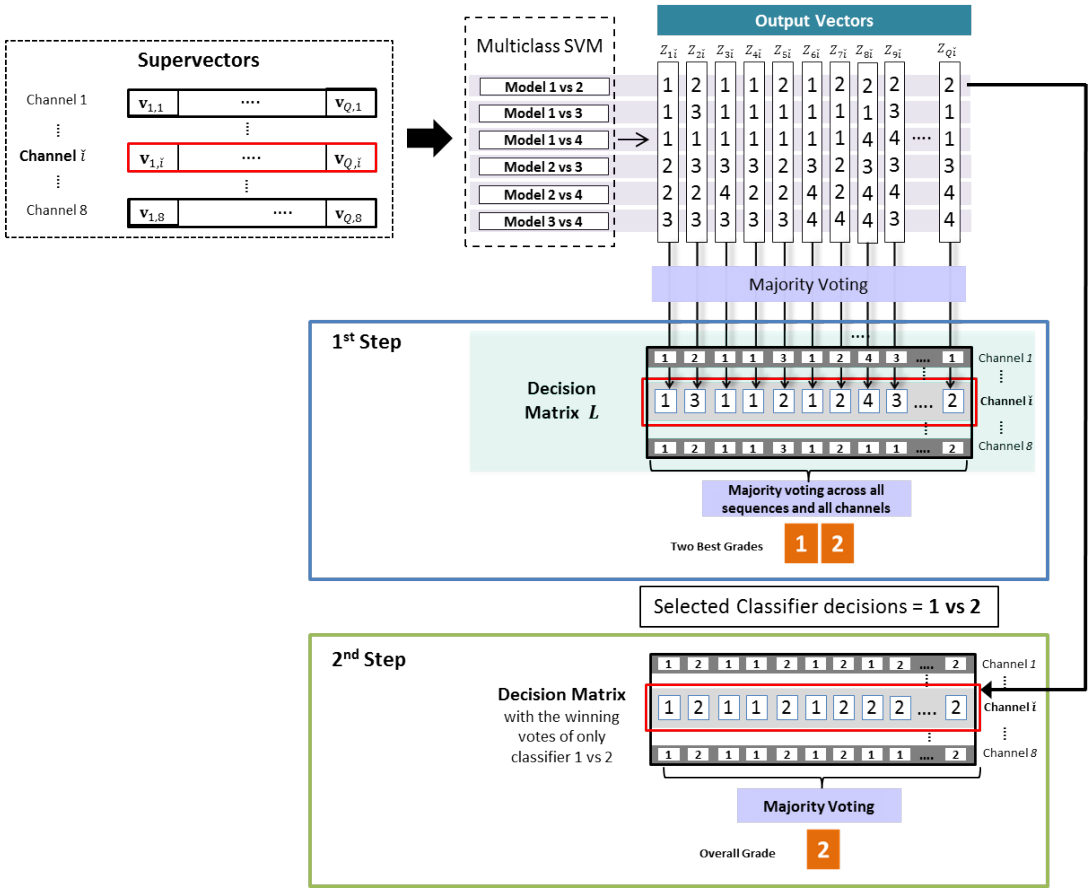


Figure 5.10: Two-step Majority Voting of assigning an overall grade.

different for each pair of SVM model. Figure 5.11 shows the process of probabilistic estimation for grading HIE. The pairwise probability $r_{\breve{a}\breve{b}}$ of a given supervector $\mathbf{v}_{q\breve{i}}$ from the model SVM($\breve{a}$ vs $\breve{b}$) can be obtained from Platt's sigmoid function in equation 5.7. Then using all the $r_{\breve{a}\breve{b}}$ provided from the 6 classifiers in the SVM multiclass classifier, the probability $U^{\breve{a}}_{q\breve{i}} = P(y = \breve{a}|\mathbf{v}_{q\breve{i}}), \breve{a} = 1, \ldots, \breve{k}$, of this supervector belonging to the grade a is calculated. Different methods have been proposed in the past to combine the pairwise probabilities $r_{\breve{a}\breve{b}}$ to obtain the individual class probabilities $U^{\breve{a}}_{q\breve{i}}$ (Wu et al. 2004). In this study, the method proposed by (Price et al. 1994) is used, as it gave the best results. According to this method the probability of a supervector $\mathbf{v}_{q\breve{i}}$ to be from the class $\breve{a}$ can be calculated as

$$U^{\breve{a}}_{q\breve{i}} = \frac{1}{\sum_{\breve{a}:\breve{a}\neq\breve{b}} \frac{1}{r_{\breve{a}\breve{b}}} - (\breve{k} - 2)} \tag{5.8}$$

In this case the output vector is 4-dimensional vector of probabilities $U_{q\breve{i}} = [U^1_{q\breve{i}} \ldots U^4_{q\breve{i}}]^T$, representing the probabilities of each of the four HIE grades

estimated from the supervector. These probabilities are averaged across all sequences and all the channels, to get the overall grade probabilities of each recording. The grade having the maximum probability, is assigned to the EEG recording.

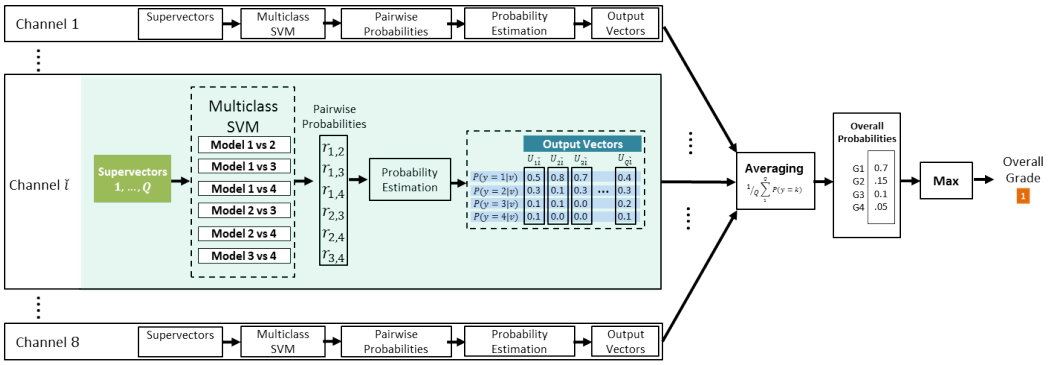


Figure 5.11: Probabilistic method of assigning the overall grade.

5.4.5 Performance Assessment

In this study, a Leave One Out (LOO) cross validation method was used to assess the performance of the developed system. Here the system was trained using the data from 53 recordings and the remaining one unseen recording used to test the system. This process was repeated until each recording has been used as a test subject. It is well known that the LOO routine produces an almost unbiased performance assessment of the developed system and the performance achieved with this system will be similar to that obtained over an unseen dataset of infinite size (Vapnik & Kotz 1982). An overall mean accuracy of the system was reported which is the ratio of correctly classified recordings to the total number of recordings in the dataset. It should be noted that for every iteration of the LOO validation, a new UBM is trained using the data from the 53 recordings. Moreover, in this process, PCA is first applied to reduce the dimension of the feature space, whilst retaining 98% of the energy. On average this leads to a reduction of the feature space from 55 to 23 (each iteration of the LOO may result in a different dimension reduction). This PCA transformation, determined over the pooled data, is fixed and then used in both the training (over the 53) and testing (on the left out file) of the supervector SVM. Therefore for each iteration of the LOO, the length of the supervector is constant, and does not depend on the grade.

5.4.6 Software Implementation

The above mentioned automated HIE grading system's main algorithm was implemented in MATLAB from feature extraction to post-processing. For classification, the SVM^{perf} version 3.0 was used which is available free from *https://www.cs.cornell.edu/people/tj/svm_light/svm_perf.html*. Apart from other performance enhancements as compared to SVM^{light} this version of SVM allows faster computation using the linear kernel. This SVM library was written in the 'C' language by Thorsten Joachims, Cornell University (Joachims 2005).

A typical training cycle for training the system using data of 53 recordings took around one hour whereas testing the unseen recording claimed approximately 10-15 minutes. The machine used for these experiments was Intel Xeon E5345 with a processing power of 2.33GHz and 8GB of RAM memory.

5.5 Results

5.5.1 Comparison with other methods

The comparison of the performance achieved by different classification approaches is presented in Table 5.3. The best overall accuracy is 87% attained by the proposed system. The results obtained are better than the 77.8% reported in (Stevenson et al. 2013). Moreover, in order to see the improvement in the performance of grading the HIE-EEG using the proposed system, a comparison was also made with a simple (PCA+GMM) classifier which used the same parameters and a simple SVM classifier with RBF kernel. It can be seen that accuracy drops to 61% with the simple GMM method and 81% with SVM alone. This indicates that the proposed combination of generative and discriminative techniques can significantly improve the accuracy.

More recently an accuracy of 89% of classifying three HIE grades (without grade 4) has been reported by (Matic et al. 2014) on a dataset of 34 neonates. These authors have similarly extracted long-term features from a long EEG segment. In contrast to our work they combined the information from the one hour EEG segment at the feature level. The increased accuracy is due to larger population (47% as compared to 33% in our study) of grade 3 and 4 recordings which are the easiest to classify for any system. Moreover, the authors combined the grade 3 and 4 recordings to one group (grade 3).

Table 5.4 presents the confusion matrix for the best obtained accuracy. It can be seen that 7 out of 54 recordings were misclassified. Most confusion occurred

Table 5.3: Comparison of the proposed method with the other techniques

Method	**(Stevenson et al. 2013)**	**GMM alone**	**SVM alone**	**Proposed Method (GMM supervector)**	**(Matic et al. 2014)***
Overall Accuracy	77.8%	61%	81%	87%	89%

* Separate dataset of 34 neonates

between grades 1 and 2 with four grade 2 recordings wrongly classified as grade 1 and one grade 1 recording misclassified as grade 2. A higher accuracy of 76% for grade 2 recordings was reported by Matic et al. (2014), compared to 71% reported in this study (Table 5). However, the accuracy of classifying mild encephalopathy (grade 1) was found to be higher in our study, 95% compared to 91%, presented by (Matic et al. 2014). The major cause of the decreased accuracy of the system for classifying these grades is the similar morphology of the EEG signals in both grades, as can be seen in Figure. 5.1. Two recordings from grade 3 were misclassified as grade 2. All recordings of grade 4 were correctly classified.

The last row of the confusion matrix in Table 5.4 shows the precision of the system, which is defined as the ratio between the number of correctly assigned decisions and the number of total decisions assigned to a specific grade. This metric is important to see how much the system confuses a specific grade given the data of other grades. It can be seen that the system is 100% precise for classifying the data of grades 3 and 4, which essentially means that if the test data is from grade 1 and 2 then the system will never classify it as grade 3 or 4. However, precision of the system for grade 2 is significantly lower than other grades. It is interesting to note that in all the other reported automated systems the grade 2 recordings remain the most difficult to classify and the precision has not increased more than 77%, whereas demographics of precision and accuracy for all the other grades is very similar. However, this is an important grade of encephalopathy to accurately classify as the trials of therapeutic hypothermia have shown that therapeutic hypothermia is most beneficial for neonates with moderate encephalopathy.

Table 5.4: Confusion Matrix of the system's output and actual assigned grade by the EEGer. The smaller text in brackets show the accuracy of individual grade.

Actual Grade	System's Output				Total	Incorrect
	1	2	3	4		
1	21(95%)	1	0	0	22	1
2	4	10(71%)	0	0	14	4
3	0	2	10(83%)	0	12	2
4	0	0	0	6(100%)	6	0
Precision(%)	84	76.9	100	100		

5.5.2 Comparison of different post-processing methods

The effect of the different post processing techniques on the overall accuracy is presented in Table 5.5. In this work, the best results were obtained with the novel two-step voting post-processing technique. It was observed that for misclassified files, the actual correct grade was often the second guessed grade. In addition, the confidence level of the wrong decision was low. This observation led to the development of the two-step voting technique which has been validated. The proposed technique not only improved the accuracy of the system but also the certainty of the overall decisions.

Apart from voting based methods, three different probabilistic methods were also examined in this study. The best overall accuracy of 85.4% was obtained using the method proposed by (Price et al. 1994). The probabilistic estimates are straight forward for this method whereas others need internal optimizations e.g. gradient descent to find the optimal class probability. In this work, the optimization algorithms proposed in (Wu et al. 2004) are used. It was noted that both of these methods were sensitive to data imbalance. As the number of data points in each class were not same, therefore they did not achieve higher accuracy.

5.5.3 Effect of different feature groups on the overall accuracy

Apart from presenting the overall accuracy of the automated system using all the 55 features, it is important to show the importance of features contributing towards the HIE EEG classification. However, due to many non-linear transformations of the original feature space it is difficult to see the contribution

Table 5.5: Comparison of different post processing techniques. (It must be noted that not all the misclassifications were the same for these methods.)

	Voting		Probabilistic Estimation		
	Majority voting	**Two-step voting**	**(Price et al. 1994)**	**(Hastie & Tibshirani 1998)**	**(Refregier & Vallet 1991)**
Overall Accuracy	83.3%	87%	85.4%	83.3%	77.7%

of each individual feature to the overall accuracy of the system. Therefore to work around this problem and show the importance of different features, the original 55 features set is categorized into 3 overlapping groups depending on the information which is captured from the EEG signal (spectral, structural and energy as described in Table 5.2). Figure 5.12 shows the overall accuracy and grade-wise accuracy when only one feature group was used and also when one feature group was removed from the original set. It can be seen that when only energy and structural feature groups are used, the accuracy of grades 3 and 4 increase whereas grade 1's accuracy decreases. However, using only-spectral-features increases grade 1's accuracy while decreasing grade 3 and 4's accuracy. The lowest accuracy was achieved using structural features only. The last three categories in the Figure 5.12 show the effect of removing one feature group. The system accuracy degrades the most when structural features are removed from the original feature set.

The spectral and energy feature groups yield the best performance when used on their own. However, the importance of structural features is evident when it is removed from the original set, as it decreases the overall accuracy and particularly in grade 1 and 4. Together it shows that grade 1 and 2 have more frequency related activity present in the EEG and therefore spectral and structural features are the best choice. Grade 3 and 4 can be distinguished from others by looking at the energy of the signal and thus the features from the energy group play an important role in classifying these grades. The overall accuracy using a pair of feature groups is higher than using only one feature group at a time. Specifically, structural features add the most information when combined with both spectral and energy features meaning they do not contain sufficient information on their own for grades classification but provide vital information in combination with others. It should be noted that the misclassifications in all these experiments were not the same.

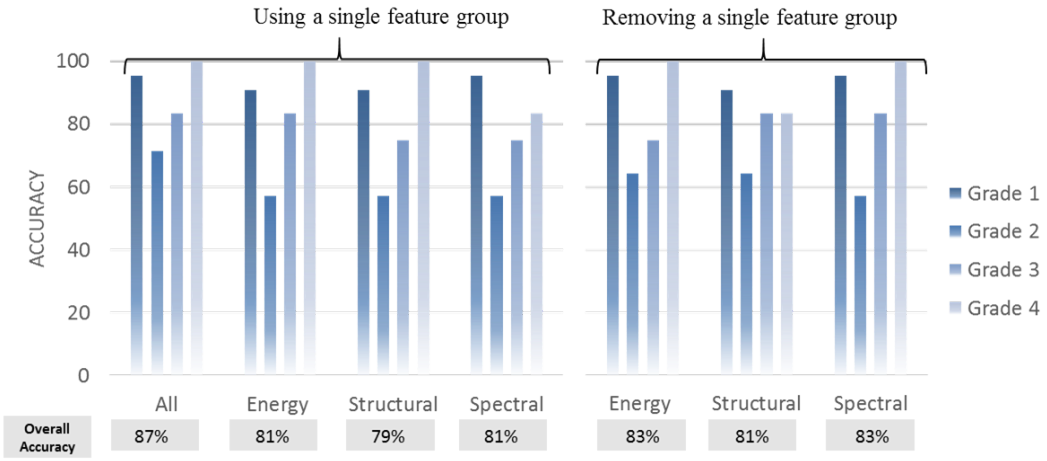


Figure 5.12: Contribution of each feature group to the overall and individual grade's accuracy.

5.5.4 Performance with the 'unknown' label

One of the advantages of the developed system is the availability of a confidence measure for the produced decision. The confidence level can be extracted from the ratio of the number of winning votes of the winning grade to the total number of votes. Presenting the decision along with a confidence level can be particularly useful in a clinical environment. In this manner, clinicians can take into account the significance/worth of a particular decision made by the system. Therefore, we have also investigated how the confidence of the computed decisions can be used in the decision making process. A threshold on the confidence level is applied to mark a decision as uncertain. As the confidence level threshold is increased, the percentage of overall decisions, which are now suspected as being uncertain, naturally increases. These uncertain decisions can be assigned an 'unknown' label and then ignored in the process of grading. Figure 5.13 shows that the accuracy of the system improves almost linearly with the increase in the certainty threshold and the resultant increase in excluded 'unknown' decisions. The circle marker represents the original result as given in Table 5.4, when no data is excluded. With the increase in the certainty threshold, and exclusion of 50% of the now 'unknown' decisions, the performance was improved to 96% as indicated by the square on Figure 5.13. Clinicians have access to other sources of information about the neonate and hence the lack of certain decision from the system may be more acceptable than the availability of uncertain or wrong decisions.

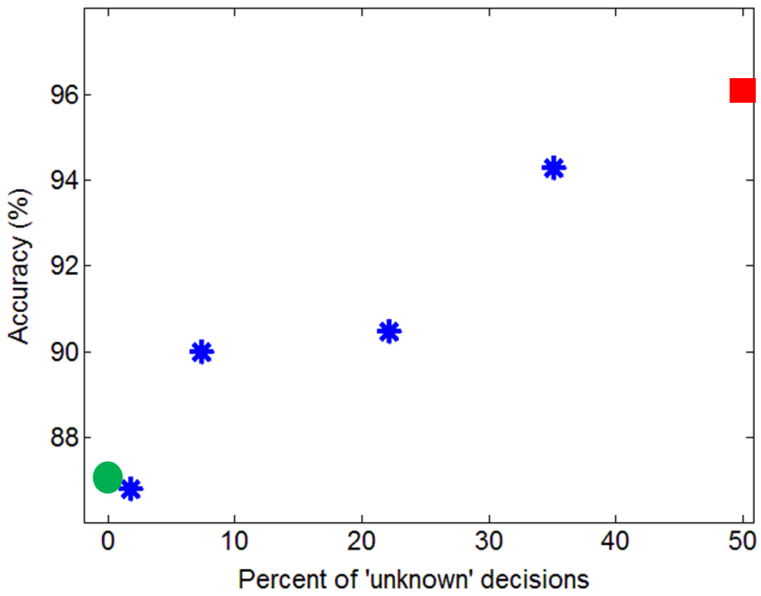


Figure 5.13: Improved accuracy of the system at the expense of not accommodating the uncertain decision. The circle marker represents the original result as given in Table 5 when no data is excluded. The red square represents the 96% accuracy at the cost of excluding 'unknown' decisions.

5.5.5 Analysis of inter-grader agreement

Inter-grader agreement is an important measure for assessing the degree of agreement between graders. In this study, two graders independently graded the HIE-EEG recordings in the dataset. As can be seen in Table 5.6, the inter-grader agreement was 92% meaning there was a disagreement on the grade in 5 recordings. It was found that the reasons for most of the disagreements among the graders were outside the guidelines provided in Table 5.1, such as the presence of runs of sharp waves and spikes. It was also observed that there were some HIE-EEG features that were overlooked by one grader and were of particular interest to the other. These included the presence of sleep-wake cycles, asymmetry and asynchrony.

Moreover, four out of five of disagreements between graders were in grade 2 and 3 files. This was a unexpected finding, given that the most confused grades for the automated system were 1 and 2. This fact emphasizes the importance of the developed technology as a decision support tool that can provide complementary information to the healthcare professional, improving in this case the classification of EEG grades 2 and 3.

The automated systems are highly dependent on the precision of the annotations of the neurophysiologists. In fact, comparing the output of an au-

tomated system using annotations of different graders, it is possible to assess which grader was more consistent during the annotation process. For this purpose, an experiment was carried out in which the system was trained and tested in matched and mismatched conditions using the same or a different grader's annotations for training and testing. The best accuracy (87%) was achieved by training and testing the system with the grader 1's annotations. The accuracy dropped down to 79.6% when annotations of grader 2 were used for both training and testing. The system scored reasonably well (85.1%) when it was trained on the annotations of grader 2 and tested on the annotations of grader 1; the accuracy however reduced when trained with grader 1's annotation and tested with grader 2's annotations. This means that, the way the automated system was creating the model from the training data was closer to the grader 1's annotations, and hence, whenever it was tested against grader 1 annotations, it resulted in better performance.

Table 5.6: Comparison of the system's performance (Accuracy %) with the annotations of different graders.

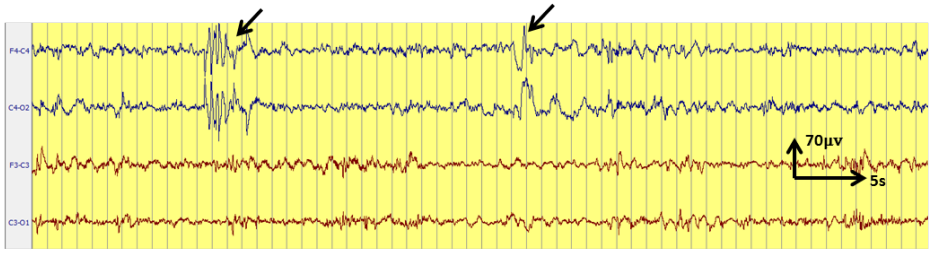
			Test	
			G-1	**G-2**
Train	**Annotations**	**G-1**	100	92.5
		G-2	92.5	100
	Automated System	**G-1**	87	81.4
		G-2	85.1	79.6

* G-1= Grader 1, G-2= Grader 2

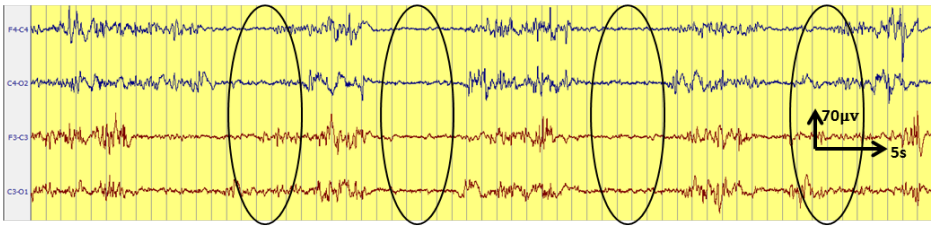
5.5.6 Analysis of the miscalssifications

Out of seven misclassified recordings, there were two that had a very high confidence level and were constantly misclassified through all the experiments. These files were presented to the graders for a further review. It was reported that in one file, the majority of the HIE-EEG was of grade 1 but the neurophysiologists graded it as grade 2 because there were spikes and sharp waves present in some parts of the recording. Moreover, periods of discontinuity (runs of low voltage) and high amplitude delta waves were also present. Although such activities did not occur very frequently, their presence even to a small degree can cause the EEG grade annotation to increase. Since the proposed system determines the grade over 8 channels and 1 hour of EEG, these possibly brief

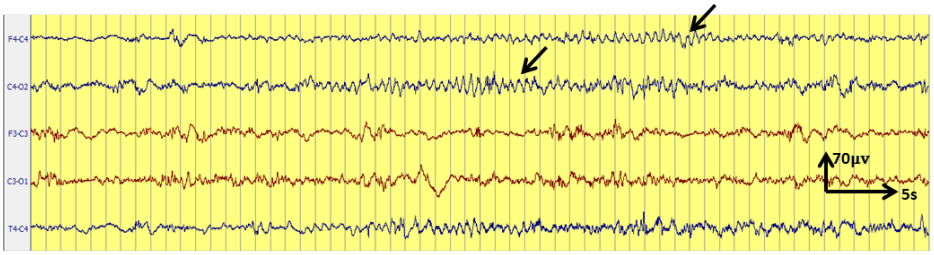
but influential events can be easily missed. Furthermore, our system currently makes decisions based on the majority of a grade's sequences, therefore such recordings could be misclassified. Some examples of similar misclassifications are shown in Figure 5.14. In Figure 5.14, a the system wrongly classified grade 2 EEG as grade 1. In this case the whole recording was misclassified because although the background EEG indeed looked like grade 1 to the annotators, it was downgraded to grade 2, due to the frequent presence of sharp waves. In Figure 5.14b, grade 3 EEG was misclassified as grade 2. In this case the annotators focused on the IBI which although less than 10sec (required for grade 2), the pattern was however persistent, asymmetric, asynchronous and lacked variability. In Figure 5.14c, grade 2 EEG was incorrectly classified as grade 1. The upgrade was determined because of the presence of asymmetry relating to the sharp components (runs of sharp waves). The reason for the second certain misclassified recording was the presence of persistent muscle artefacts. If a large part of the EEG is corrupted with such artefacts then the system could result in a wrong grade. Clearly any system that would be deployed in a NICU would require some sort of integrated artefact detection/rejection.



(a) Grade 2 EEG misclassified as grade 1; the arrows indicate the runs of sharp waves.



(b) Grade 3 EEG misclassified as grade 2; the ellipses indicate the IBI



(c) Grade 2 EEG misclassified as grade 1; the arrows point the asymmetry relating to the sharp components

Figure 5.14: Some representative examples of misclassifications.

5.5.7 Future of the automated HIE classification system

In order to show how the final automated system's output would look like, a complete EEG recording of a neonate lasting approximately 24 hours was graded using the proposed system. Figure 5.15 presents a smoothed probabilistic trace, using the method presented in section 5.4.4, that can be visualized in the clinical environment. The EEG recording was started at approximately 6 hours after birth. The bottom four plots show the probabilities of the individual grades. The top plot shows an overall probability of the severity of the HIE grade computed from those individual probabilities. For better visualization purposes the overall probability is scaled so that it is highest at the centre of a grade's interval. The overall probability is biased towards the second best grade otherwise.

Two 1-hour segments from this recording were annotated by the graders, taken at 12 and 24 hours after the birth. The system correctly classified these segments and the evolution of the HIE injury in the whole recording can be seen to migrate from grade 3 to grade 2. The confidence level of making a decision for a specific grade was computed from all the decisions made by the system for that grade's recordings in the dataset and these followed roughly the same pattern of accuracy as described in Table 5.4. These confidence level values are utilized in the overall probability plot in Figure 5.15 in the form of a shadowed line under the probability trace. For instance, as the probability trace migrates from grade 3 to 2 the shadow broadens representing the lower confidence of its grade 2 decisions.

5.6 Conclusions

An automated system for grading the HIE using background EEG is presented in this chapter. The system is based on a cross-disciplinary approach employing a GMM supervector SVM that was originally developed for speaker identification. This method of modelling the sequence of feature vectors extracted from short discrete windows has significantly improved the accuracy compared to the system proposed in (Stevenson et al. 2013). Novel post-processing techniques have been shown to increase the performance in comparison to conventional majority voting. The availability of a confidence measure for the produced decision allowed the use of a new label 'unknown' for the decision which are uncertain. The level of uncertainty can be varied according to the clinical needs. With this new 'unknown' label the overall accuracy can be increased to 96%. The ef-

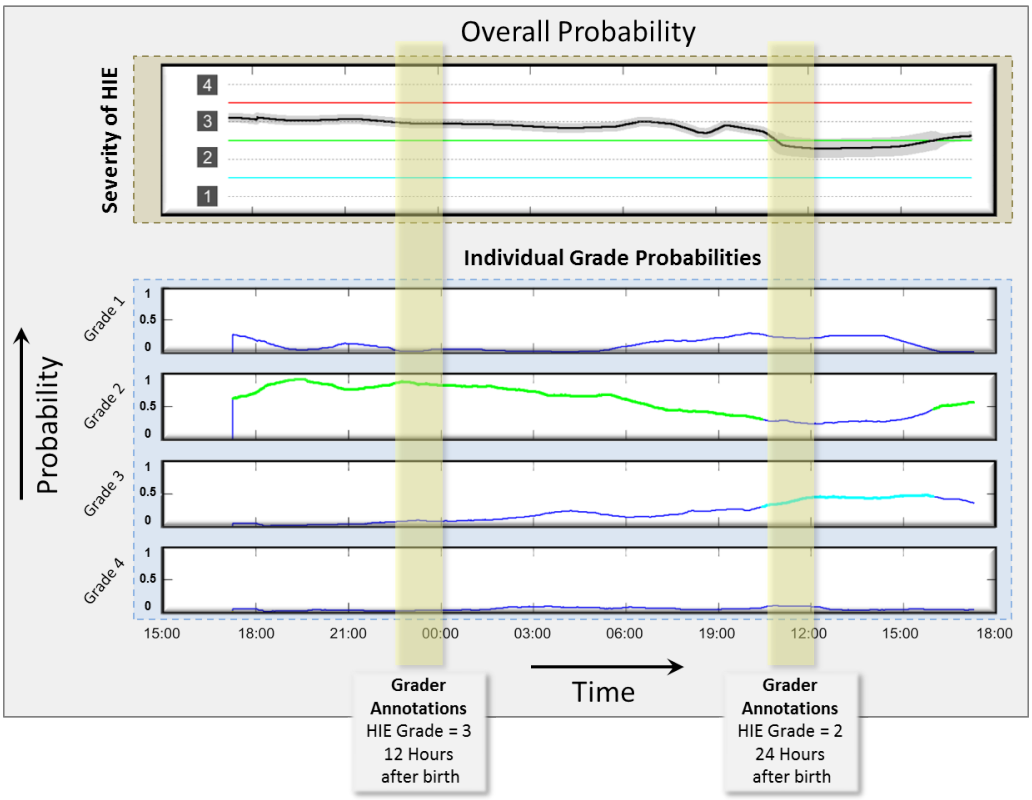


Figure 5.15: Probability trace of a 24 hour recording using the proposed automated system.

fect of different feature groups on the overall accuracy was also examined. It is noted that every feature group plays an important role to discriminate different types of background activities present in all the grades. Analysis of misclassified recordings suggest that there are brief influential events which may increase the grade of a EEG recording. Therefore, it is important to develop parallel systems that can detect such brief events and may help in improving the performance of the automated HIE grading system.

References

Ambalavanan, N., Carlo, W. A., Shankaran, S., Bann, C. M., Emrich, S. L., Higgins, R. D., Tyson, J. E., O'Shea, T. M. et al. (2006). Predicting outcomes of neonates diagnosed with hypoxemic-ischemic encephalopathy. *Pediatrics*, 118(5):2084–2093.

Azzopardi, D. V., Strohm, B., Edwards, A. D., Dyet, L., Halliday, H. L., Juszczak, E., Kapellou, O., Levene, M. et al. (2009). Moderate hypothermia to treat perinatal asphyxial encephalopathy. *New England Journal of Medicine*, 361(14):1349–1358.

Berger, R. & Garnier, Y. (1999). Pathophysiology of perinatal brain damage. *Brain Research Reviews*, 30(2):107–134.

Boylan, G. B., Murray, D. M., & Rennie, J. M. (2008). The normal EEG and aEEG. *In Neonatal Cerebral Investigation*. Cambridge University Press, pp. 83–91.

Campbell, W., Sturim, D., Reynolds, D., & Solomonoff, A. (2006). SVM based speaker verification using a GMM supervector kernel and NAP variability compensation. *In Proceedinds of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2006*, vol. 1. pp. 97–100.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Doyle, O., Temko, A., Murray, D., Lightbody, G., Marnane, W., & Boylan, G. (2010). Predicting the neurodevelopmental outcome in newborns with hypoxic-ischaemic injury. *In Proceedings of IEEE Annual International Conference of Engineering in Medicine and Biology Society, (EMBC) 2010*. IEEE, pp. 1370–1373.

Faul, S., Temko, A., & Marnane, W. (2009). Age-independent seizure detection. *In Proceedings of IEEE Annual International Conference of Engineering in Medicine and Biology Society, (EMBC) 2009*, vol. 2009. pp. 6612–6615.

Gauvain, J. & Lee, C.-H. (1994). Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Transactions on Speech and Audio Processing*, 2(2):291–298.

Gotman, J., Flanagan, D., Zhang, J., & Rosenblatt, B. (1997). Automatic seizure detection in the newborn: methods and initial evaluation. *Electroencephalography and Clinical Neurophysiology*, 103(3):356–362.

Greene, B. R., Faul, S., Marnane, W. P., Lightbody, G., Korotchikova, I., & Boylan, G. B. (2008). A comparison of quantitative EEG features for neonatal seizure detection. *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology*, 119(6):1248–1261.

Hastie, T. & Tibshirani, R. (1998). Classification by pairwise coupling. *The Annals of Statistics*, 26(2):451–471.

Hsu, C.-W. & Lin, C.-J. (2002). A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, 13(2):415–425.

Hu, H., Xu, M.-X., & Wu, W. (2007). GMM supervector based SVM with spectral features for speech emotion recognition. *In Proceedinds of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2007*. pp. 413–416.

Joachims, T. (2005). A support vector method for multivariate performance measures. *In Proceedings of the 22nd international conference on Machine learning*. ACM, pp. 377–384.

Korotchikova, I., Stevenson, N. J., Walsh, B. H., Murray, D. M., & Boylan, G. B. (2011). Quantitative EEG analysis in neonatal hypoxic ischaemic encephalopathy. *Clinical Neurophysiology*, 122(8):1671–1678.

Matic, V., Cherian, P. J., Koolen, N., Naulaers, G., Swarte, R. M., Govaert, P., Huffel, S. V., & Vos, M. D. (2014). Holistic approach for automated background EEG assessment in asphyxiated full-term infants. *Journal of Neural Engineering*, 11(6):066007.

Murray, D. M., Boylan, G. B., Ryan, C. A., & Connolly, S. (2009). Early EEG findings in hypoxic-ischemic encephalopathy predict outcomes at 2 years. *Pediatrics*, 124(3):e459–e467.

Platt, J. C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *In Advances in Large Margin Classifiers*, Neural Information Processing Series. MIT Press, pp. 61–74.

Price, D., Knerr, S., Personnaz, L., Dreyfus, G., & Dreyfus, L. P. G. (1994). Pairwise neural network classifiers with probabilistic outputs. *In Advances in Neural Information Processing Systems 7*, vol. 7. MIT Press, pp. 1109–1116.

Refregier, P. & Vallet, F. (1991). Probabilistic approach for multiclass classification with neural networks. *In Proceedings of the International Conference on Artificial Neural Networks (ICANN) 1991*. pp. 1003–1006.

Reynolds, D. (2009). Gaussian mixture models. *In Encyclopedia of Biometrics*. Springer US, pp. 659–663.

Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 10(1–3):19–41.

Stevenson, N. J., Korotchikova, I., Temko, A., Lightbody, G., Marnane, W. P., & Boylan, G. B. (2013). An automated system for grading EEG abnormality in term neonates with hypoxic-ischaemic encephalopathy. *Annals of Biomedical Engineering*, 41(4):775–785.

Temko, A., Thomas, E., Marnane, W., Lightbody, G., & Boylan, G. (2011). EEG-based neonatal seizure detection with support vector machines. *Clinical Neurophysiology*, 122(3):464–473.

Thomas, E. M., Temko, A., Lightbody, G., Marnane, W. P., & Boylan, G. B. (2010). Gaussian mixture models for classification of neonatal seizures using EEG. *Physiological Measurement*, 31(7):1047.

Vapnik, V. N. & Kotz, S. (1982). *Estimation of dependences based on empirical data*, vol. 40. Springer-verlag New York.

Volpe, J. J. (2008). *Neurology of the Newborn*. Elsevier Health Sciences.

Wu, T.-F., Lin, C.-J., & Weng, R. C. (2004). Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 5:975–1005.

You, C. H., Lee, K.-A., & Li, H. (2009). An SVM kernel with GMM-supervector based on the Bhattacharyya distance for speaker recognition. *IEEE Signal Processing Letters*, 16(1):49–52.

Zhuang, X., Zhou, X., Hasegawa-Johnson, M. A., & Huang, T. S. (2010). Real-world acoustic event detection. *Pattern Recognition Letters*, 31(12):1543–1551.

Chapter 6

Classification of HIE using heart rate variability

This chapter presents an automated system for classification of HIE injury into mild and moderate-severe grades using Heart Rate (HR) signal. The system is similar to the one presented in Chapter 5. Sequences of different short term features from time and frequency domains are used to create a statistical model based long-term features (Supervectors). SVM is then used to classify the grade of one hour long recordings.

The chapter is organized in four sections. First the motivation for using heart rate signal for HIE grading and the basis of the proposed system are presented. Then the details of the proposed method are explained with a detailed discussion on the short term HR features used in this work. This is followed by the presentation of the obtained results. Lastly, two systems based on the fusion of EEG and ECG signals are presented with a discussion on their achieved performance and limitations.

6.1 Motivation of using heart rate for HIE classification

HIE is generally graded soon after the birth into mild, moderate or severe injury. Infants with moderate-severe HIE can develop a secondary phase of injury, thought to occur from 6-48 hours after birth and which has been linked with poor neurodevelopmental outcome (Shankaran et al. 2005, Jensen et al. 2003). Therapeutic Hypothermia (TH) is used to treat these neonates, which aims to alleviate this secondary phase of injury. For TH to be effective, it is usu-

ally commenced within 6 hours after birth, prior to the potential development of secondary injury; however, the sooner this is commenced the better (Gunn et al. 1998). Thus, in this narrow window of 6 hours, clinical staff need the most information to assess the injury and then initiate the treatment.

EEG is the most common and reliable method for classifying the grade of the HIE injury. However, it requires expertise to first acquire the EEG signal and then presence of an expert neurophysiologist to interpret it. This expertise is limited to specialized units only and not readily available in many Neonatal Intensive Care Units (NICU). In addition, EEG is rarely available immediately after birth particularly at remote maternity units where neither the expertise to interpret EEG nor the equipment may be available. Therefore, it was important to find new ways to classify HIE using other physiological signals that are readily available after birth and can provide some degree of information to at least start the TH treatment.

The electrocardiogram (ECG) is a method of monitoring heart which is conveniently available after birth and is used to monitor neonates around the world. Heart rate is a common marker to assess many functions of the body. Although, a number of methods exist to estimate heart rate for example, photoplethysmography, pulseoximetry, seismocardiography, however ECG remains the most accurate and reliable method. Heart rate from an ECG signal is usually estimated by calculating the number of R peaks (beats) in a minute. The variability in the interval between RR peaks is used to measure the heart rate variability (HRV) which can help in observing the sympathetic and parasympathetic nervous systems. It is reported in previous studies on neonates that the HRV changes after hypoxia and is shown to be statistically correlated with HIE injury (Aliefendioglu et al. 2012, Selig et al. 2011). Moreover, it is shown in a recent study that different HRV features have the ability to significantly discriminate mild and moderate-severe HIE (Goulding et al. 2015).

Figure 6.1 shows examples of HRV during mild and moderate-severe HIE with the related EEG of the similar grade. HRV is inversely related to the HIE grade, the greater the HIE grade the lower the HRV. It is reported that this phenomenon occurs due to injury in the parts of the brain which are responsible for the initiation of the HR rhythm (Volpe 2008, Novak et al. 1995). Moreover, it could be the result of an immature nervous system which is the result of premature birth (Selig et al. 2011). Therefore, there are premises to believe that HRV can be used to grade the severity of HIE injury and will be useful for centres with no access to EEG.

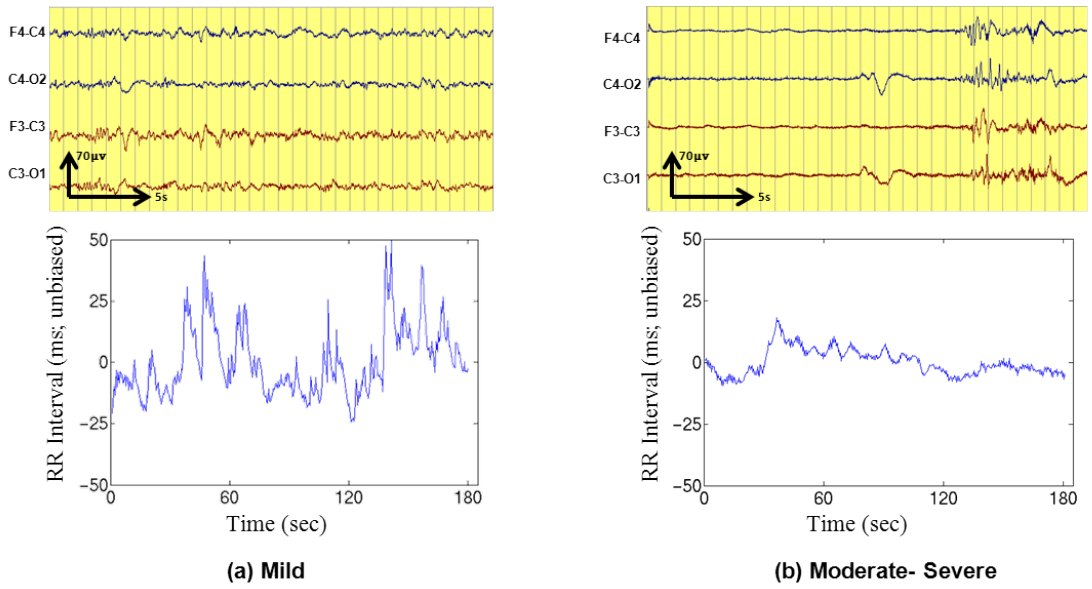


Figure 6.1: Examples of the Mild and moderate - severe HIE grade on EEG and its related HRV in RR interval. The HRV signals are 3 minute long. The signals are normalized by subtracting the mean HRV in this time duration. The corresponding EEG samples shown are 1 minute long.

To the best of the author's knowledge, there is only one previous study on developing a tool for the classification of HIE injury using HRV (Matic et al. 2013). They derived 12 features from the HRV signal and then used a simple Linear Discriminant Classifier (LDC) to classify two hour ECG recordings into mild and moderate-severe HIE. The dataset included 36 hours of ECG recordings from 18 neonates.

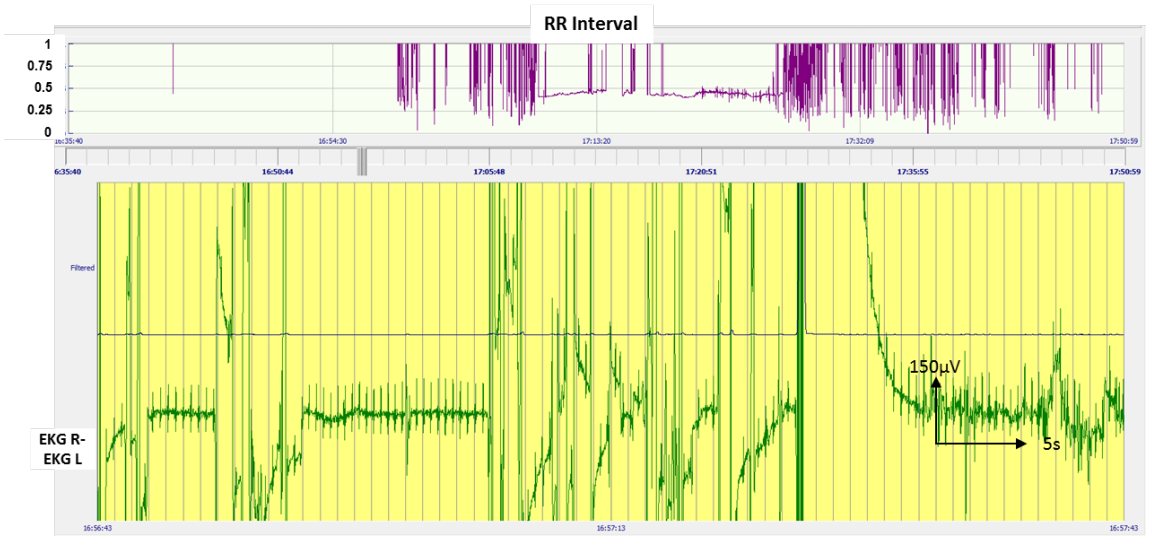
In this study, the use of HRV signal for classifying HIE into one of two grades (mild and modereate-severe) is analysed. The HRV signal is less complex and does not have a similar dynamic nature to EEG. It is shown in (Doyle et al. 2010) that HRV is not very useful for detecting short events such as seizures, because the changes in HRV are very subtle. Thus, classifying the HRV requires a sequence of data that is large enough to capture the slower time scale components in order to discriminate between HIE grades. In this regard, different HRV features extracted from different lengths of HRV signal were examined. Moreover, different classification techniques are investigated. Specifically, the supervector approach as discussed in the previous chapter is investigated for the HRV signal because it enables the extraction of a long-term summary from a sequence of the short-term HRV feature vectors.

6.2 Dataset

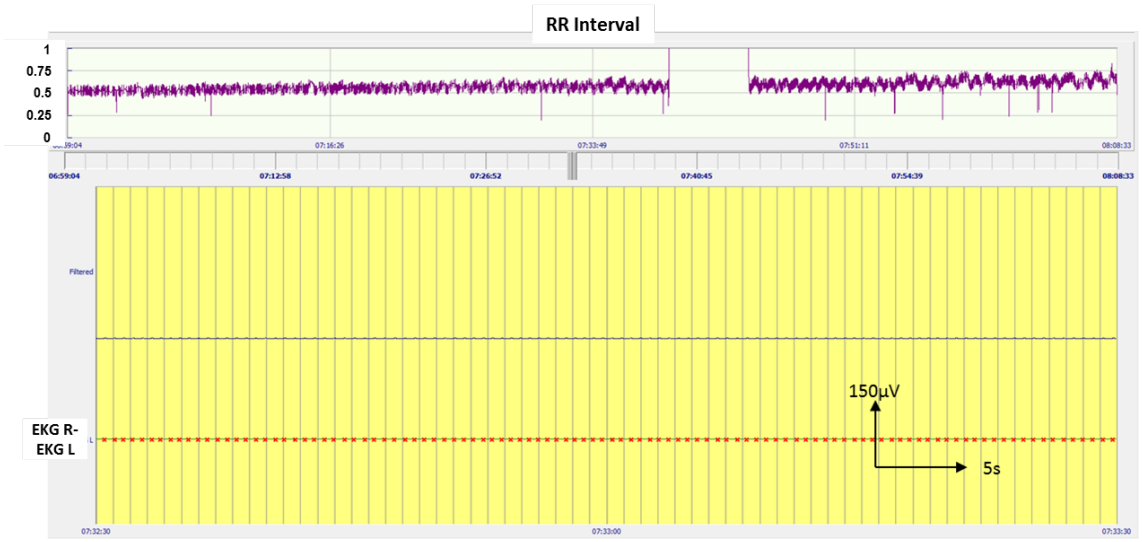
The dataset used in this work consists of 54 1-hour recordings of physiological signals of neonates with HIE injury. The data was collected in the NICU of the Cork University Maternity Hospital, Cork, Ireland. Written informed parental consent and approval from the Clinical Ethics Committee of the Cork Teaching Hospitals was obtained for all neonates. The neonates in this cohort were not treated with TH. Two electrodes were placed, one on each shoulder of the neonate for recording the ECG. The ECG recordings were graded based upon the background EEG. Each EEG recording had continuous presence of a specific grade. The EEG grade of the one hour recording was defined with the consensus of two independent EEGers using the guidelines outlined in (Murray et al. 2009). This same dataset has earlier been used for grading HIE using EEG (Chapter 5).

This dataset was collected with the main focus on EEG quality, so the ECG was not monitored and maintained adequately during the recordings. Six out of 54 recordings were excluded due to the extremely bad quality or absence of an ECG trace. Some examples of such recordings are shown in Figure 6.2. As can be seen, either a big part of the recording was corrupted by the artifacts (Figure 6.2a) or the sensitivity of the ECG signal was very low to correctly detect the R peaks (Figure 6.2b). Major artifacts that were clearly visible, such as movement and electrode displacement were manually annotated and removed. Minor artifacts however, were not removed. Figure 6.3 shows some examples of artifacts and their annotations. Moreover, individual R peaks were also not corrected. Lastly, Figure 6.3c shows an example of a recording where the sensitivity of the ECG signal changed during the last part. As the Pan-Tompkins method used in our current work to detect the R peaks was not adaptive, therefore it was not able to correctly detect the R peaks in that region of the recording and hence was manually removed.

Based on the quality of ECG, this dataset can be seen as overly pessimistic, if compared to the real world situations. The developed system is expected to perform better if the quality of ECG is maintained.



(a) Large chunk of recording corrupted by artifacts

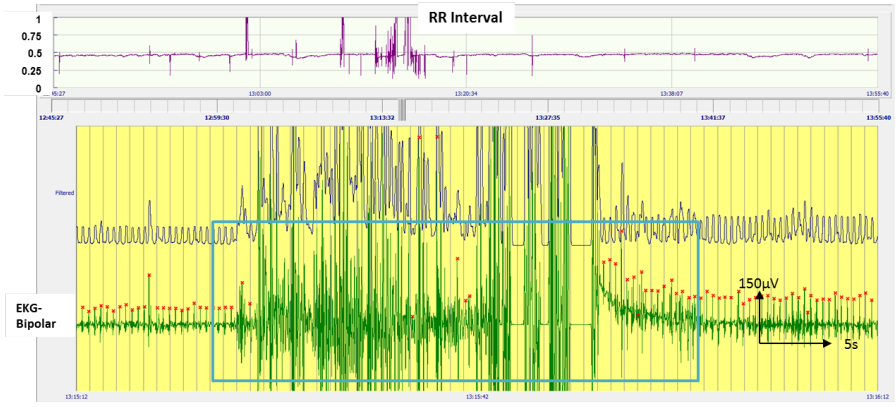


(b) Very low sensitivity of ECG signal

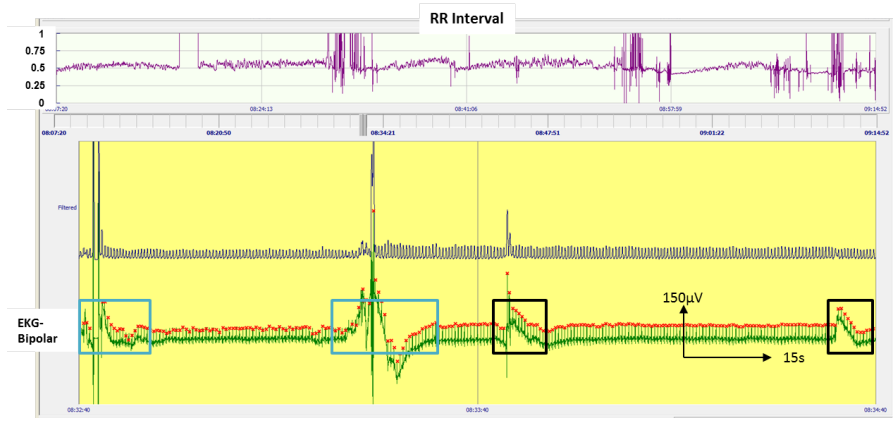
Figure 6.2: Examples of recordings that were excluded from the dataset. (Top plot in each figure shows the RR interval of the whole recording.)

6.3 Overall system

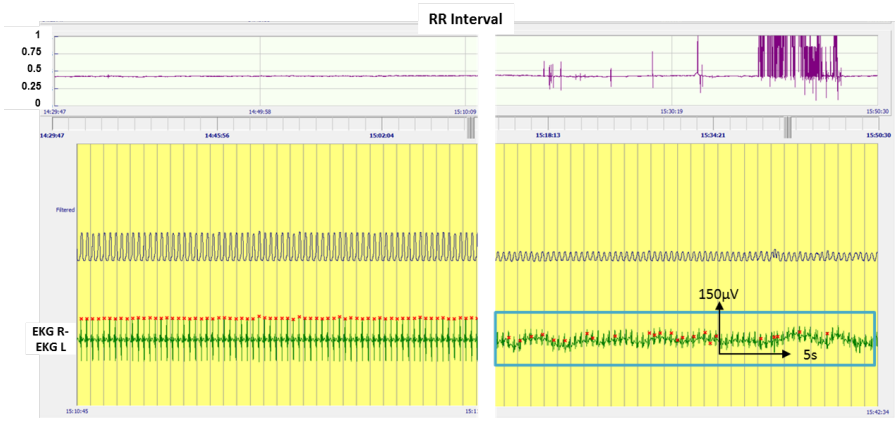
Figure 6.4 shows an outline of the complete HRV based HIE grading system. First, instantaneous HR is extracted from the raw ECG signal. Then, features are extracted from the HR signal. Sequences of these feature vectors are used to create a statistical model. Parameters of this statistical model are then used as input to a discriminant classifier. The output of the classifier for each sequence is then converted into probability of moderate-severe grade of HIE. Details of each individual block are as follows.



(a) Blue box = Major artifacts (Annotated)



(b) Blue boxes = Major artifacts (Annotated), Black boxes = Minor artifacts (Not Annotated)



(c) An example of recording where first part has different voltage sensitivity. R peaks were missed by the Pan-Tompkins algorithm in the second part. Blue box = Annotated part

Figure 6.3: Examples of major artifacts that was manually annotated in the recordings. (Top plot in each figure shows the RR interval of the whole recording.)

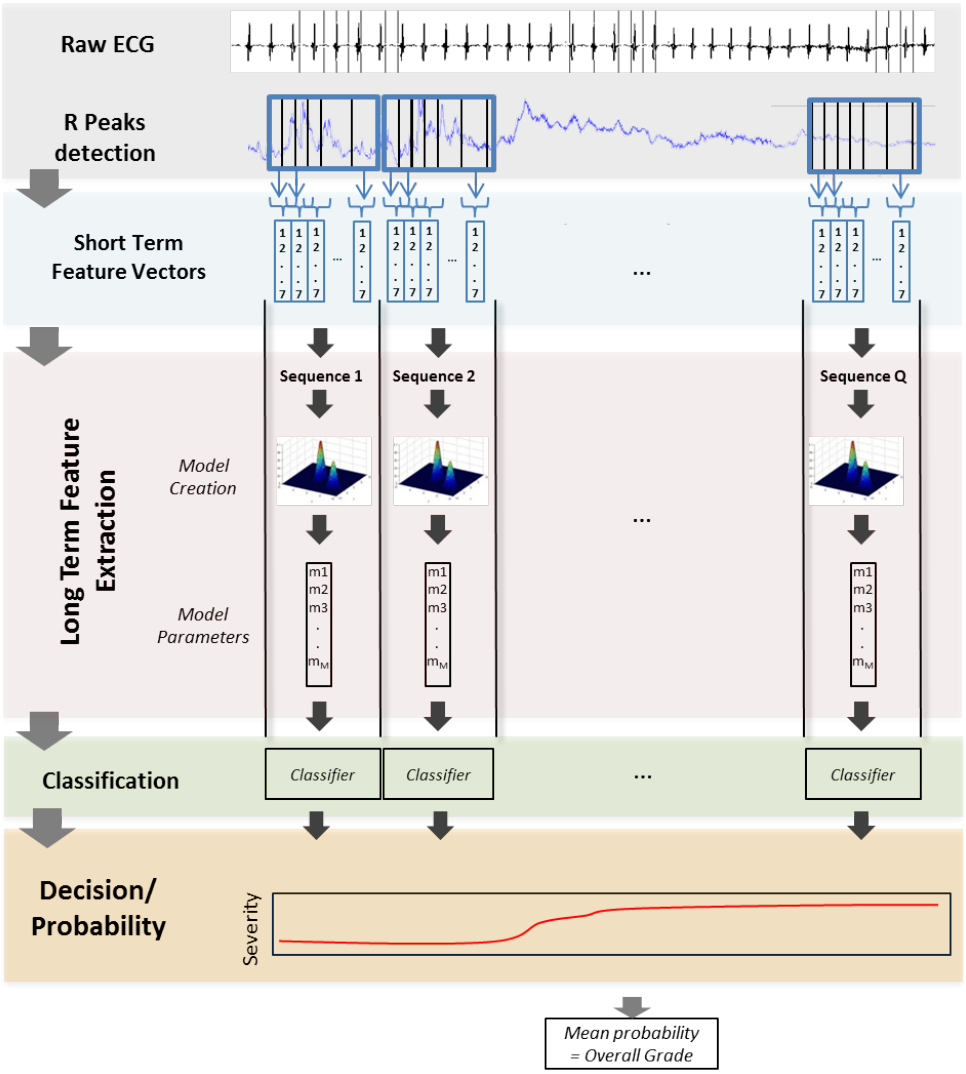


Figure 6.4: Overall system for classifying HRV recordings.

6.3.1 Pre-processing and feature extraction

The R-peaks were extracted from the raw ECG signals using the Pan Tompkins method (Pan & Tompkins 1985). The Pan-Tompkins algorithm can be divided into four stages (Figure 6.5): 1) Preprocessing stage: here the signal is filtered using a bandpass filter to attenuate the noise. 2) Differentiator stage; here the information about the slope of QRS complex is extracted. 3) Squaring stage: here the differentiated signal is squared to remove negative points and accentuate the differentiated signal. 4) A moving average filter is then used to include the information about both the slope and width of the QRS complex. A threshold is applied on the resulting signal to get the timing of the R-peaks.

The R-peaks obtained from the ECG are, by nature, not uniformly sampled in time. The methods to transform the time-domain signals to the frequency-

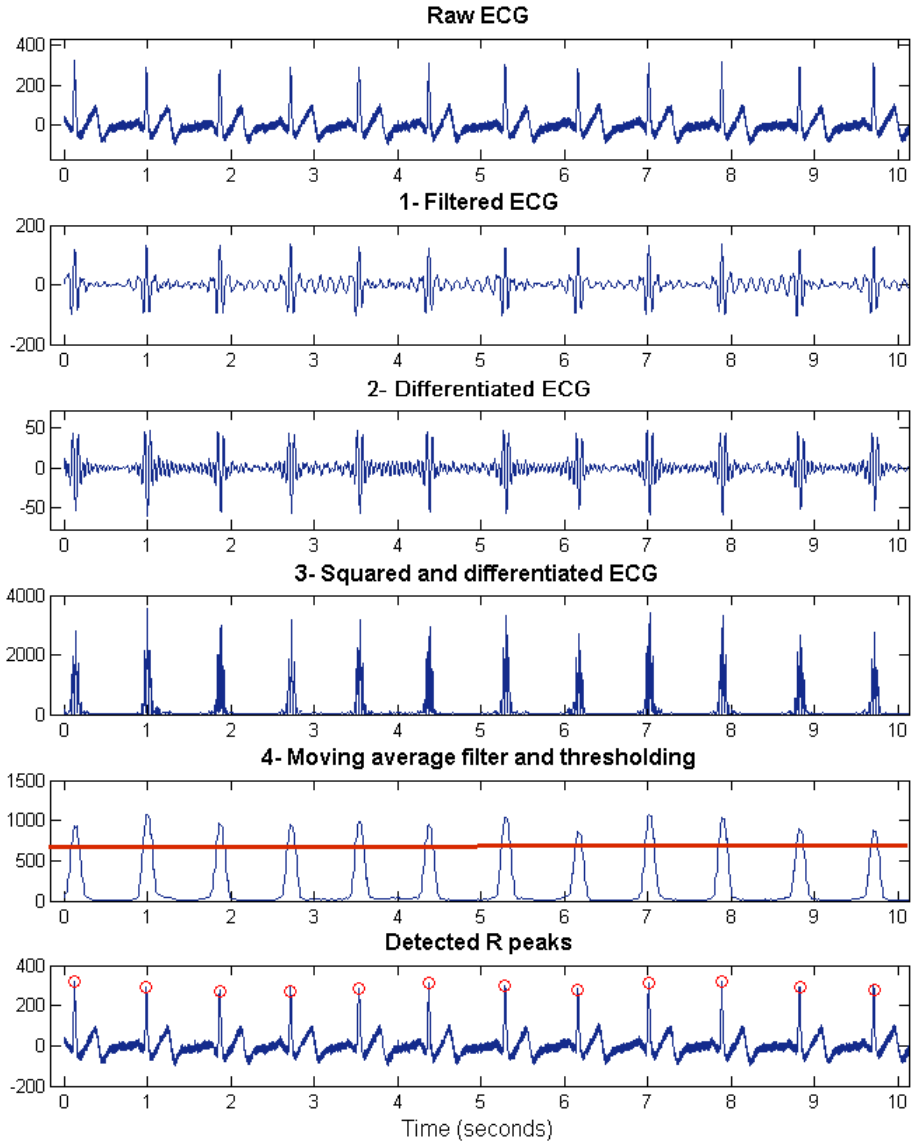


Figure 6.5: An example of an ECG signal (of an adult) going through different stages of Pan-Tompkins algorithm for RR peaks extraction.

Table 6.1: HRV features used in this study

Domain	Features
Frequency	Power in very low frequency (VLF) band (0.008-0.04 Hz)
	Power in low frequency (LF) band (0.04-0.2 Hz)
	Power in high frequency (HF) band (0.2-2 Hz)
	Ratio of LF/HF
Time	Mean NN interval
	Standard deviation NN interval
	Triangular interpolation of NN interval (TINN) histogram
	Skewness
	Kurtosis
	Entropy
	1st derivative of standard deviation between NN interval
	2nd derivative of standard deviation between NN interval

domain (e.g. FFT) can not be directly applied on the unevenly sampled signals. Therefore, the R-peaks were uniformly re-sampled to 256 Hz using hermite-spline quadratic interpolation. The RR interval after these transformations is regarded as the normalized RR interval (NN interval). Thereafter, 12 HRV features from the time and frequency domains were extracted from a fixed length window (epochs) with 30 second overlap. The features are outlined in the Table 6.1.

The frequency domain features are extracted by applying the Fast Fourier Transform (FFT) on the uniformly sampled HR signal. The mean power in three frequency bands is used as features (Very Low Frequency (VLF)= 0.008-0.04Hz, Low Frequency (LF) = 0.04-0.2Hz and High Frequency (HF)= 0.2-2 Hz). These frequency bands are specially chosen for neonates. The respiration activity caused by the Respiratory Sinus Arrhythmia (RSA) is usually observed on the HF band and is generally considered to be modulated by the parasympathetic autonomic nervous system (Pomeranz et al. 1985). The LF band is considered to represent the Mayer waves of the blood pressure changes. The VLF band corresponds to the fluctuations in the thermoregulation of the body (Malliani et al. 1991, Rosenstock et al. 1999). The LF and VLF bands have been utilized in many studies for neonates including studies on HIE induced infants, and have been shown to correlate with brain injuries. The ratio of LF/HF provides a measure of the balance between parasympathetic and sympathetic nervous systems. It is shown in (Goulding et al. 2015, Volpe 2008) that HIE may

cause decreased blood pressure and reduced thermal and respiratory control of the body which could all get represented from the HR signal. This phenomenon can also be observed in Figure 6.7, which shows the line plots of the normalized histograms generated from the features of two groups of HIE grades. It can be seen that the neonates with the mild-moderate HIE have lower VLF and LF values. However, the HF band and the subsequent LF/HF ratio feature does not fully distinguish the two groups.

Different time domain features are also extracted from the HR signal. The mean NN interval is extracted by taking an average of the NN intervals timings in a given epoch. The standard deviation provides the measure of variability of the NN interval in a given window in time. The entropy quantifies the amount of disorder and captures the changing dynamics of HR. Skewness and kurtosis are common statistical features that provide the information about the shape of the distribution of NN intervals.

Three features from the commonly used geometrical representations of HR signal were also extracted (Triangular interpolation of the NN interval (TINN), first derivative of the standard deviation (SD1) and second derivative of the standard deviation SD2). TINN provides a measure of the baseline width of the NN intervals distribution measured as a base of the triangle (Camm et al. 1996). It can be seen in Figure 6.6 that the distribution of HR sequence from mild HIE group is broader as compared to the HR signal from the moderate-severe HIE grade which is captured by the TINN feature. The Poincaré plot is another graphical representation of NN interval variability. It is created by plotting each NN-interval point against the successive NN interval. Figure 6.6 shows the Poincaré plot of mild and moderate severe HR signals. A decrease in the spread of the poincare plot along x and y axis can be clearly seen in the moderate-severe HIE induced HR. Two features from this plot are extracted in this work: the first derivative of the standard deviation which captures the spread across the x-axis and the second derivative of the standard deviation which captures the spread across the y-axis.

Although the statistical significance of these features to characterize HIE grades have been discussed in a number of previous studies (Goulding et al. 2015, Matic et al. 2013, Aliefendioglu et al. 2012), we wanted to present them as a graphical representation to show the strength of each feature to separate the mild and moderate-severe HIE. Figure 6.7 shows the histograms of the HR features using all the data of these grades of HIE. The features were first normalized to zero mean and unit variance. These histograms are normalized

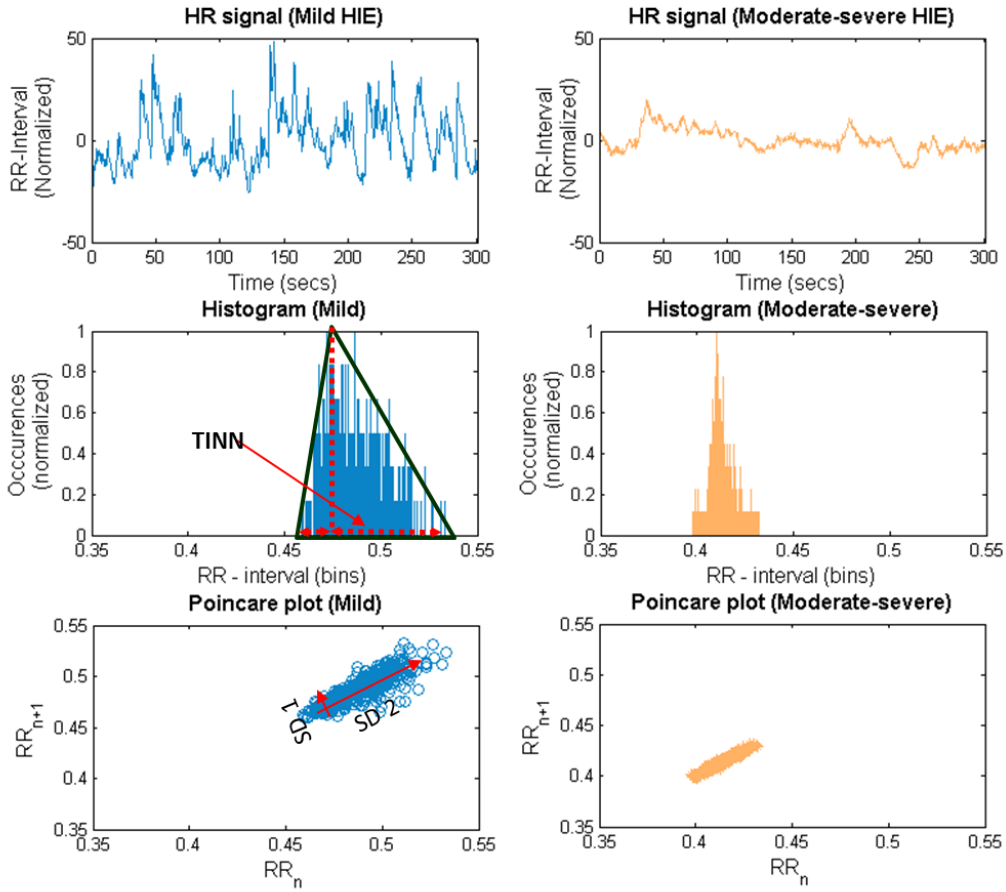


Figure 6.6: Representation of HR signal in frequency domain, histogram and Poincaré plot.

along the Y axis using the maximum value of occurrence to decrease the effect of data imbalance. The most distinguishing features from the frequency domain as seen from this figure are the power in VLF and LF bands which both decrease in moderate-severe HIE induced HR. Other features that show good separation ability are variance and histogram based standard deviation, entropy, TINN and the derivatives of standard deviations.

In essence, these features provide a generalized picture of both time and frequency domain representation of the HRV. The performance of these features for HIE grading has been discussed in several recent clinical studies . More details on the derivation of above mentioned features can be seen in (Doyle 2010).

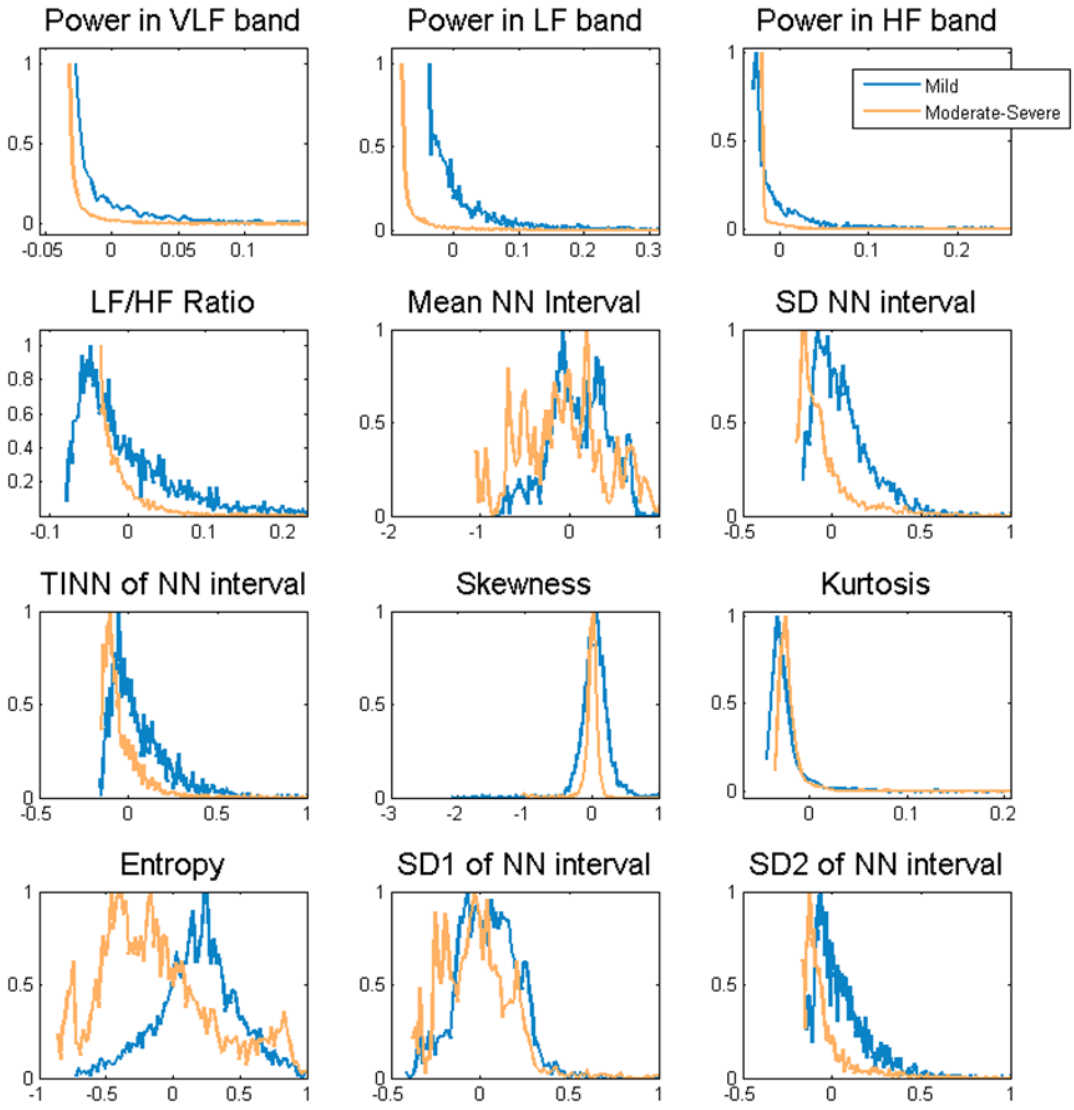


Figure 6.7: Histograms of the features used in this study.

6.3.2 Long term feature extraction (Supervector)

The process of the long term feature extraction is similar to the one described in Chapter 5. Therefore for HR signal, this process will be briefly described.

Figure 6.8 shows the process of extracting long term feature vectors. First, a GMM-UBM (general HR model) is created with all the available data of both grades. The system was tested with {2,4,8,16} number of Gaussian components and the performance for each of them is reported in the results section. Principle Component Analysis (PCA) is used to decorrelate the original feature space which allows the use of a diagonal covariance matrix. All the variance of the original space is retained. Once the UBM is trained, maximum a-posteriori

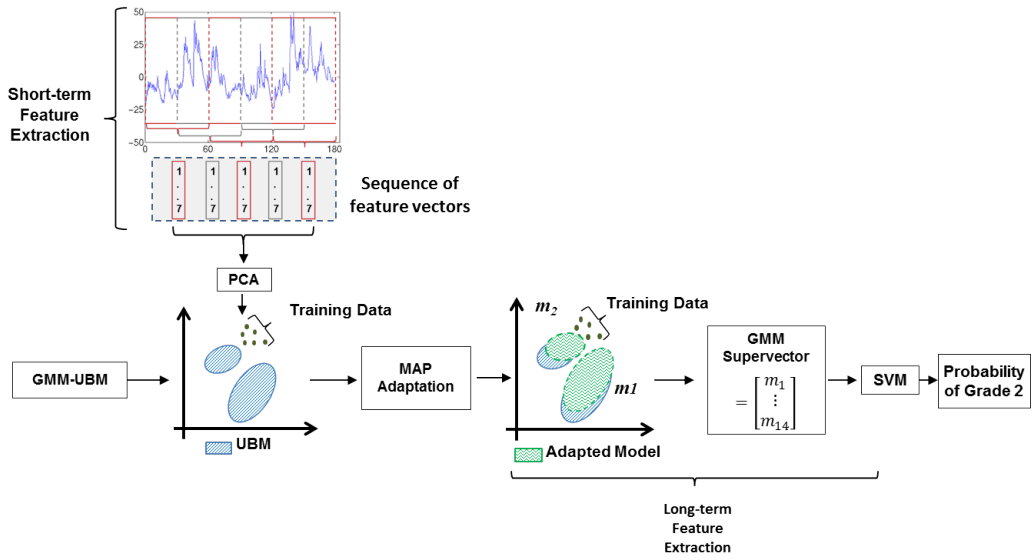


Figure 6.8: The process of creating a long-term feature vector (supervector). The Gaussian components shown here are just for illustration purposes.

(MAP) adaptation is used to adapt the parameters of this UBM using the training data (Reynolds 2009, Gauvain & Lee 1994). Sequences of short term feature vectors are used to adapt the UBM. In this work only the means of the UBM model are adapted. A supervector is then created by concatenating all the means of the adapted model. These supervectors are then fed to the SVM as feature vectors representing the long sequences of the short-term HR features. Different lengths of sequences were tested in this study and overall accuracy using each length is reported in the Results section.

6.3.3 Classification

The standard linear SVM is used as a classifier. During the training phase, supervectors of the two grades are used to create an SVM model. Three-fold cross validation over the training data is used to find the regularization parameter C for the linear SVM.

In the testing phase, the test supervectors are passed through the trained SVM model to get one decision per sequence of the HRV signal. Output of the SVM is then converted to probabilities using the method outlined in (Platt 1999).

6.3.4 Performance assessment

In order to assess the performance of the proposed system, the Leave One Out (LOO) cross validation method was used. Here our system was trained using the data from 48 recordings categorized into two classes; mild (grade 1) and moderate-severe (grade 2). The remaining one unseen recording was used to test the system. A separate UBM was created each time which does not contain the data from the unseen recording. The mean probability of grade 2 of all the recordings was calculated from the individual sequence based probabilities. Three different classifiers are compared in this study; GMM alone, SVM alone and Supervector based classifier.

6.3.5 Software Implementation

The implementation of this system is similar to the one described in section 5.4.6.

The training and testing time were however very short as compared to the EEG based system because there was only one ECG channel to process. On average it took 10-15 minutes to complete the training cycle for one round of LOO and 3-5 minutes to test an unseen recording.

6.4 Results

6.4.1 Features performance

It was empirically noted that classifiers perform better when a subset of features was used. It is also clear from the histograms in Figure 6.7 that not all the features from the dataset have good discrimination ability. For this reason, a simple feature selection criteria was executed. ROC area, for each feature was calculated by varying the threshold from 0-1 (features are normalized to have zero mean and unit variance) using all the available data. Features with ROC area higher then 60% were selected. This resulted in a subset of 6 features as highlighted in Table 6.2. Although performance of some features, e.g. kurtosis and skewness, is poor in discriminating the HIE grades. However in a separate study, these features have shown significant discrimination power in detecting artifacts from the ECG signal (Gholinezhadasnefestani et al. 2015). Therefore, usability of such features can be explored to develop a parallel artifact rejection block.

Table 6.2: ROC area achieved by each feature

Features	ROC area (%)
Power in VLF band	**60.7**
Power in LF band	**66.8**
Power in HF band	59.3
Ratio of LF/HF	54.1
Mean NN interval	58.3
Standard deviation NN interval	**69.1**
TINN	**66.1**
Skewness	54.6
Kurtosis	47.9
Entropy	**69.6**
1st derivative of standard deviation between NN interval	58.4
2nd derivative of standard deviation between NN interval	**68.8**

6.4.2 Performance comparison of different classifiers

Table 6.3 shows the performances of different classification methods using all and the subset of features. It can be seen, that overall the subset of features achieved a higher ROC area. The presented results are the best achieved by each system using different parameters. For the GMM based system, $M = \{2, 4, 8, 16\}$ Gaussian components were tried along with different feature extraction window lengths of {1, 3, 5} minutes. The highest ROC area of 76.5% was achieved using 4 Gaussian components and a 3 minute window. For the SVM based system, three window lengths {1,3,5} minutes for features extraction were tested. A radial basis function kernel was used inside the SVM. The highest ROC area of 68.3% was obtained using a 1 minute window length.

The best ROC area of 79.9% among all systems was achieved using the proposed supervector based approach. For this system, 2 different kinds of parameters were tested; the number of Gaussian components {2,4,8,16} and the number of epochs (sequence length) to create the supervector. The best results were obtained with 2 Gaussian components and a sequence of 7 epochs. Although, the subset of features resulted in better performance; however our feature selection criteria was very coarse and did not produce the best set of features. It is well known that some features on their own do not provide enough separability between classes but could increase the performance if used with other features. Therefore, the best performing system was also experimented by adding another significantly important HRV feature (mean NN interval). This

Table 6.3: Performance (ROC area) of different methods in classifying the 2 grades of HIE

Method	**Roc Area (%)**
GMM alone	
All features	74.5
Subset	76.5
SVM alone	
All features	65.5
Subset	68.3
Supervector	
All features	74.2
Subset	79.9
Subset with mean NN interval	**81.0**

feature on its own did not perform well as shown in Table 6.2; however when added into the subset of best performing features, increased the ROC area to 81%. This simple experiment suggests that it is necessary to implement a more sophisticated feature selection algorithm to further increase the classification performance of the system.

Figure 6.9 shows the comparison of the ROC areas of the proposed system compared to other basic classification approaches. It can be seen that, most of the improvement seen in the supervector approach is obtained in the sensitivity metric meaning that it has better performance for classifying grade 2 recordings. However, individually the supervector based classifier can be thought of as being more conservative towards grade 2 as it performs better in the specificity region as compared to the sensitivity region.

6.4.3 Comparison of different sequence lengths

Table 6.2 presents the results of supervector based system with different lengths of sequences. It can be seen that 7 epochs corresponding to four minutes of ECG, gave the best results. A shift of 1 epoch (30 second) was used for each of the results for the proposed approach. This duration of sequence length in time is similar to the one defined in a clinical study (5 min) by (Goulding et al. 2015).

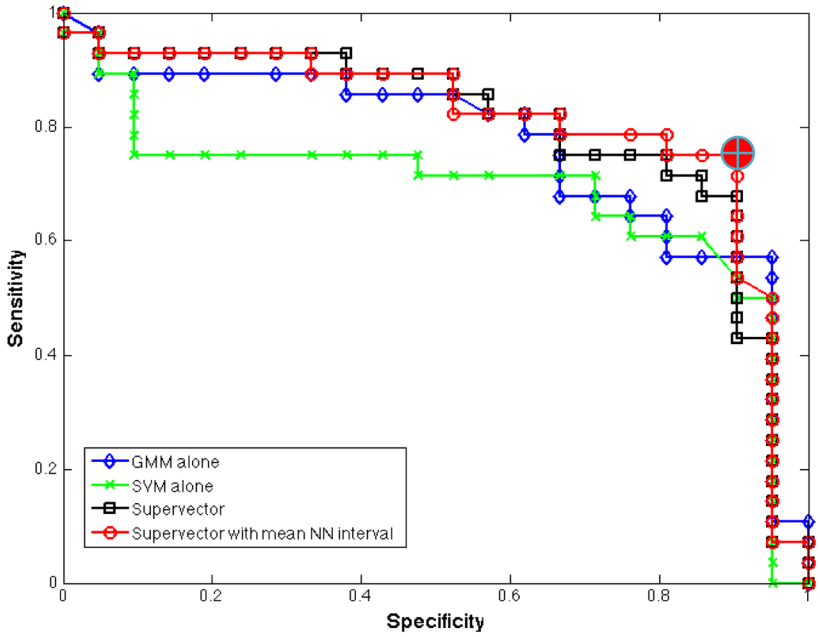


Figure 6.9: Comparison of ROC area (AUC) obtained by different methods across all neonates. AUC for Supervector=79.9%, Supervector with meant NN interval feature=81%, GMM=76.5% and SVM =68.3%.

Table 6.4: Comparison of different sequence lengths on the overall performance of the system

Sequence length (epochs)	ROC Area (%)
3	79.6
5	80.3
7	**81.0**
9	79.9

6.4.4 Confusion matrix

The confusion matrix of the system is presented in Table 6.5. This matrix is obtained with the best operating point obtained from the ROC curve of the supervector approach (indicated by circle marker in Figure 6.9). The best point on the ROC curve is the point which has the minimum Euclidean distance from the ideal point i.e. (1,1). Out of 48 recordings, 9 were misclassified by the proposed system. Two misclassifications were in grade 1 whereas 7 errors were made in grade 2. This corresponds to an overall accuracy of 81.6% achieved by the proposed system. These results are better than the system reported in

(Matic et al. 2013). The overall accuracy obtained by their system was 80% on a smaller dataset of 19 neonates and 36 recordings.

The last row of Table 6.5 shows the precision of the system which is defined as the ratio between the number of correctly assigned decisions and the number of total decisions assigned to a specific grade. It can be seen that, the accuracy of classifying grade 1 is high but its precision is low; this means that many neonates that required cooling were not able to get the treatment. On a positive note, however, the system was 91% precise in predicting the treatment for the ones who needed it.

Table 6.5: Confusion matrix of the proposed system's performance

		System output		**Accuracy (%)**
		1	**2**	
Actual Grade	**1**	19	2	90.4
	2	7	21	75
Precision (%)		73	91.3	

6.5 Fusion of EEG and HRV

In the previous chapter, it was shown that the supervector based system achieved an accuracy of 87% for classifying 4 different grades of HIE using EEG. If just two grades (1 = mild and 2 = moderate-severe) are assumed, then the EEG based system can achieve an accuracy of 95% as compare to 81.6% achieved by the HRV based system. Although, the number of misclassifications were high in the HRV based system, there is still some benefit in combining the information from both modalities to make the decisions of the automated system more reliable. In this regard, a simple preliminary study was conducted to start moving in the direction of a multi-modal HIE classification system.

Development of information fusion systems is a very broad area and has increasingly becoming popular due to the availability of huge digital data banks in this era. The techniques reported in the past can be divided into three basic types; feature, classifier and decision level fusion.

The most straight forward method is to combine the features from different modalities at the feature level. A recent example of multi-modal feature level fusion of HRV, EEG and clinical features to grade HIE was proposed in (Temko

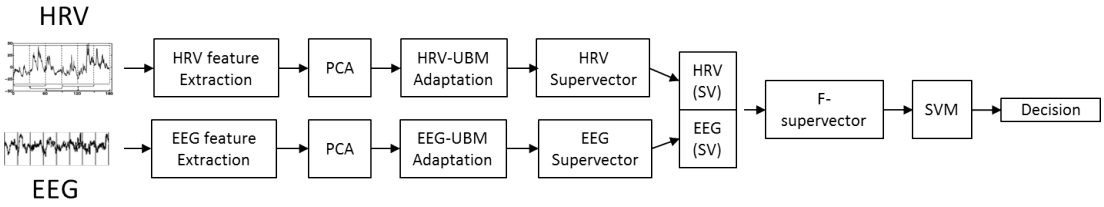
et al. 2015). In the classifier level techniques, features from different modalities are used together to create a single model. An example of such a technique is the information bottleneck based multi-stream approach for speaker diarization proposed in (Vijayasenan et al. 2009). Another method of combining two streams of speech features using hidden Markov models was introduced in (Ketabdar et al. 2006). In this method, a single GMM model is created by combining two separate GMMs trained from the data of the individual modalities. In the decision level techniques data is classified separately by the expert classifiers of each individual modality and then the output of the classifiers is combined using simple probability, fuzzy logic or voting based fusion methods (Stylianou et al. 2005, Temko et al. 2006, Butko et al. 2008).

Some of the biggest challenges, in feature and classifier level techniques are artifacts, variable dimensionality and the resolution of data coming from different modalities. As the signals are usually collected from multiple sources, the artifacts are usually different and do not correlate in time. The number of features extracted from the signals of different modalities are usually variable which pose a challenge in fusing the information to create a single classifier model. Lastly, each source of data may require a different window size for effective feature extraction. This results in different resolutions of data that do not synchronise in time. Decision level techniques do not suffer from the above mentioned problems but due to late integration, the information sharing for better classification is reduced.

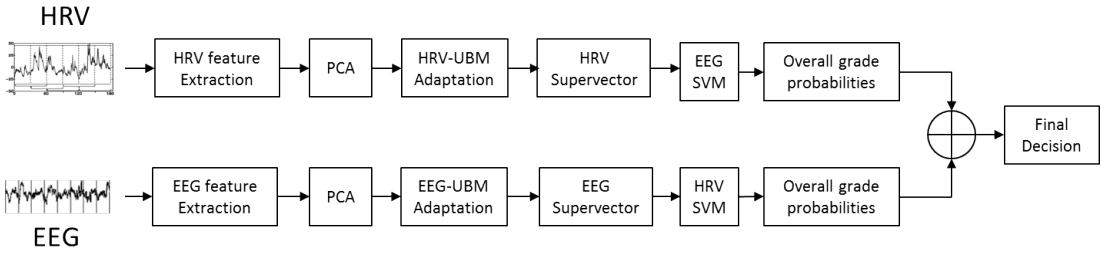
6.5.1 Proposed methods of fusion

In this work, two approaches are examined to fuse HRV and EEG signals: 1) Using supervectors (classifier level) and 2) Fusing the overall probabilities from two different systems (decision level).

Figure 6.10a shows an overview of the supervector based HIE classification system. EEG features were extracted from 8 second epochs of EEG with 4 second overlap. Sequences of 20 epochs (same as the best performing EEG based system) of EEG which corresponds to 84 seconds were then created to adapt the EEG-UBM model. HRV features were extracted from 60 second epochs of HRV signal with 4 second overlap. Seven consecutive epochs were used to create a sequence for HRV-UBM adaptation. The start and end time of both HRV and EEG sequences were synchronised. Two separate UBMs were trained, one using only the EEG data and the other using only the HRV data. Then two supervectors were created, one by using a sequence of EEG features and the second from



(a) Supervector based fusion system



(b) Final probabilities based fusion system

Figure 6.10: Overview block diagrams of two fusion based HIE classification systems

the sequence of HRV features. These two supervectors were then combined and a fused-supervector (f-supervector) was created. As there are 8 EEG channels and only one ECG channel, therefore the ECG channel was replicated for all the EEG channels. The f-supervectors are then fed to a linear SVM for classification. Artifacts of both signals are different not only morphology and etiology wise but also they also do not co-exist at the same time points. Therefore, in situations when an artifact has appeared in the HRV signal, then the decision was made only with EEG data using the EEG specialist classifier. This process is kept automated.

The second approach is a decision level technique. Two completely separate best performing systems are used to produce the overall probabilities of the HIE grade. These systems are completely trained and tested on a single modality. For example, the EEG based system produces the probability of grade 2 using only EEG data, whereas HRV based system produces probabilities based on the HRV data. These overall probabilities are then combined using simple averaging and a global probability measure is produced at the end. The benefit of decision level techniques is the flexibility to give some weight to the decision of a particular system. For example, we know that the EEG based system is more powerful in determining the grade of HIE therefore higher weights can be assigned to its probabilities as compared to the probabilities from the HRV based

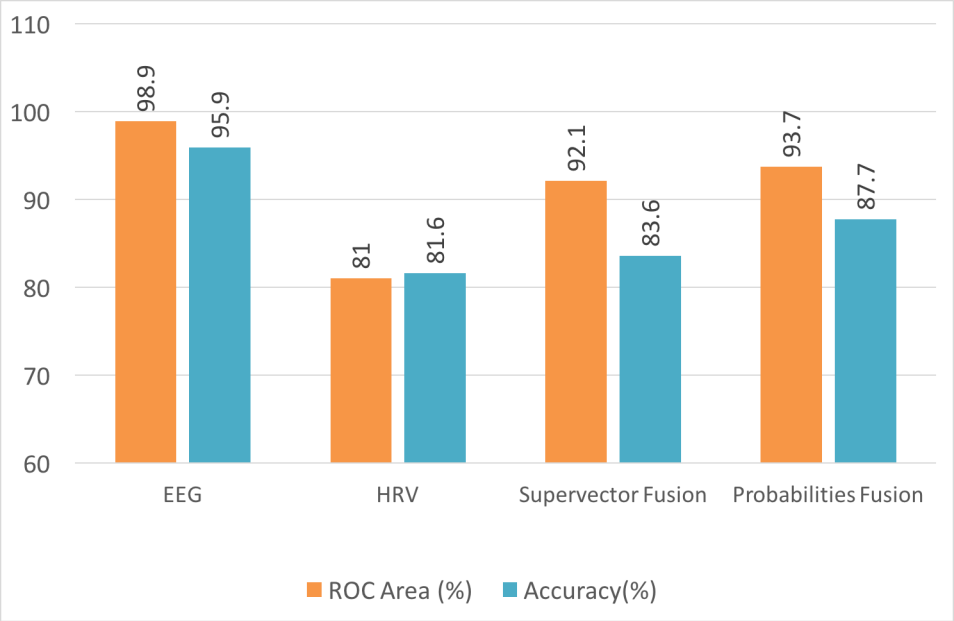


Figure 6.11: Comparison of ROC area and overall accuracy using different modalities and fusion based approaches.

system. Moreover, any particular system does not has to deal with the artifacts of the other modality.

6.5.2 Results and discussion

Figure 6.11 presents a comparison of performance (in terms of ROC area and overall accuracy at the best operating point on the ROC curve) using single modality and fusion based systems. The EEG based system achieved the highest ROC area of 98.9% and an overall accuracy of 95.9%. The fusion based methods were not able to further improve the performance of the EEG based system. The final-probabilities-based-fusion method attained an overall accuracy of 87.7% as compared to 83.6% obtained by the supervector based fusion method. Although, having 4% difference in accuracy, both systems were able to achieve similar ROC areas.

The confusion matrix of the fusion based system along EEG based system is shown in Table 6.6. The EEG based system correctly classified all grade 2 recordings whereas it made two misclassifications in grade 1 recordings. The statistics of fusion based systems are very similar to the confusion matrix of HRV based system presented in Table 6.5. Two misclassifications were made in grade 1 recordings by both systems whereas six were made in grade 2 recordings by the supervector based fusion system and 4 by the final probabilities based

Table 6.6: Confusion matrices of fusion based systems compared with the EEG based system.

		EEG		**F1**		**F2**	
		1	**2**	**1**	**2**	**1**	**2**
Actual	**1**	19	2	19	2	19	2
Grade	**2**	0	28	6	22	4	24

* F1= Supervector based fusion
F2= Final probabilities based fusion

Table 6.7: Common misclassifications and unique new decisions made by HRV alone, EEG alone and fusion based HIE classification systems

	Total Misclassified	Common misclassifications to				Unique new correct decisions to			
		F 1	**F 2**	**HRV**	**EEG**	**F 1**	**F 2**	**HRV**	**EEG**
F 1	8	-	6	7	1	-	0	2	1
F 2	6	6	-	6	1	2	-	3	1
HRV	9	7	6	-	1	1	0	-	1
EEG	2	1	1	1	-	7	5	8	-

* F1= Supervector based fusion
F2= Final probabilities based fusion

system.

In order to get further insight on the lower accuracy produced by the fusion based system, the common misclassifications and the number of unique decisions produced by each individual modality and the fusion based systems were examined (Table 6.7). As expected, the fusion based systems were heavily effected by the weak discrimination capability of HRV. Seven out of eight misclassifications made by the supervector based fusion system were similar to the misclassifications made by the HRV based system and only one was similar to the EEG based system. Likewise, all the six misclassifications made by the final probabilities based fusion system were similar to the HRV based system. Moreover, only 1 new correct decision was made by the fusion based system when compared to the EEG-based system. The HRV based system also produced only 1 new correct decision.

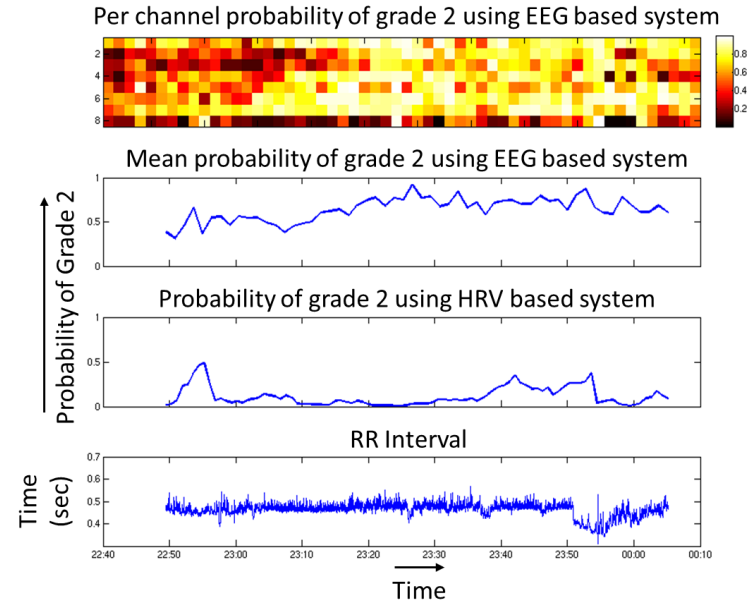
These initial results suggests that the final probabilities based systems are more preferable as compared to early fusion of the two modalities. Moreover,

the lower accuracy of fusion based systems, indicates that there is not much complementary information present in the HRV signal that can help increase the accuracy of classifying HIE grades. Another contributing factor to lower accuracy with systems that use the HRV signal, could be the disassociation of EEG to HRV at different time points. It is reported that the correlation of HRV with EEG could increase/decrease at different time points after birth (Goulding et al. 2015). This was indeed observed in our experiments. Some examples of such recordings are shown in Figure 6.12a, 6.12b and 6.13. It can be seen in Figure 6.12a that the probability of grade 2 is higher when based on the EEG. It is evident however, from the probability trace of HRV based system and the RR-interval plot that the HRV does not correlate with the EEG grade. This recording was correctly classified by the EEG based system, however the HRV based systems misclassified it. Figure 6.12b shows the probabilities of a misclassified grade 1 recording. The EEG grade and the resulting probability was lower in most of the recording, the HRV plot and probability of the HRV based system is however completely the opposite. Last, Figure 6.13 shows an example of a grade 1 recording that was correctly classified by both systems. However, it can be seen that the HRV and subsequent probabilities from the HRV based system did not correlate with the EEG at all time points. The EEG based system classified this file with a high grade 1 probability, whereas the HRV based system barely classified it correctly.

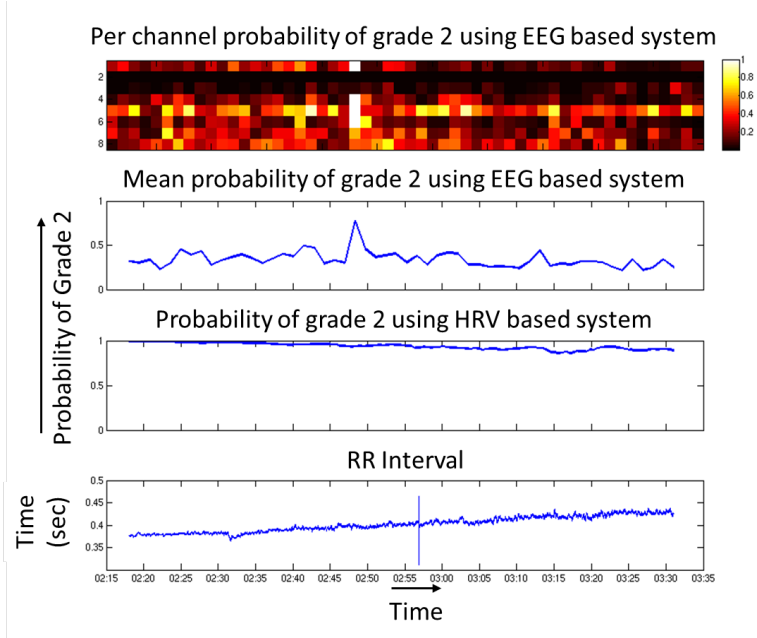
6.6 Conclusion

A novel system of grading HIE injury using HRV in neonates is proposed. This was a proof of concept study to show the importance of HRV for the classification of HIE. The results are promising for grading the HIE severity which could help the clinical staff in making a decision for grading HIE when EEG is not available. Also, if future HRV studies assessing HRV within 6 hours after birth show consistent results, HRV may have the potential to identify infants that should be treated with TH. The dataset for this study was not collected primarily for HRV analysis so the HRV quality was not in good condition. It is expected that the results could be further improved with a better quality dataset.

Two systems for the fusion of EEG and HRV signals are also presented. Although the preliminary results does not show any further improvement for classifying HIE into mild and moderate-severe grades, the misclassified examples do indicate a number of key directions need to be investigated in the future



(a) Grade 2 recording misclassified by HRV based systems.



(b) Grade 1 recording misclassified by the HRV based systems.

Figure 6.12: Examples of misclassified recordings (1 hour long) showing dissociation of HRV with the EEG grade. The top plot in each figure shows the probability of grade 2 of each EEG sequence for every channel using EEG based system. The mean probability of the EEG based system of all the channels is shown in the top second plot. The third plot presents the probability trace obtained by the HRV based system. The last plot shows the raw RR interval acquired using Pan-Tompkins method.

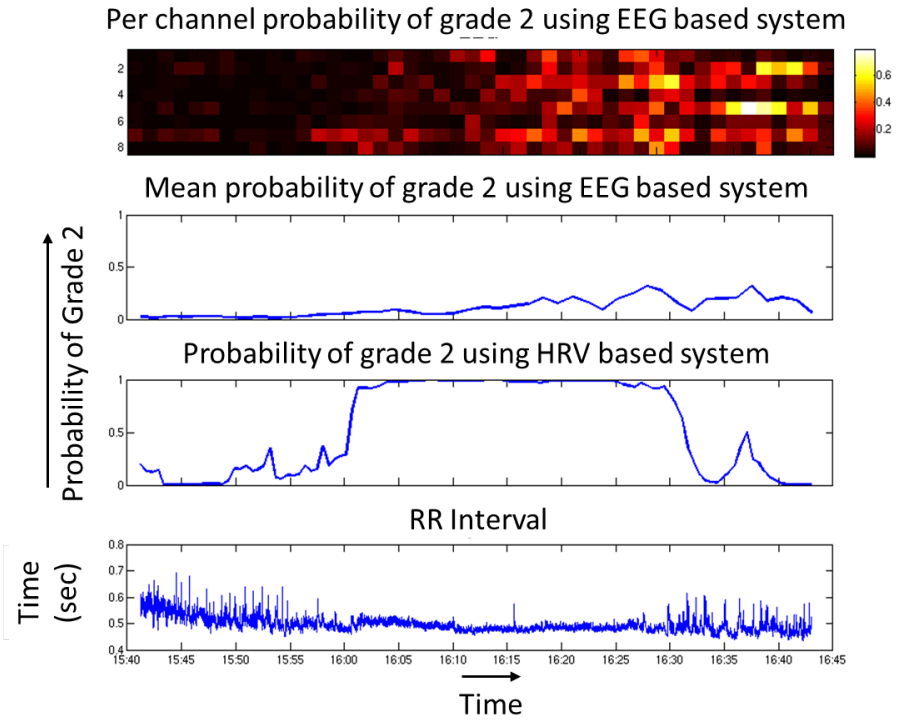


Figure 6.13: Example of a one hour grade 1 recording that was correctly classified by both systems. HRV based systems classified the recording with very low probability due to uncorrelated HRV segment in the middle of recording. The top plot shows the probability of grade 2 of each EEG sequence for every channel using the EEG based system. The mean probability of EEG based system of all the channels is shown in top second plot. The third plot presents the probability trace obtained by the HRV based system. The last plot shows the raw RR interval acquired using Pan-Tompkins method.

to allow HRV to become a better biomarker for HIE classification and make its information integrable with the EEG and other physiological signals.

References

Aliefendioglu, D., Dogru, T., Albayrak, M., Dibekmısırlıoglu, E., & Sanli, C. (2012). Heart rate variability in neonates with hypoxic ischemic encephalopathy. *Indian Journal of Pediatrics*, 79(11):1468–1472.

Butko, T., Temko, A., Nadeu, C., & Canton, C. (2008). Inclusion of video information for detection of acoustic events using the fuzzy integral. *In Machine Learning for Multimodal Interaction*. Springer, pp. 74–85.

Camm, A. J., Malik, M., Bigger, J., Breithardt, G., Cerutti, S., Cohen, R., Coumel, P., Fallen, E. et al. (1996). Heart rate variability. Standards of measurement, physiological interpretation, and clinical use. *European Heart Journal*, 17(3):354–381.

Doyle, O. (2010). *Statistical pattern recognition of physiological signals for newborn monitoring*. Ph.D. thesis, University College Cork, Cork, Ireland.

Doyle, O., Temko, A., Marnane, W., Lightbody, G., & Boylan, G. (2010). Heart rate based automatic seizure detection in the newborn. *Medical Engineering & Physics*, 32(8):829–839.

Gauvain, J. & Lee, C.-H. (1994). Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Transactions on Speech and Audio Processing*, 2(2):291–298.

Gholinezhadasnefestani, S., Temko, A., Stevenson, N., Boylan, G., Lightbody, G., & Marnane, W. (2015). Assessment of quality of ECG for accurate estimation of Heart Rate Variability in newborns. *In Proceedings of IEEE Annual International Conference of Engineering in Medicine and Biology Society, (EMBC) 2015*. IEEE, pp. 5863–5866.

Goulding, R. M., Stevenson, N. J., Murray, D. M., Livingstone, V., Filan, P. M., & Boylan, G. B. (2015). Heart rate variability in hypoxic ischaemic en-

cephalopathy: correlation with EEG grade and two-year neurodevelopmental outcome. *Pediatric Research*, 77(5):681–687.

Gunn, A. J., Gunn, T. R., Gunning, M. I., Williams, C. E., & Gluckman, P. D. (1998). Neuroprotection with prolonged head cooling started before postischemic seizures in fetal sheep. *Pediatrics*, 102(5):1098–1106.

Jensen, A., Garnier, Y., Middelanis, J., & Berger, R. (2003). Perinatal brain damage—from pathophysiology to prevention. *European Journal of Obstetrics & Gynecology and Reproductive Biology*, 110, Supplement:S70 – S79.

Ketabdar, H., Bourlard, H., & Bengio, S. (2006). Hierarchical multi-stream posterior based speech recognition system. *In Machine Learning for Multimodal Interaction*. Springer, pp. 294–306.

Malliani, A., Pagani, M., Lombardi, F., & Cerutti, S. (1991). Cardiovascular neural regulation explored in the frequency domain. *Circulation*, 84(2):482–492.

Matic, V., Cherian, P. J., Widjaja, D., Jansen, K., Naulaers, G., Van Huffel, S., & De Vos, M. (2013). Heart rate variability in newborns with hypoxic brain injury. *In Oxygen Transport to Tissue XXXV*. Springer, pp. 43–48.

Murray, D. M., Boylan, G. B., Ryan, C. A., & Connolly, S. (2009). Early EEG findings in hypoxic-ischemic encephalopathy predict outcomes at 2 years. *Pediatrics*, 124(3):e459–e467.

Novak, V., Novak, P., deMarchie, M., & Schondorf, R. (1995). The effect of severe brainstem injury on heart rate and blood pressure oscillations. *Clinical Autonomic Research*, 5(1):24–30.

Pan, J. & Tompkins, W. J. (1985). A real-time QRS detection algorithm. *IEEE Transactions on Biomedical Engineering*, 32(3):230–236.

Platt, J. C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *In Advances in Large Margin Classifiers*, Neural Information Processing Series. MIT Press, pp. 61–74.

Pomeranz, B., Macaulay, R., Caudill, M. A., Kutz, I., Adam, D., Gordon, D., Kilborn, K. M., Barger, A. C. et al. (1985). Assessment of autonomic function in humans by heart rate spectral analysis. *American Journal of Physiology-Heart and Circulatory Physiology*, 248(1):H151–H153.

Reynolds, D. (2009). Gaussian mixture models. *In Encyclopedia of Biometrics*. Springer US, pp. 659–663.

Rosenstock, E., Cassuto, Y., & Zmora, E. (1999). Heart rate variability in the neonate and infant: analytical methods, physiological and clinical observations. *Acta Paediatrica*, 88(5):477–482.

Selig, F. A., Tonolli, E. R., Silva, E. V. C. M. d., & Godoy, M. F. d. (2011). Heart rate variability in preterm and term neonates. *Arquivos brasileiros de cardiologia*, 96(6):443–449.

Shankaran, S., Laptook, A. R., Ehrenkranz, R. A., Tyson, J. E., McDonald, S. A., Donovan, E. F., Fanaroff, A. A., Poole, W. K. et al. (2005). Whole-body hypothermia for neonates with hypoxic–ischemic encephalopathy. *New England Journal of Medicine*, 353(15):1574–1584.

Stylianou, Y., Pantazis, Y., Calderero, F., Larroy, P., Severin, F., Schimke, S., Bonal, R., Matta, F. et al. (2005). GMM-based multimodal biometric verification. *In Proceedings of eINTERFACE 05 Summer Workshop on Multimodal Interfaces*. pp. 1–5.

Temko, A., Doyle, O., Murray, D., Lightbody, G., Boylan, G., & Marnane, W. (2015). Multimodal predictor of neurodevelopmental outcome in newborns with hypoxic-ischaemic encephalopathy. *Computers in Biology and Medicine*, 63:169–177.

Temko, A., Macho, D., & Nadeu, C. (2006). Improving the performance of acoustic event classification by selecting and combining information sources using the fuzzy integral. *In Machine Learning for Multimodal Interaction*, vol. 3869 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, pp. 357–368.

Vijayasenan, D., Valente, F., & Bourlard, H. (2009). An information theoretic approach to speaker diarization of meeting data. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(7):1382–1393.

Volpe, J. J. (2008). *Neurology of the Newborn*. Elsevier Health Sciences.

Chapter 7

Conclusions and Future Work

Brain injuries around the time of birth may impair the neural functions of a baby for the rest of life or in severe cases may lead to death. EEG is considered one of the best method for monitoring sick neonates but its interpretation and availability hinders its wide usage in the NICUs. Automated systems for detecting abnormal brain functions and grading brain injuries could help revolutionize neurocritical cot-side monitoring.

This work presented automated systems for three classification tasks

- *Neonatal seizure detection*
- *Grading the severity of HIE injury using EEG*
- *Grading the severity of HIE injury using ECG*

In this context, the main objective of this thesis was to investigate the methods for the exploration of temporal and contextual information of neonatal EEG at the classifier level. This chapter will look at the main contributions in obtaining these objectives and the conclusions drawn from the presented work. Lastly, the future directions to carry out further research in this area are suggested.

7.1 Conclusions and main contributions

The work executed in this thesis mainly focused on the development of a fundamental architecture for the dynamic classification of different neonatal physiological signals. To this end, the motivations and the need for a system which can better utilize the contextual information present in EEG and HR signals is highlighted and has been shown to improve the performance over the static

classification systems. This section, will summarize the main conclusions and contributions of each chapter.

Chapter 2 provided a brief overview of the medical background of seizures and hypoxic-ischemic encephalopathy. Moreover, different neonatal EEG patterns that characterize the abnormal brain function are presented. Artifacts are one of the challenging problem for any automated system, therefore a brief summary of some major artifacts was also presented.

A theoretical background of the different classification methods utilized in this work were presented in Chapter 3. The SVM is used as the primary classifier for the above mentioned automated systems. Different kernel functions were examined that enables the SVM to classify sequences of feature vectors and consequently allowing it to explore the contextual information present at both short and longer time scale.

A novel way of incorporating the contextual information of neonatal seizures at the classifier-level is presented in Chapter 4. A DTW based kernel function is used which enabled the SVM to exploit the temporal evolution of seizures from sequences of short-term feature vectors. Detailed working of the DTW is presented along with toy and real examples. It was reported in (Temko et al. 2011b) that short seizures of length less than 1 minute are frequently missed. Therefore, the DTW based technique was specifically tested on a dataset that consisted of only short-seizure and non seizures events. A new approach of using a classifier to make per channel annotations was devised in this work. This approach was used to create a dataset of only short seizures and non-seizure events. It is shown that using the DTW based method, the ROC area for the classification of short seizures events improved from 80.8% to 81.9%. Moreover the fusion of static-SVM and dynamic-SVM, further improved the ROC area to 86.1%. The same fusion based system was also validated on a larger dataset of 267 hours of EEG recordings from 17 neonates. A comparison of the performance of static, DTW and fusion based classifier was presented. Results showed promising improvement in the number of detected seizure events at significantly low false detection rates. Most importantly, a 12% improvement in the detection of short seizures events was achieved using the fusion based system which were previously being missed by the static-SVM based system.

An EEG-based automated system for classifying the severity of HIE injury into one of the four grades was presented in Chapter 5. The system is based on a cross-disciplinary approach employing a supervector kernel for SVM. Complete details of creating a supervector using Gaussian Mixture Models (GMM)

and the Maximum a Posteriori (MAP) adaptation technique were presented. This method of modelling the sequences of feature vectors extracted from short discrete windows has shown significant improvement over the system proposed by (Stevenson et al. 2013). An overall accuracy of 87% is achieved to classify 1 hour recordings of 54 neonates. With a new label 'unknown' assigned to the recordings with lower confidence levels, an accuracy of 96% is attained. Different post-processing techniques to grade one hour long files were proposed. In comparison to conventional majority voting, the novel 2-step majority voting technique has shown to increase both the accuracy and confidence/quality of decisions produced. A novel method to convert the output of a multi-class classifier into a continuous probabilistic trace of the grade of the background EEG was also presented. The probability trace, shown as an example in Figure 5.15, could be integrated in a cot-side monitor along with the probability trace of the in-house developed neonatal seizure detection system (Temko et al. 2015b, 2013) and other physiological measures. The proposed automated HIE grading system can provide significant assistance to healthcare professionals in assessing the severity of HIE. This represents a practical and user friendly implementation which acts as a decision support system in the clinical environment.

A novel system of grading HIE injury in neonates using HR was proposed in Chapter 6. The study has highlighted the importance of HR for the classification of HIE into at least two grades of severity (mild and moderate-severe). These two grades are important because an infant is usually treated with Therapeutic Hypothermia (TH), once the encephalopathy develops from the mild to the moderate grade. The proposed system uses a supervector approach similar to the EEG based system for HIE grading. The system was validated on the 1 hour ECG recordings of 48 neonates. The best ROC area achieved by our system was 81% as compared to 67% and 70.4% attained by epoch based SVM and GMM methods respectively. Promising results indicate that this system can be used for grading brain injury in neonates in the centres where the EEG is not readily available. The dataset for this study was not collected primarily for HR analysis so the ECG quality was not in good condition. It is expected that the performance could be further improved with a better quality dataset.

A preliminary study on the combination of EEG and ECG signals was carried out to grade HIE recordings. Two fusion techniques are investigated in this work. The first approach (early integration) combines the long term statistical features of HR and EEG extracted by adapting the two separate UBM models and then uses SVM to classify the HIE grade. The second approach (late

integration) uses two completely independent systems for the classification of each signal and then the probabilistic outputs of both systems are fused to get a grade of the HIE. Both schemes did not improve over the EEG-only based system. However the preliminary results favour the usage of a late integration approach. Initial investigations of the misclassifcations show that the EEG does not correlate with the HR at all times.

7.2 Future work

There are a number of key areas that still need attention to advance the state of the art in automated systems for monitoring brain injuries and their translation to the clinical environment. The following are a few future work directions that can be extracted from this work.

As mentioned above, this study provided the foundation for developing a contextually aware dynamic classification system for neonatal brain monitoring. A basic classifiers fusion approach is applied in this work. It is shown in (Temko et al. 2015c) that ensemble of many classifiers can lead to better results for detecting seizures. Classifier fusion is a promising method and therefore other sophisticated techniques such as adaboost (Freund & Schapire 1997) and random forests (Breiman 2001) classifier can be utilised in future for further improvements.

There are certain aspects of creating a complete dynamic classification system which were not the focus of this work. For example, the length of the sequences to be classified was fixed in this study. In reality, however, the duration of events such as seizures is variable. In order to define the length of these events and then extract features accordingly, a separate segmentation block is needed. In speech recognition for example, voice activity detectors based on the energy of the signal are deployed before the classification stage to extract and separate the meaningful chunks of the speech signal (Ramirez et al. 2007). Another method used in speaker identification is the HMM based 'diarization' which segments an audio stream according to the speakers identity (Tranter & Reynolds 2006). It is expected that the performance of the dynamic classifiers may further increase with the addition of such EEG events segmentation block.

Despite a good improvement in the detection rate using the proposed fusion based classifier for neonatal seizure detection, the system did not significantly increase the performance when the post-processing stage was included. This behavior does not come by surprise. Indeed the post-processing proposed in

(Temko et al. 2011a) is developed for static-SVM. As it is evident from the results of fusion without post-processing, DTW based system provides very important complementary information which helped in increasing the performance not only at the epoch level but also at the events level. However, the moving average filter is a very coarse method to explore the contextual information at decision level and it does not accentuate the benefits of the proposed fusion. Therefore, a different post-processing technique is needed for further improvement. Moreover, a number of sequential kernels as presented in chapter 3 can also be tested for the seizure detection task. Lastly, this work only investigated the temporal information in the EEG, however it is reported in other works that using the spatial (channel) information could also increase the seizure detection rate (Deburchgraeve et al. 2008, Greene et al. 2008). Therefore another step could be the incorporation of methods that allow both the temporal and spatial classification of EEG sequences.

In the area of automated HIE grading using EEG, there are certain areas that can be improved in the presented HIE grading automated system. It is noted that there are some brief temporal activities e.g. spikes, sharp waves and certain spatial characteristics such as asynchrony and asymmetry which are not detected by our system. Inclusion of parallel detectors using techniques such as matching pursuit (Mallat 2008), offer promising results for detecting such short temporal events and may considerably increase the accuracy. The presence of a sleep-wake cycle is a major feature in deciding the grade of HIE-EEG. It has been shown that incorporating the information about the sleep wake cycling may increase the performance of the automated system (Stevenson et al. 2013). The detection of sleep wake cycle requires the analysis of EEG at much larger intervals. The proposed system however, works with a window length of approximately 84 seconds which is not sufficient to detect the sleep-wake cycles. Inclusion of a parallel sleep-wake cycle detector could also improve the accuracy of the system. Lastly, the feature set used in the current study is the same as for the seizure detection. This feature set can either be expanded or reduced to better suit this problem area.

HR based automated HIE grading is still at its early stages. One of the key areas to be addressed is the identification of new HRV features. This study only utilized some of the most popular features identified in the medical literature to grade HIE. This feature set is not exhaustive and needs to be extended to more sophisticated features that can better represent the HRV signal. Moreover, as shown through results, it is inevitable to develop a feature selection routine.

Therefore future work is required to further investigate and also expand the features set for the HRV and deploy a feature selection algorithm. Artifacts were one of the biggest challenge in this work. Efforts towards the development of an automated artifact detection and rejection algorithm for ECG and HRV signals is well under-way. This could help in increasing the classification rate of HIE grades (Gholinezhadasnefestani et al. 2015).

Combining the EEG with other physiological signals is one of the most important areas that will be the focus of attention in coming years. In this regard, the presented work has highlighted some important aspects to be considered in the future to combine the information from the EEG and HRV signals. The investigation of the time points where HRV correlates the most with the EEG grade may help to design a system that can assign a weight to the information coming from the two domains. It is shown in (Temko et al. 2015a) that including a feature selection routine for both EEG and HRV increased the performance for predicting the outcome of neonates suffered from HIE. Therefore, such a routine for a fusion based system may increase the performance. And last but not least, more robust information theory and machine learning approaches for data fusion can be investigated to develop a full multi-modal and multi-stream HIE grading system (Vijayasenan et al. 2009, Valente 2009).

References

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Deburchgraeve, W., Cherian, P., De Vos, M., Swarte, R., Blok, J., Visser, G. H., Govaert, P., & Van Huffel, S. (2008). Automated neonatal seizure detection mimicking a human observer reading EEG. *Clinical Neurophysiology*, 119(11):2447–2454.

Freund, Y. & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119 – 139.

Gholinezhadasnefestani, S., Temko, A., Stevenson, N., Boylan, G., Lightbody, G., & Marnane, W. (2015). Assessment of quality of ECG for accurate estimation of Heart Rate Variability in newborns. *In Proceedings of IEEE Annual International Conference of Engineering in Medicine and Biology Society, (EMBC) 2015*. IEEE, pp. 5863–5866.

Greene, B. R., Marnane, W. P., Lightbody, G., Reilly, R. B., & Boylan, G. B. (2008). Classifier models and architectures for EEG-based neonatal seizure detection. *Physiological Measurement*, 29(10):1157–1178.

Mallat, S. (2008). *A wavelet tour of signal processing: the sparse way*. Academic press, 3rd ed.

Ramirez, J., Górriz, J. M., & Segura, J. C. (2007). Voice activity detection. fundamentals and speech recognition system robustness. *In Robust Speech Recognition and Understanding*. I-Tech Education and Publishing, pp. 1–19.

Stevenson, N. J., Korotchikova, I., Temko, A., Lightbody, G., Marnane, W. P., & Boylan, G. B. (2013). An automated system for grading EEG abnormality in term neonates with hypoxic-ischaemic encephalopathy. *Annals of Biomedical Engineering*, 41(4):775–785.

Temko, A., Boylan, G., Marnane, W., & Lightbody, G. (2013). Robust neonatal EEG seizure detection through adaptive background modeling. *International Journal of Neural Systems*, 23(04):1350018.

Temko, A., Doyle, O., Murray, D., Lightbody, G., Boylan, G., & Marnane, W. (2015a). Multimodal predictor of neurodevelopmental outcome in newborns with hypoxic-ischaemic encephalopathy. *Computers in Biology and Medicine*, 63:169–177.

Temko, A., Marnane, W., Boylan, G., & Lightbody, G. (2015b). Clinical implementation of a neonatal seizure detection algorithm. *Decision Support Systems*, 70:86–96.

Temko, A., Nadeu, C., Marnane, W., Boylan, G. B., & Lightbody, G. (2011a). EEG signal description with spectral-envelope-based speech recognition features for detection of neonatal seizures. *IEEE Transactions on Information Technology in Biomedicine, IEEE Transactions on*, 15(6):839–847.

Temko, A., Sarkar, A., & Lightbody, G. (2015c). Detection of seizures in intracranial EEG: UPenn and Mayo clinic's seizure detection challenge. *In Proceedings of IEEE Annual International Conference of Engineering in Medicine and Biology Society, (EMBC) 2015*. IEEE, pp. 6582–6585.

Temko, A., Thomas, E., Marnane, W., Lightbody, G., & Boylan, G. (2011b). EEG-based neonatal seizure detection with support vector machines. *Clinical Neurophysiology*, 122(3):464–473.

Tranter, S. & Reynolds, D. (2006). An overview of automatic speaker diarization systems. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1557–1565.

Valente, F. (2009). A Novel Criterion for Classifiers Combination in Multistream Speech Recognition. *IEEE Signal Processing Letters*, 16(7):561–564.

Vijayasenan, D., Valente, F., & Bourlard, H. (2009). An information theoretic approach to speaker diarization of meeting data. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(7):1382–1393.